CSE599i: Online and Adaptive Machine Learning **Winter 2018**

Lecture 9: Linear Bandits and Thompson Sampling

Lecturer: Kevin Jamieson *Scribes: Tanner Fiez, Liyuan Zheng, Eunsol Choi, Yue Zhang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

We begin these notes by giving an overview on the background of the linear bandit problem and specify the learning model for the problem.

1.1 Linear Bandit Background

In the linear bandit problem a learning agent chooses an arm at each round and receives a stochastic reward. The expected value of this stochastic reward is an unknown linear function of the arm choice. As is standard in bandit problems, a learning agent seeks to maximize the cumulative reward over an n round horizon. The stochastic bandit problem can be seen as a special case of the linear bandit problem when the set of available arms at each round is the standard basis e_i for the Euclidean space $\mathbb{R}^d$, i.e. the vector e_i is a vector with all 0s except for a 1 in the ith coordinate. As a result each arm is independent of the others and the reward associated with each arm depends only on a single parameter as is the case in stochastic bandits.

The underlying algorithmic approach to solve this problem uses the optimism in the face of uncertainty (OFU) principle. The OFU principle solves the exploration-exploitation tradeoff in the linear bandit problem by maintaining a confidence set for the vector of coefficients of the linear function that governs rewards. In each round the algorithm chooses an estimate of the coefficients of the linear function from the confidence set and then takes an action so that the predicted reward is maximized. The problem reduces to constructing confidence sets for the vector of coefficients of the linear function based on the action-reward pairs observed in the past time steps.

The linear bandit problem was first studied by Auer et al. (2002) [1] under the name of linear reinforcement learning. Since the introduction of the problem, several works have improved the analysis and explored variants of the problem. The most influential works include Dani et al. (2008) [2], Rusmevichientong et al. (2010) [3], and Abbasi et al. (2011) [4]. In each of these works the set of available arms remains constant, but the set is only restricted to being a bounded subset of a finite-dimensional vector space. Variants of the problem formulation have also been widely applied to recommendation systems following the work of Li et al. (2010) [5] within the context of web advertisement.

An important property of this problem is that the arms are not independent because future arm choices depend on the confidence sets constructed from past choices. In the literature, several works including [5] have failed to recognize this property leading to faulty analysis. This fine detail requires special care which we explore in depth in Section 2.

1.2 Learning Model

In each round t, an agent chooses an arm X_t from a decision set $D = D_t \subseteq \mathbb{R}^d$ and receives a reward given by $Y_t = \langle X_t, \theta_* \rangle + \eta_t$ (e.g., with $\theta_* \in \mathbb{R}^d$, an unknown parameter to the agent, and η_t, a random noise parameter satisfying $E[\eta_t | X_{1:t}, \eta_{1:t-1}] = 0$ with tail constraints). The agent seeks to maximize the expected cumulative reward over n rounds given by $\sum_{t=1}^{n} \langle X_t, \theta_* \rangle$. The optimal strategy is then to select the arm at each round t that maximizes the expected immediate reward such as $x_t^* = \arg\max_{x \in D_t} \langle x, \theta_* \rangle$. The expected cumulative reward of the optimal strategy is therefore given by $\sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle$.

We now introduce the notion of regret in the linear bandit setting. In general the notion of regret is used to compare the performance of the optimal strategy with a bandit strategy. In stochastic problems we often consider two notions of averaged regret: expected regret and the pseudo regret. The expected regret is the expectation of the regret with respect to the action that is optimal on the sequence of reward realizations. Formally we can denote this as follows:

$$\begin{aligned} R_n &= \mathbb{E}\left[\max_{x_t \in D_t} \sum_{t=1}^{n} (\langle x_t, \theta_* \rangle + \eta_t) - \sum_{t=1}^{n} (\langle X_t, \theta_* \rangle + \eta_t)\right] \\ &= \mathbb{E}\left[\max_{x_t \in D_t} \sum_{t=1}^{n} \langle x_t, \theta_* \rangle - \sum_{t=1}^{n} \langle X_t, \theta_* \rangle\right] \\ &= \mathbb{E}\left[\max_{x_t \in D_t} \sum_{t=1}^{n} \langle x_t - X_t, \theta_* \rangle\right]. \end{aligned}$$

The pseudo-regret is a weaker notion of regret in which the performance of a bandit strategy is compared to the optimal strategy in expectation. We will consider the pseudo-regret exclusively from here on. The pseudo-regret for the linear bandit is as follows:

$$\begin{aligned} R_n &= \max_{x_t \in D_t} \mathbb{E}\left[\sum_{t=1}^{n} (\langle x_t, \theta_* \rangle + \eta_t) - \sum_{t=1}^{n} (\langle X_t, \theta_* \rangle + \eta_t)\right] \\ &= \max_{x_t \in D_t} \mathbb{E}\left[\sum_{t=1}^{n} \langle x_t, \theta_* \rangle + \eta_t\right] - \mathbb{E}\left[\sum_{t=1}^{n} \langle X_t, \theta_* \rangle + \eta_t\right] \\ &= \sum_{t=1}^{n} \langle x_t^*, \theta_* \rangle - \mathbb{E}\left[\sum_{t=1}^{n} \langle X_t, \theta_* \rangle\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n} \langle x_t^* - X_t, \theta_* \rangle\right]. \end{aligned}$$

In this case the pseudo-regret has the same expected value as the regret. Since the noise term is uncontrollable there is little purpose to bound it, hence why often the pseudo-regret is considered instead of the regret. From this point forward we will refer to the pseudo-regret simply as the regret.

In order to prove upper bounds on the regret we will make some assumptions. First, we assume that $\{D_t\}_{t=1}^{\infty}$ lies in a bounded set. We also assume that η_t is conditionally R-sub-Gaussian where $R \geq 0$ is a fixed constant. This assumption is equivalent to following condition: $\forall\ \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda \eta_t} | X_{1:t}, \eta_{1:t-1}] \leq \exp(\frac{\lambda^2 R^2}{2})$.

2 Motivating Linear Bandits

A unique property of the linear bandit framework is the dependence between arm plays. Special attention must be paid to this critical detail in order to develop proper analysis and algorithms for the linear bandit problem. We devote this section to a motivating example that illustrates how the analysis for linear bandits diverges from that of stochastic bandits. Specifically, this example demonstrates that blindly applying techniques from stochastic bandits will lead to erroneous analysis for linear bandits.

2.1 Bounding stochastic processes with correlated stopping times

In this subsection we will give an example of bounding a stochastic process of i.i.d. random variables with correlated stopping times. In the next subsection we will attempt to apply the same analysis to the linear bandit problem in which the random variables are not i.i.d. to highlight why the standard analysis techniques from stochastic bandits cannot be applied to linear bandits. To help clarify the analysis that follows we provide definitions for a stochastic process and the stopping time of a stochastic process.

Definition 1. *Stochastic Process. Given a probability space $(\Omega, \mathcal{F}, P)$ where Ω is a sample space, $\mathcal{F}$ is a set of events, and P is a mapping from an event to a probability, a stochastic process is a sequence of random variables $\boldsymbol{Z} = \{Z_t : t \in T\}$ where T is the index set.*

Definition 2. *Stopping Time. Given a stochastic process $\boldsymbol{Z} = \{Z_n : n \geq 1\}$, a stopping time with respect to $\boldsymbol{Z}$ is a random time τ such that for each $n \geq 1$ the event $\{\tau = n\}$ is completely determined by at most the total information known up to time n, i.e. $\{Z_1, \ldots, Z_n\}$.*

Consider a stochastic process $\mathbf{Z} = \{Z_1, \ldots, Z_n\}$ of i.i.d. random variables defined as

$$Z_i = \begin{cases} 1, & w.p.\ \frac{1}{2} \\ -1, & w.p.\ \frac{1}{2} \end{cases}.$$

Given a stopping time τ the empirical mean of the stochastic process is defined as $\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i$. We will now consider the expectation of the stochastic process when there exists a fixed stopping time and when the stopping time is correlated with the stochastic process.

