STOCHASTIC CONTINUUM ARMED BANDIT PROBLEM OF FEW LINEAR PARAMETERS IN HIGH DIMENSIONS

HEMANT TYAGI, SEBASTIAN STICH AND BERND GÄRTNER

Department of Computer Science,
Institute of Theoretical Computer Science,
ETH Zürich, CH-8092, Switzerland

Abstract. We consider a stochastic continuum armed bandit problem where the arms are indexed by the $\ell_2$ ball $B_d(1+\nu)$ of radius $1+\nu$ in $\mathbb{R}^d$. The reward functions $r : B_d(1+\nu) \to \mathbb{R}$ are considered to intrinsically depend on $k \ll d$ unknown linear parameters so that $r(x) = g(Ax)$ where $A$ is a full rank $k \times d$ matrix. Assuming the mean reward function to be smooth we make use of results from low-rank matrix recovery literature and derive an efficient randomized algorithm which achieves a regret bound of $O(C(k, d)n^{\frac{1+k}{2} + (\log n)^{\frac{1}{2}}})$ with high probability. Here $C(k, d)$ is at most polynomial in $d$ and $k$ and $n$ is the number of rounds or the sampling budget which is assumed to be known beforehand.

1. Introduction

In the continuum armed bandit problem, a player is given a set of strategies $S$—typically a compact subset of $\mathbb{R}^d$. At each round $t = 1, \ldots, n$, the player chooses a strategy $x_t$ from $S$ and then receives a reward $r_t(x_t)$. Here $r_t : S \to \mathbb{R}$ is the reward function chosen by the environment at time $t$ according to the underlying model. The model we consider in this work is stochastic i.e. the reward functions are assumed to be sampled in an i.i.d manner from an underlying distribution at each round. The player selects strategies across different rounds with the goal of maximizing the total expected reward. Specifically, the performance of the player is measured in terms of regret defined as the difference between the total expected reward of the best fixed (i.e. not varying with time) strategy and the expected reward of the sequence of strategies played by the player. If the regret after $n$ rounds is sub-linear in $n$, this implies as $n \to \infty$ that the per-round expected reward of the player asymptotically approaches that of the best fixed strategy.

The problem faced by the player at each round is the classical “exploration-exploitation dilemma”. On one hand if the player chooses to focus his attention on a particular strategy which he considers to be the best (“exploitation”) then he might fail to know about other strategies which have a higher expected reward. However if the player spends too much time collecting information (“exploration”) then he might fail to play the optimal strategy sufficiently often. Some applications of continuum armed bandit problems are in: (i) online auction mechanism design \cite{1, 2} where the set of feasible prices is representable as an interval and, (ii) online oblivious routing \cite{3} where $S$ is a flow polytope.

E-mail address: htyagi@inf.ethz.ch, sstich@inf.ethz.ch, gaertner@inf.ethz.ch.

Key words and phrases. Bandit problems, continuum armed bandits, functions of few variables, online optimization, low-rank matrix recovery.

The project CG Learning acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 255827.
For a $d$-dimensional strategy space, if the only assumption made on the reward functions is on their degree of smoothness then any algorithm will incur worst-case regret which depends exponentially on $d$. To see this, let $S = [-1,1]^d$ and consider a time invariant reward function that is zero in all but one orthant $O$ of $S$. This curse of dimensionality is avoided by reward functions possessing more structure, two popular cases being linear reward functions (see for example [4, 5]) and convex reward functions (see for example [6, 7]) for which the regret is polynomial in $d$ and sub-linear in $n$.

Low dimensional models for high dimensional reward functions. Recently there has been work in the online optimization literature where the reward functions are assumed to be low-dimensional or in other words have only a few degrees of freedom compared to the ambient dimension. In [8, 9] the authors consider the linear stochastic bandit problem in the setting that the unknown parameter (of dimension $d$) is $k$-sparse with $k \ll d$. In [10] the authors consider both stochastic and adversarial versions of continuum armed bandits where the $d$-variate reward functions are assumed to depend on an unknown subset of the coordinate variables of size $k \ll d$. They derive nearly optimal regret bounds with the rate of regret depending only on $k$. In [11] the authors consider the problem of Bayesian optimization of high dimensional functions by again assuming the functions to depend on only a few relevant variables. Considering the function to be a sample from a high dimensional Gaussian process they provide an algorithm with strong theoretical guarantees in terms of regret bounds. This model is generalized in [12] where the authors consider the underlying function to effectively vary along a low-dimensional subspace. Assuming the noise-less setting they adopt a Bayesian optimization framework and derive bounds on simple regret.

We consider the setting where the reward function $r_t : B_d(1 + \nu) \to \mathbb{R}$ at each time $t$ depends on an unknown collection of $k \ll d$ linear parameters implying $r_t(x) = g_t(Ax)$ where $A \in \mathbb{R}^{k \times d}$ is full rank. This model can be seen as a generalization of [10] where the reward functions were modeled as $r(x_1, \ldots, x_d) = g(x_{i_1}, \ldots, x_{i_k})$. Thus in the special case where each row of $A$ has a single 1 and 0’s otherwise, we arrive at the setting of [10]. There has also been significant effort in other fields to develop tractable algorithms for approximating $d$ variate functions (with $d$ large) from point queries by assuming the functions to intrinsically depend on a few variables or parameters (cf. [13, 14, 15, 16] and references within). In particular the authors in [17, 18] considered the problem of approximating functions of the form $f(x) = g(Ax)$ from point queries.

Very recently and independently a work parallel to ours [19] considered the same bandit problem as ours i.e. they also assume the $d$-variate reward functions to depend on $k \ll d$ unknown linear parameters. Although they consider the mean reward function to reside in a RKHS (Reproducible Kernel Hilbert space) and adopt a Bayesian optimization framework, the scheme they employ is similar to ours. We comment on their results in more detail in the concluding remarks section towards the end.

Other related Work. The continuum armed bandit problem was first introduced in [20] for the case $d = 1$ where an algorithm achieving a regret bound of $o(n^{(2\alpha+1)/(3\alpha+1)+\eta})$ for any $\eta > 0$ was proposed for local Hölder continuous mean reward functions with exponent $\alpha \in (0,1)$. In [2] a lower bound of $\Omega(n^{1/2})$ was proven for this problem. This was then improved upon in [7] where the author derived upper and lower bounds of $O(n^{\frac{\alpha+1}{2\alpha+1}} (\log n)^{\frac{\alpha}{2\alpha+1}})$ and $\Omega(n^{\frac{\alpha+1}{2\alpha+1}})$ respectively. In [21] the author considered a class of mean reward functions defined over a compact convex subset of $\mathbb{R}^d$ which have (i) a unique maximum $x^*$, (ii) are three times continuously differentiable and (iii) whose gradients are well behaved near $x^*$. It was shown that a modified version of the Kiefer-Wolfowitz algorithm achieves a regret bound of $O(n^{1/2})$ which is also optimal. In [22] the $d = 1$ case was treated, with the mean reward function assumed to only satisfy a local Hölder condition around the
maxima $x^*$ with exponent $\alpha \in (0, \infty)$. Under these assumptions the authors considered a modification of Kleinberg’s CAB1 algorithm [7] and achieved a regret bound of $O(n^{1+\alpha/2} (\log n)^{1+\alpha/2})$ for some known $0 < \beta < 1$. In [23, 24] the authors studied a very general setting for the multi-armed bandit problem in which $S$ forms a metric space, with the reward function assumed to satisfy a Lipschitz condition with respect to this metric.

