**University of Alberta**

**Library Release Form**

**Name of Author**: Yasin Abbasi-Yadkori

**Title of Thesis**: Forced-Exploration Based Algorithms for Playing in Bandits with Large Action Sets

**Degree**: Master of Science

**Year this Degree Granted**: 2009



Yasin Abbasi-Yadkori
4071 - 33A Street
Edmonton, Alberta
Canada, T6T-1R4

**Date**: _____

**University of Alberta**

# Forced-Exploration Based Algorithms for Playing in Bandits with Large Action Sets

by

**Yasin Abbasi-Yadkori**

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

in

Computing Science

Department of Computing Science

Edmonton, Alberta
Spring 2009

**University of Alberta**

**Faculty of Graduate Studies and Research**

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Forced-Exploration Based Algorithms for Playing in Bandits with Large Action Sets** submitted by Yasin Abbasi-Yadkori in partial fulfillment of the requirements for the degree of **Master of Science** in Computing Science.

_____

Csaba Szepesvári

_____

Dale Schuurmans

_____

Edit Gombay

**Date**: _____

# Abstract

We study the bandit problem when the payoff function (a) is linear in actions or (b) satisfies some smoothness conditions. For each structure, we upper bound the regret of the Forced Exploration algorithm. The main objective of this thesis is to show what we can achieve by using the Forced Exploration algorithm.

For the linear case, we propose an algorithm, called FEL, and prove that its regret at time $T$ is upper bounded by $d\sqrt{T} + \tilde{O}\left(\frac{\sqrt{T}}{d\nu_d^2}\right)$, where $d$ is the number of features and $\nu_d$ is the smallest eigenvalue of the dispersion matrix. Our experiments support our upper bound and show that FEL outperforms some alternative algorithms when there are correlations between the payoffs of the actions. When there are no such correlations, the UCT Algorithm of Kocsis and Szepesvari (2006) outperforms FEL.

For the smooth payoff case, we propose an algorithm with a regret bounded by $O(T^{\frac{2+\alpha}{2+2\alpha}})$, where $\alpha \in \mathbb{N}$ is the differentiability of the payoff function that must be known to the algorithm. The state of the art result for this problem is due to Auer et al. (2007) who propose an algorithm with a regret bounded by $O(T^{\frac{1+\alpha-\alpha\beta}{1+2\alpha-\alpha\beta}})$ where $\beta$ is a problem-dependent parameter. The advantage of our algorithm is that it allows a flexible combination of known payoff structures with a non-parametric approach.

# Acknowledgements

# Table of Contents

# Chapter 1

# Introduction

Sequential decision making problems such as web advertising (Pandey et al., 2007), design of clinical trials (Thompson, 1933), sequential design of experiments (Lai and Robbins, 1985), and online pricing (Kleinberg, 2005) are often formulated as bandit problems. Pick a set of actions $\mathcal{A}$. In a bandit problem at time $t$ a learner takes some action (also called an arm) $A_t \in \mathcal{A}$ and receives a reward $Y_t$ such that

$$Y_t = h(A_t) + Z_t,$$

where $h$ is an unknown function and $Z_t$ is a zero-mean noise. The objective is to minimize the regret,

$$R(T) = T \max_{a \in \mathcal{A}} h(a) - \sum_{t=1}^{T} h(A_t), \tag{1.1}$$

where $T$ is the time horizon.

The bandit problem has been studied for various action spaces. The most widely studied version of the bandit problem is the $K$-armed bandit problem where the action space is finite, $\mathcal{A} = \{1, 2, \ldots, K\}$ (Robbins, 1952). Here we are interested in problems where the action space is large, or infinite. Before describing our approach we provide a brief overview of the relevant literature. In particular, a literature review for the $K$-armed bandit problem is presented in Section 1.1, while Section 1.2 provides a literature review for the case when $\mathcal{A}$ is very large or continuous.

## 1.1 $K$-armed Bandit Problems

In this chapter, we describe the $K$-armed bandit problem and a few algorithms proposed to solve it. We use $i = \{1, \ldots, K\} = \mathcal{A}$ to denote the actions in the action space. Further, we let $h_i$ denote the mean payoff of action $i$, $Y_{i,t}$ be its random payoff at time $t$, and $h^* = \max_i h_i$ be the mean payoff of the optimal action. Let $\bar{Y}_{i,T_i(t)}$ be the average of the first $T_i(t)$ payoffs of action $i$, where $T_i(t)$ is the number of times we have played action $i$ up to time $t$.

Let $f_0 := 0$, $S_i(0) := 0$ and $T_i(0) := 0$ for $i = 1, \ldots, K$
**for** $t := 1, 2, \ldots$ **do**
  **if** $f_{t-1} < f_t^*$ **then**
    {Exploration:}
    $i_t \sim P$ {Draw a random action from $\mathcal{A}$ according to distribution $P$}
    $f_t := f_{t-1} + 1$
  **else**
    {Exploitation:}
    $i_t := \text{argmax}_i \bar{Y}_{i,T_i(t)}$
  **end if**
  Play $i_t$ and receive payoff $Y_{i_t,t}$
  **for** $i := 1, \ldots, K$ **do**
    $T_i(t) := T_i(t-1) + \mathbb{I}_{\{i=i_t\}}$
    $S_i(t) := S_i(t-1) + \mathbb{I}_{\{i=i_t\}} Y_{i_t,t}$
    $\bar{Y}_{i,T_i(t)} := S_i(t)/T_i(t)$
  **end for**
**end for**

Table 1.1: The Forced Exploration Algorithm for the $K$-armed bandit problem. Its input is the exploration schedule $(f_t^*)$, an increasing sequence of natural numbers.

Perhaps, the simplest solution to the $K$-armed bandit problem is the method of Forced Exploration, shown here in Table 1.1, where we deterministically divide time steps into exploration and exploitation steps (to our best knowledge it was Robbins (1952) who described this idea first under the name "forced sampling"). In an exploration step we draw a random action from the action space according to a distribution $P$. Here the only constraint on $P$ is that $P(i) > 0$ must hold for all the actions $i \in \mathcal{A}$. A typical choice is to use the uniform distribution in which case $P(i) = 1/K$, $i \in \mathcal{A}$. In an exploitation step we choose the action with the maximum average payoff, i.e., $\text{argmax}_i \bar{Y}_{i,T_i(t)}$. A similar algorithm is the so-called $\epsilon$-greedy method, shown here in Table 1.2, where at each time step $t$ we explore with probability $\epsilon_t$ and otherwise take the greedy action $\text{argmax}_i \bar{Y}_{i,T_i(t)}$. Auer et al. (2002) analyze the $\epsilon$-greedy Algorithm and show that it achieves an $O(\log T)$ regret when $\epsilon_t = cK/(d^2 t)$, where $c$ and $d$ are tuning parameters:

**Theorem 1.** *Pick any number of actions $K > 1$ and any reward distributions $(P_1, \ldots, P_K)$ supported in $[0, 1]$. Let $\Delta_i = h^* - h_i$. If $\epsilon$-greedy is run with input parameters $(c, d)$ such that*

$$0 < d \leq \min_{i:h_i<h^*} \Delta_i$$

*and*

$$\frac{c}{5d^2} \geq 1 \text{ and } \frac{c}{2} \geq 1,$$

*then the probability that after $t \geq cK/d$ of plays it chooses an action $j$ with $\Delta_j > 0$ is at*

**Inputs:** $c > 0$ and $0 < d < 1$.
Let $S_i(0) := 0$ and $T_i(0)$ for $i = 1, \ldots, K$
Define the sequence $\epsilon_t \in (0, 1]$, $t = 1, 2, \ldots$ by

$$\epsilon_t := \min\left\{1, \frac{cK}{d^2 t}\right\}$$

**for** $t := 1, 2, \ldots$ **do**
  $r_t \sim u_{[0,1]}$ {Draw a random number from $[0, 1]$ according to the uniform distribution}
  **if** $r_t < 1 - \epsilon_t$ **then**
    $i_t := \text{argmax}_i \bar{Y}_{i,T_i(t)}$
  **else**
    $i_t \sim P$ {Draw a random action from $\mathcal{A}$ according to distribution $P$}
  **end if**
  Play $i_t$ and receive payoff $Y_{i_t,t}$
  **for** $i := 1, \ldots, K$ **do**
    $T_i(t) := T_i(t-1) + \mathbb{I}_{\{i=i_t\}}$
    $S_i(t) := S_i(t-1) + \mathbb{I}_{\{i=i_t\}} Y_{i_t,t}$
    $\bar{Y}_{i,T_i(t)} := S_i(t)/T_i(t)$
  **end for**
**end for**

Table 1.2: The $\epsilon$-greedy Algorithm for the $K$-armed bandit problem

*most*

$$\frac{c}{d^2 t} + 2\left(\frac{c}{d^2}\log\frac{(t-1)d^2 e^{1/2}}{cK}\right)\left(\frac{cK}{(t-1)d^2 e^{1/2}}\right)^{c/(5d^2)} + \frac{4e}{d^2}\left(\frac{cK}{(t-1)d^2 e^{1/2}}\right)^{c/2}.$$

Thus the theorem essentially bounds the probability of choosing a suboptimal action (i.e. an action $j$ with $\Delta_j > 0$). Since the regret up to time $T$ can be bounded as the sum of such probabilities, which by this theorem can be bounded as $O(\sum_{t=1}^{T} 1/t)$, we get that $\epsilon$-greedy in this case achieves a regret whose expectation can be bounded by $O(\log T)$.

It seems that it is impossible to achieve an $O(\log T)$ regret using the $\epsilon$-greedy approach with no knowledge of some problem-dependent parameters. However, a proof of this claim remains for future work.

In a more sophisticated approach, Lai and Robbins (1985) assume that $Y_{i,t} \sim p_i(\cdot) = p(\cdot; \theta_i)$ for some unknown $\theta_i \in \Theta \subset \mathbb{R}^d$ and a known parametric family of density functions $p(\cdot; \cdot)$. Let $p^*$ be the probability density of the payoff of the optimal action. Lai and Robbins (1985) introduce the principle of *Optimism in the Face of Uncertainty* (OFU) to build a confidence interval around the unknown model and play optimistically with respect to this model. The expected regret of their Upper Confidence Index (UCI) algorithm is upper

Let $S_i(0) := 0$, $T_i(0) := 0$ and $\bar{Y}_{i,0} := 0$ for $i = 1, \ldots, K$
**for** $t := 1, 2, \ldots$ **do**
  $i_t := \text{argmax}_i \left( \bar{Y}_{i,T_i(t-1)} + \sqrt{\frac{2\log(t-1)}{T_i(t-1)}} \right)$
  Play $i_t$ and receive payoff $Y_{i_t,t}$
  **for** $i := 1, \ldots, K$ **do**
    $T_i(t) := T_i(t-1) + \mathbb{I}_{\{i=i_t\}}$
    $S_i(t) := S_i(t-1) + \mathbb{I}_{\{i=i_t\}} Y_{i_t,t}$
    $\bar{Y}_{i,T_i(t)} := S_i(t)/T_i(t)$
  **end for**
**end for**

Table 1.3: The UCB1 Algorithm for the $K$-armed bandit problem. By convention we let $c/0 = +\infty$. This makes the algorith select each action once at the beginning.

bounded by $(1/D(p_j||p^*) + o(1)) \log(T)$, where

$$D(p_j||p^*) = \int p_j(x) \log \frac{p_j(x)}{p^*(x)}\, dx$$

is the Kullback-Leibler divergence between $p_j$ and $p^*$. They also prove a lower bound for this problem that matches their upper bound up to sublogarithmic ($o(\log T)$) terms, showing that their upper bound is asymptotically unimprovable.

More recently, Auer et al. (2002) introduced the so-called Upper Confidence Bounds algorithm which they call UCB1 and which is shown here in Table 1.3. At time $t$, UCB1 chooses the action with index $\text{argmax}_i\{\bar{Y}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}\}$, where $c_{t,T_i(t)}$ is a so-called bonus term for action $i$. Auer et al. (2002) propose to use

$$c_{t,T_i(t)} = \sqrt{\frac{2\log t}{T_i(t)}}$$

and prove the following theorem:

**Theorem 2.** *For all $K > 1$, if policy UCB1 is run on $K$ actions having arbitrary reward distributions $P_1, \ldots, P_K$ with support in $[0, 1]$, then the expected regret of UCB1 after any number $t$ of plays is at most*

$$\left[ 8 \sum_{i:h_i<h^*} \left( \frac{\log t}{\Delta_i} \right) \right] + \left( 1 + \frac{\pi^2}{3} \right) \left( \sum_{j=1}^{K} \Delta_j \right),$$

*where $h_1, \ldots, h_K$ are the expected values of $P_1, \ldots, P_K$ and $\Delta_i = h^* - h_i$.*

UCB1 is based on the UFO principle in the sense that it constructs a confidence set around the unknown model and plays optimistically with respect to this model. UCB1 has several advantages over other algorithms: it doesn't require a knowledge of the parameters of the problem, it is very easy to implement, and its regret grows only at a logarithmic rate. Further, it is a very simple algorithm. Auer et al. (2002) also provide a finite-time

analysis for the first time. However, their analysis looses *optimality* (the multipliers of the leading term in the lower and upper bounds will not match anymore, i.e., in the above bound compared to the result of Lai and Robbins (1985), the leading term will be bigger than $1/D(p_j||p^*)$).

Another variant of UCB1, called UCB-Tuned, is also proposed by Auer et al. (2002). The variation is in the form of the bonus term. At time $t$, UCB-Tuned uses the following bonus term:

$$\sqrt{\frac{\log t}{T_j(t)}\min\{1/4,\, V_j(T_j(t),t)\}},$$

where

$$V_j(s,t) = \left(\frac{1}{s}\sum_{\tau=1}^{s} Y_{j,\tau}^2\right) - \bar{Y}_{j,s}^2 + \sqrt{\frac{2\log t}{s}}$$

is an upper confidence bound for the variance of action $j$. Auer et al. (2002) mention that UCB-Tuned performs substantially better than UCB1 in all of their experiments. Auer et al. (2002) have compared the performance of $\epsilon$-greedy with UCB-Tuned on a wide range of problems and have concluded that if the parameters of $\epsilon$-greedy were tuned appropriately then it almost always outperforms UCB-Tuned. However, they have found that the performance of $\epsilon$-greedy rapidly degrades if the parameters are not appropriately tuned or the payoffs of the suboptimal actions differed a lot from the optimal value. The nice property of UCB-Tuned is that it performs uniformly well on all problems. However, Auer et al. (2002) did not provide an analysis for the regret of UCB-Tuned.

More recently, Audibert et al. (2008) analyzed the regret of a refinement of this algorithm. Their algorithm, called UCB-V, is implemented as follows: Assume that $Y_{k,t} \in [0,1]$. Let $c > 1$ and $\epsilon = \{\epsilon_{s,t}\}_{s\geq 0, t\geq 0}$ be nonnegative real numbers such that for any fixed $s$ the function $\epsilon_{s,t}$ is nondecreasing. Further, define

$$B_{k,s,t} = \bar{Y}_{k,s} + \sqrt{\frac{2V_{k,s}\epsilon_{s,t}}{s}} + c\frac{3\epsilon_{s,t}}{s},$$

and

$$V_{k,s} = \frac{1}{s}\sum_{\tau=1}^{s}(Y_{k,\tau} - \bar{Y}_{k,s})^2.$$

At time $t$ UCB-V chooses $\mathrm{argmax}_i\, B_{i,T_i(t-1),t}$. Audibert et al. (2008) prove a logarithmic upper bound for the regret of UCB-V and experimentally show that when the variance of the sub-optimal actions are low, UCB-V has major advantage over UCB1.

## 1.2 Bandit Problems with Large Action Sets

When the action space is very large, we need to make some assumptions about its structure, or it is impossible to achieve a nontrivial (sublinear) regret bound when the number of time steps is reasonably small. In fact, an assumption on the mean reward as a function of the

actions is always necessary when the action space is infinite. Thus, in these cases it makes sense to restrict the problem in some way. For example, in an ad allocation problem, the value of each ad might be linear in some features, making the problem manageable even if the number of ads was huge.

In this section, we first consider the case when the payoff function is linear in the actions and the actions belong to a Euclidean space (this will be made precise in a moment). Then we drop the linearity assumption and consider a more general case when the payoff function satisfies some smoothness conditions only.

Remember that the payoff $Y_t$ at time $t$ is assumed to satisfy

$$Y_t = h(A_t) + Z_t,$$

where $A_t$ is an action determined by some algorithm based on the past actions and payoffs and $Z_t$ is some "noise". In particular, throughout this section we make the following assumptions:

**Assumption A1** The noise sequence $(Z_t)$ satisfies $\mathbb{E}[Z_t|A_t] = 0$, no matter how the action sequence $(A_t)$ is chosen based on the past payoffs. Further, $|Z_t| \leq 1$ holds with probability one.

**Assumption A2** The reward function is uniformly bounded and in particular $\|h\|_\infty \leq 1$.

### 1.2.1 Stochastic Linear Bandit Problems

To the best of our knowledge, the linear bandit problem was first introduced by Auer (2002) (cf. Section 1.2.2). In this problem, we assume that

$$h(a) = \theta_*^T a,$$

where $\theta_*$ is an unknown parameter vector, $\theta_*^T$ denotes the transpose of $\theta_*$, and $a \in \mathbb{R}^d$.

The state of the art result in this problem is due to Dani et al. (2008) who studied the Confidence Ellipsoid Algorithm, shown here in Table 1.4. This algorithm (in a closed form) was proposed, but not analyzed by Auer (2002). Dani et al. (2008) have shown that the regret of this algorithm at time $T$ is bounded by $\tilde{O}(d\sqrt{T})$ [1]. The algorithm implements the OFU principle: it builds a high probability confidence set, $B_t$, around the true parameter vector:

$$B_t = \{\theta \in \mathbb{R}^d \mid (\theta - \hat{\theta}_t)^T C_t (\theta - \hat{\theta}_t) \leq \beta_t\}$$

where $\beta_t = O(d \log^2 t)$, $\hat{\theta}_t = \operatorname{argmin}_\theta \sum_{s=1}^t (Y_s - \theta^T A_s)^2$ is the least squares estimate of $\theta_*$, and where $C_t = \sum_{s=1}^t A_s A_s^T$ is the correlation matrix built from the actions chosen in the

[1] We say that $a_n = \tilde{O}(b_n)$, where $a_n \geq 0$, $b_n > 0$ are two sequences if $\exists m \geq 0, C > 0$ such that $a_n \leq C b_n \log^m b_n$.