1. Considering a fixed stopping time $\tau = t \in n$, where n is the index set of the stochastic process, the expectation is trivial:
$$\mathbb{E}[\frac{1}{t}\sum_{i=1}^{t} Z_i] = 0.$$

2. Considering a stopping time that is correlated with the stochastic process $\tau = \min\{t : \sum_{i=1}^{t} Z_i = 1\}$:
$$\mathbb{E}[\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i] = \mathbb{E}[\frac{1}{\tau}] = \sum_{t=1}^{\infty}\frac{1}{t}P(\tau = t) = \frac{1}{1}P(Z_1 = 1) + \sum_{2}^{\infty}\frac{1}{t}P(\tau = t) > \frac{1}{2}.$$
Also note that $\mathbb{E}[\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i] < 1$. In fact, $\mathbb{E}[\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i] = \mathbb{E}[\frac{1}{\tau}] \approx 0.575$, which indicates the empirical mean is biased.

Now we consider bounding the stochastic process by first developing a bound for a fixed stopping time and then extending the analysis to the correlated stopping time t. The analysis to bound the stochastic process under a fixed stopping time is as follows:

$$\begin{aligned}
P(\frac{1}{t}\sum_{i=1}^{t} Z_i > \epsilon) &= P(\exp(\lambda\sum_{i=1}^{t} Z_i) > \exp(\lambda t\epsilon)) && \text{for any } \lambda > 0 \\
&\leq \exp(-\lambda t\epsilon)\mathbb{E}[\exp(\lambda\sum_{i=1}^{t} Z_i)] && \text{Chernoff Bound} \\
&= \exp(-\lambda t\epsilon)\prod_{i=1}^{t}\mathbb{E}[\exp(\lambda Z_i)] && \text{independence of } Z_i\text{'s} \\
&\leq \exp(-\lambda t\epsilon)(\exp(\lambda^2/2))^t && \text{Hoeffding's Lemma} \\
&= \exp(-\lambda t\epsilon + \lambda^2 t/2).
\end{aligned}$$

Letting $\lambda = \epsilon$ we get a bound on the stochastic process with a fixed stopping time

$$P(\frac{1}{t}\sum_{i=1}^{t} Z_i > \epsilon) \leq e^{-t\epsilon^2/2}.$$

Letting $\delta = e^{-t\epsilon^2/2}$ we get with probability at least $1-\delta$,

$$\frac{1}{t}\sum_{i=1}^{t} Z_i \leq \sqrt{\frac{2\log(\frac{1}{\delta})}{t}}$$

We can now develop a bound for the stochastic process when the stopping time is correlated with the stochastic process using the union bound. The intuition here is, as definition $\tau = \min\{t : \sum_{i=1}^{t} Z_i = 1\}$, τ is not deterministic. So we derive a bound holds for any arbitrary τ simultaneously.

$$\begin{aligned}
P(\bigcup_{\tau=1}^{\infty}\{\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i > \epsilon_\tau\}) &\leq \sum_{\tau=1}^{\infty} P(\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i > \epsilon_\tau) \\
&\leq \sum_{\tau=1}^{\infty} e^{-\tau\epsilon_\tau^2/2} \\
&= \sum_{\tau=1}^{\infty} \frac{\delta}{2\tau^2} \\
&= \delta\frac{1}{2}\sum_{\tau=1}^{\infty}\frac{1}{\tau^2} \\
&\leq \delta.
\end{aligned}$$

Now letting $\epsilon_\tau = \sqrt{\frac{2\log(2\tau^2/\delta)}{\tau}}$ we get with probability at least $1-\delta$

$$\frac{1}{\tau}\sum_{i=1}^{\tau} Z_i \leq \sqrt{\frac{2\log(2\tau^2/\delta)}{\tau}}.$$

As we presented above, we were able to bound the stochastic process with i.i.d. random variables with correlated stopping times. Now, we will apply the same technique to the linear bandits, where arms are correlated and show that it in fact the analysis method is not applicable demonstrating that new techniques that must be developed for the problem.

2.2 Bounding linear bandits with correlated stopping times

We now apply the analysis of the previous section that was used to bound a stochastic process of i.i.d. random variables to the linear bandit problem. In doing so, we elucidate the complications that arise in the linear bandit framework due to the dependence between arm plays. We briefly review the linear bandit learning model for clarity purposes.

In each round of the linear bandit problem, a learning agent plays an arm $x_i \in \mathbb{R}^d$ and receives a reward $y_i \in \mathbb{R}$ that is a linear function of a parameter vector $\theta_* \in \mathbb{R}^d$ and a noise parameter $\eta_i \in \mathbb{R}$. Formally, the reward the learning agent receives from playing an arm is given by $y_i = x_i^T\theta_* + \eta_i$ where η_i is drawn from a zero-mean sub-Gaussian distribution meaning $\mathbb{E}[\eta_i] = 0$ and $\mathbb{E}[e^{\lambda\eta_i}] \leq e^{\lambda^2/2}$. Suppose each arm chosen at specific round is independent of previous choices, meaning that the sequence of arm choices can be chosen before the start of the game. Given a sequence of arm choices, observed rewards, and noise parameters $\{x_i, y_i, \eta_i\}_{i=1}^t$ we denote the stacked sequences of each as $X \in \mathbb{R}^{t\times d}, Y \in \mathbb{R}^t$, and $\eta \in \mathbb{R}^t$ respectively.

Using this information we can derive a least-squares estimate of θ_* given as follows

$$\hat{\theta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\theta_* + \eta) = \theta_* + (X^TX)^{-1}X^T\eta.$$

Thus the difference between the parameter estimate $\hat{\theta}$ and the true parameters θ_* is

$$\hat{\theta} - \theta_* = (X^T X)^{-1} X^T \eta.$$

Now we will consider some fixed arm choice $x \in \mathbb{R}^d$ and denote $w^T = x^T (X^T X)^{-1} X^T$. Note that for all t, w_t is a deterministic vector. In the linear bandit problem we are interested in bounding $x^T(\hat{\theta} - \theta_*) = x^T (X^T X)^{-1} X^T \eta = w^T \eta$. In the analysis that follows we will apply the methods from the previous section in an attempt to develop a bound, and most importantly we will identify where these methods fail for the linear bandit problem. As in the previous analysis we begin with a fixed stopping time and derive a bound. For an arbitrary constant μ,

$$\begin{aligned}
P(x^T(\hat{\theta} - \theta_*) > \mu) &= P(w^T \eta > \mu) \\
&\leq \exp(-\lambda\mu)\mathbb{E}[\exp(\lambda w^T \eta)], \quad \text{let } \lambda > 0 && \text{Chernoff Bound} \\
&= \exp(-\lambda\mu)\mathbb{E}[\exp(\lambda \sum_{i=1}^{t} w_i \eta_i)] \\
&= \exp(-\lambda\mu) \prod_{i=1}^{t} \mathbb{E}[\exp(\lambda w_i \eta_i)] && \text{independence of } w_i\eta_i \qquad (1) \\
&\leq \exp(-\lambda\mu) \prod_{i=1}^{t} \exp(\lambda^2 w_i^2/2) && \text{sub-Gaussian assumption} \\
&= \exp(-\lambda\mu) \exp(\frac{\lambda^2}{2} ||w||_2^2) \\
&\leq \exp(-\frac{\mu^2}{2||w||_2^2}) && \lambda = \frac{\mu}{||w||_2^2} \\
&= \exp(-\frac{\mu^2}{2x^T(X^T X)^{-1} x}) = \delta,
\end{aligned}$$

where in the final step we made use of the following equality

$$||w||_2^2 = x^T (X^T X)^{-1} X^T X (X^T X)^{-1} x = x^T (X^T X)^{-1} x.$$

Thus with probability at least $1 - \delta$,

$$x^T(\hat{\theta} - \theta_*) \leq \sqrt{2x^T (X^T X)^{-1} x \log(\frac{1}{\delta})}. \qquad (2)$$

We now consider a bound when the stopping time is correlated with the process using the bound for a fixed stopping time and applying the union bound:

$$\begin{aligned}
P(\bigcup_{t=1}^{\infty} \{x^T(\hat{\theta}_t - \theta_*) > \mu_t\}) &\leq \sum_{t=1}^{\infty} P(x^T(\hat{\theta}_t - \theta_*) > \mu_t) \\
&\leq \sum_{t=1}^{\infty} \exp(-\frac{\mu_t^2}{2x^T (X_t^T X_t)^{-1} x}) \\
&= \sum_{t=1}^{\infty} \frac{\delta}{2t^2} \leq \delta
\end{aligned}$$

Now letting $\mu_t = \sqrt{2x^T (X_t^T X_t)^{-1} x \log(\frac{2t^2}{\delta})}$ we get with probability at least $1 - \delta$,

$$x^T(\hat{\theta}_t - \theta_*) \leq \sqrt{2x^T (X_t^T X_t)^{-1} x \log(\frac{2t^2}{\delta})}. \qquad (3)$$

Using the methods we developed for the stochastic process of i.i.d. random variables we developed a bound for the linear bandit problem. When the sequence of arm choices $\{x_i\}_{i=1}^t$ are selected before the start of the game, this analysis is correct and (3) holds for correlated stopping times. However, when x_t depends on $\{x_i, y_i\}_{i=1}^{t-1}$, which is usually the case in linear bandit, our claim is erroneous. We now detail where the analysis breaks down before giving an explicit counterexample contradicting the preceding result.