**Our Contributions.** Our main contribution is to derive an algorithm namely CAB-LP$(d,k)$ which achieves an upper bound of $O(C(k,d)n^{1+\beta} (\log n)^{\beta/2})$ on the regret after $n$ rounds. Here $C(k,d)$ depends at most polynomially on the number of linear parameters $k$ and the dimension $d$ and $n$ denotes the sampling budget which we assume to be known to the algorithm. The main idea of our algorithm is to first use a fraction of the budget for estimating the unknown $k$-dimensional subspace spanned by the rows of the linear parameter matrix $A$. After obtaining this estimate we then employ the CAB1 algorithm [7] which is restricted to play strategies only from the estimated subspace. To derive sub-linear regret bounds we show that a careful allocation of the sampling budget is necessary between the two phases.

**Organization of the paper.** The rest of the paper is organized as follows. In Section 2 we state the problem formally. Next we explain the main intuition behind our approach along with our main results in Section 3. In Section 4 we provide a formal analysis of our approach and derive regret bounds. Finally we provide concluding remarks in Section 5.

## 2. Problem Setup

We assume that a set of strategies $S$ is available to the player. For our purposes $S$ is considered to be the $\ell_2$-ball of radius $1 + \nu$ for some $\nu > 0$, denoted as $B_d(1 + \nu)$. At each time $t = 1, \ldots, n$ the environment chooses a reward function $r_t : B_d(1 + \nu) \to \mathbb{R}$. Upon playing the strategy $x_t$, the player receives the reward $r_t(x_t)$. Here the number of rounds $n$ (sampling budget) is assumed to be known to the player. We consider the setting where each $r_t$ depends on $k \ll d$ unknown linear parameters $a_1, \ldots, a_k \in \mathbb{R}^d$ with $k$ assumed to be known to the player. In particular, denoting $A = [a_1 \ldots a_k]^T \in \mathbb{R}^{k \times d}$ we assume that $r_t(x) = g_t(Ax)$.

The reward functions $g_t$ are considered to be samples from some fixed but unknown probability distribution over functions $g : B_k(1 + \nu) \to \mathbb{R}$. We then have the expected reward function as $\bar{g}(u) = \mathbb{E}[g(u)]$ where $u \in B_k(1 + \nu)$. More specifically we consider the model:

$$r_t(x) = \bar{g}(Ax) + \eta_t; \quad t = 1, 2, \ldots, n$$

where $(\eta_t)_{t=1}^n$ is i.i.d Gaussian noise with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \sigma^2$. We assume $\bar{g}$ to be sufficiently smooth — in particular to be two times continuously differentiable. Specifically, we assume for some constant $C_2 > 0$ that

$$\sup_{|\beta| \leq 2} \| D^\beta \bar{g} \|_{\infty} \leq C_2; \quad D^\beta \bar{g} = \frac{\partial |\beta| \bar{g}}{\partial y_1^{\beta_1} \ldots \partial y_k^{\beta_k}}, \quad |\beta| = \beta_1 + \cdots + \beta_k. \quad (2.1)$$

Note that the individual samples $g_t$ need not necessarily be smooth. An important assumption that we make on the mean reward function is on the conditioning of the following matrix:

$$H^r := \int_{\mathbb{R}^{d-1}} \nabla \bar{r}(x) \nabla \bar{r}(x)^T dx = A^T \int_{\mathbb{R}^{d-1}} \nabla \bar{g}(Ax) \nabla \bar{g}(Ax)^T dx \cdot A, \quad (2.2)$$

where the second equality follows from the identity $\nabla \bar{r}(x) = A^T \nabla \bar{g}(Ax)$. We assume that:

$$\sigma_1(H^r) \geq \sigma_2(H^r) \geq \cdots \geq \sigma_k(H^r) \geq \alpha > 0 \quad (2.3)$$

where $\sigma_i(H^r)$ denotes the $i^{th}$ singular value of $H^r$. In particular the parameter $\alpha$ determines the tractability of our algorithm. This is explained in detail in Section 4.4. We also note that conditions of the form [23, 24] are actually necessary to formulate a tractable algorithm, as explained in [17].
For example when $k = 1$, if we only make smoothness assumptions on $\bar{g}$, then one can construct $\bar{g}$ so that $\Omega(2^d)$ many samples are needed to distinguish between $\bar{r}(x) \equiv 0$ and $\bar{r}(x) \equiv \bar{g}(a^T x)$.

Lastly we assume without loss of generality $A$ to be row orthonormal so that $A A^T = I$. Indeed if this is not the case then through SVD (singular value decomposition) of $A$ we obtain $A = U \sum \Sigma V^T$ where $U, \Sigma, V^T$ are unitary, diagonal and row orthonormal matrices respectively.

Therefore we obtain

$$\bar{r}(x) = \bar{g}(A x) = \bar{g}(U \Sigma V^T x) = \bar{g}(V^T x)$$

where $\bar{g}'(y) = \bar{g}(U \Sigma y)$ for $y \in B_k(1 + \nu)$. Hence within a scaling of the parameter $C_2$ by a factor depending polynomially on $k, \sigma_1(A)$ we can assume $A$ to be row-orthonormal.

**Regret after $n$ rounds.** After $n$ rounds of play the cumulative expected regret is defined as:

$$R(n) = \sum_{i=1}^{n} \mathbb{E}[r_t(x^*) - r_t(x_t)] = \sum_{i=1}^{n} [\bar{g}(A^* x) - \bar{g}(A x_t)], \quad (2.4)$$

where $x^*$ is the optimal strategy belonging to the set

$$\arg\max_{x \in B_d(1+\nu)} \mathbb{E}[r_t(x)] = \arg\max_{x \in B_d(1+\nu)} \bar{g}(A x) \quad (2.5)$$

Here $x_1, x_2, \ldots, x_n$ is the sequence of strategies played by the algorithm. The goal of the algorithm is to minimize regret i.e. ensure $R(n) = o(n)$ so that $\lim_{n \to \infty} R(n)/n = 0$.

### 3. Main idea and Results

The main idea behind our algorithm is to proceed in two phases namely : (i) **PHASE 1** where we use a fraction of the sampling budget $n$ to recover an estimate of the ($k$ dimensional) subspace spanned by the rows of $A$ and then (ii) **PHASE 2** where we employ a standard continuum armed bandit algorithm that plays strategies from the previously estimated $k$ dimensional subspace.

Intuitively we can imagine that the closer the estimated subspace is to the original one, the closer will the regret bound achieved by the CAB algorithm be to the one it would have achieved by playing strategies from the unknown $k$-dimensional subspace. However one should be careful here since spending too many samples from the budget $n$ on **PHASE 1** can lead to regret which is $\Theta(n)$. On the other hand if the recovered subspace is a bad estimate then it can again lead to $\Theta(n)$ regret since the optimization carried out in **PHASE 2** would be rendered meaningless.

Hence it is important to carefully divide the sampling budget between the two phases in order to guarantee a regret bound that is sub-linear in $n$. We now describe these two phases in more detail and outline the above idea formally.