$C_1 := \sum_{i=1}^{d} b_i b_i^T$ where $\{b_1, \ldots, b_d\}$ is a barycentric spanner for $\mathcal{A}$, $\hat{\theta}_1 := 0$
**for** $t := 1, 2, \ldots$ **do**
  $\beta_{t,\delta} := \max\left(128 d \log(t) \log(t^2/\delta),\ \left(\frac{8}{3}\log\left(\frac{t^2}{\delta}\right)\right)^2\right)$
  $B_t := \{\theta : (\theta - \hat{\theta}_t)^T C_t (\theta - \hat{\theta}_t) \le \beta_{t,\delta}\}$
  $A_t := \mathrm{argmax}_{a \in \mathcal{A}} \max_{\theta \in B_t} (\theta^T a)$
  Play $A_t$ and receive payoff $Y_t$
  $C_{t+1} := C_t + A_t A_t^T$
  $\hat{\theta}_{t+1} := C_{t+1}^{-1} \sum_{s=1}^{t} Y_s A_s$
**end for**

Table 1.4: The Confidence Ellipsoid Algorithm for stochastic linear bandit problems. The algorithm has a single parameter, $0 < \delta < 1$, where $1 - \delta$ gives the desired confidence level

past. Following the OFU principle, the algorithm at time $t$ chooses the action

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in B_t} \theta^T a. \tag{1.2}$$

The regret of Confidence Ellipsoid algorithm depends on a parameter $\Delta$, which is defined in the following way: Define an extremal point of $\mathcal{A}$ as a point that is not a proper convex combination of members of $\mathcal{A}$ and let $\mathcal{S}$ be the set of extremal points of $\mathcal{A}$. Define the set of suboptimal extremal points as

$$\mathcal{S}_- = \left\{a \in \mathcal{S} : \theta_*^T a < \theta_*^T a_*\right\},$$

where $a_*$ is the optimal action:

$$a_* = \operatorname*{argmax}_{a \in \mathcal{A}} \theta_*^T a.$$

Define the "gap", $\Delta$, as

$$\Delta = \theta_*^T a_* - \sup_{a \in \mathcal{S}_-} \theta_*^T a.$$

The Confidence Ellipsoid Algorithm achieves a polylogarithmic regret when $\Delta > 0$, but generally its regret is $\tilde{O}(d\sqrt{T})$. For example, when $\mathcal{A}$ is finite or a polytope, $\Delta > 0$, while if it is a ball then $\Delta = 0$. The next two theorems are the main upper bounds proven by Dani et al. (2008):

**Theorem 3** (Problem Dependent Upper Bound)**.** *Fix $0 < \delta < 1$ and consider a run of the Confidence Ellipsoid Algorithm on a stochastic linear bandit problem, where $\mathcal{A} \subset [-1, 1]^d$ is convex, $h(a) = \theta_*^T a$, and the gap $\Delta = \Delta(\mathcal{A}, \theta_*)$ is positive. Recall (cf. Table 1.4) that*

$$\beta_T = \max\left(128 d \log(T) \log(T^2/\delta),\ \left(\frac{8}{3}\log\left(\frac{T^2}{\delta}\right)\right)^2\right).$$

*Then $\exists T_0 > 0$ such that*

$$\mathbb{P}\left(\forall T \ge T_0,\ R(T) \le \frac{8 d \beta_{T,\delta} \log T}{\Delta}\right) \ge 1 - \delta.$$

**Theorem 4** (Problem Independent Upper Bound)**.** *Consider the same setting as in Theorem 3, except that $\Delta = \Delta(\mathcal{A}, \theta_*) > 0$ is not required. Then $\exists T_0 > 0$ such that*

$$\mathbb{P}\left(\forall T \geq T_0,\, R(T) \leq \sqrt{8dT\,\beta_{T,\delta}\log T}\right) \geq 1 - \delta.$$

Dani et al. (2008) discuss the efficiency of the Confidence Ellipsoid Algorithm and point out that the optimization problem (1.2) is NP-hard and so the Confidence Ellipsoid Algorithm is not practical for large values of $d$. Dani et al. (2008) also propose another algorithm that is computationally more efficient, but they only show a regret of $\tilde{O}(d^{3/2}\sqrt{T})$ for this algorithm.

In addition to the above upper bounds, Dani et al. (2008) claim a lower bound of $\Omega(d\sqrt{T})$ [2]. This is for the case when $\sup_{a\in\mathcal{A}} \|a\|_2 = \sqrt{d}$. However, we believe that their lower bound analysis seems to have a gap: The claim that the proof technique of Lemma 15 can be used to prove Lemma 16 is not correct and Lemma 16 doesn't hold as stated.

### 1.2.2 Associative Linear Bandits

Auer (2002) proposed the so-called SupLinRel algorithm for the adversarial Associative Reinforcement Learning problem. In this problem at each time step $t$ the learner has to pick one of a finite number ($K$) of actions. However, before this choice an adversary is presenting the $d$-dimensional side information vectors $s_{1t}, \ldots, s_{Kt}$ to the learner. If the learner chooses action $i$ then he will receive a random payoff with mean $\theta_*^T s_t$, where $\theta_*$ is an unknown parameter vector. SupLinRel works by constructing a confidence ball for $\theta_*$ and uses the OFU principle to decide which action to choose, much like the Confidence Ellipsoid Algorithm. The algorithm has a single parameter that should be chosen by the user.

Auer (2002) proves the following theorem (Theorem 6 in Auer (2002)):

**Theorem 5.** *Let $\mathcal{A} = \{1, \ldots, K\}$. Assume that $\|\theta_*\|_2 \leq 1$ and $r_t \in [0, 1]$. Fix $0 < \delta < 1$ and time $T > 0$. When algorithm SupLinRel is run with parameter $\delta/(1 + \log T)$ then with probability $1 - \delta$ the regret of the algorithm up to time $T$ is bounded by*

$$R(T) \leq 44\,[1 + \log(2KT\log(T/\delta))]^{3/2}\,\sqrt{dT} + 2\sqrt{T}.$$

Note that the theorem in (Auer, 2002) has $\log T$ instead of $\log(T/\delta)$, which is a misprint. It seems possible that SupLinRel can actually be used to simulate the behavior of a Confidence Ellipsoid Algorithm (Auer, 2007). This opens up the possibility of an $\tilde{O}(\sqrt{dT})$ regret bound for the class of problems satisfying the conditions stated in the theorem. Note that here $\sup_{a\in\mathcal{A}} \|a\|_2 = 1$ and $\|\theta\|_2 \leq 1$. However, this construction is beyond the scope of this thesis.

[2] Let $(a_n)$, $(b_n)$ be two nonnegative sequences. We say that $a_n = \Omega(b_n)$ if $\exists C > 0$ such that $a_n \geq C b_n$.

**Inputs:** $T > 0$ (horizon), $0 < \zeta \leq 1$ (exponent in (1.3)).
$t := 1$
**while** $t \leq T$ **do**
  $n := \left\lceil \left( \frac{t}{\log t} \right)^{\frac{1}{2\zeta+1}} \right\rceil$
  Initialize UCB1 with strategy set $\{1/n, 2/n, \ldots, 1\}$
  **for** $s := t, t+1, \ldots, \min(2t-1, T)$ **do**
    Get action $A_s$ from UCB1
    Play $A_s$ and receive $Y_t$
    Feed $Y_t$ back to UCB1
  **end for**
  $t := 2t$
**end while**

Table 1.5: CAB1 algorithm for continuum-armed bandit problems. Note that unlike the previous algorithms, this algorithm needs to know the time horizon $T > 0$.

### 1.2.3 Continuum-armed Bandit Problems

In this section, we drop the linearity assumption of Section 1.2.1 and study the more general case when the payoff function satisfies some smoothness conditions. This problem is called the nonparametric or continuum-armed bandit problem.

Kleinberg (2004) considered the case when $\mathcal{A} = [0, 1]$. He assumes that the target function is uniformly locally Holder with constant $L$, exponent $\zeta \leq 1$, and neighborhood size $\nu > 0$ in the sense that for all $a$, $a' \in \mathcal{A}$ with $|a - a'| \leq \nu$,

$$|h^*(a) - h^*(a')| \leq L \left| a - a' \right|^{\zeta}. \tag{1.3}$$

Under this assumption, he proves a lower bound of $O(T^{\frac{\zeta+1}{2\zeta+1}})$. He also proposes an algorithm, CAB1, shown in Table 1.5, and proves the following theorem:

**Theorem 6.** *For known $\zeta$, the regret of algorithm CAB1 is $O(T^{\frac{\zeta+1}{2\zeta+1}} \log^{\frac{\zeta}{2\zeta+1}}(T))$.*

The idea behind CAB1 is simple, making this result very elegant: divide the action space into *appropriate* number of intervals and play UCB1 on these intervals. However, the uniformly locally Holderness assumption turns out to be quite restrictive: As Auer et al. (2007) point out, if $\zeta > 1$, $h^*$ must be a constant function.

Another related work is due to Cope (2006) who considered the case when the action space is a convex compact subset of $\mathbb{R}^d$. He assumes that the target function is unimodal, three times differentiable, and it satisfies the inequalities

$$C_1 \left\| a - a^* \right\|_2^2 \leq (a - a^*)^T \left[ \frac{\partial}{\partial a_i} h^*(a) \right]_{i=1}^{D},$$
$$\left\| \left[ \frac{\partial}{\partial a_i} h^*(a) \right]_{i=1}^{D} \right\|_2 \leq C_2 \left\| a - a^* \right\|_2,$$

**Input:** $n$
Divide $[0,1]$ into $n$ subintervals $I_k$ with $I_k = [\frac{k-1}{n}, \frac{k}{n})$ $(1 \leq k < n)$ and $I_n = [\frac{n-1}{n}, 1]$
$t := 0$, $t_k := 0$
**for** $i := 1, \ldots, n$ **do**
  Choose from interval $I_i$ a point uniformly at random and receive reward $Y$
  $\hat{b}_i := Y$, $t_i := t_i + 1$, $t := t + 1$
**end for**
**for** $t := 1, 2, \ldots$ **do**
  $k := \text{argmax}_i \left( \hat{b}_i + \sqrt{\frac{2 \log(t-1)}{t_i}} \right)$
  Choose $A_t$ from $I_k$ uniformly at random and receive reward $Y_t$
  $\hat{b}_k := \frac{t_k \hat{b}_k + Y_t}{t_k + 1}$
  $t_k := t_k + 1$
**end for**

Table 1.6: The UCBC algorithm for continuum-armed bandit problems

where $a^*$ is the optimal action, $a \in \mathcal{A}$ is an arbitrary action, and $C_1, C_2 > 0$ are two constants. Cope uses Kiefer-Wolfowitz method and proves that under the above assumptions, its regret satisfies $\mathbb{E}[R(T)] = O(T^{1/2})$.

The state of the art result for the nonparametric bandit problem is due to Auer et al. (2007) who propose the UCBC algorithm shown in Table 1.6. UCBC is very similar to CAB1 of Kleinberg (2004), but uses different number of intervals depending on the information that it has about the regularity of the mean payoff function.

Auer et al. (2007) make the following assumptions:

**Assumption A3** There exists constants $L \geq 0$, $\zeta > 0$, $\nu > 0$ such that for any point $a_* \in [0,1]$ with $\limsup_{a \to a_*} h(a) = h_* = \sup_{a \in [0,1]} h(a)$, and all $a \in [0,1]$ such that $|a - a_*| < \nu$,

$$h(a_*) - h(a) \leq L |a_* - a|^{\zeta}. \tag{1.4}$$

We note that this assumption requires continuity only at the maxima and is thus considerably weaker than Kleinberg's assumption. While (1.4) constraints $h(a)$ from below in the vicinity of $a_*$, the next assumption in some sense is concerned by restricting the size of the actions that are competing with an optimal action.

**Assumption A4** There exist constants $M \geq 0$, $\beta > 0$ such that for all $\epsilon > 0$,

$$m(\{a : h_* - \epsilon < h(a) \leq h_*\}) \leq M \epsilon^{\beta}$$

holds, where $m$ denotes the Lebesgue measure.

With $\beta = 0$ and $M = 1$, Assumption A4 holds for any function. In this case, Auer et al. (2007) prove the following result for the UCBC Algorithm:

**Theorem 7.** *Pick $T > 0$, $n = \left(\frac{T}{\log T}\right)^{\frac{1}{2\zeta+1}}$. Then*

$$\mathbb{E}\left[R(T)\right] \leq \left(4L + \frac{c}{L} T^{\frac{1+\zeta}{1+2\zeta}} (\log T)^{\frac{\zeta}{1+2\zeta}}\right).$$

*Further if $n = \left(\frac{T}{\log T}\right)^{\frac{1}{3}}$ and if $T$ is sufficiently large, then*

$$\mathbb{E}\left[R(T)\right] \leq 4LT^{\max\{1-\frac{\zeta}{3}, \frac{2}{3}\}} (\log T)^{\frac{1}{3}} + \frac{c}{L} T^{\frac{2}{3}} (\log T)^{\frac{2}{3}}.$$

Note that the bound in the first part is generally tighter than the one in the second part. However the first part requires the knowledge of $\zeta$ while the second doesn't require this.

Auer et al. (2007) also prove that under Assumptions A3 and A4, for known $\zeta$ and $\beta$, with $n = \left(\frac{T}{\log T}\right)^{\frac{1}{1+2\zeta-\zeta\beta}}$, we get

$$\mathbb{E}\left[R(T)\right] \leq \left(4L + \frac{4cML^{\beta-1}}{2^{1-\beta}-1}\right) T^{\frac{1+\zeta-\zeta\beta}{1+2\zeta-\zeta\beta}} (\log T)^{\frac{\zeta}{1+2\zeta-\zeta\beta}},$$

which is unimprovable (see Theorem 3 in Auer et al. (2007)). Finally, for an important special case, Auer et al. (2007) prove the following theorem:

**Theorem 8.** *If $h$ has a finite number of maxima $a_*$ with $\limsup_{a \to a_*} h(a) = h_*$ and $h$ has continuous second derivatives which are non-vanishing at all these $a_*$ then UCBC with $n = \left(\frac{T}{\log T}\right)^{\frac{1}{4}}$ achieves*

$$\mathbb{E}\left[R(T)\right] \leq O(\sqrt{T \log T}).$$

# Chapter 2

# Playing in Parametric Bandit Problems

In this chapter, we propose an algorithm for the stochastic linear bandit problem and analyze its regret. In particular, we let $\mathcal{A} \subset \mathbb{R}^d$ satisfy Assumption A7 (cf. p.13) and

$$h(a) = \theta_*^T a$$

for some $\theta_* \in \mathbb{R}^d$ unknown to the algorithm. The algorithm, that we call FEL (Forced Exploration for Linear bandit problems), is shown in Table 2.1. The algorithm has two parameters: an increasing sequence $(f_t^*)$ and a distribution $P$. We require the distribution $P$ to be such that with $A \sim P(.)$ the matrix $\mathbb{E}\left[AA^T\right]$ is non-singular (remember that the actions $A$ belong to a Euclidean space now). The sequence $(f_t^*)$ determines the number of exploration steps the algorithm is forced to take up to time $t$. By exploration, we mean drawing a random action from $\mathcal{A}$ according to distribution $P$. When FEL is exploiting, it takes

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \theta_t^T a,$$

where $\theta_t$ is the least squares estimate of the parameter vector based only on the information gathered during the exploration steps up to time $t$. Using only the exploration information makes the analysis simpler. Analyzing this algorithm when it uses the information gathered during the exploitation phases remains future work (in Chapter 4, we will see empirically that using this information can substantially improve the performance of FEL in some problems.)

Define the optimal action for parameter $\theta \in \mathbb{R}^d$ as

$$a(\theta) = \operatorname*{argmax}_{a \in \mathcal{A}} \theta^T a \tag{2.1}$$

and the instantaneous regret as

$$r(\theta) = \theta_*^T (a_* - a(\theta)). \tag{2.2}$$

Let $C_0 := \mathbf{I}$, $y_0 := 0$, $\theta_0 := 0$, $f_0 := 0$ $\{C_0 \in \mathbb{R}^{d\times d}$, and $y_0, \theta_0 \in \mathbb{R}^d\}$
**for** $t := 1, 2, \ldots$ **do**
  **if** $f_{t-1} < f_t^*$ **then**
    {Exploration:}
    $A_t \sim P$ {Draw a random action from $\mathcal{A}$ according to distribution $P$}
    Take $A_t$ and receive payoff $Y_t$
    $C_t := C_{t-1} + A_t A_t^T$
    $y_t := y_{t-1} + Y_t A_t$
    $\theta_t := (\mathbf{I} + C_t)^{-1} y_t$
    $f_t := f_{t-1} + 1$
  **else**
    {Exploitation:}
    $A_t := \text{argmax}_{a\in\mathcal{A}}\, \theta_{t-1}^T a$
    Take $A_t$ and receive payoff $Y_t$
    $C_t := C_{t-1}$, $y_t := y_{t-1}$, $\theta_t := \theta_{t-1}$, $f_t := f_{t-1}$
  **end if**
**end for**

Table 2.1: FEL algorithm for stochastic linear bandit problems. Note that $\mathbf{I}$ denotes the identity matrix, making the algorithm estimate the unknown parameter using ridge regression. Note that $f_t$, unlike $C_t$ and $\theta_t$, is not random.

The main idea of FEL and its regret analysis in Section 2.2 (cf. p.25) is that under various conditions (cf. Section 2.3 on p.27) the following assumption holds:

**Assumption A5** The regret function, as defined by (2.2), satisfies

$$r(\theta) \le c \left\|\theta - \theta_*\right\|_2^2 + c' \left\|\theta - \theta_*\right\|_2^3,$$

for some $c, c' > 0$.

In Section 2.3, we will show that this assumption holds for a few interesting action spaces. We also make the following assumptions:

**Assumption A6** The probability distribution $P$ is such that if $A \sim P(.)$ then the matrix $\mathbb{E}\left[AA^T\right]$ is non-singular.

**Assumption A7** There exists $B > 0$ such that for any $a \in \mathcal{A}$, $\left\|a\right\|_2 \le B$.

This means that function $r(\theta)$ can be bounded from above by $\left\|\theta - \theta_*\right\|_2^2$ around $\theta_*$. This fact will be used in Section 2.2 to show that the regret of FEL with $f_t^* = d\sqrt{t}$ up to time $T$ is upper bounded by $d\sqrt{T} + \tilde{O}\left(\frac{\sqrt{T}}{d\nu_d^2}\right)$, where $\nu_d$ is the smallest eigenvalue of the dispersion matrix $\mathbb{E}\left[A_1 A_1^T\right]$.

## 2.1 Analysis of Least Squares Solution

Let $\theta_t$ be the estimate of $\theta_*$ at time step $t$ produced by the FEL algorithm, which is essentially the ridge regression estimator. As said before we assume that $r(\theta) \le c \left\|\theta - \theta_*\right\|_2^2 +$

$c' \|\theta - \theta_*\|_2^3$. So, given a bound on $\|\theta - \theta_*\|_2$, we can bound the instantaneous regret. In this section, we provide such a bound for $\|\theta - \theta_*\|_2$. This result might be well-known in the literature, though we were not able to find it. First we state our assumptions and then the main theorem will be presented.