Specifically the following step in our linear bandit analysis from (1) is not applicable.

$$\exp(-\lambda\mu)\mathbb{E}[\exp(\lambda\sum_{i=1}^{t} w_i\eta_i)] = \exp(-\lambda\mu)\prod_{i=1}^{t}\mathbb{E}[\exp(\lambda w_i\eta_i)]$$

Before, when $\{x_i\}_{i=1}^t$ were chosen prior to the start of the game, we had that $w_i\eta_i$ are independent so that $\mathbb{E}_\eta[\prod_i \exp(w_i\eta_i)] = \prod_i \mathbb{E}_\eta[\exp(w_i\eta_i)]$ holds. This equality is correct when w_i are constants. For example, consider that we are doing batch learning and the data is fixed beforehand. We could then treat $w^T = x^T(X^TX)^{-1}X^T$ as a constant and all of the analysis would be applicable. However, if we are sampling and collecting data adaptively the choice of arms depends on previous choices. This is because each arm choice provides information about θ_* so with each arm choice we gain information about every other available arm as well. It is clear that due to this each $w_i\eta_i$ are simply not independent, leading to the problematic analysis.

We now present an explicit example to prove that the preceding analysis cannot be applied to the linear bandit problem. Consider that there are d identical arms $(e_1, \ldots, e_d) \in \mathbb{R}^d$, where d is still the dimension of arms but since each arm is orthogonal to others, we have d arms. $\theta_* = 0, \eta_i = \begin{cases} 1, & \text{w.p. } 0.5 \\ -1, & \text{w.p. } 0.5 \end{cases}$ and the learning agent plays according to the procedure given in Algorithm 1.

Algorithm 1 Correlated stopping time

```
t = 1
for i = 1, 2, ..., d do
    while sum of rewards ≠ 1 do keep sampling arm x_t = e_i
        t ← t + 1
    end while
end for
τ = t − 1
```

The least-squares estimate of the parameters at time s is then given by

$$\hat{\theta_i} = \frac{\sum_{t=1}^{s}\mathbf{1}\{x_t = e_i\}\eta_t}{\sum_{t=1}^{s}\mathbf{1}\{x_t = e_i\}} = \frac{1}{T_i}$$

where $T_i = \sum_{t=1}^{s}\mathbf{1}\{x_t = e_i\}$. Set $x = \mathbf{1}$, i.e. a d-dimensional vector of ones, so $x^T(\hat{\theta}_\tau - \theta_*) = \sum_{i=1}^{d}\frac{1}{T_i}$, and

$$x^T(X_t^TX_t)^{-1}x = x^T(\sum_{i=1}^{d} e_ie_i^TT_i)^{-1}x = x^T(\sum_{i=1}^{d} e_ie_i^T\frac{1}{T_i})x = \sum_{i=1}^{d}\frac{1}{T_i}$$

where T_i is defined as number of times the i-th arm being pulled according to Algorithm 1.

$$\frac{x^T(\hat{\theta_\tau} - \theta_*)}{\sqrt{x^T(X_t^TX_t)^{-1}x}} = \sqrt{\sum_{i=1}^{d}\frac{1}{T_i}} \gtrapprox \sqrt{0.575d}$$

Comparing with the previous analysis, specifically (2), we see that when the arm plays are correlated, the right hand side of the above display scales like $\sqrt{d}$ but no such factor is present in (2). Thus, we cannot hope to achieve the same bound without a d dependence as in (2 when the arm plays are dependent. In the next section, we present a bound that resolves this issue.

We have now shown the additional difficulties that arise in the analysis of linear bandits due to the dependence between arms. In the sections that follow we present the novel techniques that are used to handle the dependence in order to develop correct analysis and discuss and prove the best results for the problem setting to date.

3 Linear Bandits

We will focus on the state-of-the-art methods for the linear bandit problem developed in Abbasi et al. (2011) [4]. The results in this work improved the previously best regret bound give in Dani et al. (2008) [2] by roughly a factor of $\sqrt{\log(n)}$, where n is the horizon, and experiments show that the empirical performance gain is even greater. The improved analysis hinges on the construction of smaller confidence sets by handling the dependence between arms through a novel tail inequality for vector-valued martingales which is proven using the theory of self-normalized processes. The tail inequality enables bounds to be derived that hold uniformly in time due the self-normalized form and the property that it holds for stopped martingales. This immediately saves a factor of $\log(n)$ in the regret by avoiding a union bound. The generality of this allows for improvements of any deviation bound that uses a union bound over time. We will elaborate further on this key innovation in Section 3.3. Finally, the confidence sets constructed in [4] save some worst case quantities from the bounds in [2] by replacing them with empirical quantities that are always smaller.

3.1 Main Results

The best results for the linear bandit problem at present come from [4]. We briefly outline the main results in the paper and compare these results with prior work in this section. In the sections that follow we provide more detail on the algorithmic approach as well as the anytime and problem dependent regret bounds.

As we noted previously, the stochastic bandit problem can be considered as a special case of the linear bandit problem. Due to this, the methods developed for the linear bandit problem are general enough to be applied to the stochastic bandit problem. The approach to constructing confidence bounds in [4] results in a modification of the UCB algorithm for the d-armed bandit problem from Auer et al. (2002) [6]. The modified UCB algorithm has regret of $O(d\log(1/\delta)/\Delta)$ with probability $1-\delta$, where Δ is the gap between the expected rewards of the two best arms. This bound does not depend on n which seemingly contradicts the seminal work of Lai and Robbins in [7], but since the algorithm takes δ as an input letting $\delta = 1/n$ the expected regret bound yields the same result as in [6] of $O((d\log n)/\Delta)$. This results says that the modified UCB algorithm achieves with high probability constant regret.

In the case of the linear bandit problem, [4] modifies the ConfidenceBall algorithm developed in [2] by using the new approach to constructing confidence sets outlined in section 1.1. The resulting algorithm reduces the regret bound from at most $O(d\log(n)\sqrt{n\log(n/\delta)})$ to at most $O(d\log(n)\sqrt{n}+\sqrt{dn\log(n/\delta)})$ with at least probability $1-\delta$. This is roughly an improvement of a multiplicative factor $\sqrt{\log(n)}$. Moreover, the regret of the problem dependent bound from [2] is reduced from $O(\frac{d^2}{\Delta}\log(n\delta)\log^2(n))$ to $O(\frac{\log(1/\delta)}{\Delta}(\log(n)+d\log\log n)^2)$. Here Δ is a generalized definition of the gap when the decision set is possibly infinite. Specifically, Δ can be interpreted as the gap between the values of the best and second best extremal points of the decision set. In the case that the decision set is finite, Δ is exactly the same as in the stochastic bandit problem. It is worth noting that for some decision sets, e.g., a sphere, $\Delta = 0$ in which case the problem dependent bound is not applicable.