(1) **PHASE 1** (Subspace recovery phase.) In this phase we use the first $n_1 < n$ samples from our budget to generate an estimate $\hat{A} \in \mathbb{R}^{k \times d}$ of $A$ such that the row space of $\hat{A}$ is close to that of $A$. In particular we measure this closeness in terms of the Frobenius norm implying that we would like $\| A^T A - \hat{A}^T \hat{A} \|_F$ to be sufficiently small. Denoting the total regret in this phase by $R_1$ we then have that:

$$R_1 = \sum_{t=1}^{n_1} [\bar{r}(x^*) - \bar{r}(x_t)] = O(n_1). \quad (3.1)$$

We can see that $n_1$ should necessarily be $o(n)$ otherwise the total regret would be dominated by $R_1$ leading to linear regret. Furthermore $x_t \in B_d(1+\nu)$ denotes the strategy played at time $t$. 
(2) **PHASE 2** (Optimization phase.) Say that we have in hand an estimate \( \hat{A} \) from **PHASE 1**. We now employ a standard CAB algorithm that is restricted to play strategies from the row space of \( \hat{A} \). Let us denote \( n_2 = n - n_1 \) to be the duration of this phase and \( \mathcal{P} \subset B_d(1 + \nu) \) where
\[
\mathcal{P} := \left\{ \hat{A}^T y \in \mathbb{R}^d : y \in B_k(1 + \nu) \right\}.
\]
The CAB algorithm will play strategies only from \( \mathcal{P} \) and therefore will strive to optimize against the optimal strategy \( x^{**} = \hat{A}^T y^{**} \in \mathcal{P} \) where
\[
y^{**} = \arg\max_{y \in B_k(1 + \nu)} \bar{g}(A\hat{A}^T y).
\]
Furthermore we also observe that the total regret incurred in this phase can be written as:
\[
\sum_{t=n_1+1}^{n} [\bar{r}(x^*) - \bar{r}(x_t)] = \sum_{t=n_1+1}^{n} [\bar{r}(x^*) - \bar{r}(x_t)] + \sum_{t=n_1+1}^{n} [\bar{r}(x^{**}) - \bar{r}(x_t)].
\]
Note that \( R_2 \) represents the expected regret incurred by the CAB algorithm against the optimal strategy from \( \mathcal{P} \). In particular, we will obtain \( R_2 = o(n - n_1) \).

Next, the term \( R_3 \) captures the offset between the actual optimal strategy \( x^* \subset B_d(1 + \nu) \) and \( x^{**} \subset \mathcal{P} \). In particular \( R_3 \) can be bounded by making use of: (i) the Lipschitz continuity of the mean reward \( \bar{g} \) and, (ii) the bound on the subspace estimation error : \( \| A^T A - \hat{A}^T \hat{A} \|_F \). This is shown precisely in the form of the following Lemma, the proof of which is presented in the appendix.

**Lemma 1.** We have that \( R_3 \leq \frac{n_2 C_2 \sqrt{k(1 + \nu)}}{\sqrt{2}} \| A^T A - \hat{A}^T \hat{A} \|_F \) where \( n_2 = n - n_1 \) and \( C_2 > 0 \) is the constant defined in (2.1).

**Main results.** Our main result is to derive a randomized algorithm namely CAB-LP \((d, k)\) which achieves a regret bound of \( O(C(k, d)n^{\frac{1}{1+k}}(\log n)^{\frac{1}{2+k}}) \) after \( n \) rounds with \( C(k, d) \) being at most polynomial in the number of parameters \( k \) and the dimension \( d \). We state this formally in the form of the following theorem below.

**Theorem 1.** Under the notations and assumptions mentioned we have that algorithm CAB-LP \((d, k)\) achieves a total regret of
\[
R_1 + R_2 + R_3 = O \left( \frac{k^{13}d^2\sigma^2(\log k)^4}{\alpha^4} (\max\{d, \alpha^{-1}\})^2 \left( \frac{n}{\log n} \right)^{\frac{3}{1+k}} + n^{\frac{1}{1+k}}(\log n)^{\frac{1}{2+k}} \right)
\]
after \( n \) rounds with high probability.

Note that the first term in (1) represents the regret incurred in the subspace recovery phase (**PHASE 1**) while the second term captures the regret incurred in the optimization phase (**PHASE 2**). Also note that the regret bound has a nearly optimal dependence on \( n \). To see consider the scenario where \( A \) is known to the algorithm. In this case we would not have the subspace recovery phase implying a regret bound of \( O(n^{\frac{1}{1+k}}(\log n)^{\frac{1}{2+k}}) \). This nearly matches the lower bound of \( \Omega(n^{\frac{1}{1+k}}) \) derived in [25] for stochastic continuum armed bandits with \( k \)-variate Lipschitz continuous reward functions. In fact it seems to be possible to remove the \( \log n \) term completely by using recent results for finite-armed bandits. We discuss this in detail in Section 5. Lastly we also note the dependence of our regret bound on the parameter \( \alpha \). As explained in Section 4.4, \( \alpha \) typically decreases as \( d \to \infty \). Hence in order to obtain regret bounds that are at most polynomial in \( d \) we

---

1This theorem is stated again in Section 4 for completeness.
would like $\alpha$ to be polynomial in $d^{-1}$. To this end, Proposition 3 in Section 4.2 which was proven in [20] describes a fairly general class of functions for which $\alpha$ is $\Theta(d^{-1})$.

4. Analysis

We now provide a thorough analysis of the two phase scheme discussed in the previous section. We start by first describing a low-rank matrix recovery scheme which is used for obtaining an estimate of the unknown subspace represented by the row-space of $A$.

4.1. Analysis of sub-space recovery phase. We first observe that the Taylor expansion of $\bar{r}$ around any $x \in B_d(1 + \nu)$ along the direction $\phi \in \mathbb{R}^d$ give us:

$$\bar{r}(x + \epsilon \phi) - \bar{r}(x) = \epsilon \langle \phi, \nabla \bar{r}(x) \rangle + \frac{1}{2} \epsilon^2 \phi^T \nabla^2 \bar{r}(\xi) \phi$$

(4.1)

for any $\epsilon > 0$ and $\xi = x + \theta \epsilon \phi$ with $0 < \theta < 1$. In particular by using $\nabla \bar{r}(x) = A^T \bar{g}(Ax)$ in (4.1) we obtain:

$$\langle \phi, A^T \bar{g}(Ax) \rangle = \frac{\bar{r}(x + \epsilon \phi) - \bar{r}(x)}{\epsilon} - \frac{1}{2} \epsilon \phi^T \nabla^2 \bar{r}(\xi) \phi.$$

(4.2)

We now introduce the sampling scheme by stating the choice of $x$ and sampling direction $\phi$ in (4.2). We first construct

$$\mathcal{X} := \left\{ x_j \in S^{d-1} ; \; j = 1, \ldots, m_\mathcal{X} \right\}.$$

(4.3)

This is the set of samples at which we consider the Taylor expansion of $\bar{r}$ as in (4.1). In particular, we form $\mathcal{X}$ by sampling uniformly at random from $S^{d-1}$. Next, we construct the set of sampling directions $\Phi$ for $i = 1, \ldots, m_\Phi$, $j = 1, \ldots, m_\mathcal{X}$ and $l = 1, \ldots, d$ where:

$$\Phi := \left\{ \phi_{i,j} \in B_d(\sqrt{d/m_\Phi}) : [\phi_{i,j}]_l = \pm \frac{1}{\sqrt{m_\Phi}} \text{ with probability } 1/2 \right\}.$$

(4.4)

Note that we consider $m_\Phi$ random sampling directions for each point in $\mathcal{X}$. Hence we have that the total number of samples collected so far is

$$|\mathcal{X}| + |\Phi| = m_\mathcal{X} + m_\mathcal{X} m_\Phi = m_\mathcal{X} (m_\Phi + 1).$$

Now note that at each time $1 \leq t \leq m_\mathcal{X}(m_\Phi + 1)$ upon choosing the strategy $x_t$ we obtain the reward $r_t(x_t) = \bar{r}(x_t) + \eta_t$ where $\eta_t$ is i.i.d Gaussian noise. Therefore by first sampling at points $x_1, \ldots, x_{m_\mathcal{X}} \in \mathcal{X}$ and then sampling at $x_j + \epsilon \phi_{1,j}, \ldots, x_j + \epsilon \phi_{m_\Phi,j}$ for each $x_j$ we have from (4.2) the following for $i = 1, \ldots, m_\Phi$ and $j = 1, \ldots, m_\mathcal{X}$.