For the convenience of the reader we collect the most important quantities used in the algorithm and in the proofs below in the following definition:

**Definition 1.** *Let $C_f$ and $C_{f,1}$, be constants such that for all $t$, $2\sum_{i=1}^{d}\left(1+\frac{2}{f_t\nu_d^2}\right)^2 \le C_f d$ and $C_{f,1} f_t^* \le f_t$. Let*

$$\begin{aligned}
H_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1}\le f_s^*\}} A_s Z_s, \\
C_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1}\le f_s^*\}} A_s A_s^T, \quad \text{(unnormalized empirical dispersion matrix)} \\
S_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_s\le f_s^*\}} A_s Y_s, \\
\tilde{C}_t &= \mathbf{I} + C_t, \\
\theta_t &= \tilde{C}_t^{-1} S_t \qquad\qquad \text{(estimate of } \theta_*\text{)}
\end{aligned}$$

*Further, let*

$$C_t = \mathcal{U}_t \Lambda_t \mathcal{U}_t^T$$

*be the SVD decomposition of $C_t$, where*

$$\Lambda_t = \mathrm{diag}(\lambda_{t1}, \dots, \lambda_{td}),$$

*with $\lambda_{t1} \ge \cdots \ge \lambda_{td}$ and $\mathcal{U}_t \mathcal{U}_t^T = \mathbf{I}$. Finally, let*

$$D_t = (\Lambda_t + \mathbf{I})^{-1}\Lambda_t - \mathbf{I}$$

*and let $\nu_1 \ge \dots \nu_d > 0$ be the eigenvalues of the dispersion matrix $\mathbb{E}\left[A_1 A_1^T\right]$. In what follows, for a positive definite matrix $C$ we let $\lambda_{max}(C)$ ($\lambda_{min}(C)$) denote its maximum (respectively minimum) eigenvalue.*

**Theorem 9.** *Let Assumptions A1, A2 (cf. p.6), A6 and A7 (cf. p.13) hold. Let $\theta_t$, $\nu_d$, $C_f$, and $C_{f,1}$ be as defined by Definition 1. Let $c_1 = \nu_d^2/(16d^2)$ and*

$$c_2 = \frac{16(1+\log(4d))}{(1+\nu_d/2)^2 \min\{1, 72d/\nu_d^2\}}.$$

*Then for $t$ big enough such that*

$$f_t^* \ge \frac{1}{C_{f,1}} \max\left\{ \frac{2}{c_1}\left(-(\log c_2 - 2\log t) + \log\frac{1}{c_1}\right),\ \frac{48d^2}{\nu_d^2}\left(\log\frac{24d^2}{\nu_d^2} - \frac{1}{3}\log\frac{16d^3\log(d(d+1))}{\nu_d^4}\right)\right\}$$

*it holds that*

$$\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] \le \frac{256(1+\log(4d))B^2}{C_{f,1} f_t^* \nu_d^2 \min\{1, 72d/\nu_d^2\}} + 2\|\theta_*\|_2^2\left(\frac{C_f d}{\nu_d^2 C_{f,1}^2 f_t^{*2}} + \frac{32d^3\log ed(d+1)}{C_{f,1}^3 f_t^{*3}\nu_d^4}\right).$$

First we prove the following lemma:

**Lemma 10.** *Let $\theta_t$, $D_t$, $\lambda_{max}$, $\tilde{C}_t^{-1}$ and $H_t$ be as defined by Definition 1. Then it holds that*

$$\|\theta_t - \theta_*\|_2^2 \leq 2\lambda_{max}(D_t^2)\,\|\theta_*\|_2^2 + 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2.$$

*Proof.* First let's decompose the error $\theta_t - \theta_*$. We have

$$S_t = \sum_{s=1}^{t} A_s Y_s = \sum_{s=1}^{t} A_s(A_s^T\theta_* + Z_s) = C_t\theta_* + H_t.$$

Further, we have

$$\begin{aligned}\theta_t &= \tilde{C}_t^{-1}S_t \\ &= (\mathbf{I} + \mathcal{U}_t\Lambda_t\mathcal{U}_t^T)^{-1}S_t \\ &= (\mathcal{U}_t(\Lambda_t + \mathbf{I})\mathcal{U}_t^T)^{-1}S_t \\ &= \mathcal{U}_t(\Lambda_t + \mathbf{I})^{-1}\mathcal{U}_t^T S_t.\end{aligned}$$

By substituting $S_t$ in the above equation we get

$$\theta_t = \mathcal{U}_t(\Lambda_t + \mathbf{I})^{-1}\Lambda_t\mathcal{U}_t^T\theta_* + \mathcal{U}_t(\Lambda_t + \mathbf{I})^{-1}\mathcal{U}_t^T H_t. \tag{2.3}$$

It holds that

$$\mathcal{U}_t(\Lambda_t + \mathbf{I})^{-1}\Lambda_t\mathcal{U}_t^T\theta_* - \theta_* = \mathcal{U}_t[(\Lambda_t + \mathbf{I})^{-1}\Lambda_t - \mathbf{I}]\mathcal{U}_t^T\theta_*. \tag{2.4}$$

Also we have

$$(\Lambda_t + \mathbf{I})^{-1}\Lambda_t = \mathrm{diag}(\dots, \frac{\lambda_{ti}}{1+\lambda_{ti}}, \dots).$$

Hence

$$D_t = \mathrm{diag}(\dots, -\frac{1}{1+\lambda_{ti}}, \dots). \tag{2.5}$$

By (2.3), (2.4) and (2.5), we get

$$\theta_t - \theta_* = \mathcal{U}_t D_t\mathcal{U}_t^T\theta_* + \tilde{C}_t^{-1}H_t.$$

Hence,

$$\begin{aligned}\|\theta_t - \theta_*\|_2^2 &\leq 2\left(\left\|\mathcal{U}_t D_t\mathcal{U}_t^T\theta_*\right\|_2^2 + \left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right) \\ &= 2\theta_*^T\mathcal{U}_t D_t^2\mathcal{U}_t^T\theta_* + 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2 \\ &= 2\left\|\mathcal{U}_t^T\theta_*\right\|_{D_t^2}^2 + 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2 \\ &\leq 2\lambda_{max}(D_t^2)\left\|\mathcal{U}_t^T\theta_*\right\|_2^2 + 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2 \\ &= 2\lambda_{max}(D_t^2)\,\|\theta_*\|_2^2 + 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2,\end{aligned}$$

finishing the proof. □

Hence we can bound $\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right]$ by bounding $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$ and $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$. We will bound $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$ and $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$ in the following subsections. In particular, Lemma 11 bounds $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$ and Lemma 19 bounds $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$ by appropriate quantities, leading to the desired result.

### 2.1.1 Bounding $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$

The next lemma upper bounds $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$:

**Lemma 11.** *Let Assumptions A1, A2 (cf. p.6), A6 and A7 (cf. p.13) hold. Let $H_t$ be as defined by Definition 1. Let $c_1 = \nu_d^2/(16d^2)$ and*

$$c_2 = \frac{16(1+\log(4d))}{(1+\nu_d/2)^2 \min\{1, 72d/\nu_d^2\}}.$$

*Then for*

$$f_t^* \geq \frac{2}{c_1 C_{f,1}}\left(-(\log c_2 - 2\log t) + \log\frac{1}{c_1}\right),$$

*we have*

$$\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right] \leq \frac{128B^2(1+\log(4d))}{f_t \nu_d^2 \min\left\{1, \frac{72d}{\nu_d^2}\right\}}.$$

*Further, if*

$$f_t^* \geq \frac{1}{C_{f,1}} \max\left\{4\left(\frac{2d}{\nu_d}\right)^2 \log t,\ \frac{128B^2}{\nu_d^2}\left(1+\frac{2d}{9}\right)\log^2 t\right\},$$

*then*

$$\left\|\tilde{C}_t^{-1}H_t\right\|_2^2 \leq \frac{1}{4}$$

*holds w.p. $1 - 4d/t$.*

We upper bound $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right]$ in two steps. First we upper bound $\|H_t\|_2$ and then we lower bound the eigenvalues of $\tilde{C}_t$.

**Lemma 12.** *Let Assumptions A1 and A7 (cf. p.6, p.13) hold. Let $0 \leq \delta \leq 1$ and $H_t$ be as defined as in Definition 1. Then, for any $t \geq 1$, $0 < \delta < 1$,*

$$\|H_t\|_2 \leq 2B\sqrt{f_t x} + \frac{2\sqrt{2d}\, B\, x}{3}. \tag{2.6}$$

*where $x = \log\frac{2d}{\delta}$.*

*Proof.* Fix $t \geq 1$, $0 < \delta < 1$ and let $x = \log(2d/\delta)$. Fix $1 \leq i \leq d$. Let $\mathcal{F}_s = \sigma(A_1, Y_1, \ldots, A_s, Y_s)$ be the $\sigma$-algebra defined by the history up to time $s$. Applying Bernstein's inequality (cf. Theorem 36, Appendix B) to $(\mathbb{I}_{\{f_{s-1}\leq f_s^*\}} Z_s A_{si}, \mathcal{F}_s)$, we get that for any $V_{ti} > 0$, the probability that both

$$|H_{ti}| \geq \sqrt{2V_{ti}x} + \frac{2Bx}{3} \quad \text{and} \quad \sum_{s=1}^{t} \mathbb{E}\left[(\mathbb{I}_{\{f_{s-1}\leq f_s^*\}} Z_s A_{si})^2 | \mathcal{F}_{s-1}\right] \leq V_{ti} \tag{2.7}$$

hold is at most $\delta/d$, Let $\sigma_i^2 = \mathbb{E}\left[A_{si}^2\right]$. Note that since $\|a\|_2 \le B$ holds for any $a \in \mathcal{A}$,

$$\sum_{i=1}^{d} \sigma_i^2 = \mathbb{E}\left[\sum_{i=1}^{d} A_{si}^2\right] \le B^2. \tag{2.8}$$

Choose $V_{ti} = f_t \sigma_i^2$. Since $\sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1} \le f_s^*\}} = f_t$ ($\sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1} \le f_s^*\}}$ is the number of times we explored, and $f_t$ in the algorithm just counts the number of exploration steps),

$$\sum_{s=1}^{t} \mathbb{E}\left[(\mathbb{I}_{\{f_{s-1} \le f_s^*\}} Z_s A_{si})^2 | \mathcal{F}_{s-1}\right] \le V_{ti}$$

holds w.p.1, thanks to the independence $(A_{si})$ and the boundedness of $(Z_s)$. Thus,

$$|H_{ti}| \ge \sqrt{2V_{ti}x} + \frac{2Bx}{3}$$

holds with probability at most $\delta/d$. Thus, outside of an event of probability at most $\delta$,

$$|H_{ti}| \le \sqrt{2V_{ti}x} + \frac{2Bx}{3}$$

holds for all $1 \le i \le d$. We continue the calculation on the event when all these inequalities hold. Squaring both sides and using $(|a| + |b|)^2 \le 2(a^2 + b^2)$ we get that

$$\sum_{i=1}^{d} H_{ti}^2 \le 4 f_t \left(\sum_{i=1}^{d} \sigma_i^2\right) x + 2d \left(\frac{2Bx}{3}\right)^2.$$

Using (2.8), we get

$$\sum_{i=1}^{d} H_{ti}^2 \le 4 f_t B^2 \, x + 2d \left(\frac{2Bx}{3}\right)^2.$$

Using $\sqrt{|a| + |b|} \le \sqrt{|a|} + \sqrt{|b|}$, we get

$$\|H_t\|_2 \le 2B\sqrt{f_t x} + \frac{2\sqrt{2d}\, B\, x}{3},$$

which is the desired inequality. □

*Remark* 13. If $Z_s$ is an i.i.d. sequence then using McDiarmid's inequality (cf. Theorem 37, Appendix B) it is possible to prove that for all $0 < \delta < 1$, $t \ge 1$,

$$\|H_t\|_2 \le B\sqrt{2 f_t \log \frac{1}{\delta}}, \tag{2.9}$$

which is tighter than the above bound in three ways: instead of $\log(2d/\delta)$ the bound has $\log(1/\delta)$ and (2.9) does not have the second term that we have in (2.6). Finally, the leading constant in the McDiarmid-based bound is smaller by a factor of $\sqrt{2}$.

We lower bound the eigenvalues of $\tilde{C}_t$ by using a matrix perturbation result stated as Corollary 4.10 in (Stewart and Sun, 1990). First let us define the following matrix norms

for a matrix $M = [m_{ij}]_{i,j=1}^{n,k}$:

$$\begin{aligned} \|M\|_1 &= \max_{1\le j\le k} \sum_{i=1}^{n} |m_{ij}|, \\ \|M\|_\infty &= \max_{1\le i\le n} \sum_{j=1}^{k} |m_{ij}|, \\ \|M\|_2 &= \sigma(M). \end{aligned} \tag{2.10}$$

Here $\sigma(M)$ denotes the largest singular value of $M$.

**Theorem 14** (Stewart and Sun (1990), Corollary 4.10)**.** *Let $M$ be a symmetric matrix with eigenvalues $\nu_1 \ge \nu_2 \ge \ldots \ge \nu_d$ and $\tilde{M} = M + E$ denote a symmetric perturbation of $M$ such that the eigenvalues of $\tilde{M}$ are $\tilde{\nu}_1 \ge \tilde{\nu}_2 \ge \ldots \ge \tilde{\nu}_d$. Then,*

$$\max_i\{|\tilde{\nu}_i - \nu_i|\} \le \|E\|_2.$$

Now, we lower bound the eigenvalues of the unnormalized empirical dispersion matrix.

**Lemma 15.** *Let $0 < \delta < 1$. Let $\lambda_{t1} \ge \ldots \lambda_{td} > 0$ and $\nu_1 \ge \ldots \nu_d > 0$ be as defined in Definition 1. Then there exists a time $t_0 > 0$ such that for any fixed $t > t_0$, with probability at least $1 - \delta$,*

$$\max_i\{|f_t\nu_i - \lambda_i|\} \le d\sqrt{2f_t \log d(d+1)/\delta}.$$

*Proof.* Let $E_t = C_t - \mathbb{E}[C_t]$ and let $e_{ij}$ be the $(i,j)$-th element of $E_t$. By the Hoeffding-Azuma inequality (cf. Theorem 35, Appendix B), w.p. at least $1-\delta$, it holds simultaneously for any $1 \le i, j \le d$ that

$$|e_{i,j}| \le \sqrt{2f_t \log d(d+1)/\delta}.$$

By Theorem 2.11 of (Stewart and Sun, 1990) we have

$$\|E_t\|_2^2 \le \|E_t\|_1 \|E_t\|_\infty \le (d \max_{i,j} |e_{ij}|)^2.$$

Hence,

$$\|E_t\|_2 \le d\sqrt{2f_t \log d(d+1)/\delta}. \tag{2.11}$$

By Theorem 14 and Inequality (2.11), we get

$$\max_i\{|f_t\nu_i - \lambda_i|\} \le d\sqrt{2f_t \log d(d+1)/\delta},$$

finishing the proof. □

**Lemma 16.** *Fix $0 < \delta < 1$ and let $\nu_i$ and $\lambda_{ti}$ be as defined in Definition 1. If $t$ is such that*

$$f_t \ge 2\left(\frac{2d}{\nu_i}\right)^2 \log\left(\frac{d(d+1)}{\delta}\right)$$

*then*

$$\lambda_{ti} \ge \frac{f_t}{2}\nu_i$$

*holds w.p. at least $1 - \delta$.*

*Proof.* From Lemma 15,

$$\mathbb{P}\left(|\lambda_{ti} - f_t\nu_i| \le d\sqrt{2f_t \log d(d+1)/\delta},\ i = 1, \dots, d\right) \ge 1 - \delta$$

holds for any $t$, $0 < \delta < 1$. In order to have $\lambda_{ti} \ge \frac{f_t}{2}\nu_i$, we only need the following inequality to hold:

$$f_t\nu_i - d\sqrt{2f_t \log d(d+1)/\delta} \ge \frac{f_t\nu_i}{2}.$$

With a series of rearrangement we get that this inequality is equivalent to

$$f_t \ge \frac{8d^2}{\nu_i^2} \log d(d+1)/\delta,$$

thus proving the result. □

We also need the following lemma:

**Lemma 17.** *Fix $a, b, c, C > 0$. Let $Z$ be a random variable such that (i) $\mathbb{P}(Z > b) = 0$ and (ii) for any $\epsilon \in [0, a]$, $\mathbb{P}(Z > \epsilon) \le C\exp(-c\epsilon)$. Then*

$$\mathbb{E}[Z] \le \frac{1 + \log C}{c} + (b - a)\exp(-ca).$$

*Proof.* We may assume that $Z \ge 0$ since $\mathbb{E}[Z] \le \mathbb{E}[\max(Z, 0)]$ and $\mathbb{P}(\max(Z, 0) > \epsilon) \le \mathbb{P}(Z > \epsilon)$. Then

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^\infty \mathbb{P}(Z > \epsilon)\, d\epsilon \\
&\le \int_0^x 1 d\epsilon + \mathbb{I}_{\{x<a\}} C \int_x^a \exp(-c\epsilon) d\epsilon + \int_a^b C\exp(-ca) d\epsilon \\
&\le x + \mathbb{I}_{\{x<a\}} C \int_x^\infty \exp(-c\epsilon) d\epsilon + C(b-a)\exp(-ca) \\
&= x + \mathbb{I}_{\{x<a\}} \frac{C\exp(-cx)}{c} + C(b-a)\exp(-ca) \\
&\le \frac{1 + \log C}{c} + C(b-a)\exp(-ca),
\end{aligned}$$

where the last inequality follows from the choice $x = \log(C)/c$. □

Now, we are ready to upper bound $\mathbb{E}\left[\left\|\tilde{C}_t^{-1} H_t\right\|_2^2\right]$.

*Proof of Lemma 11.* Fix $0 < \delta < 1$ and let $\delta' = \delta/2$. By Lemma 16, with probability at least $1 - \delta'$ it holds that $\lambda_{td} \ge f_t\nu_d/2 \ge 0$ when

$$f_t \ge 2\left(\frac{2d}{\nu_d}\right)^2 \log\left(\frac{d(d+1)}{\delta'}\right). \tag{2.12}$$

By Lemma 12, with probability $1 - \delta'$,

$$\|H_t\|_2 \le 2B\sqrt{f_t \log\frac{4d}{\delta}} + \frac{2\sqrt{2d}\, B}{3} \log\frac{4d}{\delta}. \tag{2.13}$$

Hence, since $\left\|\tilde{C}_t^{-1}H_t\right\|_2 \leq \lambda_{max}(\tilde{C}_t^{-1})\left\|H_t\right\|_2 = \lambda_{min}(\tilde{C}_t)^{-1}\left\|H_t\right\|_2$, and $\lambda_{min}(\tilde{C}_{td}) = 1+\lambda_{td}$, we get from (2.13) that

$$\begin{aligned}\left\|\tilde{C}_t^{-1}H_t\right\|_2 &\leq \frac{2B}{1+f_t\nu_d/2}\left(\sqrt{f_t\log\frac{4d}{\delta}}+\frac{\sqrt{2d}}{3}\log\frac{4d}{\delta}\right) \\ &\leq \frac{4B}{1+f_t\nu_d/2}\max\left\{\sqrt{f_t\log\frac{4d}{\delta}},\ \frac{\sqrt{2d}}{3}\log\frac{4d}{\delta}\right\}\end{aligned} \tag{2.14}$$

holds w.p. $1-\delta$ when (2.12) holds. Let $x = \sqrt{\log(4d/\delta)}$. Hence

$$\frac{\left\|\tilde{C}_t^{-1}H_t\right\|_2^2}{\left(\frac{4B}{1+f_t\nu_d/2}\right)^2} \leq \max\left\{x^2f_t,\ x^4\left(\frac{2d}{9}\right)\right\} \tag{2.15}$$

holds w.p. $1-4d\exp(-x^2)$ when

$$f_t \geq 4\left(\frac{2d}{\nu_d}\right)^2 x^2. \tag{2.16}$$

Now, we apply Lemma 17 to bound the expected value of $Z = \left\|\tilde{C}_t^{-1}H_t\right\|_2^2 / \left(\frac{4B}{1+f_t\nu_d/2}\right)^2$. Lemma 17 requires a deterministic upper bound for $Z$. This is obtained as follows:

$$\left\|\tilde{C}_t^{-1}H_t\right\|_2 \leq \frac{\left\|H_t\right\|_2}{1+\lambda_{dt}} \leq \left\|H_t\right\|_2 \leq Bt. \tag{2.17}$$

Hence,

$$Z \leq \frac{t^2(1+f_t\nu_d/2)^2}{16} = b. \tag{2.18}$$

Let

$$\begin{aligned}&\epsilon = \max\{x^2f_t, x^4(2d/9)\}, \\ &a = \frac{f_t^2\nu_d^2}{16d^2}a_0, \quad \text{where } \ a_0 = \max\left\{1,\ \frac{\nu_d^2}{72d}\right\}.\end{aligned}$$

We show that if $\epsilon \leq a$, then Inequality (2.16) holds. Assume $\epsilon \leq a$. If $a_0 = 1$, then it follows from definition of $\epsilon$ that $x^2f_t \leq f_t^2\nu_d^2/(16d^2)$, which gives Inequality (2.16). If $a_0 = \nu_d^2/(72d)$, then it follows from definition of $\epsilon$ that $x^4(2d/9) \leq (f_t^2\nu_d^4)/((2^7)(3^2)d^3)$, which gives Inequality (2.16). Hence, by (2.15) and (2.16), if $\epsilon \leq a$ then

$$\mathbb{P}\left(Z > \epsilon\right) \leq 4d\exp(-x^2).$$

Next, we need to find $C$ and $c$ such that $4de^{-x^2} \leq Ce^{-c\epsilon}$. Using $C = 4d$, we get that this is equivalent to $x^2 \geq c\epsilon$. This last inequality is satisfied if we had $x^2 \geq c\max\left\{x^2f_t,\ x^4\left(\frac{2d}{9}\right)\right\}$, or if $x^2 \geq cx^2f_t$ and $x^2 \geq cx^4(2d/9)$. After rearrangements, using (2.16), we get that the choice $c = \frac{1}{f_t}\min\left\{1,\ \frac{72d}{\nu_d^2}\right\}$ makes these inequalities true. Hence we have the following values

to be used in Lemma 17:

$$\begin{aligned} C &= 4d, \\ a &= \frac{f_t^2 \nu_d^2}{16d^2} \max\left\{1, \frac{\nu_d^2}{72d}\right\}, \\ b &= \frac{t^2(1+f_t\nu_d/2)^2}{16}, \\ c &= \frac{1}{f_t} \min\left\{1, \frac{72d}{\nu_d^2}\right\}. \end{aligned}$$

Now, by Lemma 17, we get the following upper bound for $Z = \left\|\tilde{C}_t^{-1} H_t\right\|_2^2 / \left(\frac{4B}{1+f_t\nu_d/2}\right)^2$:

$$\mathbb{E}\left[\frac{\left\|\tilde{C}_t^{-1} H_t\right\|_2^2}{\left(\frac{4B}{1+f_t\nu_d/2}\right)^2}\right] \leq \frac{1+\log C}{c} + (b-a)\exp(-ca) \leq \frac{1+\log C}{c} + b\exp(-ca).$$

We have that $\exp(-ca) = \exp\left(-\frac{\nu_d^2}{16d^2} f_t\right)$. Let $c_1 = \nu_d^2/(16d^2)$ and

$$c_2 = \frac{16(1+\log(4d))}{(1+\nu_d/2)^2 \min\{1, 72d/\nu_d^2\}}.$$

By Proposition 33, if

$$f_t \geq \frac{2}{c_1}\left(-(\log c_2 - 2\log t) + \log\frac{1}{c_1}\right). \tag{2.19}$$

then $b\exp(-ca) \leq \frac{1+\log C}{c}$. Hence under Condition (2.19), we have that

$$\mathbb{E}\left[\left\|\tilde{C}_t^{-1} H_t\right\|_2^2\right] \leq \frac{128B^2(1+\log(4d))}{f_t\nu_d^2 \min\left\{1, \frac{72d}{\nu_d^2}\right\}},$$

proving the first part of the theorem.