3.2 Algorithmic Approach

The approach to solving the linear bandit problem utilizes the OFU principle that is the standard approach in the stochastic bandit problem. The main idea of this approach in the linear bandit problem is that at each time t a confidence set $C_{t-1} \subseteq \mathbb{R}^d$ is maintained in which θ_* lies with high probability using the history of arm choices and observed rewards up until time t given by the sequences $X_t, \dots, X_{t-1}$ and $Y_t, \dots, Y_{t-1}$ respectively. At time t the agent makes the most optimistic estimate of θ which is $\tilde{\theta}_t = \arg\max_{\theta \in C_{t-1}} (\max_{x \in D_t} \langle x, \theta \rangle)$ in order to make an arm selection of $X_t = \arg\max_{x \in D_t} \langle x, \tilde{\theta}_t \rangle$ that maximizes the reward according to the optimistic estimate $\tilde{\theta}$. The previous notation was used for clarity, but we can more compactly consider the process of choosing an estimate of θ and an arm by considering the joint optimization

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta) \in D_t \times C_{t-1}}{\arg\max} \langle x, \theta \rangle.$$

This joint optimization forms the decision step of the algorithm used in the linear bandit problem. Following the decision the agent plays the arm given by the strategy and observes a reward which is then used to update the confidence sets before the end of the round. The complete algorithm is given in Algorithm 2 where for right now we have abstracted away how the confidence sets are constructed as that is the main technical problem to solve.

Algorithm 2 OFUL Algorithm

1: **for** $t := 1, 2, \dots$ **do**
2: $(X_t, \tilde{\theta}_t) = \arg\max_{(x,\theta) \in D_t \times C_{t-1}} \langle x, \theta \rangle$
3: Play X_t and observe reward Y_t
4: Update C_t
5: **end for**

3.3 Self-Normalized Tail Inequality for Vector-Valued Martingales

In this section we will go through the methods used to develop the self-normalized tail inequality for vector-valued martingales. This inequality is crucial to developing the confidence sets used in [4].

3.3.1 Preliminaries

Before we get delve into the theorem for the self-normalized tail inequality for vector valued martingales we review some probability and measure theory for clarity purposes.

Definition 3. *σ-algebra. Given a set X then the σ-algebra $\mathcal{F}$ is a nonempty collection of subsets of X such that the following properties hold*

1. *X is in $\mathcal{F}$.*
2. *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$. Equivalently, $\mathcal{F}$ is closed under complement.*
3. *If $A_1, A_2, \dots$ is a countable collection of sets in $\mathcal{F}$ then their union $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$. Equivalently, $\mathcal{F}$ is closed under countable unions.*

Note that several other properties of a σ-algebra immediately follow from these properties including $\emptyset \in \mathcal{F}$ and $\mathcal{F}$ is closed under countable intersections.

Definition 4. *$\mathcal{F}$-Measurable. Let the pair $(X, \mathcal{F})$ where $\mathcal{F}$ is a σ-algebra of subsets of X denote a measurable space. Then a map $f : (X, \mathcal{F}) \to \mathbb{R}$ is $\mathcal{F}$-measurable if, and only if, $f^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}^*$. Here we are using $\mathcal{B}^*$ to denote the the extended Borel sets of $\mathbb{R}^* = \{x \in \mathbb{R} | x \neq 0\}$ which is the set of unions of sets from the Borel sets $\mathcal{B}$ with subsets of $\{-\infty, \infty\}$. Recall a Borel set is any set in a topological space that*

can be formed from open sets through the operations of countable union, countable intersection, and relative complement.

Definition 5. *Filtration. A family of σ-algebras $\mathcal{F}_t$ is defined to be a filtration if $\mathcal{F}_{t_1} \subset \mathcal{F}_{t_2}$ whenever $t_1 \leq t_2$. A stochastic process $\{X_t\}$ is said to be adapted to filtration $\{\mathcal{F}_t\}$ if $X_t \in \mathcal{F}_t$ for every t.*

Definition 6. *Martingale. A stochastic process $\{X_t\}$ adapted to filtration $\{\mathcal{F}_t\}$ is defined to be a martingale if*

1. $\mathbb{E}[|X_t|] < \infty \; \forall \; t$.
2. $\mathbb{E}[X_t|\mathcal{F}_s] = X_s \; \forall \; s < t$.

3.3.2 Assumptions

We assume the decision sets $\{D_t\}_{t=1}^{\infty}$ can be arbitrary. As a result of this assumption, the sequence of actions $X_t \in D_t$ is also arbitrary. This creates complicated stochastic dependencies so we will drop any assumptions on $\{X_t\}_{t=1}^{\infty}$ and instead develop a more general result.

We consider the σ-algebra $\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_{t+1}, \eta_1, \eta_2, \ldots, \eta_t)$ so that X_t is $\mathcal{F}_{t-1}$-measurable and η_t is $\mathcal{F}_t$-measurable. Relaxing this assumption we assume that $\{\mathcal{F}_t\}_{t=0}^{\infty}$ is any filtration of σ-algebras such that for any $t \geq 1$, X_t is $\mathcal{F}_{t-1}$-measurable and η_t is $\mathcal{F}_t$-measurable which means that $Y_t = \langle X_t, \theta_* \rangle + \eta_t$ is $\mathcal{F}_t$-measurable. Under these assumptions the sequence $\{S_t\}_{t=0}^{\infty}, S_t = \sum_{s=1}^{t} \eta_t X_t$, is a martingale with respect to $\{\mathcal{F}_t\}_{t=0}^{\infty}$. This property is needed for the construction of the confidence sets for θ_*.

3.3.3 Result

Given the assumptions of section 3.3.2 we have the following result that is key in the construction of the confidence sets.

Theorem 1. *Self-Normalized Bound for Vector-Valued Martingales. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R > 0$ i.e.*

$$\forall \; \lambda \in \mathbb{R} \qquad \mathbb{E}[e^{\lambda \eta_t} | X_{1:t}, \eta_{1:t-1}] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_{t-1}$-measurable. Assume that V is a $d \times d$ positive definite matrix. For any $t \geq 0$ define

$$\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^T \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$

$$\|S_t\|^2_{\overline{V}_t^{-1}} \leq 2R^2 \log\left(\frac{det(\overline{V}_t)^{1/2} det(V_t)^{1/2}}{\delta}\right).$$

Here we are using the notation $\|S_t\|_{\overline{V}_t^{-1}}$ to denote the weighted norm $\sqrt{S_t^T \overline{V}_t^{-1} S_t}$. The name self-normalized bound comes about because this norm is the deviation of the martingale and it is weighted by the matrix $\overline{V}_t^{-1}$ which is also derived from the martingale.

We omit the proof of this theorem in these notes, but we refer the interested reader to Appendix A of [4] for the complete proof. We also refer the reader to Pena et al. (2004) [8] and Pena et al. (2008) [9] where much of the theory from self-normalized processes is contained that builds the bulk of the proof. The outline of the proof is that a supermartingale is defined which allows for a self-normalized bound for vector-valued martingales to be developed using the method of mixtures technique which then enables a stopping time construction that leads to the theorem.

3.4 Construction of Confidence Sets

In this section we will examine the method of constructing confidence sets developed in [4] as well as compare the results with those from [2] and [3]. The construction of confidence sets is the crux of any upper confidence based algorithm because the regret of an algorithm is directly tied to the size of the confidence sets. The reduction in the confidence set provided in [4] is primarily due to the inequality provided in Theorem 1.

3.4.1 Results

Let $\hat{\theta}_t$ be the ℓ^2-regularized least-squares estimate of θ_* with regularization parameter $\lambda > 0$ given by

$$\hat{\theta}_t = \arg\min_{\theta} \|\mathbf{X}_{1:t}\theta - \mathbf{Y}_{1:t}\| + \lambda\|\theta\|_2^2 = (\mathbf{X}_{1:t}^T\mathbf{X}_{1:t} + \lambda I)^{-1}\mathbf{X}_{1:t}^T\mathbf{Y}_{1:t}$$

where we are denoting $\mathbf{X}_{1:t}$ as a matrix with rows $X_1^T, X_2^T, \ldots, X_t^T$ and $\mathbf{Y}_{1:t}$ as the vector $(Y_1, \ldots, Y_t)^T$. The following theorem says that with high probability θ_* lies with high probability in an ellipsoid with center at $\hat{\theta}_t$.