$$\langle \phi_{i,j}, A^T \bar{g}(Ax_j) \rangle = \frac{r_{m_\mathcal{X}+ij}(x_j + \epsilon \phi_{i,j}) - r_j(x_j)}{\epsilon} + \frac{\eta_j - \eta_{i,j}}{\epsilon} - \frac{1}{2} \epsilon \phi_{i,j}^T \nabla^2 \bar{r}(\xi_{i,j}) \phi_{i,j}.$$

(4.5)

We sum up (4.5) over all $j$ for each $i = 1, \ldots, m_\Phi$. This yields $m_\Phi$ equations that can be summarized in the following succinct form:

$$\Phi(X) = y + N + H.$$

(4.6)

Here $X = A^T G$ where $G := [\nabla \bar{g}(Ax_1) \mid \nabla \bar{g}(Ax_2) \mid \cdots \mid \nabla \bar{g}(Ax_{m_\mathcal{X}})]_{k \times m_\mathcal{X}}$. Note that $X \in \mathbb{R}^{d \times m_\mathcal{X}}$ has rank at most $k$. Next, $\Phi(X) := [(\Phi_1, X), \ldots, (\Phi_{m_\Phi}, X)] \in \mathbb{R}^{m_\Phi}$ where

$$\Phi_i = [\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,m_\mathcal{X}}] \in \mathbb{R}^{d \times m_\mathcal{X}}$$

(4.7)

represents the $i^{th}$ measurement matrix and $\langle \Phi_i, X \rangle = \text{Tr}(\Phi_i^T X)$ represents the $i^{th}$ measurement of $X$. The measurement vector is represented by $y = [y_1 \ldots y_{m_\Phi}] \in \mathbb{R}^{m_\Phi}$ where

$$y_i = \frac{1}{\epsilon} \sum_{j=1}^{m_\mathcal{X}} (r_{m_\mathcal{X}+ij}(x_j + \epsilon \phi_{i,j}) - r_j(x_j)).$$

(4.8)

\[\text{The above sampling scheme was considered first in [17] and later in [18] for the problem of approximating functions of the form } f(x) = g(Ax) \text{ from point queries.}\]
Lastly \( N = [N_1 \ldots N_{m_A}] \) and \( H = [H_1 \ldots H_{m_A}] \) represent the noise terms with
\[
N_i = \frac{\epsilon}{\epsilon} \sum_{j=1}^{m_X} (\eta_j - \eta_{i,j}) \quad \text{(Stochastic noise)},
\]
\[
H_i = - \frac{\epsilon}{2} \sum_{j=1}^{m_X} \phi_{i,j}^T \nabla^2 \tilde{r}(\xi_{i,j}) \phi_{i,j} \quad \text{(Noise due to non-linearity of mean reward function)}.
\]

Importantly, we observe that (4.6) represents (noisy) linear measurements of the matrix \( X \) which has rank \( k \ll d \). Hence by employing a standard solver for recovering low-rank matrices from noisy linear measurements we can hope to recover an approximation \( \hat{X} \) to the unknown matrix \( X \). Furthermore we note that information about the linear parameter matrix \( \Phi \) and \( \delta < k \) has rank at most \( k \) in the sense of \( \| \cdot \|_F \). This implies that for all matrices \( X_k \) of rank at most \( k \):
\[
(1 - \delta_k) \| X_k\|_F^2 \leq \| \Phi(X_k)\|_2^2 \leq (1 + \delta_k) \| X_k\|_F^2
\]
holds true for some isometry constant \( \delta_k \in (0,1) \). In general, any \( \Phi \) that satisfies (4.9) is said to have \( \delta_k \)-RIP. In our case since \( \Phi \) is a Bernoulli random measurement operator, it can be verified via standard covering arguments and concentration inequalities [27, 28] that \( \Phi \) satisfies \( \delta \)-RIP for \( 0 < \delta_k < \delta < 1 \) with probability at least \( 1 - 2 \exp(-m_\Phi q(\delta) + k(d + m_X + 1)u(\delta)) \) where
\[
q(\delta) = \frac{1}{144} \left( \delta^2 - \frac{\delta^3}{9} \right), \quad u(\delta) = \log \left( \frac{36\sqrt{2}}{\delta} \right).
\]

An estimate of the low-rank matrix \( X \) from the measurement vector \( y \) can be obtained through convex programming. For our purposes we consider the following nuclear norm minimization problem also known as the matrix Dantzig selector (DS) [29],
\[
\hat{X}_{DS} = \text{argmin} \| M\|_* \quad \text{s.t.} \quad \| \Phi^*(y - \Phi(M))\| \leq \lambda. \tag{4.10}
\]
Here \( \Phi^* : \mathbb{R}^{m_\Phi} \to \mathbb{R}^{d \times m_X} \) denotes the adjoint of the linear operator \( \Phi : \mathbb{R}^{d \times m_X} \to \mathbb{R}^{m_\Phi} \). Furthermore for any matrix, \( \| \cdot \|_* \) and \( \| \cdot \|_F \) denote its nuclear norm (sum of singular values) and operator norm (largest singular value) respectively. By making use of the error bound for matrix DS presented as Theorem 1 in [29] we obtain the following result on the performance of the matrix DS tuned to our problem setting. The proof is deferred to the appendix.

**Lemma 2.** Let \( \hat{X}_{DS} \in \mathbb{R}^{d \times m_X} \) denote the solution of (4.10) and let \( \hat{X}_{DS}^{(k)} \) be the best rank \( k \) approximation to \( \hat{X}_{DS} \) in the sense of \( \| \cdot \|_F \). Then for some constant \( \gamma > 2\sqrt{\log 12} \)
\[
0 < \delta_{4k} < \delta < \sqrt{2} - 1 \text{ we have that}
\]
\[
\| \hat{X}_{DS}^{(k)} - X \|_F \leq (C_0k)^{1/2} \left( \frac{C_2\epsilon d m_X k^2}{\sqrt{m_\Phi}} + \frac{8\gamma \sigma \sqrt{m_X m_\Phi m}}{\epsilon} \right) \left( 1 + \delta \right)^{1/2}
\]
with probability at least \( 1 - 2 \exp(-m_\Phi q(\delta) + 4k(d + m_X + 1)u(\delta)) - 4 \exp(-cm) \). Here \( m = \max \{ d, m_X \} \). Furthermore the constants \( C_0, c > 0 \) depend on \( \delta \) and \( \gamma \) respectively.
Approximating row-space($A$). Let’s say we have\footnote{Of course in practice we will not be able to solve (4.11) exactly, but will instead obtain a solution that can be made to come arbitrarily close to the actual solution. This difference will hence appear as an additional error term in the error bound of Lemma 2.} in hand $\hat{X}^{(k)}_{DS} \in \mathbb{R}^{d \times m_X}$ as the best rank $k$ approximation of the solution to (4.10). We can now obtain an estimate $\hat{A}$ of row-space($A$) by setting $\hat{A}^T$ to be equal to the $(d \times k)$ left singular vector matrix of $\hat{X}^{(k)}_{DS}$. The quality of this estimation as measured by $\| \hat{A}^T \hat{A} - A^T A \|_F$ was quantified in Lemma 2 of \cite{26} for the noiseless case ($\sigma = 0$). We adapt this result to our setting ($\sigma > 0$) and state it below. The proof is presented in the appendix.