Now we bound $\mathbb{P}\left(\left\|\tilde{C}_t^{-1} H_t\right\|_2^2 \leq 1/4\right)$. Recall that we defined $x = \sqrt{\log(4d/\delta)}$. Choose $\delta > 0$ such that $\exp(-x^2) = 1/t$. Hence, $x = \sqrt{\log t}$. Hence, by (2.14),

$$\frac{\left\|\tilde{C}_t^{-1} H_t\right\|_2^2}{2\left(\frac{2B}{1+f_t\nu_d/2}\right)^2} \leq f_t \log t + \left(\frac{2d}{9}\right)\log^2 t$$

holds w.p. $1 - 4d/t$ when

$$f_t \geq 4\left(\frac{2d}{\nu_d}\right)^2 \log t.$$

By Proposition 33, we can show that if

$$f_t \geq \frac{128B^2}{\nu_d^2}\left(1 + \frac{2d}{9}\right)\log^2 t \tag{2.20}$$

then

$$2\left(f_t \log t + \frac{2d}{9}\log^2 t\right)\frac{16B^2}{f_t^2\nu_d^2} \leq \frac{1}{4}.$$

Hence, if Inequality (2.20) holds, then

$$\left\|\tilde{C}_t^{-1} H_t\right\|_2^2 \leq \frac{1}{4}$$

holds w.p. $1 - 4d/t$.

□

### 2.1.2 Bounding $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$

In this section, we bound $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$. First consider the following lemma:

**Lemma 18.** *Let $D_t$, $\nu_i$, $C_f$ and $\lambda_{ti}$ be as defined in Definition 1. Then w.p. at least $1-\delta$,*

$$\lambda_{max}(D_t^2) \leq \frac{C_f d}{\nu_d^2 f_t^2} + \frac{8}{\nu_d^2 f_t^2} \sum_{i=1}^{d} \left( \frac{\lambda_{ti}}{\nu_i f_t} - 1 \right)^2$$

*holds if $f_t \geq (2d/\nu_d)^2 (2 \log d(d+1)/\delta)$.*

*Proof.* Assume that $f_t \geq (2d/\nu_d)^2 (2 \log d(d+1)/\delta)$. Then by Lemma 16 we have $1 + \lambda_{ti} \geq \frac{f_t}{2}\nu_i$ w.p. at least $1-\delta$. Hence,

$$\frac{1 + \lambda_{ti} - f_t \nu_i}{f_t \nu_i} \geq -\frac{1}{2}.$$

If $x \geq -1/2$ then $1/(1+x) \leq 1 + 2x$. Hence, on this event,

$$\begin{aligned} \frac{1}{1+\lambda_{ti}} &= \frac{1}{f_t \nu_i} \frac{1}{1 + \frac{1+\lambda_{ti} - f_t\nu_i}{f_t \nu_i}} \\ &\leq \frac{1}{f_t \nu_i} \left( 1 + 2 \frac{1 + \lambda_{ti} - f_t \nu_i}{f_t \nu_i} \right). \end{aligned} \tag{2.21}$$

By (2.5) and (2.21), we get

$$\begin{aligned} \lambda_{max}(D_t^2) &= \max_i \frac{1}{(1+\lambda_{ti})^2} \\ &\leq \sum_{i=1}^{d} \frac{1}{(1+\lambda_{ti})^2} \\ &\leq \frac{1}{\nu_d^2 f_t^2} \sum_{i=1}^{d} 2 \left( \left( 1 + \frac{2}{f_t \nu_d^2} \right)^2 + 4 \left( \frac{\lambda_{ti}}{f_t \nu_i} - 1 \right)^2 \right) \\ &\leq \frac{C_f d}{\nu_d^2 f_t^2} + \frac{8}{\nu_d^2 f_t^2} \sum_{i=1}^{d} \left( \frac{\lambda_{ti}}{f_t \nu_i} - 1 \right)^2. \end{aligned}$$

□

Finally, we upper bound $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$ in the following lemma:

**Lemma 19.** *Let $D_t$, $\nu_d$, $C_f$ be as defined by Definition 1. Then if*

$$f_t^* \geq \frac{48 d^2}{C_{f,1} \nu_d^2} \left( \log \frac{24 d^2}{\nu_d^2} - \frac{1}{3} \log \frac{16 d^3 \log(d(d+1))}{\nu_d^4} \right),$$

*then*

$$\mathbb{E}\left[\lambda_{max}(D_t^2)\right] \leq \frac{C_f d}{\nu_d^2 f_t^2} + \frac{32 d^3 \log ed(d+1)}{f_t^3 \nu_d^4}.$$

*Further, if*

$$f_t^* \geq \frac{2 \left\| \theta_* \right\|_2}{C_{f,1} \nu_d} \sqrt{(2 + C_f) d}$$

*then*

$$\lambda_{max}(D_t^2) \leq \frac{1}{4 \left\| \theta_* \right\|_2^2}$$

*holds w.p. $1 - d(d+1) e^{-\frac{\nu_d^2 f_t}{8 d^2}}$.*

*Proof.* Define $L_t = \frac{\nu_d^2 f_t^2}{8}\left(\lambda_{max}(D_t^2) - \frac{C_f d}{\nu_d^2 f_t^2}\right)$ and $F_t = \sum_{i=1}^d \left(\frac{d\sqrt{2f_t}}{f_t\nu_i}\right)^2$. Hence,

$$\begin{aligned} L_t &\le \sum_{i=1}^d \left(\frac{\lambda_{ti}}{f_t\nu_i} - 1\right)^2 && \text{(Lemma 18)} \\ &\le \log\left(\frac{d(d+1)}{\delta}\right)\sum_{i=1}^d \left(\frac{d\sqrt{2f_t}}{f_t\nu_i}\right)^2 && \text{(Lemma 15)} \\ &\le F_t \log\left(\frac{d(d+1)}{\delta}\right) \end{aligned}$$

holds w.p. at least $1-\delta$ if

$$f_t \ge 2\left(\frac{2d}{\nu_d}\right)^2 \log\frac{d(d+1)}{\delta}.$$

Note that here we used that Lemma 18 holds on the event set where the conclusion of Lemma 15 holds. Substitute $\epsilon = \log(d(d+1)/\delta)$. Hence if

$$\epsilon \le \frac{\nu_d^2 f_t}{8d^2} \tag{2.22}$$

then

$$\mathbb{P}\left(L_t > F_t\epsilon\right) \le d(d+1)\exp(-\epsilon), \tag{2.23}$$

or $\mathbb{P}\left(L_t > u\right) \le d(d+1)\exp(-\frac{u}{F_t})$ if $\frac{u}{F_t} \le \frac{\nu_d^2 f_t}{8d^2}$.

Now we want to apply Lemma 17 for $L_t$ to bound its expected value from the above high probability bound. Since Lemma 17 requires $L_t$ to be deterministically upper bounded, we now demonstrate such a bound. We have

$$\begin{aligned} \lambda_{max}(D_t^2) &= \max_i \frac{1}{(1+\lambda_{ti})^2} \\ &= \frac{1}{(1+\min_i \lambda_{ti})^2} \le 1. \end{aligned}$$

Hence,

$$L_t \le \frac{\nu_d^2 f_t^2}{8} - \frac{C_f d}{8} \le \frac{\nu_d^2 f_t^2}{8}. \tag{2.24}$$

Now by (2.22), (2.23) and (2.24), we obtain the following values to be used in Lemma 17:

$$C = d(d+1),\ c = \frac{1}{F_t},\ a = \frac{F_t\nu_d^2 f_t}{8d^2},\ b = \frac{\nu_d^2 f_t^2}{8}.$$

Hence,

$$\begin{aligned} \mathbb{E}\left[L_t\right] &\le \frac{1+\log C}{c} + (b-a)\exp(-ca) \\ &\le \frac{1+\log(d(d+1))}{(1/F_t)} + \frac{\nu_d^2 f_t^2}{8}\exp\left(-\frac{\nu_d^2 f_t}{8d^2}\right). \end{aligned}$$

Let

$$\frac{1+\log(d(d+1))}{(1/F_t)} \ge \frac{\nu_d^2 f_t^2}{8}\exp\left(-\frac{\nu_d^2 f_t}{8d^2}\right).$$

By Proposition 33, this inequality holds when

$$f_t \geq \frac{48d^2}{\nu_d^2}\left(\log\frac{24d^2}{\nu_d^2} - \frac{1}{3}\log\frac{16d^3\log(d(d+1))}{\nu_d^4}\right).$$

Hence we have seen that if $t$ is big enough then

$$\mathbb{E}\left[L_t\right] \leq 2F_t \log ed(d+1),$$

where

$$F_t = \frac{2d^2}{f_t}\sum_{i=1}^{d}\frac{1}{\nu_i^2} \leq \frac{2d^3}{f_t\nu_d^2}.$$

Hence

$$\mathbb{E}\left[L_t\right] = \mathbb{E}\left[\frac{\nu_d^2 f_t^2}{8}\left(\lambda_{max}(D_t^2) - \frac{C_f d}{\nu_d^2 f_t^2}\right)\right] \leq \frac{4d^3\log ed(d+1)}{f_t\nu_d^2}.$$

Reorder the above equation to get

$$\begin{aligned}\mathbb{E}\left[\lambda_{max}(D_t^2)\right] &\leq \frac{8}{\nu_d^2 f_t^2}\left[\frac{4d^3\log ed(d+1)}{f_t\nu_d^2}\right] + \frac{C_f d}{\nu_d^2 f_t^2} \\ &= \frac{C_f d}{\nu_d^2 f_t^2} + \frac{32d^3\log ed(d+1)}{f_t^3\nu_d^4}.\end{aligned}$$

Now we prove the second part of the lemma. We know that if $u \leq \nu_d^2 f_t F_t/(8d^2)$, then

$$\mathbb{P}\left(L_t < u\right) \geq 1 - d(d+1)\exp\left(-\frac{u}{F_t}\right).$$

Let $u = \nu_d^2 f_t F_t/(8d^2)$. Hence,

$$\mathbb{P}\left(L_t < \frac{\nu_d^2 f_t F_t}{8d^2}\right) \geq 1 - d(d+1)e^{-\frac{\nu_d^2 f_t}{8d^2}}.$$

Hence,

$$\lambda_{max}(D_t^2) \leq \frac{1}{\nu_d^2 f_t^2}\left(\frac{\nu_d^2 f_t F_t}{d^2} + C_f d\right)$$

holds w.p. at least $1 - d(d+1)e^{-\frac{\nu_d^2 f_t}{8d^2}}$. Now notice that if

$$f_t \geq \frac{2\left\|\theta_*\right\|_2}{\nu_d}\sqrt{(2+C_f)d}$$

then

$$\lambda_{max}(D_t^2) \leq \frac{1}{4\left\|\theta_*\right\|_2^2}$$

holds w.p. $1 - d(d+1)e^{-\frac{\nu_d^2 f_t}{8d^2}}$. □

### 2.1.3 Putting all together

*Proof of Theorem 9.* By Lemmas 10, 11 and 19, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\theta} - \theta_*\right\|_2^2\right] &\leq 2\mathbb{E}\left[\lambda_{max}(D_t^2)\right]\left\|\theta_*\right\|_2^2 + 2\mathbb{E}\left[\left\|\tilde{C}_t^{-1}H_t\right\|_2^2\right] \\ &\leq \frac{256(1+\log(4d))B^2}{f_t\nu_d^2\min\{1, 72d/\nu_d^2\}} + 2\left\|\theta_*\right\|_2^2\left(\frac{C_f d}{\nu_d^2 f_t^2} + \frac{32d^3\log ed(d+1)}{f_t^3\nu_d^4}\right).\end{aligned}$$

Using $C_{f,1}f_t^* \leq f_t$ gives the desired result. □

*Remark* 20. Assume that $B = 1$ and that $\nu_d$ behaves like $1/d$. This makes the bound of Theorem 9 in the order of $\tilde{O}(d/\sqrt{t})$. We show that this is the best upper bound that we can achieve by using the least squares solution (the same applies to ridge regression). Assume action space is $\mathcal{A} = \{e_1, \ldots, e_d\}$, where $e_i$ is the $i$th unit vector. Then it is easy to see that the least squares solution satisfies $\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] \approx d^2\sigma^2/f_t^*$, where $\sigma^2$ is the variance of the noise. Then use the exploration rate of $f_t^* = d\sqrt{t}$ and get $\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] = d\sigma^2/\sqrt{t}$.

## 2.2 FEL Analysis

First, we prove the following lemma:

**Lemma 21.** *Let Assumptions A2 and A5 hold. Then it holds that*

$$\mathbb{E}\left[r(\theta_t)\right] \leq \mathbb{E}\left[(c + c')\left\|\theta_t - \theta_*\right\|_2^2\right] + 2\mathbb{P}\left(\left\|\theta_t - \theta_*\right\|_2^2 \geq 1\right).$$

*Proof.* By Assumption A5 (cf. p.13),

$$r(\theta_t) \leq c\left\|\theta_t - \theta_*\right\|_2^2 + c'\left\|\theta_t - \theta_*\right\|_2^3. \tag{2.25}$$

Let $Z_t = \left\|\theta_t - \theta_*\right\|_2$. Because $r(\theta_t) \leq 2$, we have that

$$r(\theta_t) \leq \min\{2,\, cZ^2 + c'Z^3\}.$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[r(\theta)\right] &\leq \mathbb{E}\left[\min\{2,\, cZ^2 + c'Z^3\}\right] \\
&= \mathbb{P}\left(Z^2 < 1\right)\mathbb{E}\left[\min\{2,\, cZ^2 + c'Z^3\}|z^2 < 1\right] + \mathbb{P}\left(Z^2 \geq 1\right)\mathbb{E}\left[\min\{2,\, cZ^2 + c'Z^3\}|Z^2 \geq 1\right] \\
&\leq \mathbb{P}\left(Z^2 < 1\right)\mathbb{E}\left[\min\{2,\, (c + c')Z^2\}|Z^2 < 1\right] + \mathbb{P}\left(Z^2 \geq 1\right)\mathbb{E}\left[\min\{2,\, (c + c')Z^3\}|Z^2 \geq 1\right] \\
&\leq \mathbb{P}\left(Z^2 < 1\right)\mathbb{E}\left[\min\{2,\, (c + c')Z^2\}|Z^2 < 1\right] + 2\mathbb{P}\left(Z^2 \geq 1\right) \\
&\leq \mathbb{P}\left(Z^2 < 1\right)\mathbb{E}\left[(c + c')Z^2|Z^2 < 1\right] + 2\mathbb{P}\left(Z^2 \geq 1\right) \\
&\leq \mathbb{E}\left[(c + c')Z^2\right] + 2\mathbb{P}\left(Z^2 \geq 1\right).
\end{aligned}$$

□

The main result of our analysis is the following theorem:

**Theorem 22.** *Let Assumptions A1, A2 (cf. p.6), A5, A6 and A7 (cf. p.13) hold. Let*

$\nu_d > 0$ *be the smallest eigenvalue of* $\mathbb{E}\left[A_1 A_1^T\right]$. *Let*

$$G_1 = \frac{256(1+\log(4d))B^2}{C_{f,1} d\ \nu_d^2 \min\left\{1, \frac{72d}{\nu_d^2}\right\}},$$
$$G_2 = \frac{4C_f d\,\|\theta_*\|_2^2}{C_{f,1}^2 d^2 \nu_d^2},$$
$$G_3 = \frac{64\log(ed(d+1))\,\|\theta_*\|_2^2}{C_{f,1}^3 \nu_d^4}.$$

*Further let* $c_1 = \nu_d^2/(16d^2)$ *and*

$$c_2 = \frac{16(1+\log(4d))}{(1+\nu_d/2)^2 \min\{1, 72d/\nu_d^2\}}.$$

*Let*

$$f_t^* \geq \frac{1}{C_{f,1}} \max\left\{ \frac{2}{c_1}\left(-(\log c_2 - 2\log t) + \log\frac{1}{c_1}\right),\ 4\left(\frac{2d}{\nu_d}\right)^2 \log t,\ \frac{128B^2}{\nu_d^2}\left(1+\frac{2d}{9}\right)\log^2 t, \right. \tag{2.26}$$
$$\frac{48d^2}{\nu_d^2}\left(\log\frac{24d^2}{\nu_d^2} - \frac{1}{3}\log\frac{16d^3\log(d(d+1))}{\nu_d^4}\right),$$
$$\left.\frac{2\,\|\theta_*\|_2}{\nu_d}\sqrt{(2+C_f)d},\ \frac{8d^2}{\nu_d^2}\left(\log t - \log\frac{4}{d+1}\right)\right\}.$$

*Finally, let* $c$ *and* $c'$ *be the constants of Assumption A5. Then the expected regret of FEL with* $f_t^* = d\sqrt{t}$ *up to time* $T$ *satisfies*

$$\mathbb{E}\left[R(T)\right] \leq f_T + t_1 + 16d\log T + 1 + (c+c')\left[2G_1\sqrt{T} + G_2 + G_2\log T + 2G_3\right].$$

*Proof.* By Theorem 9, for $t \geq t_1$, it holds that

$$\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] \leq \frac{256(1+\log(4d))B^2}{C_{f,1} f_t^* \nu_d^2 \min\{1, 72d/\nu_d^2\}} + 2\,\|\theta_*\|_2^2\left(\frac{C_f d}{\nu_d^2 C_{f,1}^2 f_t^{*2}} + \frac{32d^3\log ed(d+1)}{C_{f,1}^3 f_t^{*3}\nu_d^4}\right). \tag{2.27}$$