Theorem 2. *Confidence Ellipsoid. Assume the same as in Theorem 1, let $V = I\lambda, \lambda > 0$, define $Y_t = \langle X_t, \theta_t \rangle + \eta_t$ and assume that $\|\theta_*\| \leq S$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0, \theta_*$ lies in the set*

$$C_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq R\sqrt{2\log(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta})} + \lambda^{1/2}S \right\}.$$

Furthermore, if for all $t \geq 1, \|X_t\|_2 \leq L$ then with probability at least $1 - \delta$, for all $t \geq 0, \theta_*$ lies in the set

$$C_t' = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq R\sqrt{d\log(\frac{1 + tL^2/\lambda}{\delta})} + \lambda^{1/2}S \right\}.$$

It is important to note that the $\sqrt{d}$ term is necessary in the bound above. Concretely, this factor is the cost of selecting plays adaptively—which in turn creates a dependence between arm choices—as opposed to selecting plays independently according to a fixed sequence chosen a priori. To see this recall that in Section 2 we showed that if $\mathbf{X}_{1:t}$ is chosen adaptively we have $x^T(\hat{\theta} - \theta_*) \geq \sqrt{d}\|x\|_{(\mathbf{X}_{1:t}^T\mathbf{X}_{1:t})^{-1}}$. However, if $\mathbf{Z}_{1:t}$ is chosen independently we have $x^T(\hat{\theta} - \theta_*) \leq c\|x\|_{(\mathbf{Z}_{1:t}^T\mathbf{Z}_{1:t})^{-1}}$ where c is independent of d. This fact is especially key in pure exploration bandits. For more discussion on this topic we refer the interested reader to Soare et al. [10].

The bounds given above from [4] can be compared with those from [2] and [3]. In [4] under the same conditions with probability at least $1 - \delta$ the bound is

$$\text{for all } t \text{ large enough } \|\hat{\theta}_t - \theta_*\|_{\overline{V}_t} \leq R\max\left\{\sqrt{128d\log(t)\log(\frac{t^2}{\delta})}, \frac{8}{3}\log(\frac{t^2}{\delta})\right\},$$

where large enough means that t satisfies $0 < \delta < t^2e^{-1/16}$. In [3] the bound says that for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\forall\ t \geq 2, \qquad \|\hat{\theta}_t - \theta_*\|_{\overline{V}_t} \leq 2\kappa^2 R\sqrt{\log t}\sqrt{d\log(t) + \log t^2/\delta} + \lambda^{1/2}S.$$

This bound is improvement over [2] but is still not as tight as that from [4]. It is worth noting that the bound in Theorem 2 seems to require computing the determinant of a matrix, which can be computationally expensive. However, this operation can be sped up using the matrix determinant lemma, which exploits that the matrix whose determinant is needed can be obtained via a rank one update, allowing for the determinant to be updated in linear time.

We do not include the proof of this theorem in these notes, but refer the reader to Appendix B of [4] for the complete proof. The proof primarily relies on a Determinant-Trace-Inequality and the Self-Normalized Bound for Vector-Valued Martingales from Theorem 1.

3.5 Regret Analysis of the OFUL Algorithm

In this section we will give a bound on the regret of the OFUL algorithm using the confidence sets constructed in Theorem 2. We will assume that the expected rewards are bounded, which can be viewed as a bound on θ_* as well as a bound on the decision set D_t.

3.5.1 Results

Theorem 3. *Regret of the OFUL algorithm. Assume that for all t and all $x \in D_t, \langle x, \theta_* \rangle \in [-1, 1]$. Then with probability at least $1-\delta$, the regret of the OFUL algorithm satisfies*

$$\forall\ n \geq 0, R_n \leq 2\sqrt{2dn\log(1+nL^2/\lambda d)}(\lambda^{1/2}S + R\sqrt{d\log(\frac{1+nL^2/\lambda}{\delta})}).$$

As we have stated before this bound from [4] is an improvement of approximately a factor of $\sqrt{\log(n)}$ over that provided in [2]. The improvement comes from the construction of a tighter confidence set as discussed in Section 3.4 that was enabled by the inequality in Theorem 1.

Proof.

Lemma 1. *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V a $d \times d$ positive definite matrix and define $\overline{V}_t = V + \sum_{s=1}^{t} X_s X_s^T$. Then we have*

$$\log(\frac{\det(\overline{V}_n)}{\det(V)}) \leq \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}}.$$

Further, if $\|X_t\|_2 \leq L$ for all t then

$$\sum_{t=1}^{n} \min\{1, \|X_t\|^2_{\overline{V}_{t-1}^{-1}}\} \leq 2(\log\det(\overline{V}_n) - \log\det V) \leq 2(d\log((trace(V)) + nL^2)/d) - \log\det V),$$

and if $\lambda_{\min}(V) \geq \max(1, L^2)$ then

$$\sum_{i=1}^{n} \|X_t\|^2_{\overline{V}_{t-1}^{-1}} \leq 2\log(\frac{\det(\overline{V}_n)}{\det(V)}).$$

Proof. We first prove the first inequality by decomposing $\det(\overline{V})$. From the definition of $\overline{V}$ we have

$$\det(\overline{V}_n) = \det(\overline{V}_{n-1} + X_n X_n^T).$$

A consequence of Sylvester's determinant theorem gives

$$\det(\overline{V}_{n-1} + X_n X_n^T) = \det(\overline{V}_{n-1})(1 + X_n^T \overline{V}_{n-1}^{-1} X_n).$$

Now we can use the same techniques to decompose $\det(\overline{V}_{n-1})$ as follows

$$\det(\overline{V}_{n-1}) = \det(\overline{V}_{n-2} + X_{n-1} X_{n-1}^T) = \det(\overline{V}_{n-2})(1 + X_{n-1}^T \overline{V}_{n-2}^{-1} X_{n-1}).$$

Thus we have a recursive relationship that will terminate when $\overline{V}_0 = V$ which gives us the result

$$\begin{aligned}\det(\overline{V}_n) &= \det(V)\prod_{t=1}^{n}(1 + X_t \overline{V}_{t-1}^{-1} X_t^T) \\ &= \det(V)\prod_{t=1}^{n}(1 + \|X_t\|^2_{\overline{V}_{t-1}^{-1}}).\end{aligned}$$

We can bound $\log(\det(\overline{V}_t))$ by making use of the inequality $\log(1+t) \leq t$.

$$\log(\det(\overline{V}_t)) = \log(\det(V)) + \log(\sum_{t=1}^{t} 1 + \|X_t\|^2_{\overline{V}^{-1}_{t-1}}))$$
$$\leq \log(\det(V)) + \sum_{t=1}^{t} \|X_t\|^2_{\overline{V}^{-1}_{t-1}}$$

Now the inequality $x \leq 2\log(1+x)$, which holds when $x \in [0,1]$, can be combined with the result $\det(\overline{V}_n) = \det(V)\prod_{t=1}^{n}(1+\|X_t\|^2_{\overline{V}^{-1}_{t-1}})$ to get

$$\sum_{t=1}^{n} \min\{1, \|X_t\|^2_{\overline{V}^{-1}_{t-1}}\} \leq 2\sum_{t=1}^{n} \log(1+\|X_t\|^2_{\overline{V}^{-1}_{t-1}}) = 2(\log(\det(\overline{V}_t)) - \log(\det(V))).$$

Now note that the trace of $\overline{V}_n$ is bounded by $\text{trace}(V) + nL^2$ if $\|X_t\|_2 \leq L$ which follows by simply applying the trace operator to the definition of $\overline{V}_n$. The inequality $\det(A)^{1/n} \leq \frac{1}{n}\text{tr}(A)$ represents the fact that the geometric mean is less than the arithmetic mean. Applying this inequality we can give the following upper bound on $\det(\overline{V}_n)$.

$$\det(\overline{V}_n) = \prod_{i=1}^{d} \lambda_i \leq (\frac{\text{trace}(V) + tL^2}{d})^d.$$