Lemma 3. For a fixed $0 < \rho < 1$, $m_X \geq 1$, $m_\Phi < m_X d$ let

$$a_1 = C_2 d k^2, \quad b_1 = \frac{\sqrt{(1 - \rho)\alpha}}{C_0^{1/2}(1 + \delta)^{1/2}(\sqrt{k} + 2)}.$$

For any $0 < f < 1$ we then have for the choice

$$\epsilon \in \left( \frac{f b_1 - \sqrt{f^2 b_1^2 - 32 \gamma \sigma a_1 \sqrt{m_X m}}}{2a_1 \sqrt{m_X/m_\Phi}}, \quad \frac{f b_1 + \sqrt{f^2 b_1^2 - 32 \gamma \sigma a_1 \sqrt{m_X m}}}{2a_1 \sqrt{m_X/m_\Phi}} \right)$$

(4.11)

that $\| \hat{A}^T \hat{A} - A^T A \|_F \leq \frac{2f}{1 - \epsilon}$ holds true with probability at least

$$1 - 2 \exp(-m_\Phi q(\delta) + 4k(d + m_X + 1)u(\delta)) - 4 \exp(-cm) - k \exp\left(-\frac{m_X \rho^2}{2k C_2} \right).$$

We see in the above lemma that the step size parameter $\epsilon$ cannot be chosen to be arbitrarily small. In particular for $\epsilon$ too small the stochastic noise will become prominent while for large $\epsilon$, the noise due to higher order Taylor’s terms of the mean reward function will start to dominate.

Handling stochastic noise. A point of obvious concern in Lemma 3 is the condition required on the step size parameter $\epsilon$ in (4.11). This condition is well defined if $f b_1^2 - 32 \gamma \sigma a_1 \sqrt{m_X m} > 0$. This would not have been a problem in the noiseless case where $\sigma = 0$. A natural way to guarantee the well-posedness of (4.11) is by re-sampling and averaging the points in the sets $X$ and $\Phi$. Indeed if we consider each sampling point (both in $X$ and $\Phi$) to be re-sampled $N$ times and then average the corresponding reward values the variance of the stochastic noise will be reduced by a factor of $N$. By choosing a sufficiently large value of $N$, we can clearly ensure that $f b_1^2 - 32 \gamma \sigma a_1 \sqrt{m_X m} > 0$ holds true. This is made precise in the following proposition which also states a bound on the total regret $R_1$ suffered in this phase.

Proposition 1. Consider that each point $x$ in $X$ and $\Phi$ is re-sampled $N$ times after which the reward value at $x$ is estimated as the average of the $N$ observed reward values. It follows that if $N > C' k^6 d^2 \sigma^2 m_X m$ for some constant $C' > 0$ (depending on $\rho, C_0, \delta, C_2, \gamma$) and with $m = \max \{d, m_X\}$ then (4.11) in Lemma 3 is well defined and consequently the total regret in PHASE 1 is

$$R_1 = O(n_1) = O(N m_X (m_\Phi + 1)) = O\left( \frac{k^6 d^2 \sigma^2 m_X^2 m_\Phi m}{C_2 d k^2} \right).$$

Proof. First note that (4.11) in Lemma 3 is well defined when

$$f b_1^2 - 32 \gamma \sigma a_1 \sqrt{m_X m} > 0 \iff \sigma < \frac{f b_1^2}{32 \gamma \sqrt{m_X m} C_2 d k^2}. \quad \textnormal{\footnote{In the absence of external stochastic noise (i.e. $\sigma = 0$) we can actually take $\epsilon$ to be arbitrarily small as shown in Lemma 2 of \cite{26}.}}$$
Recall from Section 3 that the total regret incurred in this phase can be written as:

\[
\sigma < \frac{f^2b_1^2}{32\gamma \sqrt{m_X m C_2 dk^2}} = \frac{C \alpha f^2}{(\sqrt{k} + \sqrt{2})^2 \sqrt{m_X m dk^2}} \tag{4.12}
\]

where \( C = \frac{(1-\rho)}{32\gamma C_0(1+\delta)C_2} \) is a constant. Upon re-sampling \( N \) times and subsequent averaging of reward values we have that the variance \( \sigma \) changes to \( \sigma / \sqrt{N} \). Replacing \( \sigma \) with \( \sigma / \sqrt{N} \) in (4.12) we obtain the stated condition on \( N \). Lastly, we note that as a consequence of re-sampling the duration of PHASE 1 i.e. \( n_1 \) is \( Nm_X(m_\Phi + 1) \) implying the stated bound on \( R_1 \). \( \square \)

4.2. Analysis of optimization phase. We now analyze PHASE 2 i.e. the optimization phase of our scheme. This phase runs during time steps \( t = n_1 + 1, n_1 + 2, \ldots, n \) where \( n_1 = Nm_X(m_\Phi + 1) \).

Given an estimate \( \hat{A} \) of the row space of \( A \) we now consider optimizing only over points lying in the row space of \( \hat{A} \). In particular consider \( \mathcal{P} \subseteq B_d(1 + \nu) \) where

\[
\mathcal{P} := \left\{ \hat{A}^T y \in \mathbb{R}^d : y \in B_k(1 + \nu) \right\}.
\]

We employ a standard CAB algorithm that plays points only from \( \mathcal{P} \) and therefore strives to optimize against the optimal strategy \( x^{**} = \hat{A}^T y^{**} \in \mathcal{P} \) where

\[
y^{**} \in \arg\max_{y \in B_k(1 + \nu)} \tilde{g}(\hat{A}\hat{A}^T y).
\]

Recall from Section 3 that the total regret incurred in this phase can be written as:

\[
\sum_{t=n_1+1}^{n} [\tilde{r}(x^*) - \tilde{r}(x_t)] = \sum_{t=n_1+1}^{n} [\tilde{r}(x^*) - \tilde{r}(x^{**})] + \sum_{t=n_1+1}^{n} [\tilde{r}(x^{**}) - \tilde{r}(x_t)] = R_3 + R_2
\]

where \( R_2 \) is the regret incurred by the CAB algorithm and \( R_3 \) is the regret incurred on account of not playing strategies from the row space of \( A \).

Bounding \( R_2 \). In order to bound \( R_2 \) we employ the CAB1 algorithm of [7] with the UCB-1 algorithm [30] as the finite armed bandit algorithm. A straightforward generalization of Theorem 3.1 of [7] to \( k \) dimensions then yields

\[
R_2 = O(n^{\frac{1+k}{2\nu}}(\log n)^{\frac{1}{2\nu}}) = O(n^{\frac{1+k}{2\nu}}(\log n)^{\frac{1}{2\nu}}). \tag{4.13}
\]

Bounding \( R_3 \). The term \( R_3 \) can be bounded from above by a straightforward combination of Lemma 4 with Lemma 3. Hence we state this in the form of the following proposition without proof.

Proposition 2. For a fixed \( 0 < \rho < 1 \), \( m_X \geq 1 \), \( m_\Phi < m_X d \) let \( a_1 \) and \( b_1 \) be as defined in Lemma 3. Denoting \( \tau \) to be the error bound stated in Lemma 2 we have for any \( 0 < f < 1 \) that if \( \epsilon \) is chosen to satisfy (4.11) then it implies that \( R_3 \leq \frac{n m_X \sqrt{(1+\nu)\sqrt{2f}}}{1-f} \) holds true with probability at least

\[
1 - 2 \exp(-m_\Phi q(\delta) + 4k(d + m_X + 1)u(\delta)) - 4 \exp(-cm) - k \exp\left(-\frac{m_X \alpha_\rho^2}{2kC_2^2}\right).
\]
4.3. **Bounding the total regret.** Finally, we have all the results sufficient to bound the total regret. Indeed by using bounds on $R_1, R_2, R_3$ from Proposition 1 \((1.13)\) and Proposition 2 respectively we have that:

\[
R_1 + R_2 + R_3 = O\left(\frac{k^9d^2\sigma^2}{\alpha^2} \frac{m_\chi^2m_\Phi m}{f^4} + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}} + n_2 \sqrt{kf}\right).
\]  \hfill (4.14)

holds with probability at least

\[
1 - 2\exp(-m_\Phi q(\delta) + 4k(d + m_\chi + 1)u(\delta)) - 4\exp(-cm) - k \exp\left(-\frac{m_\chi \alpha \beta^2}{2kC_2}\right).
\]  \hfill (4.15)

In order to bound the overall regret we need to choose the values of: $m_\chi, m_\Phi$ and $f$ carefully. We state these choices precisely in the following theorem which is also our main theorem that provides a bound on the overall regret achieved by our scheme.