By Lemma 21, we have that

$$\mathbb{E}\left[r(\theta_t)\right] \leq \mathbb{E}\left[(c+c')\,\|\theta_t - \theta_*\|_2^2\right] + 2\mathbb{P}\left(\|\theta_t - \theta_*\|_2^2 \geq 1\right). \tag{2.28}$$

Let $X_1 = 2\lambda_{max}(D_t^2)\,\|\theta_*\|_2^2$, $X_2 = 2\left\|\tilde{C}_t^{-1}H_t\right\|_2^2$ and $Z = \|\theta_t - \theta_*\|_2$. By Lemma 10, we have that

$$Z^2 \leq X_1 + X_2.$$

We have that

$$\{Z^2 \geq 1\} \subset \{X_1 + X_2 \geq 1\} \subset \left\{X_1 \geq \frac{1}{2}\right\} \cup \left\{X_2 \geq \frac{1}{2}\right\}.$$

Hence, by the second parts of Lemmas 11 and 19,

$$\mathbb{P}\left(Z^2 \geq 1\right) \leq \mathbb{P}\left(X_1 \geq \frac{1}{2}\right) + \mathbb{P}\left(X_2 \geq \frac{1}{2}\right)$$
$$\leq \frac{4d}{t} + d(d+1)\exp\left(-\frac{\nu_d^2 f_t}{8d^2}\right).$$

Hence,

$$\mathbb{E}\left[r(\theta_t)\right] \leq (c+c')\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] + \frac{4d}{t} + d(d+1)\exp\left(-\frac{\nu_d^2 f_t}{8d^2}\right)$$
$$\leq (c+c')\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] + \frac{8d}{t}. \tag{2.29}$$

The last step holds when

$$f_t \geq \frac{8d^2}{\nu_d^2}\left(\log t - \log\frac{4}{d+1}\right).$$

Inequality (2.29), (2.28) and (2.27) gives that

$$\mathbb{E}\left[r(\theta_t)\right] \leq (c+c')\left(\frac{256(1+\log(4d))B^2}{C_{f,1}f_t^*\nu_d^2\min\{1, 72d/\nu_d^2\}} + 2\|\theta_*\|_2^2\left(\frac{C_f d}{\nu_d^2 C_{f,1}^2 f_t^{*2}} + \frac{32d^3\log d(d+1)}{C_{f,1}^3 f_t^{*3}\nu_d^4}\right)\right) + \frac{16d}{t}$$
$$\leq (c+c')\left[\frac{G_1 d}{f_t^*} + \frac{G_2 d^2}{f_t^{*2}} + \frac{G_3 d^3}{f_t^{*3}}\right] + \frac{16d}{t}.$$

By the choice of $f_t^* = d\sqrt{t}$, we get the final result as follows:

$$\mathbb{E}\left[R(T)\right] \leq \sum_{t=1}^{T}(\mathbb{I}_{\{f_t < f_t^*\}} + (1 - \mathbb{I}_{\{f_t < f_t^*\}})\mathbb{E}\left[r_t\right])$$
$$\leq f_T + \sum_{t=1}^{T}(1 - \mathbb{I}_{\{f_t < f_t^*\}})\mathbb{E}\left[r(\theta_t)\right]$$
$$\leq f_T + \sum_{t=1}^{T}\mathbb{E}\left[r(\theta_t)\right]$$
$$\leq f_T + t_1 + 16d\log T + 1 + (c+c')\left[2G_1\sqrt{T} + G_2 + G_2\log T + 2G_3\right].$$

In the second line above we used that in a time step when FEL is exploiting, $r_t = r(\theta_t)$. □

*Remark* 23. After hours and hours of tedious calculations using Proposition 33 several times one can show that if

$$t \geq \max\Bigg\{\left(\frac{16}{cC_{f,1}d}\right)^2\left(\frac{1}{4}\log\frac{1}{c_1c_2} + \log\frac{8}{cC_{f,1}d}\right), \left(\frac{16d}{\nu_d^2 C_{f,1}}\right)^2,$$
$$\left(\frac{64B\sqrt{2+4d/9}}{\nu_d\sqrt{dC_{f,1}}}\log\frac{32B\sqrt{2+4d/9}}{\nu_d\sqrt{dC_{f,1}}}\right)^4,$$
$$\left(\frac{48d}{\nu_d^2}\right)^2\left(\log\frac{24d^2}{\nu_d^2} - \frac{1}{3}\log\frac{16d^3\log(d(d+1))}{\nu_d^4}\right)^2,$$
$$\frac{4\|\theta_*\|_2^2(2+C_f)}{d\nu_d^2}, \left(\frac{32d}{\nu_d^2}\right)^2\left(-\frac{1}{2}\log\frac{4}{d+1} + \log\frac{16d}{\nu_d^2}\right)\Bigg\},$$

then condition (2.26) is satisfied.

## 2.3 Results For Various Action Sets

In this section, we consider some cases when the action set $\mathcal{A}$ is such that the regret function will satisfy Assumption A5 (cf. p.13).

**Strictly convex action sets**

Let us assume that the action set is the 0-level set of some strictly convex, sufficiently smooth function, $c : \mathbb{R}^d \to \mathbb{R}$:

$$\mathcal{A} = \{a \in \mathbb{R}^d \,:\, c(a) \le 0\,\}. \tag{2.30}$$

Note that $a(\theta)$ is the solution of the following constrained, parametric linear optimization problem:

$$\begin{aligned} -\theta^T a &\to \min \\ \text{s.t. } c(a) &\le 0. \end{aligned} \tag{2.31}$$

Since the linear objective function is unbounded on $\mathbb{R}^d$, the solution always lies on the boundary of the set $\mathcal{A}$, i.e., $c(a(\theta)) = 0$ holds for any $\theta$. We make the following assumption:

**Assumption A8** The function $c$ is four-times differentiable in a neighborhood of $\theta_*$ and if $\lambda(\theta)$ denotes the the Langrange multiplier underlying the solution of the optimization problem (2.31) when the parameter is $\theta$ then $\lambda(\theta)$ is differentiable in the same neighborhood[1].

Then with the help of the Karush-Kuhn-Tucker (KKT) Theorem (cf. Theorem 32, Appendix A) and the Implicit Function Theorem (cf. Theorem 31, Appendix A) one gets that

$$-\theta + \lambda(\theta) D_a c(a(\theta)) = 0.$$

If we take derivative with respect to $\theta$ and reorder the result, we get that

$$D_\theta a(\theta) = -\frac{1}{\lambda(\theta)} (D_a^2 c(a(\theta)))^{-1} \left(D_a c(a(\theta)) D_\theta \lambda(\theta) + \mathbf{I}\right),$$

where $D_z$ is the derivative operator with respect to argument $z$.[2] It also follows from this argument that if the fourth order partial derivatives of $c$ exist and are continuous then $a(\cdot)$ is three times differentiable.

Let $f(a;\theta) = -\theta^T a$. Note that $r(\theta) = \theta_*^T a_* - f(a(\theta);\theta_*)$. We want to show that $D_\theta f(a(\theta);\theta_*)\big|_{\theta=\theta_*} = 0$. Using the chain rule,

$$D_\theta f(a(\theta);\theta_*) = D_a f(a(\theta);\theta_*) D_\theta a(\theta).$$

By the KKT conditions,

$$D_a f(a(\theta_*);\theta_*) = \lambda(\theta_*) D_a c(a(\theta_*)).$$

Further, by the complementary condition of the KKT theorem, $\lambda(\theta) c(a(\theta)) = 0$. Hence,

$$0 = D_\theta(\lambda(\theta_*) c(a(\theta_*))) = c(a(\theta_*)) D_\theta \lambda(\theta_*) + \lambda(\theta_*) D_\theta c(a(\theta_*)) = \lambda(\theta_*) D_\theta c(a(\theta_*))$$

[1] Note that the differentiability of $\lambda$ could be proven using arguments like in Chapter 12 of Nocedal and Wright (2006). These proofs are considerably technical and go beyond the scope of this thesis.

[2] That $\lambda(\theta) > 0$ follows from the KKT Theorem.

since $c(a(\theta_*)) = 0$. Hence,

$$D_\theta f(a(\theta);\theta_*)\big|_{\theta=\theta_*} = \lambda(\theta_*) D_a c(a(\theta_*)) D_\theta a(\theta_*) = \lambda(\theta_*) D_\theta c(a(\theta_*)) = 0.$$

Resorting to the Taylor-series expansion of $r(\theta)$, we get the following result:

**Theorem 24.** *Assume that the action set is given by* (2.30)*, where $c$ is a function that is strictly convex. Further, let Assumption A8 (cf. p.28) hold. Then the regret function $r$ satisfies Assumption A5 (cf. p.13).*

Note that if $\mathcal{A}$ is a sphere, or more generally an ellipsoid then $\mathcal{A}$ will satisfy the conditions of this theorem. In particular, when $\mathcal{A}$ is the unit sphere and the length of $\theta_*$ is one, $r(\theta) = \|(\theta/\|\theta\|_2) - \theta_*\|_2^2$, which can be used to show that in Assumption A5 (cf. p.13) in this case $c = 1$ can be chosen to be independent of $d$.

We suspect that the above result holds for very general action sets. In particular, it is not very difficult to see that the statement continues to hold when the set is described by a number of convex constraints which are all active in a neighborhood of $a_*$, such as in the example constructed for the lower bound proof in (Dani et al., 2008). One may believe based on the proof of this result that smoothness of $a(\cdot)$ is important. In the next section, we will look at the case when $a(\cdot)$ can have jumps, showing that smoothness is not essential. However, it remains for future work to fully characterize the cases when the subquadratic growth of $r$ holds.

**Polytopes**

In this section we assume that the action set $\mathcal{A}$ is a polytope (an intersection of a finite number of half-spaces). In this case without the loss of generality one can define function $a(\cdot)$ such that its range is the vertex set of the polytope. Then $r(\theta_*)$ becomes a piecewise constant function. We want to show that it is constant in a small neighborhood of $\theta_*$. Let $F$ be the unique $i$-face of the polygon for the largest $i = 0, 1, \ldots, d-1$ that contains $a(\theta_*)$ and which is perpendicular to $\theta_*$. (If there is no such $i$-face with $i \geq 1$ then we take $F$ to be the vertex $a(\theta_*)$.) By an elementary argument it follows that there exist a neighborhood $U$ of $\theta_*$ such that if $\theta \in U$ then $a(\theta)$ is on $F$. Since $F$ is perpendicular to $\theta_*$, for any $a, a' \in F$, $\theta_*^T a = \theta_*^T a'$. Hence, we have the following result:

**Theorem 25.** *Assume that $\mathcal{A}$ is a polytope. Then the regret function $r$ is zero in a neighborhood of $\theta_*$ and thus satisfies Assumption A5 (cf. p.13).*

Note that if the action set is a polytope, the forcing schedule can be changed to e.g. $f_t = c\log^2(t)$, which by an argument similar to the above one, but which exploits that the regret function is constant in a small neighborhood of $\theta_*$, gives a regret bound of order $O(c\log^2 T)$ (the probability of choosing a suboptimal vertex in the exploitation step will

decay at least as fast $1/T$, while the exploration steps contribute to the $\log^2 T$ growth of the regret). Note that in order to optimize the scaling of the regret with the dimension $d$, one should choose $c$ to be proportional to $\sqrt{d}$. Clearly, with this approach any regret slightly faster than $\log(t)$ can be achieved at the price of increasing the transient period.

### 2.3.1 Generalized Linear Payoff

Now, assume that the reward function takes the form: $h(a;\theta) = \theta^T \phi(a)$, i.e., the reward function takes the form of a generalized linear function (the function is linear in the parameters, but not in the actions). Here $\phi : \mathcal{A} \to \mathbb{R}^d$ and now the action space does not need to be a subset of a Euclidean space (or it could be a subset of a Euclidean space of dimension, say $s \neq d$). This case is interesting from the point of view of practical applications where the expected relatedness of the actions can be expressed with the help of some features $\phi$ (e.g., the actions could be grouped based on their expected similarity which can be expressed via the help of features; see Section 4.)

In order to make a connection to the linear payoff case, assume that the decision maker chooses a random action $A$ from some distribution $\mathcal{P}$. Then the expected immediate reward of this random action is $\mathbb{E}[h(A;\theta)] = \theta^T \mathbb{E}[a]$. If $\mathcal{P} = \sum_k p_k \delta_{a_k}$ then $\mathbb{E}[h(A;\theta)] = \theta^T \sum_k p_k \phi(a_k)$. Hence, if one defines $\tilde{\mathcal{A}}$ as the convex hull of $\mathcal{A}$ then we can view the problem as one defined with action set $\tilde{\mathcal{A}}$ and with a linear reward function $\bar{h}(\tilde{a};\theta) = \theta^T \tilde{a}$. Thus, we can apply Algorithm 2.1 to this problem. Note that the optimization problem $\operatorname{argmax}_{\tilde{a}\in\tilde{\mathcal{A}}} \theta^T \tilde{a}$ will have solutions on the boundary of $\tilde{\mathcal{A}}$. This means that in the exploitation steps, the algorithm can always select a non-randomized action.

If the action set $\mathcal{A}$ is finite, $\tilde{\mathcal{A}}$ becomes a polytope in which case the associated regret function satisfies Assumption A5 (cf. p.13). More generally, if this growth condition is satisfied, the algorithm's regret will be of order $d\sqrt{T}$ in time $T$, where $d$ is the dimension of the parameter space. Thus, the dimension (or cardinality) of the action space does not play a direct role in the regret of the algorithm (as expected). (Of course, these remarks apply to any linear bandit algorithm whose regret can be bounded in terms of the dimension of the parameter space.)

# Chapter 3

# Non-parametric Bandits

In this section, we drop the linearity assumption of Chapter 2 and study the more general case when the payoff function satisfies some smoothness conditions but is otherwise unrestricted. In particular, no parametric form for the payoff function is assumed, hence we call this the non-parametric case.

Table 3.1 shows the FEC Algorithm (Forced Exploration for Continuum-armed bandit problems), which is a modified version of the FEL Algorithm of Chapter 2. The main ideas are *(i)* using $d_t^*$ basis functions to estimate $h^*$; *(ii)* gradually increasing $d_t^*$ over time; *(iii)* and using a deterministic schedule for the exploration. The advantage of this algorithm is that it allows a flexible combination of known payoff structures with a non-parametric approach. In the next section we show that the regret of FEC at time $T$ satisfies $\mathbb{E}[R(T)] = O(T^{\frac{2+\alpha}{2+2\alpha}})$ assuming that the mean payoff function is (in some sense that will be made precise) smooth up to order $\alpha$. Unfortunately, this bound is not as good as the bound for UCBC shown in Theorem 7. In particular, for $\zeta = \alpha$, the difference between the exponents is $\alpha/((2+2\alpha)(1+2\alpha))$. Thus, although the exponent that we get for the regret is larger than the exponent for UCBC, the difference becomes negligible as $\alpha$ gets large. The difference is the result that in the parametric case our algorithm has a regret of order $\tilde{O}(d\sqrt{T})$ instead of $\tilde{O}(\sqrt{dT})$.

## 3.1 FEC Analysis

In this section, we analyze the regret of FEC. Let $D > 0$ and $d > 0$ be integers. Let $\mathcal{P}_{d,D}$ be the class of all polynomials on domain $\mathcal{A} \subset \mathbb{R}^D$ whose coordinatewise degree does not exceed $d$, $\phi_d : \mathcal{A} \to [-1,1]^d$ be the basis functions that spans $\mathcal{P}_{d,D}$. Let $(d_t)$ be an appropriate increasing sequence of integers and, by slightly abusing the notation, let us abbreviate $\phi_{d_t}$ with $\phi_t$. Further, let

$$h_t^* = \tilde{\theta}_t^T \phi_t \tag{3.1}$$

**Input:** A sequence of bases functions: $\mathcal{B} = \{b_1, b_2, \dots\}$,
where $b_d : \mathcal{A} \to [-1, 1]^d$, sequences $(f_t^*)$, $(d_t^*)$, distributions $(P_d)$
**Initialization:** Let $f_0 := 0$, $d_0 = 1$, $C_0 := 0$, $y_0 := 0$, $\theta_0 := 0$
$\{C_0 \in \mathbb{R}^{d_0 \times d_0}$, and $y_0, \theta_0 \in \mathbb{R}^{d_0}\}$
**for** $t := 1, 2, \dots$ **do**
  **if** $f_{t-1} < f_t^*$ **then**
    {Exploration:}
    $A_t \sim P_{d_t}$ {Draw a random action from $\mathcal{A}$ according to distribution $P_{d_t}$}
    Take $A_t$ and observe $Y_t$
    $C_t := C_{t-1} + \phi_t(A_t)\phi_t(A_t)^T$
    $y_t := y_{t-1} + Y_t\phi_t(A_t)$
    $\theta_t := (\mathbf{I} + C_t)^{-1} y_t$
    $f_t := f_{t-1} + 1$
  **else**
    {Exploitation:}
    $A_t := \text{argmax}_{a \in \mathcal{A}}\, \theta_{t-1}^T \phi_t(a)$
    Take $A_t$ and receive payoff $Y_t$
    $C_t := C_{t-1}$, $y_t := y_{t-1}$, $\theta_t := \theta_{t-1}$, $f_t := f_{t-1}$
  **end if**
  **if** $d_{t-1} < d_t^*$ **then**
    {Changing to a new basis:}
    $d_t := d_t^*$
    $\phi_t := b_{d_t}$
    {Reset $C_t, y_t, \theta_t$:}
    $C_t := \sum_{s=1}^{t} \phi_t(A_s)\phi_t(A_s)^T$
    $y_t := \sum_{s=1}^{t} Y_s\phi_t(A_s)$
    $\theta_t := (\mathbf{I} + C_t)^{-1} y_t$
  **else**
    $d_t := d_{t-1}$
  **end if**
**end for**

Table 3.1: FEC algorithm for continuum-armed bandit problems. Note that if $b_{d+1}$ is an extension of $b_d$ then the resetting of $C_t, y_t, \theta_t$ can be done in an efficient manner.

be the closest point to $h^*$ in $\mathcal{P}_{d_t,D}$ with respect to the supremum norm[1],

$$h_t^* = \underset{h \in \mathcal{P}_{d_t,D}}{\operatorname{argmin}} \left\| h^* - h \right\|_\infty , \tag{3.2}$$

which, for the sake of simplicity, we will assume to exist. Further, let

$$\Psi_t = \sup_{a \in \mathcal{A}} \left\| \phi_t(a) \right\|_2 ,$$
$$\zeta_t = \left\| \tilde{\theta}_t \right\|_2 .$$

Define the instantaneous regret as

$$r_t^* = h^*(a^*) - h^*(A_t)$$

and the instantaneous regret with respect to $h_t^*$ as

$$\tilde{r}_t = h_t^*(a_t^*) - h_t^*(A_t),$$

where $a_t^* = \operatorname{argmax}_{a \in \mathcal{A}} h_t^*(a)$ and $A_t$ is the action chosen by FEC at time $t$ (to be defined below). Further, define

$$r_t(\theta) = h_t^*(a_t^*) - h_t^*(a_t(\theta)),$$

where $a_t(\theta) = \operatorname{argmax}_{a \in \mathcal{A}} \theta^T \phi_t(a)$. Define the vector norms $\|v\|_2 = \sqrt{\sum_i v_i^2}$, $\|v\|_\infty = \max_i |v_i|$ and function norms $\|f\|_2 = \sqrt{\int_{\mathcal{A}} f^2(a) da}$, $\|f\|_\infty = \sup_{a \in \mathcal{A}} |f(a)|$. It is easy to see that $\|v\|_\infty \leq \|v\|_2$ and $\|f\|_2 \leq \|f\|_\infty$.