It immediately follows that

$$\log(\det(\overline{V}_n)) \leq d\log((\text{trace}(V) + tL^2)/d),$$

which completes the proof of the second inequality in the Lemma. We can also bound the sum $\sum_{t=1}^{n} \|X_t\|^2_{\overline{V}^{-1}_{t-1}}$ as a function of $\log\det(\overline{V}_t)$ given that $\lambda_{\min}(V)$ is large enough. From the inequality

$$\|X_t\|^2_{\overline{V}^{-1}_{t-1}} \leq \lambda^{-1}_{\min}(\overline{V}_{t-1})\|X_{t-1}\|^2 \leq L^2/\lambda_{\min}(V),$$

we can observe that if $\lambda_{\min}(V) \geq \max(1, L^2)$, then

$$\log(\frac{\det(\overline{V}_n)}{\det(V)}) \leq \sum_{t=1}^{n} \|X_t\|^2_{\overline{V}^{-1}_{t-1}} \leq 2\log(\frac{\det(\overline{V}_n)}{\det(V)}).$$

This proves the first and third inequalities in the Lemma. □

We can now make use of the Lemma in order to bound the instantaneous regret.

$$r_t = \langle \theta_*, x_* \rangle - \langle \theta_*, X_t \rangle.$$

Because we choose $(X_t, \tilde{\theta}_t)$ optimistically according to

$$(X_t, \tilde{\theta}_t) = \underset{(x,\theta) \in D_t \times C_{t-1}}{\arg\max} \langle x, \theta \rangle.$$

we have

$$r_t \leq \langle \tilde{\theta}_t, X_t \rangle - \langle \theta_*, X_t \rangle.$$

Using the linearity of the inner product as well as adding and subtracting $\langle \hat{\theta}_{t-1}, X_t \rangle$ we get

$$r_t = \langle \tilde{\theta}_t - \theta_*, X_t \rangle$$
$$= \langle \tilde{\theta}_t - \hat{\theta}_{t-1}, X_t \rangle + \langle \hat{\theta}_{t-1} - \theta_*, X_t \rangle.$$

Now we work towards a form of the regret that will allow us to make use of Theorem 2 regarding the confidence ellipsoid with several algebraic manipulations.

$$\begin{aligned}
r_t &= (\tilde{\theta}_t - \hat{\theta}_{t-1})^T \overline{V}_{t-1}^{1/2} \overline{V}_{t-1}^{-1/2} X_t + (\hat{\theta}_{t-1} - \theta_*)^T \overline{V}_{t-1}^{1/2} \overline{V}_{t-1}^{-1/2} X_t \\
&= (\overline{V}_{t-1}^{1/2}(\tilde{\theta}_t - \hat{\theta}_{t-1}))^T (\overline{V}_{t-1}^{-1/2} X_t) + (\overline{V}_{t-1}^{1/2}(\hat{\theta}_{t-1} - \theta_*))^T (\overline{V}_{t-1}^{-1/2} X_t) \\
&\leq \|\overline{V}_{t-1}^{1/2}(\tilde{\theta}_t - \hat{\theta}_{t-1})\|_2 \|\overline{V}_{t-1}^{-1/2} X_t\|_2 + \|\overline{V}_{t-1}^{1/2}(\hat{\theta}_{t-1} - \theta_*)\|_2 \|\overline{V}_{t-1}^{-1/2} X_t\|_2 \quad \text{(Cauchy-Schwarz Inequality)} \\
&= \|\tilde{\theta}_t - \hat{\theta}_{t-1}\|_{\overline{V}_{t-1}} \|X_t\|_{\overline{V}_{t-1}^{-1}} + \|\hat{\theta}_{t-1} - \theta_*\|_{\overline{V}_{t-1}} \|X_t\|_{\overline{V}_{t-1}^{-1}}.
\end{aligned}$$

Recall Theorem 2 for the confidence ellipsoid which states the following: If for all $t \geq 1$ we have $\|X_t\|_2 \leq L$ then with probability at least $1 - \delta$, for all $t \geq 0$, θ_* lies in the set

$$C_t' = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq R\sqrt{d \log(\frac{1 + tL^2/\lambda}{\delta})} + \lambda^{1/2} S \right\},$$

and equivalently

$$C_t' = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta} - \theta\|_{\overline{V}_t} \leq \sqrt{\beta_t} + \lambda^{1/2} S \right\}.$$

Applying Theorem 2 to the instantaneous regret gives

$$r_t \leq 2(\sqrt{\beta_{t-1}} + \lambda^{1/2} S) \|X_t\|_{\overline{V}_{t-1}^{-1}}.$$

Using this result along with the assumed fact that $r_t \leq 2$ we get

$$r_t \leq 2 \min\{(\sqrt{\beta_{t-1}} + \lambda^{1/2} S) \|X_t\|_{\overline{V}_{t-1}^{-1}}, 1\} \leq 2(\sqrt{\beta_{t-1}} + \lambda^{1/2} S) \min\{\|X_t\|_{\overline{V}_{t-1}^{-1}}, 1\}.$$

By Jensen's inequality, for any positive numbers $a_1, \ldots, a_n$ we have $\sum_{a=1}^n a_i \leq \sqrt{n \sum_{i=1}^n a_i^2}$, thus we have

$$R_n = \sum_{t=1}^n r_t \leq \sqrt{n \sum_{t=1}^n r_t^2}.$$

Now plugging in our last inequality for the instantaneous regret we get

$$\begin{aligned}
R_n &\leq \sqrt{n \sum_{t=1}^n 4(\sqrt{\beta_{n-1}} + \lambda^{1/2} S)^2 \min\{\|X_t\|^2_{\overline{V}_{t-1}^{-1}}, 1\}}, \\
&\leq 2(\sqrt{\beta_{n-1}} + \lambda^{1/2} S) \sqrt{n \sum_{t=1}^n \min\{\|X_t\|^2_{\overline{V}_{t-1}^{-1}}, 1\}}.
\end{aligned} \tag{4}$$

Now from Lemma 1 we had

$$\sum_{t=1}^n \min\{\|X_t\|_{\overline{V}_{t-1}}, 1\} \leq 2(\log\det(\overline{V}_n) - \log\det V) \leq 2(d \log((\text{trace}(V) + nL^2)/d) - \log\det V) \tag{5}$$

and with $V = \lambda I$ we can substitute $\text{trace}(V) = \lambda d$ as well as $\log\det V = d \log\lambda$ to get

$$\sum_{t=1}^n \min\{\|X_t\|_{\overline{V}_{t-1}}, 1\} \leq 2(d \log((\lambda d + nL^2)/d) - d \log\lambda) = 2d \log(1 + nL^2/\lambda d). \tag{6}$$

Plugging this into the regret we get the final result

$$R_n \leq 2\sqrt{2dn \log(1 + nL^2/\lambda d)}(\lambda^{1/2} S + R\sqrt{d \log(\frac{1 + nL^2/\lambda}{\delta})}). \tag{7}$$

□

3.6 Problem Dependent Bound

Many bandit papers develop problem dependent bounds, where the regret includes a term Δ defined to be the gap between the two best arms. In the case of the linear bandit problem we can consider the gap at time t given by Δ_t as the difference between the rewards of the best and second best actions in the decision set D_t. Define the set of suboptimal extremal points as $D_{t-} = \{X_t \in D_t : \langle X_t, \theta_* \rangle < \langle x_t^*, \theta_* \rangle\}$ and define the gap $\Delta_t = \sup_{X_t \in D_{t-}} \langle x_t^*, \theta_* \rangle - \langle X_t, \theta_* \rangle$ [2]. We will consider the smallest gap $\Delta = \min_{1 \leq t \leq n} \Delta_t$.

3.6.1 Results

Theorem 4. *Assume that $\lambda \geq 1$ and $\|\theta_*\| \leq S$ where $S \geq 1$. With probability at least $1 - \delta$, for all $n \geq 1$, the regret of the OFUL algorithm satisfies*

$$R_n \leq \frac{8d}{\Delta} \log(1 + nL^2/\lambda d)(\lambda^{\frac{1}{2}} S + R\sqrt{d \log(\frac{1 + nL^2/\lambda}{d})})^2$$

This bound from [4] scales like $O(\frac{d^2}{\Delta} \log^2 n)$ where the bound from [2] scales like $O(\frac{d^2}{\Delta} \log^3 n)$.