**Theorem 2.** Under the assumptions and notations used thus far we have for the choice : $f = \frac{1}{\sqrt{k}} \left(\frac{\log n}{n}\right)^{\frac{1}{k+1}}$ that the total regret achieved by our scheme is bounded as:

\[
R_1 + R_2 + R_3 = O\left(\frac{k^{13}d^2\sigma^2(\log k)^4}{\alpha^4} \left(\max\left\{d, \alpha^{-1}\right\}\right)^2 \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} m_\chi^2m_\Phi m + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right).
\]  \hfill (4.16)

with high probability.

*Proof.* We first observe that when $f = \frac{1}{\sqrt{k}} \left(\frac{\log n}{n}\right)^{\frac{1}{k+1}}$ then this results in:

\[
n_2 \sqrt{kf} = O(n^{\frac{1+k}{2}} (\log n)^{\frac{1}{2^{k+1}}}).
\]

Upon using this in \((4.14)\) we obtain:

\[
R_1 + R_2 + R_3 = O\left(\frac{k^9d^2\sigma^2}{\alpha^2} \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} m_\chi^2m_\Phi m + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right).
\]  \hfill (4.17)

In order to choose $m_\chi$ and $m_\Phi$ we simply note from \((4.15)\) that the choices

\[
m_\chi = c_1 k \alpha^{-1} \log k, \quad m_\Phi = c_2 k(d + m_\chi)
\]  \hfill (4.18)

for suitable constants $c_1, c_2 > 0$ ensure that the regret bound holds with high probability. Then plugging the above choice of $m_\chi$ and $m_\Phi$ in \((4.17)\) and noting that $m = \max\{d, m_\chi\} \leq (d + m_\chi)$ we obtain:

\[
R_1 + R_2 + R_3 = O\left(\frac{k^9d^2\sigma^2}{\alpha^2} \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} m_\chi^2(d + m_\chi)^2 + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right)
\]

\[
= O\left(\frac{k^{11}d^2\sigma^2(\log k)^4}{\alpha^4} \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} (d + k \alpha^{-1} \log k)^2 (d + k \alpha^{-1} \log k)^2 + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right)
\]

\[
= O\left(\frac{k^{13}d^2(\log k)^4\sigma^2}{\alpha^4} (d + k \alpha^{-1} \log k)^2 \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right)
\]

\[
= O\left(\frac{k^{13}d^2(\log k)^4\sigma^2}{\alpha^4} \left(\max\left\{d, \alpha^{-1}\right\}\right)^2 \left(\frac{n}{\log n}\right)^{\frac{4}{k+2}} + \frac{1+k}{2} (\log n) \frac{1}{2^{k+1}}\right).
\]

\[\square\]
Our complete scheme which we name CAB-LP\((d,k)\) (Continuum armed bandit of \(k\) linear parameters in \(d\) dimensions) is presented as Algorithm 1. Here \(p_1,p_2\in(0,1)\) denote the success probability parameters for the subspace recovery phase and the stated choice of \(m_X\) and \(m_\Phi\) can be verified easily by examining (4.15).

**Algorithm 1** Algorithm CAB-LP\((d,k)\)

**Input:** \(k,d,n,C_2,\sigma\).
Choose \(0<\delta<\sqrt{2}-1\) and \(\rho, p_1, p_2 \in (0,1)\). Choose \(\alpha\) according to model assumption on mean reward function.
Set \(f = \frac{1}{\sqrt{k}}\left(\frac{\log n}{n}\right)^{1/2} \), \(m_X = \frac{2kC_2^2}{\alpha \rho^2} \log(k/p_2)\) and \(m_\Phi = \frac{4k(d+m_X+1)n(\delta) + \log(2/p_1)}{\alpha \rho^2}\).
Choose re-sampling factor \(N\) according to Proposition 4.2
Choose step size \(\epsilon\) as in (4.11) with \(\sigma \leftarrow \sigma/\sqrt{N}\).

**PHASE 1 (Subspace recovery phase)** \(t = 1,\ldots,Nm_X(m_\Phi + 1)\)
- Create random sampling sets \(X\) and \(\Phi\) as explained in Section 4.1 so that \(|X| = m_X\) and \(|\Phi| = m_\Phi\).
- For \(t = 1,\ldots,m_X(m_\Phi + 1)\) collect rewards \((r_j(x_j))_{j=1}^{m_X}\) and \((r_{m_X+i,j}(x_j + \epsilon\phi_{i,j}))_{j=1,i=1}^{m_X,m_\Phi}\).
- Re-sample and average the reward values \(N\) times at each \(x\) and \(x + \epsilon\phi\) respectively \((x \in X, \phi \in \Phi)\). Form measurement vector \(y\) as in (4.3) with the averaged reward values.
- Obtain \(\hat{X}_{DS}^{(k)}\) as best rank-\(k\) approximation to solution of matrix DS (4.10) and set \(\hat{A}^T\) to left singular vector matrix of \(\hat{X}_{DS}^{(k)}\).

**PHASE 2 (Optimization phase)** \(t = Nm_X(m_\Phi + 1) + 1,\ldots,n\)

- Employ CAB1 algorithm [7] on \(P := \{\hat{A}^Ty \in \mathbb{R}^d : y \in B_k(1+\nu)\}\).

4.4. **Remarks on the tractability parameter \(\alpha\).** We now proceed to comment on the parameter \(\alpha\) of our scheme which also appears in our regret bounds. Recall from Section 2 that \(\alpha\) measures the conditioning of the following matrix:

\[
H^r := \int_{S^{d-1}} \nabla \bar{r}(x)\nabla \bar{r}(x)^T dx = A^T \cdot \int_{S^{d-1}} \nabla g(Ax)\nabla g(Ax)^T dx \cdot A.
\]  

(4.19)

More specifically, we assume that the mean reward function \(\bar{r}\) is such that:

\[
\sigma_1(H^r) \geq \sigma_2(H^r) \geq \cdots \geq \sigma_k(H^r) \geq \alpha > 0
\]

(4.20)

where \(\sigma_i(H^r)\) denotes the \(i^{th}\) singular value of \(H^r\). In other words \(\alpha\) measures how far away from 0 the lowest singular value of \(H^r\) is, implying that a larger \(\alpha\) indicates a well conditioned \(H^r\). A natural question that arises now is on the behaviour of \(\alpha\) - in particular on its dependence on dimension \(d\) and number of linear parameters \(k\). To this end we first note that the parameter typically decays with increase in \(d\). In fact for \(k > 1\) this would always be the case since as \(d \to \infty\) the matrix \(H^r\) would converge to a rank-1 matrix (see [26] for details).

We also note from our derived regret bounds that in case \(\alpha \to 0\) exponentially fast as \(d \to \infty\) then our regret bounds will have a factor exponential in \(d\) which is clearly undesirable. Hence it is important to define classes of functions for which \(\alpha\) provably decays polynomially as \(d \to \infty\) so that our regret bounds depend at most polynomially on dimension \(d\). We now state the following result from [26] which defines such a class of functions for which \(\alpha = \Theta(d^{-1})\).