We make the following assumptions:

**Assumption A9** The target function $h^*$ is a member of the Sobolev space $W^\alpha(L^\infty(\mathcal{A}))$ (cf. Definition 5, Appendix A).

**Assumption A10** Let $\tilde{\theta}_t$ be as defined in (3.1) and (3.2). The instantaneous regret with respect to $h_t^*$ satisfies

$$r_t(\theta) \leq c \left\| \theta - \tilde{\theta}_t \right\|_2^2 + c' \left\| \theta - \tilde{\theta}_t \right\|_2^3 ,$$

where $c, c' > 0$ are some constants.

**Assumption A11** The distributions $(P_d)$ are such that if $\nu_d$ is the minimum eigenvalue of $\mathbb{E}\left[AA^T\right]$, where $A \sim P_d$ then, $H/d \leq \nu_d \leq J$ for some constants $H$ and $J$.

**Assumption A12** We have that $\Psi_t \leq 1$ and $\zeta_t \leq \sqrt{d_t}$.

In order to prove the main result of this chapter, we need the following theorem. The theorem is stated in the form given here as Theorem 2.3 in (French et al., 2003).

[1] Also known as the "best approximation" of $h^*$ in $\mathcal{P}_{d_t,D}$.

**Theorem 26** (Jackson's theorem)**.** *Let $1 \leq p \leq \infty$ and $\alpha \geq 1$ be integers. Given $h \in W^{\alpha}(L^p([-1,1]^D))$, there exists a constant $E > 0$ such that for all integers $D$, $d \geq 1$, there exists a polynomial $P \in \mathcal{P}_{d,D}$ such that*

$$\|h - P\|_p \leq E(d+1)^{-\alpha} \|h\|_p .$$

Next, we define some constants and functions:

**Definition 2.** *Let*

$$\gamma = \frac{\alpha+2}{2\alpha+2}, \quad \beta = \frac{1}{2\alpha+2}.$$

*Let $E$ be the constant in Jackson's theorem. Let $c$ and $c'$ be as defined in Assumption A10. Let $C_{f,1}$ and $C_d$ be constants such that $f_t \geq C_{f,1} f_t^*$ and $d_t \geq C_d d_t^*$. Further, let*

$$G_1 = \frac{9(2^6)(c+c')}{C_{f,1}H^2}, \quad G_3 = \frac{4(c+c')}{C_{f,1}^2 H^2}\left(1 + \frac{2}{H^2}\right), \quad G_4 = \frac{64(c+c')}{C_{f,1}^3 H^4}.$$

*Let $B_t$ be a sequence such that*

$$2\sum_{i=1}^{d_t}\left(1 + \frac{2}{f_t \nu_{d_t}^2}\right)^2 = B_t.$$

*Note that $B_t \leq 2(1+2/H^2)d_t$. Let $\theta_t$ be the coefficients vector produced by the FEC algorithm at time step $t$ and let $H$ and $J$ be as defined in Assumption A11. Finally, let $t_1$ be a time such that if $t \geq t_1$, then*

$$\begin{aligned}
d_t^* &\geq \frac{1}{C_d}\max\left\{\frac{J^2}{72}, \left(\frac{2E}{H}\right)^{\frac{1}{\alpha-1}}\right\},\\
f_t^* &\leq \frac{4\log(4d_t^*)}{\|h^* - h_t^*\|_\infty^2},\\
f_t^* &\geq \frac{4}{C_{f,1}}\left(\frac{4d_t^*}{\nu_{d_t^*}}\right)^2\left(\log t + \log\frac{1+J/2}{3} + \log\frac{4d_t^*}{\nu_{d_t^*}}\right),\\
f_t^* &\geq \frac{48 d_t^{*2}}{C_{f,1}\nu_{d_t^*}^2}\log\frac{24 d_t^{*2}}{\nu_{d_t^*}^2},\\
d_t^* &\geq \frac{1}{C_d}\left(\frac{J^4}{16}\right)^{1/3},\\
f_t^* &\geq \frac{4}{C_{f,1}}\left(\frac{2d_t^*}{\nu_{d_t^*}}\right)^2\log t,\\
f_t^* &\geq \frac{4\log t}{C_{f,1}\left(\frac{H^2}{8C_d^2 d_t^{*2}} - \frac{2E^2}{d_t^{*2\alpha}}\right)}\left(1 + \sqrt{1 + H^2/72}\right),\\
f_t^* &\geq \frac{8 d_t^{*2}}{C_{f,1}\nu_{d_t^*}^2}\log\frac{(d_t^*+1)t}{4}.
\end{aligned}$$

The next theorem is the main result of this section:

**Theorem 27.** *Let $\mathcal{A}$ be a bounded convex set in $[-1,1]^d$. Let Assumptions A1 (cf. p.6), A9, A10, A11 and A12 hold, where $\alpha \in \mathbb{N}$ satisfies $\alpha > 2$. Let $t_1$, $G_1$, $G_3$, $G_4$, $E$ and $\beta$ be as*

*defined in Definition 2. Then the expected regret of FEC with* $f_t^* = t^{\frac{\alpha+2}{2\alpha+2}}$ *and* $d_t^* = \lfloor t^{\frac{1}{2\alpha+2}} \rfloor$ *up to time* $T$ *satisfies*

$$\begin{aligned}\mathbb{E}\left[R(T)\right] \leq t_1 &+ \left(1 + \frac{2\alpha+2}{\alpha+2} G_1(1+\log 4 + \beta \log T) + 4E\frac{\alpha+1}{\alpha+2}\right) T^{\frac{\alpha+2}{2\alpha+2}} \\ &+ G_3\left((\alpha+1)T^{\frac{1}{1+\alpha}} - \alpha\right) \\ &+ G_4(1+\beta\log T + \log(T^\beta+1))\left(\frac{2\alpha+2}{4-\alpha}T^{\frac{4-\alpha}{2\alpha+2}} + \frac{3\alpha-2}{\alpha+4}\right) + 32(\alpha+1)\left(T^{\frac{1}{2\alpha+2}} - 1\right) \\ &- \frac{\alpha}{\alpha+2}G_1(1+\log 4+\beta\log T) - \frac{2\alpha E}{\alpha+2} + 8.\end{aligned}$$

First, we prove the following lemma, which will be used to prove Theorem 27.

**Lemma 28.** *Let* $\mathcal{A}$ *be a bounded convex set in* $[-1,1]^d$*. Let Assumptions A1 (cf. p.6), A9 and A12 hold. Assume that* $\alpha > 2$*. Let* $C_{1,f}$*,* $C_d$*,* $B_t$*,* $\theta_t$*,* $t_1$*,* $\beta$ *and* $\gamma$ *be as defined in Definition 2. Then for* $t \geq t_1$*, we have*

$$\mathbb{E}\left[\left\|\theta_t - \tilde{\theta}_t\right\|_2^2\right] \leq \frac{9(2^6)\Psi_t^2(1+\log(4d_t^*))}{C_{f,1}f_t^*\nu_{d_t^*}^2} + 2\left\|\tilde{\theta}_t\right\|_2^2\left(\frac{B_t}{\nu_{d_t^*}^2 C_{f,1}^2 f_t^{*2}} + \frac{32 d_t^{*3}\log(ed_t^*(d_t^*+1))}{C_{f,1}^3 f_t^{*3}\nu_{d_t^*}^4}\right).$$

*Further, let* $t_4$*,* $t_5$*,* $t_6$ *and* $t_7$ *be as defined in Definition 2. Assume that* $\Psi_t \leq 1$ *and* $\zeta_t \leq \sqrt{d_t}$*. Then, if* $t \geq \max\{t_4, t_5, t_6, t_7\}$*, it holds that*

$$\mathbb{P}\left(\left\|\theta_t - \tilde{\theta}_t\right\|_2^2 \geq 1\right) \leq \frac{8d_t}{t}.$$

*Proof.* Let

$$\begin{aligned}H_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1} < f_s^*\}}\phi_t(A_s)Z_s, \\ C_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1} < f_s^*\}}\phi_t(A_s)\phi_t(A_s)^T,\end{aligned}$$

and let $\lambda_{d_t}$ be the smallest eigenvalue of $C_t$. Define

$$\epsilon_{ts} = Y_s - h_t^*(A_s).$$

Hence, $\epsilon_{ts} = h^*(A_s) - h_t^*(A_s) + Z_s$. Further, let

$$\begin{aligned}\tilde{H}_t &= \sum_{s=1}^{t} \mathbb{I}_{\{f_{s-1} < f_s^*\}}\phi_t(A_s)\epsilon_{ts}, \\ \tilde{C}_t &= \mathbf{I} + C_t.\end{aligned}$$

If we follow the steps that led to Lemma 10, we get

$$\left\|\theta_t - \tilde{\theta}_t\right\|_2^2 \leq 2\lambda_{max}(D_t^2)\left\|\tilde{\theta}_t\right\|_2^2 + 2\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2, \tag{3.3}$$

where $D_t = \mathrm{diag}(\ldots, -\frac{1}{1+\lambda_{ti}}, \ldots)$. First we upper bound $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2\right]$.

**Bounding $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2\right]$.** Fix $0<\delta<1$ and $\delta'=\delta/2$. By Lemma 15, with probability at least $1-\delta'$ it holds that

$$\lambda_{d_t} \geq f_t\nu_{d_t}/2 \geq 0, \tag{3.4}$$

provided that

$$f_t \geq (2d_t/\nu_{d_t})^2(2\ln d_t(d_t-1)/\delta). \tag{3.5}$$

By the definitions of $\tilde{C}_t$ and $\tilde{H}_t$, we have

$$\begin{aligned}
\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2 &= \left\|\tilde{C}_t^{-1}\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\phi_t(A_s)\epsilon_{ts}\right\|_2 \\
&= \left\|\tilde{C}_t^{-1}\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\phi_t(A_s)(h^*(A_s)-h_t^*(A_s)+Z_s)\right\|_2 \\
&= \left\|\tilde{C}_t^{-1}\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\phi_t(A_s)Z_s + \tilde{C}_t^{-1}\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\phi_t(A_s)(h^*(A_s)-h_t^*(A_s))\right\|_2 \\
&\leq \frac{1}{1+\lambda_{td}}\left[\|H_t\|_2 + \|h^*-h_t^*\|_\infty\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\|\phi_t(A_s)\|_2\right].
\end{aligned}$$

By Inequality (3.4) and the above inequality and the fact that $\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}}\|\phi_t(A_s)\|_2 \leq \Psi_t\sum_{s=1}^{t}\mathbb{I}_{\{f_{s-1}<f_s^*\}} \leq \Psi_t f_t$, when (3.5) holds, we get

$$\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2 \leq \frac{1}{1+f_t\nu_{d_t}/2}\Big[\,\|H_t\|_2 + \Psi_t f_t\,\|h^*-h_t^*\|_\infty\,\Big].$$

Further, by Lemma 12, with probability $1-\delta'$,

$$\|H_t\|_2 \leq 2\Psi_t\sqrt{f_t\log\frac{2d_t}{\delta'}} + \frac{2\sqrt{2d_t}}{3}\Psi_t\log\frac{2d_t}{\delta'}. \tag{3.6}$$

Hence, w.p. $1-\delta$,

$$\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2 \leq \frac{9\Psi_t^2}{(1+f_t\nu_{d_t}/2)^2}\max\left\{4f_t\log\frac{4d_t}{\delta},\ \frac{8d_t}{9}\log^2\frac{4d_t}{\delta},\ f_t^2\,\|h^*-h_t^*\|_\infty^2\right\} \tag{3.7}$$

when (3.5) holds. Let

$$x = \sqrt{\log\frac{4d_t}{\delta}}.$$

Hence, w.p. $1-4d_te^{-x^2}$,

$$\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2 \leq \frac{9\Psi_t^2}{(1+f_t\nu_{d_t}/2)^2}\max\left\{4f_tx^2,\ \frac{8d_t}{9}x^4,\ f_t^2\,\|h^*-h_t^*\|_\infty^2\right\} \tag{3.8}$$

when (3.5) holds, i.e., when

$$f_t \geq 4\left(\frac{2d_t}{\nu_{d_t}}\right)^2x^2. \tag{3.9}$$

Let

$$\begin{aligned}
\epsilon &= \max\left\{4f_tx^2,\ \frac{8d_t}{9}x^4,\ f_t^2\,\|h^*-h_t^*\|_\infty^2\right\}, \\
a &= f_t^2\max\left\{\left(\frac{\nu_{d_t}}{2d_t}\right)^2,\ \frac{\nu_{d_t}^4}{288d_t^3},\ \|h^*-h_t^*\|_\infty^2\right\}.
\end{aligned}$$

Note that $a = \left(\frac{f_t \nu_{d_t}}{2d_t}\right)^2$ when

$$d_t \geq \max\left\{\frac{J^2}{72}, \left(\frac{2E}{H}\right)^{\frac{1}{\alpha-1}}\right\},$$

which holds by assumption. We show that if $\epsilon \leq a$, then Inequality (3.9) holds. Assume that $\epsilon \leq a$. Hence, $4f_t x^2 \leq \epsilon \leq a = f_t^2 \nu_{d_t}^2/(4d_t^2)$, which is equivalent to Inequality (3.9).

Now, we apply Lemma 17 for $Z = \left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2 / \left(\frac{9\Psi_t^2}{(1+f_t\nu_{d_t}/2)^2}\right)$ to bound its expected value, where we will use the bound (3.7). Lemma 17 requires a deterministic upper bound for $Z$. This is obtained as follows (see the derivation of Inequality (2.17) and use the fact that $\left\|\tilde{H}_t\right\|_2 \leq t\Psi_t$):

$$Z \leq \frac{(1+f_t\nu_{d_t}/2)^2 t^2}{9\Psi_t^2}.$$

Then we find $c$ and $C$ such that $4d_t e^{-x^2} \leq Ce^{-c\epsilon}$. Choose $C = 4d_t$. We need to find $c > 0$ s.t.

$$x^2 \geq c\epsilon = c\max\left\{4f_t x^2, \frac{8d_t}{9}x^4, f_t^2\left\|h^* - h_t^*\right\|_\infty^2\right\}.$$

We solve this inequality for all cases: (1) for $x^2 \geq 4cf_t x^2$, it is enough to have $c \leq 1/(4f_t)$, (2) for $x^2 \geq \frac{8d_t c}{9}x^4$, using (3.9), it is enough to have $c \leq 18d_t/(f_t\nu_{d_t}^2)$, (3) for $x^2 \geq c\left\|h^* - h_t^*\right\|_\infty^2 f_t^*$, by $x^2 \geq \log(4d_t)$, it is enough to have $c \leq \log(4d_t)/(f_t^2\left\|h^* - h_t^*\right\|_\infty^2)$. Putting all together, we have

$$c = \min\left\{\frac{1}{4f_t}, \frac{18d_t}{\nu_{d_t}^2 f_t}, \frac{\log(4d_t)}{f_t^2\left\|h^* - h_t^*\right\|_\infty^2}\right\}.$$

Note that $c = 1/(4f_t)$ when

$$d_t \geq \frac{J^2}{72}, \quad f_t \leq \frac{4\log(4d_t)}{\left\|h^* - h_t^*\right\|_\infty^2}.$$

Now, by Lemma 17, we get the following upper bound for $\mathbb{E}\left[\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2\right]$:

$$\mathbb{E}\left[\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2\right] \leq \frac{9\Psi_t^2}{(1+f_t\nu_{d_t}/2)^2}\left(\frac{1+\log C}{c} + (b-a)e^{-ca}\right), \quad \text{where } b = \frac{(1+f_t\nu_{d_t}/2)^2 t^2}{9\Psi_t^2}.$$

We have $(1+\log C)/c \geq be^{-ca}$ when

$$4f_t(1+\log(4d_t)) \geq \frac{(1+f_t\nu_{d_t}/2)^2 t^2}{9\Psi_t^2} e^{-f_t\left(\frac{\nu_{d_t}}{4d_t}\right)^2}. \tag{3.10}$$

A tedious calculation that uses Proposition 33 show that this latter inequality follows from

$$f_t \geq 4\Psi_t^2\left(\frac{4d_t}{\nu_{d_t}}\right)^2\left(\log t + \log\frac{1+J/2}{3} + \log\frac{4d_t}{\nu_{d_t}}\right),$$

which holds by assumption. Hence,

$$\mathbb{E}\left[\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2\right] \leq \frac{288\Psi_t^2(1+\log(4d_t^*))}{C_{f,1}f_t^*\nu_{d_t^*}^2}. \tag{3.11}$$

**Bounding $\mathbb{E}\left[\lambda_{max}(D_t^2)\right]$.** If $t$ is big enough such that

$$d_t \geq \left(\frac{J^4}{16}\right)^{1/3}$$

and

$$f_t \geq \frac{48d_t^2}{\nu_{d_t}^2}\log\frac{24d_t^2}{\nu_{d_t}^2},$$

then

$$f_t \geq \frac{48d_t^2}{\nu_{d_t}^2}\left(\log\frac{24d_t^2}{\nu_{d_t}^2} - \frac{1}{3}\log\frac{16d_t^3\log(d_t(d_t+1))}{\nu_{d_t}^4}\right). \tag{3.12}$$

Now if the above inequality holds, then by Lemma 19, we have that

$$\mathbb{E}\left[\lambda_{max}(D_t^2)\right] \leq \frac{B_t}{\nu_{d_t}^2 f_t^2} + \frac{32d_t^3\log(ed_t(d_t+1))}{f_t^3\nu_{d_t}^4}.$$

By Inequalities (3.3), (3.11) and the above bound we get

$$\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2\right] \leq \frac{576\Psi_t^2(1+\log(4d_t^*))}{C_{f,1}f_t^*\nu_{d_t^*}^2} + 2\left\|\tilde{\theta}_t\right\|_2^2\left(\frac{B_t}{\nu_{d_t^*}^2 C_{f,1}^2 f_t^{*2}} + \frac{32d_t^{*3}\log(d_t^*(d_t^*+1))}{C_{f,1}^3 f_t^{*3}\nu_{d_t^*}^4}\right),$$

which finishes the proof of the first part.