Proof. According to the prove in the Theorem 3, we have that

$$\sum_{t=1}^{n} r_t^2 \leq 8d \log(1 + nL^2/\lambda d)(\lambda^{\frac{1}{2}} S + R\sqrt{d \log(\frac{1 + nL^2/\lambda}{d})})^2$$

and

$$R_n = \sum_{t=1}^{n} r_t \leq \sum_{t=1}^{n} \frac{r_t^2}{\Delta} \leq \frac{8d}{\Delta} \log(1 + nL^2/\lambda d)(\lambda^{\frac{1}{2}} S + R\sqrt{d \log(\frac{1 + nL^2/\lambda}{d})})^2$$

where the first inequality follows from the fact that either $r_t = 0$ or $\Delta \leq r_t$. □

3.7 Applications of Linear Bandits

The most prominent applications of linear bandits are in path routing and recommendation systems. In the literature several works have considered a linear bandit approach to path routing including with various information feedback structures including [11], [12], [13]. The general setup described in [14] considers a graph $G = (V, E)$ with n vertices and d edges. An arm $x \in \mathbb{R}^d$ is an incidence vector of a path, i.e. $x_e = 1$ if edge e belongs to the path and $x_e = 0$ otherwise and $D \subset \mathbb{R}^d$ is the collection of all incidence vectors of path on the graph. The parameter vector $\theta_* \in \mathbb{R}^d$ is designed such that θ_*^e is the delay on edge e. The delay of a path with incidence vector x is then given by $\langle x, \theta_* \rangle$. The problem then becomes to minimize the delay using the standard linear bandit framework.

Linear bandits have been widely applied to recommendation systems. In this application context is often considered as well to create a contextual linear bandit problem. The most significant work in this regard is [5]. However, the standard linear bandit framework that we have presented applies as well. Consider the setting of movie recommendations, an arm $x \in \mathbb{R}^d$ can be represented as characteristics of a movie that comes from a set $D \subset \mathbb{R}^d$ and the parameter vector $\theta_* \in \mathbb{R}^d$ can represent a users preferences for characteristics of a movie.

4 Sampling Methods

Here we will discuss another method that can be used to solve linear bandits. Thompson sampling (also called posterior sampling) is an algorithmic technique that is commonly used in linear bandit problems, but can also be applied to more general problems in sequential decision-making. We will first discuss its application to simple bandit problems, and then discuss its application to linear bandits and other problems.

4.1 Background

The basic idea behind Thompson sampling [15] (TS) is to assume a simple prior distribution on the parameters of the reward distribution of every arm, and choose arms based on samples from that distribution. At each time step, the algorithm draws a sample from the posterior distribution of each arm, selects the best arm according to the samples, and updates the distribution based on the observed reward. TS is a randomized probability matching algorithm, making it robust to being trapped in an early bad decision, and is empirically proven to be effective in many problems. It can be directly applied to solve linear bandit problems.

4.2 Special Case: Bernoulli Bandits

We start with the Beta-Bernoulli Bandit problem. Let's assume there are K arms, each with a reward of 1 with a probability of θ_k, and a reward of 0 with a probability of $1 - \theta_k$ when played. Thus θ_k can be interpreted as a mean reward, which is unknown but fixed over time.

Here we review briefly of the beta distribution. The pdf the beta distribution with parameters $\alpha > 0, \beta > 0$ is $p(\theta_k) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_k^{\alpha-1}(1-\theta_k)^{\beta-1}$ where Γ denotes the gamma function. It's convenient to work with this distribution because of its conjugacy properties, i.e., each arm's posterior distribution is also beta distribution that can be updated according to a simple rule.

$$(\alpha_k, \beta_k) = \begin{cases} (\alpha_k, \beta_k) & \text{if } x_i \neq k, \\ (\alpha_k, \beta_k) + (r_t, 1 - r_t) & \text{else} \end{cases}$$

Algorithm 3 Beta Bernoulli Thompson Algorithm (K, α, β) [16]

procedure BETABERNOULLI-THOMPSONSAMPLING(K, α_0, β_0)
  for t = 1,2,3,4,... **do**
    Sample model.
    for k = 1,2,..., K **do**
      Sample $\hat{\theta}_k \sim beta(\alpha_k, \beta_k)$
    end for
    Select and apply action:
    $x_t \leftarrow \text{argmax}_k \hat{\theta}_k$
    Apply x_t and observe r_t
    Update distribution:
    $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t}, \beta_{y_t}) + (r_t, 1 - r_t)$
  end for
end procedure

The important thing here is that the reward estimate $\hat{\theta}_k$ is randomly sampled from the posterior distribution, rather than the expectation $\alpha_k/(\alpha_k + \beta_k)$. To restate this, it does not sample $\hat{\theta}_k$ according to the distribution of y_t if kth arm is selected. Instead, $\hat{\theta}_k$ represents a probable success probability, not a probable observation.

In this setting, Thompson sampling generally performs better than the ϵ-greedy algorithm (the algorithm which applies the greedy action with probability 1 ϵ and otherwise selects an action uniformly at random) [16]. This is because Thomson sampling prioritizes exploration of the arms that are more likely to be optimal, and avoids arms that are known to have poor performance, unlike ϵ-greedy which has no bias in its exploration.

4.3 General Case: General Thomson Sampling

Here we move to a more general case. At each time step from $1, ..., T$, the agent applies an action $x_t \in \chi$. After applying the action x_t, the agent receives an observation $y_t \in Y$, which is generated from a conditional probability measure $q_\theta(\cdot|x_t)$. This is a distribution that is parameterized by $\theta \in \Theta$, an unknown variable associated with the instance of the bandit problem. The reward will be $r_t = r(y_t)$, where r is a function from Y to $\mathbb{R}$. The agent does not know θ, but does have a prior distribution p over Θ.

Thompson sampling starts by drawing a random sample from p (unlike the greedy algorithm, which will take $\hat{\theta}$ to be the expected value of θ). Then, it will select the action that maximizes the expected reward, which, given a finite set of observations Y, is $\mathbb{E}_{q_{\hat{\theta}}}\left[r(y_t)|x_t = x\right] = \sum_{o \in Y} q_{\hat{\theta}}(o|x) r(o)$. Then the distribution p is updated by conditioning on the observation $\hat{y}_t$. If Θ is a finite set, then this conditional distribution can be written by Bayes rule as:

$$\mathbb{P}_{p,q}(\theta = u|x_t, y_t) = \frac{p(u) q_u(y_t|x_t)}{\sum_{v \in \Theta} p(v) q_v(y_t|x_t)}$$

Algorithm 4 General Thomson Sampling (χ, p, q, r) [16]

procedure ThompsonSampling(χ, p, q, r)
  for t = 1,2,3,4,... **do**
    Sample model.
    Sample $\hat{\theta} \sim p$
    Select and apply action:
    $x_t \leftarrow \text{argmax}_{x \in \chi} \mathbb{E}_{q_{\hat{\theta}}}\left[r(y_t)|x_t = x\right]$
    Apply x_t and observe y_t
    Update distribution:
    $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot|x_t, y_t)$
  end for
end procedure

In the simplified Beta Bernoulli bandit problem, simple and efficient Bayesian inference was feasible because of its conjugacy property. However, many practical applications do not allow exact Bayesian inference to be tractable.

4.3.1 Regret bounds

The theoretical bounds for Thompson sampling has been given in recent studies [17]. The randomized setting of Thompson sampling, which differs from the UCB algorithm, that achieves exploration by adding a deterministic, non-negative bias inversely proportional to the number of plays to the observed empirical means.

Consider the N-armed stochastic bandit problem, where arm i has expected reward μ_i, and say that the arm with highest reward is arm 1. Here, the expected regret is defined as $\mathbb{E}[R(T)] = \mathbb{E}[\sum_{t=1}^{T}(\mu_1 - \mu_{x_t})]$. For details on the proof, see [17].