**Proposition 3** (Proposition 1 in [26]). Assume that \(g : B_k(1) \to \mathbb{R}\), with \(g\) being a \(C^2\) function, has Lipschitz continuous second order partial derivatives in an open neighborhood of the origin,
$U_\theta = B_k(\theta)$ for some fixed $0 < \theta < 1$, 
\[
\frac{[\partial^2 g(y_1) - \partial^2 g(y_2)]}{\|y_1 - y_2\|} < L_{i,j} \quad \forall y_1, y_2 \in U_\theta, y_1 \neq y_2, \ i, j = 1, \ldots, k.
\]

Denote $L = \max_{1 \leq i, j \leq k} L_{i,j}$. Also under the notation defined earlier, assume that $\nabla^2 g(0)$ is full rank. Then provided that $\frac{\partial g}{\partial y_i}(0) = 0; \ \forall i = 1, \ldots, k$ we have $\alpha = \Theta(1/d)$ as $d \to \infty$.

The class of functions defined in the above Proposition covers a number of function models such as sparse additive models of the form $\sum_{i=1}^k g_i(y)$ where $g_i$'s are kernel functions \[31\]. We refer the reader to Section 5 of \[26\] for further details in this regard. Finally, in light of the above discussion on $\alpha$ we arrive at the following Corollary of Theorem 2 with the help of Proposition 3.

**Corollary 1.** Assuming that the mean reward function $\bar{r}: B_d(1 + \nu) \to \mathbb{R}$ where $\bar{r}(x) = \bar{g}(Ax)$ is such that $\bar{g}$ satisfies the conditions of Proposition 3 we have that the regret bound achieved by Algorithm CAB-LP(d,k) is given by:

\[
R_1 + R_2 + R_3 = O\left(k^{13} d^6 \sigma^2 (\log k)^4 \left(\frac{n}{\log n}\right)^{\frac{1}{2}} + n^{\frac{1+k}{2+k}} (\log n)^{\frac{1}{2+k}}\right).
\]

### 5. Concluding Remarks

To summarize, we considered a stochastic continuum armed bandit problem where the reward functions reside in a high dimensional space of dimension $d$ but intrinsically depend on $k$-linear combinations of the $d$ coordinate variables. Assuming the time horizon $n$ to be known we derived a randomized algorithm that achieves a cumulative regret bound of $O(C(k, d)n^{\frac{1+k}{2+k}} (\log n)^{\frac{1}{k+2}})$ with high probability where $C(k, d)$ is at most polynomial in $k, d$. Our algorithm combines results from low rank matrix recovery literature with existing results on continuum armed bandits.

We noted earlier that recently and independently, \[19\] consider the same problem as in this paper with the difference that the mean reward functions are assumed to reside in a RKHS (Reproducing Kernel Hilbert Space). They consider the Bayesian optimization framework and present an algorithm which has the same idea as ours in the sense of first estimating the unknown subspace spanned by the linear parameters and then performing Bayesian optimization on the estimated subspace. Furthermore their algorithm also achieves this by careful allocation of the sampling budget amongst the two phases.

**Improved regret bounds.** We now mention that the regret bounds derived in this paper can possibly be sharpened by employing recent results from finite armed bandit literature. For instance, if the range of the reward functions was restricted to be $[0, 1]$ then one can simply use the INF algorithm \[32\] as a sub-routine in the CAB1 algorithm \[7\] to get rid of the $\log n$ factor and obtain a regret bound of $O(C(k, d)n^{\frac{1+k}{2+k}})$. When the range of the reward functions is $\mathbb{R}$, as is the case in our setting, it seems possible to consider a variant of the MOSS algorithm \[32\] along with proof techniques considered in a modified UCB-1 algorithm in Section 2 of \[33\] to remove the $\log n$ factor from the regret bound.

**Future work.** For future work it would be interesting to consider the setting where the time horizon $n$ is unknown to the algorithm and to prove regret bounds for the same. In particular, it would be interesting to derive algorithms which do not involve recovering an approximation of the unknown $k$-dimensional subspace spanned by the $k$ linear parameters. Lastly we mention other directions such as an adversarial version of our problem where the reward functions are chosen arbitrarily by an adversary and also a setting where the unknown matrix $A$ is allowed to change across time.
References

[1] A. Blum, V. Kumar, A. Rudra, and F. Wu. Online learning in online auctions. In Proceedings of 14th Symp. on Discrete Alg., pages 202–204, 2003.
[2] R. Kleinberg and T. Leighton. The value of knowing a demand curve: bounds on regret for online posted-price auctions. In Proceedings of Foundations of Computer Science, 2003., pages 594–605, 2003.
[3] N. Bansal, A. Blum, S. Chawla, and A. Meyerson. Online oblivious routing. In Proceedings of ACM Symposium in Parallelism in Algorithms and Architectures, pages 44–49, 2003.
[4] B. McElman and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In Proceedings of the 17th Annual Conference on Learning Theory (COLT), pages 109–123, 2004.
[5] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Proceedings of the 21st Annual Conference on Learning Theory (COLT), 2008.
[6] A.D. Flaxman, A.T. Kalai, and H.B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms, pages 385–394, 2005.
[7] R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In 18th Advances in Neural Information Processing Systems, 2004.
[8] A. Carpentier and R. Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In Proceedings of AllStats, pages 190–198, 2012.
[9] Y. Abbasi-yadkori, D. Pal, and C. Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In Proceedings of AllStats, 2012.
[10] H. Tyagi and B. Görtler. Continuum armed bandit problem of few variables in high dimensions. CoRR, abs/1304.5793, 2013.
[11] B. Chen, R. Castro, and A. Krause. Joint optimization and variable selection of high-dimensional gaussian processes. InProc. International Conference on Machine Learning (ICML), 2012.
[12] Z. Wang, M. Zoghi, F. Hutter, D. Matheson, and N. de Freitas. Bayesian optimization in high dimensions via random embeddings. In Proc. IJCAI, 2013.
[13] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. Constr. Approx., 33:125–143, 2011.
[14] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput., 15:1373–1396, 2003.
[15] R. Coifman and M. Maggioni. Diffusion wavelets. Appl. Comput. Harmon. Anal., 21:53–94, 2006.
[16] E. Greenshtein. Best subset selection, persistence in high dimensional statistical learning and optimization under \ell_1 constraint. Ann. Stat., 34:2367–2386, 2006.
[17] M. Fornasier, K. Schnass, and J Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. Found. Comput. Math., 12(2):229–262, 2012.
[18] H. Tyagi and V. Cevher. Active learning of multi-index function models. In Advances in Neural Information Processing Systems 25, pages 1475–1483, 2012.
[19] Josip Djolonga, Andreas Krause, and Volkan Cevher. High dimensional gaussian process bandits. In To appear in Neural Information Processing Systems (NIPS), 2013.
[20] R. Agrawal. The continuum-armed bandit problem. SIAM J. Control and Optimization, 33:1926–1951, 1995.
[21] E.W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. Automatic Control, IEEE Transactions on, 54:1243–1253, 2009.
[22] P. Auer, R. Ortner, and C. Szepesvari. Improved rates for the stochastic continuum-armed bandit problem. In Proceedings of 20th Conference on Learning Theory (COLT), pages 454–468, 2007.
[23] R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In Proceedings of the 40th annual ACM symposium on Theory of computing, STOC ’08, pages 681–690, 2008.
[24] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. X-armed bandits. Journal of Machine Learning Research (JMLR), 12:1587–1627, 2011.
[25] S. Bubeck, G. Stoltz, and J.Y. Yu. Lipschitz bandits without the Lipschitz constant. In Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT), pages 144–158, 2011.
[26] H. Tyagi and V. Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. preprint, 2012.
[27] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM REVIEW, 52:471–501, 2010.
[28] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. The Annals of Statistics, 28(5):1392–1338, 2000.
[29] E.J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. CoRR, abs/1001.0339, 2010.
Appendix A. Proofs