Now we prove the second part of the lemma. Define $x = \sqrt{\log(4d_t/\delta)}$. Let $\delta$ be such that $\exp(-x^2) = 1/t$. Hence, $x = \sqrt{\log t}$. Therefore, using (3.8), we get that

$$\frac{\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2}{\left(\frac{9\Psi_t^2}{(1+f_t\nu_{d_t}/2)^2}\right)} \leq 4f_t\log t + \left(\frac{8d_t}{9}\right)\log^2 t + f_t^2\left\|h^* - h_t^*\right\|_\infty^2 \tag{3.13}$$

holds w.p. $1 - 4d_t/t$ when

$$f_t \geq 4\left(\frac{2d_t}{\nu_{d_t}}\right)^2\log t.$$

By assumption $\Psi_t \leq 1$ and $\zeta_t \leq \sqrt{d_t}$. Again by Proposition 33, we can show that if

$$f_t \geq \frac{4\log t}{\left(\frac{H^2}{8d_t^2} - \frac{2E^2}{d_t^{2\alpha}}\right)}\left(1 + \sqrt{1 + H^2/72}\right),$$

then by Inequality (3.13), w.p. at least $1 - 4d_t/t$,

$$\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2 \leq \frac{1}{4}.$$

Further, by Lemma 19, if

$$f_t^* \geq \frac{2\zeta_t}{C_{f,1}\nu_{d_t}}\sqrt{2d_t + B_t} \tag{3.14}$$

then

$$\lambda_{max}(D_t^2) \leq \frac{1}{4\left\|\tilde{\theta}_t\right\|_2^2}$$

holds w.p. $1 - d_t(d_t+1)\exp\left(-\frac{\nu_{d_t}^2 f_t}{8d_t^2}\right)$. Hence, by (3.3),

$$\begin{aligned}\mathbb{P}\left(\left\|\theta_t - \tilde{\theta}_t\right\|_2^2 \geq 1\right) &\leq \mathbb{P}\left(2\left\|\tilde{C}_t^{-1}\tilde{H}_t\right\|_2^2 \geq \frac{1}{2}\right) + \mathbb{P}\left(2\left\|\tilde{\theta}_t\right\|_2^2 \lambda_{max}(D_t^2) \geq \frac{1}{2}\right) \\ &\leq \frac{4d_t}{t} + d_t(d_t+1)\exp\left(-\frac{\nu_{d_t}^2 f_t}{8d_t^2}\right) \leq \frac{8d_t}{t},\end{aligned}$$

where the last step holds if

$$f_t \geq \frac{8d_t^2}{\nu_{d_t}^2}\log\frac{(d_t+1)t}{4}.$$

□

*Proof of Theorem 27.* By the definitions of $\tilde{r}_t$ and $r_t^*$, we have

$$\begin{aligned}\tilde{r}_t &= h_t^*(a_t^*) - h_t^*(A_t), \\ r_t^* &= h^*(a^*) - h^*(A_t),\end{aligned}$$

and

$$r_t^* - \tilde{r}_t = h_t^*(A_t) - h^*(A_t) + h^*(a^*) - h_t^*(a_t^*).$$

Hence,

$$\begin{aligned}r_t^* &\leq \tilde{r}_t + \|h^* - h_t^*\|_\infty + h^*(a^*) - h_t^*(a_t^*) \\ &\leq \tilde{r}_t + \|h^* - h_t^*\|_\infty + h^*(a^*) - h_t^*(a^*) \\ &\leq \tilde{r}_t + 2\|h^* - h_t^*\|_\infty.\end{aligned}$$

By Lemma 21, we have that

$$\mathbb{E}[\tilde{r}_t] \leq \mathbb{E}\left[(c+c')\left\|\theta_t - \tilde{\theta}_t\right\|_2^2\right] + 2\mathbb{P}\left(\left\|\theta_t - \tilde{\theta}_t\right\|_2^2 \geq 1\right).$$

Hence,

$$\mathbb{E}[r_t^*] \leq \mathbb{E}\left[(c+c')\left\|\theta_t - \tilde{\theta}_t\right\|_2^2\right] + 2\mathbb{P}\left(\left\|\theta_t - \tilde{\theta}_t\right\|_2^2 \geq 1\right) + 2\|h^* - h_t^*\|_\infty.$$

Then by Lemma 28 we get the following inequality:

$$\begin{aligned}\mathbb{E}[r_t^*] \leq{}& \frac{32}{C_{f,1}f_t^*\nu_{d_t^*}^2}(c+c')\left(18\Psi_t^2(1+\log(4d_t^*))\right) \\ &+ 2\left\|\tilde{\theta}_t\right\|_2^2(c+c')\left(\frac{B_t}{\nu_{d_t^*}^2 C_{f,1}^2 f_t^{*2}} + \frac{32d_t^{*3}\log(ed_t^*(d_t^*+1))}{C_{f,1}^3 f_t^{*3}\nu_{d_t^*}^4}\right) \\ &+ 2\|h^* - h_t^*\|_\infty + \frac{16d_t}{t}\end{aligned}$$

for $t \geq \max\{t_1, \ldots, t_7\}$. By assumption, $\Psi_t \leq 1$ and $\zeta_t \leq \sqrt{d_t}$. Hence,

$$\begin{aligned}\mathbb{E}[r_t^*] \leq{}& \frac{32}{C_{f,1}f_t^*\nu_{d_t^*}^2}(c+c')\left(18(1+\log(4d_t^*))\right) \\ &+ 2d_t(c+c')\left(\frac{B_t}{\nu_{d_t^*}^2 C_{f,1}^2 f_t^{*2}} + \frac{32d_t^{*3}\log(ed_t^*(d_t^*+1))}{C_{f,1}^3 f_t^{*3}\nu_{d_t^*}^4}\right) \\ &+ 2\|h^* - h_t^*\|_\infty + \frac{16d_t}{t}.\end{aligned} \tag{3.15}$$

Then the total regret is

$$\begin{aligned}\mathbb{E}\left[R(T)\right] &= \sum_{t=1}^{T}(\mathbb{I}_{\{f_t<f_t^*\}} + (1-\mathbb{I}_{\{f_t<f_t^*\}})\mathbb{E}\left[r_t^*\right]) \\ &\leq f_T + \sum_{t=1}^{T}(1-\mathbb{I}_{\{f_t<f_t^*\}})\mathbb{E}\left[r_t^*\right] \\ &\leq f_T + \sum_{t=1}^{T}\mathbb{E}\left[r_t^*\right],\end{aligned}$$

Let $G_1$, $G_3$ and $G_4$ be as in Definition 2. By Inequality (3.15), we get

$$\begin{aligned}\mathbb{E}\left[R(T)\right] \leq\ & t_1 + \left(1 + \frac{2\alpha+2}{\alpha+2}G_1(1+\log 4+\beta\log T) + 4E\frac{\alpha+1}{\alpha+2}\right)T^{\frac{\alpha+2}{2\alpha+2}} \\ & + G_3\left((\alpha+1)T^{\frac{1}{1+\alpha}} - \alpha\right) \\ & + G_4(1+\beta\log T+\log(T^\beta+1))\left(\frac{2\alpha+2}{4-\alpha}T^{\frac{4-\alpha}{2\alpha+2}} + \frac{3\alpha-2}{\alpha+4}\right) + 32(\alpha+1)\left(T^{\frac{1}{2\alpha+2}}-1\right) \\ & - \frac{\alpha}{\alpha+2}G_1(1+\log 4+\beta\log T) - \frac{2\alpha E}{\alpha+2} + 8.\end{aligned}$$

finishing the proof □

*Remark* 29. By Proposition 33, constant $t_1$ in Definition 2 can be chosen as

$$\begin{aligned}t_1 = \max\Bigg\{ & \left(\frac{J^2}{C_d 72}\right)^{1/\beta}, \frac{1}{C_d^{1/\beta}}\left(\frac{2E}{H}\right)^{\frac{1}{\beta(\alpha-2)}}, \left(\frac{H^2}{72}\right)^{1/(3\beta)}, \left(\frac{E^2}{4\log(4C_d)}\right)^{\frac{2(\alpha+1)}{\alpha-2}}, \\ & \left[\frac{32(1+2\beta)}{C_{f,1}H(\gamma-4\beta)}\left(-\frac{\gamma-4\beta}{1+2\beta}\log\frac{3H}{4(1+J/2)} + \log\frac{16(1+2\beta)}{C_{f,1}H(\gamma-4\beta)}\right)\right]^{\frac{1}{\gamma-4\beta}}, \\ & \left(\frac{3(2^7)\beta}{H^2C_{f,1}(\gamma-4\beta)}\left[\frac{\gamma-4\beta}{4\beta} + \log\frac{3(2^6)\beta}{H^2C_{f,1}(\gamma-4\beta)}\right]\right)^{\frac{1}{\gamma-4\beta}}, \\ & \frac{1}{C_d^{1/\beta}}\left(\frac{J^4}{16}\right)^{\frac{1}{3\beta}}, \\ & \left[\frac{32}{C_{f,1}H^2(\gamma-4\beta)}\log\frac{16}{C_{f,1}H^2(\gamma-4\beta)}\right]^{1/(\gamma-4\beta)}, \\ & \frac{64C_d^2(1+\sqrt{1+H^2/72})}{C_{f,1}H^2(\gamma-2\beta)}\left(\frac{C_{f,1}E^2(\gamma-2\beta)}{2(1+\sqrt{1+H^2/72})} + \log\frac{32C_d^2(1+\sqrt{1+H^2/72})}{C_{f,1}H^2(\gamma-2\beta)}\right)^{\frac{4(\alpha+1)}{\alpha}}, \\ & \left(\frac{16(1+\beta)}{C_{f,1}H^2(\gamma-4\beta)}\left[-\frac{8\log 2}{C_{f,1}H^2} + \log\frac{8(1+\beta)}{C_{f,1}H^2(\gamma-4\beta)}\right]\right)^{\frac{1}{\gamma-4\beta}}\Bigg\}.\end{aligned}$$

*Remark* 30. The theorem relies on whether Assumptions A10, A11 (cf. p.33) hold. It remains for future work to verify that these assumptions hold for reasonable classes of functions.

# Chapter 4

# Experiments

This section has three parts. First we confirm that the regret of FEL is in the order of $\tilde{O}(d\sqrt{T})$ where $d$ is the size of the parameter vector and $T$ is time. Then we apply FEL to the ad allocation problem and show that it outperforms Confidence Ellipsoid of Dani et al. (2008) and algorithms proposed by Pandey et al. (2007) for this problem. Finally, we compare FEL with UCT of Kocsis and Szepesvari (2006) and show that FEL is not really competitive with UCT when there are no correlations between actions.

## 4.1 Scaling with $d$ and $T$

Assume that $d$ is even. Let $\mathcal{A}_d$ be the Cartesian product of $d/2$ circles, $\mathcal{A}_d = \{(a_1, \ldots, a_d) : a_1^2 + a_2^2 = a_3^2 + a_4^2 = \cdots = a_{d-1}^2 + a_d^2 = 1\}$. In this section, we empirically show that the regret of FEL on this problem scales according to $\tilde{O}(d\sqrt{T})$.

We run FEL for 1000 timesteps with $d = [2, 4, 8, 16, 32, 64]$ and repeat this experiment for 5 times. The value of the optimal action is equal to 1 in all experiments. The zero-mean noise is normally distributed with standard deviation equal to 0.1 (this noise does not satisfy the boundedness assumptions of our results, but our results in fact can be extended to this case). Figures 4.1 and 4.2 confirm that the regret of FEL scales linearly with $\sqrt{T}$ and $d$ (black line, labeled as $r = d$, shows the linear behavior).

In Chapter 2, we analyzed FEL when it uses only exploration information to estimate the parameter vector. In Figure 4.2, *FEL-U* refers to FEL when it uses all information and *FEL-NU* refers to FEL when it uses only the information gathered during exploration steps. As we can see in Figure 4.2, the performance of *FEL-U* and *FEL-NU* are almost identical in this problem. We have multiplied *FEL-U* with a small constant to make them distinguishable. So we only report the results for *FEL-U*, which gives slightly better results. The results for *FEL-NU* are reported only for $T = 1000$ and denoted by $T = 1000$, *FEL-NU* in Figure 4.2.

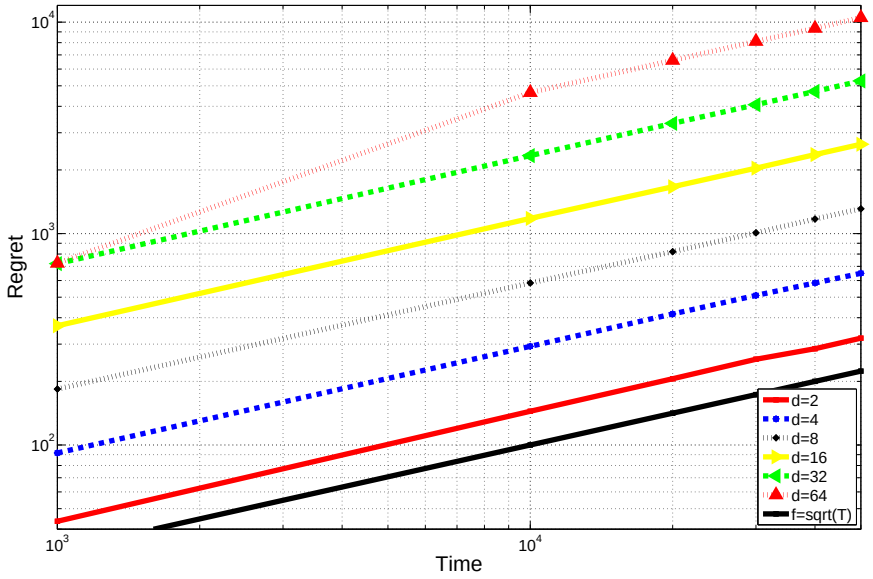


Figure 4.1: The total regret of FEL-U as a function of the time. For more explanation see the text.

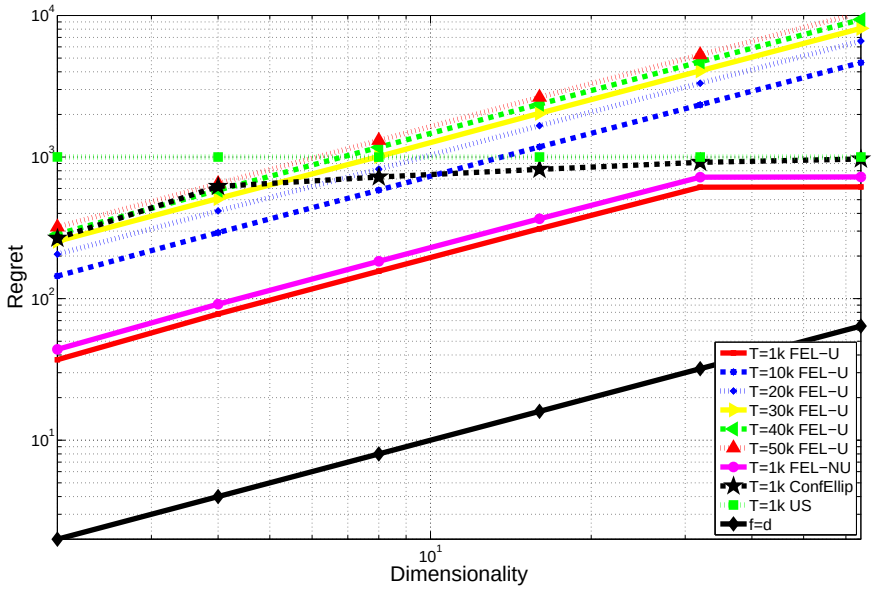


Figure 4.2: The total regret as a function of the dimensionality of the parameter vector. For more explanation see the text.

We have two interesting observations in Figure 4.2: 1) for $T = 1000$, the regret of FEL becomes almost constant for big values of $d$. Generally, for this class of problems it holds that (not shown here) if $t$ is fixed and $d \to \infty$, then the regret becomes $ct$ with some $c > 0$ that depends on the problem class. 2) The regret of Confidence Ellipsoid Algorithm is almost constant with respect to $d$. Actually Confidence Ellipsoid is still in its transient mode and its regret is $ct$. The regret of the uniform sampling, referred to as *US* , is also included in Figure 4.2. We observe that the regret of *ConfEllip* is converging to the regret of *US* as $d \to \infty$. Based on our observations (again not shown here), it takes a very long time for Confidence Ellipsoid to exit from its transient mode in this problem (something in the order of $T = 500,000$).

## 4.2 The Ad Allocation Problem

In the ad allocation problem, a website is provided with a number of ads. At each time step, the website chooses an ad to display. The website gets paid by each user-click. The objective is to maximize the number of user-clicks.

Since the number of ads is usually very large, we can not directly apply a $K$-armed bandit algorithm such as the popular UCB1 of Auer et al. (2002) to this problem. Fortunately, there are correlations between the ads and we exploit this property to achieve better performance.

The problem setting, borrowed from Pandey et al. (2007) is as follows: We have $C$ clusters indexed by $i \in \{1, \ldots, C\}$. Let us denote the optimal cluster (cluster containing the optimal action) by $i_{opt}$ (we refer to ads as actions). Each suboptimal cluster contains $N$ actions. The optimal cluster contains $N_{opt}$ actions. The outcome of an action is 0 or 1 and is distributed according to a Bernoulli distribution. Let $\mu_{opt}$ be the expected payoff of the optimal action in the optimal cluster, $\mu^s$ be the payoff of the best action among other clusters, $\mu_i$ be the payoff of the best action in cluster $i$, and $\mu_{i,j}$ be the payoff of the $j$th action of cluster $i$. Define $\Delta = \mu_{opt} - \mu^s$ as the cluster separation. The cohesiveness of cluster $i$ is defined as follows:

$$\delta_i = \frac{1}{N} \sum_{j=1}^{N} (\mu_i - \mu_{i,j}).$$

Further we let $\delta_{opt}$ denote the cohesiveness of the optimal cluster.

Following Pandey et al. (2007), we use $C = 10$ and $N = 10$ (i.e., the total number of actions is 100). The configuration is shown in Figure 4.3. For the suboptimal clusters, we use $\mu_i = 0.5$ and $\delta_i = 0.1$. The default values of the parameters of the optimal cluster are $N_{opt} = 10$, $\delta_{opt} = 0.3$, and $\mu_{opt} = 0.63$. The time horizon is 12000 and we repeat each experiment 200 times.

Pandey et al. (2007) simply put ads in clusters and use a two-stage UCB1 algorithm. In the first stage, they choose the cluster and in the second stage, they choose the ad. itself.

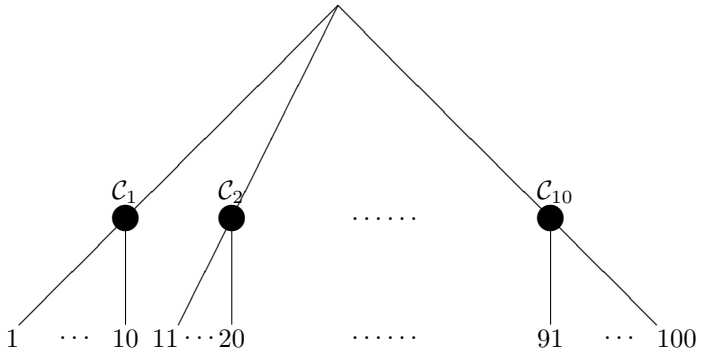


Figure 4.3: Each cluster contains 10 actions.

Pandey et al. (2007) show that this method substantially outperforms UCB1.

We use one basis for each cluster and one for each action. For example, if we had 10 clusters and 10 actions per cluster, we will have 110 bases (so $d = 110$). Denote by $\mathcal{C}_j \subset \mathcal{A}$ the set of actions belonging to cluster $j$ ($1 \leq j \leq 10$) and let $\mathcal{A} = \{1, 2, \ldots, 100\}$. Then the feature vector for each action is a vector of length 110 with two ones and 108 zeros: $\phi_i(a) = \mathbb{I}_{\{i=a, i\leq 100\}} + \mathbb{I}_{\{a\in\mathcal{C}_{i-100}, i>100\}}$. Like UCB1, we start running the algorithm by taking each action once. Then, we execute FEL with the exploration rate of $\sqrt{dt}$. Note that if we use $f_t^* = d\sqrt{t}$, then a simple calculation shows that FEL explores until $t \approx 10,000$!

Figure 4.4 summarizes our findings. Algorithms *Mean* and *Max* are introduced by Pandey et al. (2007). *FEL-U* refers to FEL when we are using all information and FEL-NU refers to FEL when we use only exploration information. *FEL-Lasso* refers to the FEL Algorithm when it uses a Lasso-like penalty term in estimating the parameter vector. Finally, *ConfEllip* refers to the Confidence Ellipsoid Algorithm of Dani et al. (2008). Due to time constraints, we repeated each experiment of *ConfEllip* for 5 times and each experiment of *FEL-Lasso* for only one time.

Figure 4.4 compares the total reward of these algorithms as (a) the cluster separation changes, (b) the number of actions in the optimal cluster changes, and (c) $(1-\delta_{opt})$ changes. As Figure 4.4 confirms, *FEL-U* outperforms both algorithms proposed by Pandey et al. (2007) in all three experiments under almost all conditions tested. The numbers that we are reporting in these figures for *Mean* and *Max* are slightly different (lower) than those that we see in (Pandey et al., 2007). We suspect that this might be due to different implementations of the underlying UCB1 algorithm.