Theorem 5. *(Two-armed stochastic bandit regret bound for Thompson sampling)*

$$\mathbb{E}[R(T)] = O\left(\frac{\ln(T)}{\Delta} + \frac{1}{\Delta^3}\right)$$

where $\Delta = \mu_1 - \mu_2$

More generally,

Theorem 6. *(N-armed stochastic bandit regret bound for Thompson sampling)*

$$\mathbb{E}[R(T)] \leq O\left(\left(\sum_{a=2}^{N} \frac{1}{\triangle_a^2}\right) \ln T\right)$$

where $\Delta_i = \mu_1 - \mu_i$

4.4 Linear Bandits with Thompson Sampling

Let $\theta \in \mathbb{R}^m$ be drawn from a normal distribution. $N(\mu_0, \sum_0)$. There is a set of actions $\mathcal{X}$, where $\mathcal{X} \in \mathbb{R}^m$. At each time step t=1,...T, an action $x_t \in \mathcal{X}$ is selected, and a reward $y_t = \theta^T x_t + w_t$ is observed, where w_t, the random noise, is sampled from a zero-mean sub-Gaussian distribution. We can use Thompson Sampling for linear bandits by keeping a distribution on θ, and updating based on each observation.

- Prior distribution: This is a multivariate Gaussian with parameters μ_0 and Σ_0.
- Updates: say that we choose action x_t and observe a reward y_t. Then we update the mean and variance by:

  $$\Sigma_{t+1} = (\Sigma_t^{-1} + x_t x_t^T/\sigma_w^2)^{-1}$$

  and

  $$\mu_{t+1} = \Sigma_{t+1}(\Sigma_t^{-1}\mu_t + x_t(y_t + \tilde{w}_t^n/\sigma_w^2))$$

  These equations come from the Bayesian updates for a Gaussian distribution.
- Sampling: we sample $\hat{\theta}$ from $N(\mu_t, \Sigma_t)$, and choose action x_t that maximizes $\hat{\theta}^T x_t$. Depending on the nature of the set of actions $\mathcal{X}$, this can be done by linear programming or searching through the possibilities (if $\mathcal{X}$ is finite).
- Ensemble model: We keep N means and update them at each iteration: $\theta_t^1, ..., \theta_t^N$, while keeping one covariance matrix Σ_t. At each step we find the x_t and n that maximizes $\theta_t^{nT} x_t$.

4.5 Regret Bounds

When working with Thompson sampling, we often use the Bayes regret rather than the standard notion of regret. The Bayes regret is defined as the expected regret, where the expectation is defined with respect to a prior distribution on the reward functions. Any asymptotic bound on Bayesian regret is also an asymptotic bound on regret, even if the prior is misspecified. A formal definition is given below.

Definitions

$f_\theta(a) = \langle \phi(a), \theta \rangle$ is the reward function for action a, given the linear bandit parameter θ.

A_t is the action selected by the algorithm at time t.

π is a distribution over the action space generated by the Thompson sampling algorithm, from which actions are selected.

T is the number of iterations.

Then, the Bayes regret is defined as:

$$\text{BayesRegret}(T, \pi) = \sum_{t=1}^{T} \mathbb{E}_\theta \left[\max_a f_\theta(a) - f_\theta(A_t)\right]$$

Bayesian regret bounds for linear bandits The Bayes regret bound for Thompson sampling on linear bandits was proved in [18] to be the following:

Proposition 1. *BayesRegret*$(T, \pi^{TS}) = O(d \log T \sqrt{T})$.

d here is the dimensionality of θ. For details of the proof, see [18].

4.5.1 Connections with UCB

UCB and Thompson sampling both have similar regret bounds. For details on the regret bounds of Thompson sampling in the linear bandits case, see [18]. An advantage of UCB compared to the Thompson sampling is that UCB can be applied to any function, while Thompson sampling requires it to be sub-Gaussian distribution. On the other hand, the Thompson Sampling algorithms does not require an upper confidence bound function, and the action selection might be more computationally tractable.

Here we briefly review the regret bound for linear bandits of UCB algorithm (LinUCB) presented in Section 3.1. The resulting bound for linear bandit problem is $O(\frac{\log(1/\delta)}{\Delta}(\log(T) + d \log\log T)^2)$ [4], where T is the time horizon.

4.5.2 Approximations

In many cases, the distributions do not have conjugate priors and thus cannot be easily updated with closed-form updates.

In these cases, approximate posterior sampling mechanisms (such as Gibbs sampling, sampling from a Laplace approximation and the bootstrap) are used. For details, please refer to [16].

4.6 Information-Directed Sampling

Upper confidence bound algorithms (UCB) and Thompson sampling are very effective in many applications. However, these methods can fail when faced with more complex information structure, such as a sparase linear model, and Russo and Van Roy [19] suggest the Information-Directed Sampling (IDS) method to address this. Here, each action is sampled to minimize the ratio between squared expected single-period regret ($\triangle_t(\pi)^2$) and a measure of information gain: the mutual information between the optimal action and the next observation. First I will introduce some notation, and define the objective. The agent chooses from a set of actions $(A_t)_{t\in\mathbb{N}}$ from a finite action set A and observes outcomes $(Y_{t,A_t})_{t\in\mathbb{N}}$. A random outcome $Y_{t,a} \in Y$ associated with each action $a \in A$ and time $t \in \mathbb{N}$. A random variable θ such that, conditioned on it, $(Y_t)t \in N$ is an iid. We denote the set of probability distributions over A by $D(A)$. There's a reward $R(y)$ with each outcome from a fixed and known function. We study the expected regret over the time up to T, which is, Over all action sampling distributions $\pi \in D(A)$, the ratio between the square of expected regret and information gain about the optimal action A^*. In particular, the policy π^{IDS} is defined by $\mathbb{E}[\text{Regret}(T)] = \mathbb{E}[\sum_{t=1}^{T}(R_{t,A^*} - R_{t,A_t})]$. This is called as ***Bayesian regret or Bayes risk***.

Information directed sampling balances two terms: $g_t(\pi)$: information gain and $\triangle_t(\pi)^2$: the expected instantaneous regret of action a at time t. $g_t(a)$ equals the expected reduction in entropy of the posterior distribution of A^*, and detail on more precise definition of both $g_t(a)$ and $\triangle_t(\pi)^2$ can be found in [19]. Then IDS is defined as $\pi_t^{IDS} \in \text{argmin}_{\pi\in D(A)}\Psi_{IDS} = \frac{\triangle_t(\pi)^2}{g_t(\pi)}$. Then the optimization is to minimize

$$\Psi(\pi) = (\pi^T \triangle)^2 / \pi^T g, \text{ subject to } \pi^T e = 1, \pi \geq 0.$$

$\Psi_t(\pi)$ is the information ratio of sampling distribution π, which roughly measures the "cost" paid for information. For more details, see [19].

4.7 Conclusion

Most popular approaches to linear bandit problems are to extend upper-confidence bound (UCB) algorithms and Thompson sampling, as both can give a strong performance guarantees. For some problem cases, however, both do not actively seek for information that can lead to poor performance. For specific examples on where UCB and Thompson sampling fails but IDS can successfully reach optimal solution, such as sparse linear bandit problems, look at [19]. Even in cases such as Bernouilli bandits where Thompson sampling is known to be very effective, [19] shows IDS can outperform UCB and linear bandit methods.

References

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

[2] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[3] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

[5] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[7] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[8] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, pages 1902–1933, 2004.

[9] Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008.

[10] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836, 2014.

[11] András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct):2369–2403, 2007.

[12] Varsha Dani, Sham M Kakade, and Thomas P Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, pages 345–352, 2008.

[13] Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. 2009.

[14] Shipra Agrawal. Linear bandits. `agrawal.wikischolars.columbia.edu/file/view/Lecture+8.pdf`, 2016.

[15] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[16] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *CoRR*, abs/1707.02038, 2017.

[17] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, 2012.

[18] Daniel Russo and Benjamin Van Roy. Learning to Optimize Via Posterior Sampling. *arXiv:1301.2609 [cs]*, January 2013. arXiv: 1301.2609.

[19] Daniel Russo and Benjamin Van Roy. Learning to optimize via information directed sampling. In *NIPS*, 2017.