A.1. Proof of Lemma 1

Proof. We can bound $R_3$ from above as follows.

$$R_3 = \sum_{t=n_1+1}^{n} [\bar{r}(x^*) - \bar{r}(x^{**})]$$ (A.1)

$$= n_2[g(\hat{A}x^*) - \bar{g}(\hat{A}\hat{A}^T\hat{A}x^{**})]$$ (A.2)

$$\leq n_2[g(\hat{A}x^*) - \bar{g}(\hat{A}\hat{A}^T\hat{A}x^*)]$$ (A.3)

$$\leq n_2C_2\sqrt{k} \| \hat{A}x^* - \hat{A}\hat{A}^T\hat{A}x^* \|$$ (A.4)

$$\leq n_2C_2\sqrt{k} (1 + \nu) \| A - \hat{A}\hat{A}^T\hat{A} \|_F$$ (A.5)

$$= n_2C_2\sqrt{k}(1 + \nu) \| A^T \hat{A} - \hat{A}^T\hat{A} \|_F.$$ (A.6)

In (A.3) we used the fact that $x^{**} = \hat{A}^T\hat{A}x^{**}$ since $x^{**} \in P$. In (A.4) we used the fact that $\bar{g}(\hat{A}\hat{A}^T\hat{A}x^{**}) \geq \bar{g}(\hat{A}\hat{A}^T\hat{A}x^*)$ since $\hat{A}^T\hat{A}x^* \in P$ and $x^{**} \in P$ is an optimal strategy. (A.5) follows from the mean value theorem along with the smoothness assumption made in (2.1). In (A.6) we used the simple inequality: $\| Bx \| \leq \| B \|_F \| x \|$. Obtaining (A.7) from (A.6) is a straightforward exercise. □

A.2. Proof of Lemma 2

Proof. We first have the following result by simply using Theorem 1 in [29] in our setting for bounding the error of the matrix Dantzig selector.

**Theorem 3.** For any $X \in \mathbb{R}^{d \times m_X}$ such that rank($X$) $\leq k$ let $\tilde{X}_{DS}$ be the solution of (4.10). If $\delta_{4k} < \delta < \sqrt{2} - 1$ and $\| \Phi^*(H + N) \| \leq \lambda$ then we have with probability at least $1 - 2e^{-m_\Phi(\delta) + 4k(d + m_X + 1)\mu(\delta)}$ that

$$\| X - \tilde{X}_{DS} \|_F^2 \leq C_0k\lambda^2$$

where $C_0$ depends only on the isometry constant $\delta_{4k}$.

What remains to be found for our purposes is $\lambda$ which is a bound on $\| \Phi^*(H + N) \|$. Firstly note that $\| \Phi^*(H + N) \| \leq \| \Phi^*(H) \| + \| \Phi^*(N) \|$. From Lemma 1 and Corollary 1 of [26] we have that:

$$\| \Phi^*(H) \| \leq \frac{C_2d_mXk^2}{2\sqrt{m_\Phi}}(1 + \delta)^{1/2}$$
holds with probability at least $1 - 2e^{-cm}$ where $c = \frac{\gamma^2}{2} - 2\log 12$ and $\gamma > 2\sqrt{\log 12}$. This can be verified using the proof technique of Lemma 1.1 of [29] by taking care of the fact that the entries of $L_1$ are correlated as they are identical copies of the same Gaussian random variable $\frac{1}{\epsilon} \sum_{j=1}^{mX} \eta_j$. Furthermore we also have that:

$$
\| \Phi^*(L_2) \| \leq \frac{2\gamma\sigma}{\epsilon} \sqrt{(1 + \delta)mXm}
$$

(A.9)

holds with probability at least $1 - 2e^{-cm}$ with constants $c, \gamma$ as defined earlier. This is verifiable easily using the proof technique Lemma 1.1 of [29] as the entries of $L_2$ are i.i.d Gaussian random variables. Combining (A.8) and (A.9) we then have that the following holds true with probability at least $1 - 4e^{-cm}$.

$$
\| \Phi^*(L_1) \| + \| \Phi^*(L_2) \| \leq \frac{4\gamma\sigma}{\epsilon} \sqrt{(1 + \delta)mXm}.\tag{A.10}
$$

Lastly, it is fairly easy to see that $\| \hat{X}^{(k)}_{DS} - X \|_F \leq 2 \| \hat{X}_{DS} - X \|_F$ where $\hat{X}^{(k)}_{DS}$ is the best rank $k$ approximation to $\hat{X}_{DS}$ (see for example the proof of Corollary 1 in [24]). Combining the above observations we arrive at the stated error bound with probability at least $1 - 2e^{-m\Phi(q)\delta + 4k(d + mX + 1)u(\delta)} - 4e^{-cm} - k \exp\left(- \frac{mX\alpha\rho^2}{2kC_2^2} \right)$.

A.3. Proof of Lemma 3.

Proof. Let $\tau$ denote the bound on $\| \hat{X}^{(k)}_{DS} - X \|_F$ as stated in Lemma 2 of [26] which gives us that if $\tau < \sqrt{(1 - \rho)mX\alpha k} \sqrt{k + \sqrt{2}}$ holds then it implies that:

$$
\| \tilde{A}^T\tilde{A} - A^TA \|_F \leq \frac{2\tau}{\sqrt{(1 - \rho)mX\alpha} - \tau}
$$

(A.11)

holds true for any $0 < \rho < 1$ with probability at least

$$
1 - 2\exp(-m\Phi(q)\delta + 4k(d + mX + 1)u(\delta)) - 4\exp(-cm) - k \exp\left(- \frac{mX\alpha\rho^2}{2kC_2^2} \right).
$$
The proof makes use of Weyl’s inequality \cite{34} and Wedin’s perturbation bound \cite{35}. Therefore upon using the value of $\tau$ we have that $\tau < f \frac{\sqrt{(1-\rho)m\alpha k}}{\sqrt{k+\sqrt{2}}}$ holds for any $0 < f < 1$ if:

$$C_0^{1/2} k^{1/2} (1 + \delta)^{1/2} \left( \frac{C_2 \epsilon d m_\chi k^2}{\sqrt{m_\Phi}} + \frac{8 \gamma \sigma \sqrt{m_\chi m_\phi m}}{\epsilon} \right) < f \frac{\sqrt{(1-\rho)m_\chi \alpha k}}{\sqrt{k+\sqrt{2}}}$$ \hspace{1cm} (A.12)

$$\Leftrightarrow C_2 d k^2 \epsilon \sqrt{\frac{m_\chi}{m_\Phi}} + \frac{8 \gamma \sigma \sqrt{m_\phi m}}{\epsilon} < f \left( \frac{1}{C_0^{1/2} (1 + \delta)^{1/2} \sqrt{k+\sqrt{2}}} \right)$$ \hspace{1cm} (A.13)

$$\Leftrightarrow a_1 \sqrt{\frac{m_\chi}{m_\Phi}} \epsilon^2 - fb_1 \epsilon + 8 \gamma \sigma \sqrt{m_\phi m} < 0.$$ \hspace{1cm} (A.14)

From (A.14) we get the stated condition on $\epsilon$. Lastly upon using $\tau < f \frac{\sqrt{(1-\rho)m\chi k}}{\sqrt{k+\sqrt{2}}}$ in (A.11) we obtain the stated bound on $\| \hat{A}^T \hat{A} - A^T A \|_F$. \hfill \Box