Further, by comparing the performance of *FEL-U* and *FEL-NU*, we observe that there is a huge advantage in using the information gathered during the exploitation phases.

Another interesting observation is the poor performance of *ConfEllip*. We explain this observation by noting that we have 110 parameters and 100 actions. So it is expected that

*ConfEllip* can't outperform simple UCB1 in this problem. The numbers that Pandey et al. (2007) report for UCB1 are higher that our results for *ConfEllip*. Having said that, one strength of the FEL Algorithm comes from the implicit regularization that is happening because $C_t$ is initialized with $\mathbf{I}$.

Finally, we observe that *FEL-Lasso* outperforms all alternative algorithms. We attribute this performance to the fact that in this problem, the parameter vector is sparse: We expect 10+(a small number) of the elements of the parameter vector be around 1 and the rest of them be almost 0. This situation is particularly suitable for a method like Lasso.

## 4.3 FEL vs. UCT

Assume that we have a large action set and no prior information about the correlations between actions is given. The dominant algorithm for such problems is UCT of Kocsis and Szepesvari (2006).

UCT is implemented in the following way: Let $m$ be the number of actions. We build a tree of height $\log m$ (so it has $m$ leaves) and put each action on a leaf. We index nodes by $n_i$ where $1 \leq i \leq 2m-1$ (for example, we have $n_2$ and $n_3$ at depth 2). For node $n_i$, we construct a UCB1 algorithm with action set $\{n_{2i-1}, n_{2i}\}$. Further, we use $C_p = \sqrt{\frac{\log T}{T_i}}$ instead of $\sqrt{\frac{2\log T}{T_i}}$ as the bonus term. In each timestep, we start from the root and decide which child to choose according to the UCB1 Algorithm of that node. This process continues until we reach a leaf where we observe a reward. Now, we update the UCB1 Algorithms of all parents of this leaf (up to the root) assuming that we have observed this reward when choosing them.

One might wonder how FEL performs when we have a large uncorrelated action set. In this section, we compare the performance of FEL with UCT on a toy problem.

The toy problem is constructed as follows: We have 16 actions with payoffs generated uniformly randomly: {0.3612, 0.5936, 0.4552, 0.4263, 0.7276, 0.4632, 0.5171, 0.4739, 0.1034, 0.9755, 0.9249, 0.0290, 0.0166, 0.9703, 0.9594, 0.9076}. Then we build a tree as described above and put these actions on the leaves. The basis functions of FEL are built in the following way: Assign a basis function to each node of the tree. This basis is 1 when taking any action below it. Otherwise, it is 0. So we will have totally 31 basis functions. We use FEL with $f_t^* = 10 \log t$. We also run UCT with $C_p = \sqrt{2}$ and with $C_p = 0.21\sqrt{2}$. The second bonus term is tuned such that it gives the best performance for UCT on this problem. The time horizon is 10,000 and we repeat each experiment 10 times.

Figure 4.5 summarizes our findings and shows that FEL is not really competitive with UCT. In this figure, *UCT-UM* (UCT-UnModified) refers to the version of UCT with $C_p =$

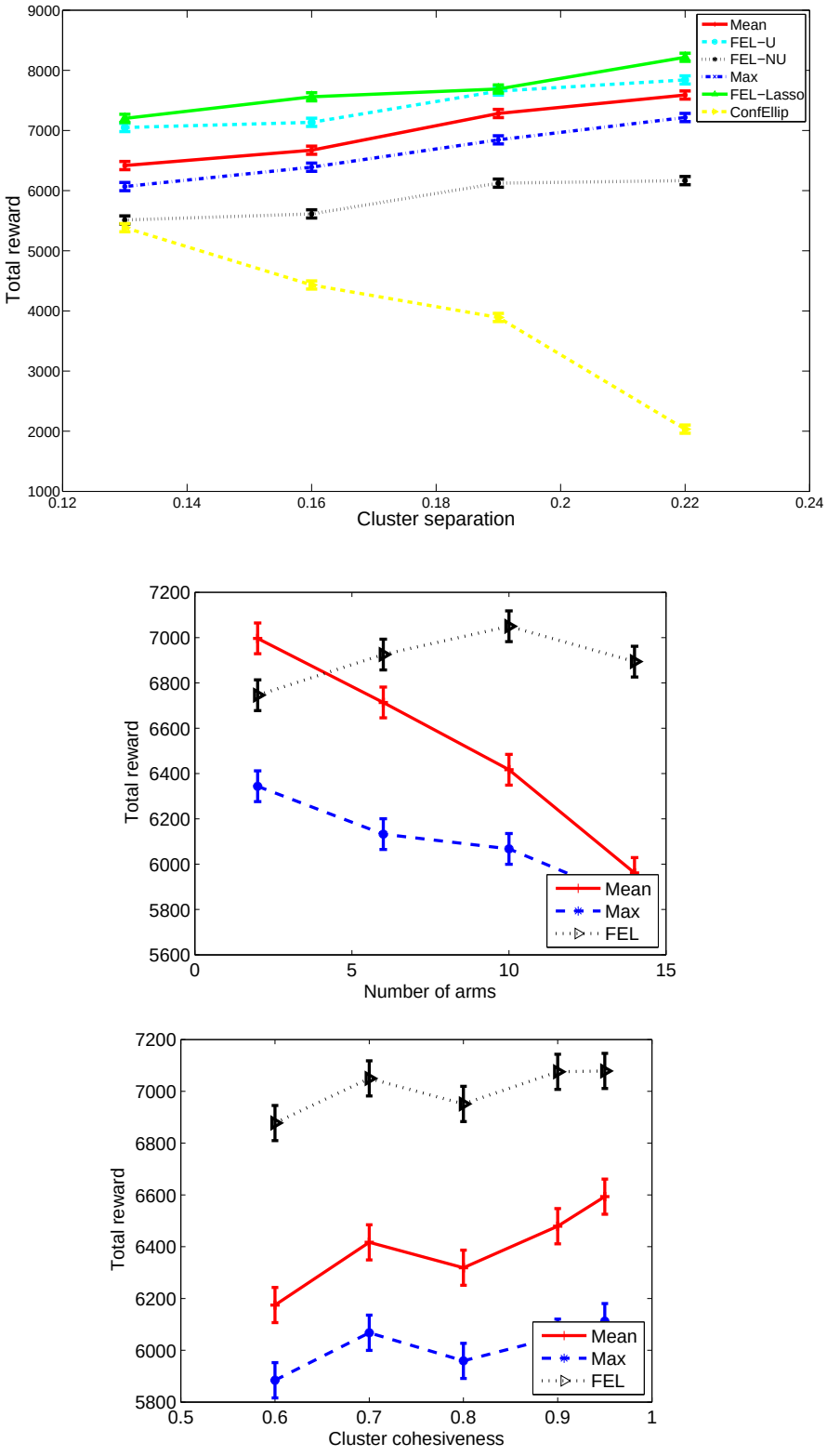


Figure 4.4: Total reward as a function of (a) Cluster separation ($\Delta$); (b) Number of actions in the optimal cluster ($N_{opt}$); (c) Optimal cluster cohesiveness ($1-\delta_{opt}$). 95 percent confidence bands are provided. *Features* is outperforming both algorithms proposed by Pandey et al. (2007) in all three domains under almost all conditions tested.

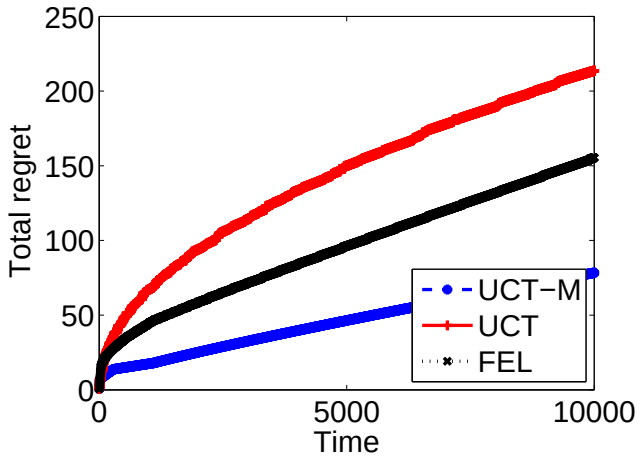


Figure 4.5: The regret of FEL compared to the regret of UCT.

$\sqrt{2}$ and $UCT$ refers to the version of UCT with $C_p = \sqrt{2}$.

# Chapter 5

# Conclusions

We studied the linear and continuum-armed bandit problems and explored the idea of how far the Forced Exploration methods (FEL and FEC) can take us. We analyzed FEL and found that its regret is bounded by $\tilde{O}(d\sqrt{T})$. We provided a simple example to show that this is the best that least squares method can do in the linear bandit problem (Remark 20). We also discovered the fact that $r(\theta)$ is approximately quadratic in the error in the neighborhood of $\theta_*$ for a few interesting action spaces. We conclude that FEL can be proven to be competitive, which is confirmed by the experiments. The experiments also confirm our conjecture of how the regret of FEL scales with $d$ and $T$ in a number of cases (Section 4.1). Our experiments go beyond the studied algorithm in two ways: 1) Learn from all data. This looks like a good idea, though we were not able to prove that this is indeed a good idea. 2) Learn with a Lasso-like penalty.

We also experimented with the Confidence Ellipsoid Algorithm and found that in one particular case the Confidence Ellipsoid Algorithm is not competitive with FEL. In this case, FEL+Lasso is even better than FEL. We provided a heuristic explanation for this observation.

We applied the techniques to a non-parametric situation and achieved a regret bounded by $O\left(T^{\frac{2+\alpha}{2+2\alpha}}\right)$, which is known to be a suboptimal rate.

However, the Forced Exploration method has the nice property that one can bound the expected instantaneous regret for $t$ big enough. This is usually impossible to do for the bandit algorithms. Further, its implementation is efficient, compared to algorithms like Confidence Ellipsoid that require solving intractable optimization problems.

Finally, note that all the bounds in this thesis contain problem-dependent constants. Although it is possible to tune the algorithm so that the performance then can be bounded in a problem-independent way (assuming a reasonable class e.g. when the expected payoffs are bounded), the resulting bound scales worse than $\sqrt{T}$ with time. This is not a weakness of the analysis, but a property of the algorithms considered. Hence the quest for a practically good performing, computationally efficient algorithm with provable uniform $O(\sqrt{T})$ regret

bound and which scales with dimension in an optimal manner is still open.

# Bibliography


A. Antos, V. Grover, and C. Szepesvari. Active learning in multi-armed bandits. In ALT-2008, pages 287–302, 2008.

J.-Y. Audibert, R. Munos, and Cs. Szepesvari. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. In Theoretical Computer Science-2008, 2008.

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

P. Auer. Personal communication, 2007.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. Machine Learning, 47(2-3):235–256, 2002.

P. Auer, R. Ortner, and Cs. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In Proceedings of the 20th Annual Conference on Learning Theory (COLT-07), pages 454–468, 2007.

N. Cesa-Bianchi and G. Lugosi. Prediction, Learning, and Games. Cambridge University Press, New York, NY, USA, 2006.

E. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. (submitted), 2006.

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

M. French, C. Szepesvari, and E. Rogers. Performance of Nonlinear Approximate Adaptive Controllers. Wiley, 2003.

R.C. Gunning and H. Rossi. Analytic Functions of Several Complex Variables. Prentice-Hall, New York, 1965.

W. Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58:13–30, 1963.

W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago, 1939.

R.D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In NIPS-2004, 2004.

R.D. Kleinberg. Online Decision Problems with Large Strategy Sets. PhD thesis, Department of Mathematics, Massachusetts Institute of Technology, 2005.

L. Kocsis and Cs. Szepesvari. Bandit based monte-carlo planning. In ECML-2006, pages 282–293, 2006.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6:4–22, 1985.

V.B. Melas. Functional Approach to Optimal Experimental Design. Springer, 2006.

J. Nocedal and S.J. Wright. Numerical Optimization. Springer, 2006.

S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. In ICML-2007, pages 721–728, 2007.

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

J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

G.W. Stewart and Ji-guang Sun. Matrix Perturbation Theory. Academic Press, 1990.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25:285–294, 1933.

# Appendix A

# Background in Calculus

**Definition 3** (Point-wise Holder function)**.** *Let $\mathcal{A} \in \mathbb{R}^d$, $h : \mathcal{A} \to \mathbb{R}$, $a_* \in \mathcal{A}$. If $\exists L, \beta > 0$ such that for all $a \in \mathcal{A}$, $h(a^*) - h(a) \leq L|a^* - a|^\beta$ then we say that function $h$ is Holder at $a^*$ with constant $L$ and exponent $\beta$.*

**Definition 4** (Uniformly locally Holder function)**.** *Let $\mathcal{A} \in \mathbb{R}^d$, $h : \mathcal{A} \to \mathbb{R}$, $a_* \in \mathcal{A}$. If $\exists \delta, L, \beta > 0$ such that for all $a$, $a' \in \mathcal{A}$ with $\|a - a'\| \leq \delta$,*

$$\|h^*(a) - h^*(a')\| \leq L\|a - a'\|^\beta$$

*then we say that function $h$ is Holder with constant $L$, exponent $\beta$, and neighborhood $\delta$.*

The next theorem is stated in the form given here as Theorem 1.8.1 in Melas (2006). The original version can be found in Gunning and Rossi (1965), Chapter 1.

**Theorem 31** (Implicit Function Theorem)**.** *Let $G : \mathbb{R}^{s+k} \mapsto \mathbb{R}^s$ be a function and fix $u_0 \in \mathbb{R}^k$ such that*

- *the equation $G(v, u_0) = 0$ has a solution $v_0$; and*
- *the function $G$ is continuous and has continuous first partial derivatives $\frac{\partial}{\partial u_i} G(v, u)$, $\frac{\partial}{\partial v_j} G(v, u)$, for $1 \leq i \leq k$ and $1 \leq j \leq s$ in the neighborhood of $(u_0, v_0)$ and*
- $\det \left[ \frac{\partial}{\partial v_j} G_i(v_0, u_0) \right]_{i,j=1}^{s} \neq 0.$

*Then there exists a neighborhood $\mathcal{U}$ of the point $u_0$ and function $g : \mathcal{U} \to \mathbb{R}^s$ such that in $\mathcal{U}$ we have (1) $G(u, g(u)) = 0$, (2) $v_0 = g(u_0)$, (3) $g$ is continuous and*

$$J(g(u), u) \frac{\partial g(u)}{\partial u_j} = -L_j(g(u), u), \quad j = 1, \ldots, k,$$

*where*

$$J(v, u) = \left[ \frac{\partial}{\partial v_j} G_i(v, u) \right]_{i,j=1}^{s}, \quad L_j(v, u) = \left[ \frac{\partial}{\partial u_j} G_i(v, u) \right]_{i=1}^{s}.$$

*Further, if $\hat{g} : \mathcal{U} \to \mathbb{R}^s$ satisfies (1) and (2) then $\hat{g} = g$.*

The following theorem can be found in Karush (1939):

**Theorem 32** (Karush-Kuhn-Tucker Theorem)**.** *Consider the following optimization problem:*

$$\begin{aligned} & f(x) \to \min! \quad s.t. \\ & g(x) \leq 0, \\ & h(x) = 0, \end{aligned}$$

*where $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable functions at $x_*$. Assume that $x_*$ is a local minimum and $\nabla g(x)$ and $\nabla h(x)$ are linearly independent at $x_*$. Then there exists constants $\mu$ and $\lambda$ such that the following holds:*

$$\begin{aligned} & \nabla f(x_*) + \mu \nabla g(x_*) + \lambda \nabla h(x_*) = 0, \\ & g(x_*) \leq 0,\ h(x_*) = 0, \\ & \mu \geq 0, \\ & \mu g(x_*) = 0. \end{aligned}$$

The following definition is stated by French et al. (2003):

**Definition 5** (Sobolev spaces)**.** *Let $1 \leq p \leq +\infty$ and $\alpha \geq 1$ be integers. Let $\mathcal{A} \subset \mathbb{R}^d$ be an open connected domain. Define $W^\alpha(L^p(\mathcal{A}))$ as the set of all measurable functions $h$ defined on $\mathcal{A}$ whose distributional derivatives $D^\omega h$, $|\omega| \leq \alpha$, lie in $L^p(\mathcal{A})$. Here $D^\omega$ is the distributional derivative of $h$ with respect to the multi-index $\omega = (\omega_1, \ldots, \omega_d)$, which means*

$$(D^\omega h)(x) = \left( \frac{\partial^{(\omega_1)}}{\partial a_1^{(\omega_1)}} \cdots \frac{\partial^{(\omega_d)}}{\partial a_d^{(\omega_d)}} h \right)(a),$$

*and*

$$|\omega| = |\omega_1| + \cdots + |\omega_d| .$$

*The semi-norm for $W^\alpha(L^p(\mathcal{A}))$ is defined by*

$$|h|_{W^\alpha(L^p(\mathcal{A}))} = \begin{cases} \sum_{|\omega|=\alpha} \|D^\alpha h\|_{L^p(\mathcal{A})}, & \text{if } 1 \leq p \leq \infty, \\ \max_{|\omega|=\alpha} \|D^\alpha h\|_{L^\infty(\mathcal{A})}, & \text{otherwise,} \end{cases}$$

*and the norm by*

$$\|h\|_{W^\alpha(L^p(\mathcal{A}))} = |h|_{W^\alpha(L^p(\mathcal{A}))} + \|h\|_{L^p(\mathcal{A})} .$$

The next proposition is stated in the form given here as Lemma 7 in Antos et al. (2008).

**Proposition 33.** *Let $q(t) = at + b$, $l(t) = \log t$, where $a > 0$. Then for any $t \geq (2/a)(-b + \log(1/a))$, $q(t) \geq l(t)$.*

# Appendix B

# Exponential Tail Inequalities

We use the following bounds throughout this thesis:

**Theorem 34** (Hoeffding's inequality (Hoeffding, 1963)). *If $X_1, \ldots, X_n$ are independent and $a_i \leq X_i \leq b_i (i = 1, 2, \ldots, n)$, then for $t > 0$*

$$\mathbb{P}\left(\bar{X} - \mu \geq t\right) \leq \exp\left(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

*where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\mu = \mathbb{E}\left[\bar{X}\right]$.*

The next theorem stated as Lemma A.7 in (Cesa-Bianchi and Lugosi, 2006) is the extension of Hoeffding's inequality to the martingales:

**Theorem 35** (Hoeffding-Azuma inequality). *Let $V_1, V_2, \ldots$ be a martingale difference sequence with respect to some sequence $X_1, X_2, \ldots$ such that $V_i \in [A_i, A_i + c_i]$ for some random variable $A_i$, measurable with respect to $X_1, \ldots, X_{i-1}$ and a positive constant $c_i$. If $S_n = \sum_{i=1}^{n} V_i$, then for any $t > 0$,*

$$\mathbb{P}\left(S_n > t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

*and*

$$\mathbb{P}\left(S_n < -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Next, we have Bernstein's inequality :

**Theorem 36** (Bernstein's inequality). *Let $(X_t, \mathcal{F}_t)$ be a bounded martingale difference series with $|X_t| \leq R$ w.p.1. Then, for any $n \geq 1$, $V > 0$, $0 < \delta < 1$, the simultaneous inequalities*

$$\sum_{t=1}^{n} X_t \quad \geq \quad \sqrt{2Vx} + \frac{2Rx}{3}, \quad \sum_{t=1}^{n} \mathbb{E}\left[X_t^2 | \mathcal{F}_{t-1}\right] \leq V$$

*hold with probability at most $\delta$, where $x = \log(1/\delta)$.*

Finally, we have McDiarmid's inequality (Shawe-Taylor and Cristianini, 2004):

**Theorem 37** (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathcal{X}$, and assume $f : \mathcal{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

*Then for all $\epsilon > 0$,*

$$\mathbb{P}\left(f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right] \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$