University of Alberta

Online Learning for Linearly Parametrized Control Problems

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 **Doctor of Philosophy**.

in

Statistical Machine Learning

Department of Computing Science

Fall 2012
Edmonton, Alberta

Abstract

In an online control problem, a learner makes an effort to control the state of an unknown environment so as to minimize the sum of the losses he suffers. In this thesis, we study several online control problems, ranging from the simple bandit problems, through classical LQ control problems, to more complex non-linear problems. The main topic is the design of algorithms for these problems and the development of finite-time performance guarantees.

A common theme of the problems is that they assume a linear parametric uncertainty. Accordingly, our methods employ a linear-in-the-parameters predictor and construct a confidence set that contains the true parameter with high probability. In particular, the algorithms always use the parameter that gives rise to the lowest expected loss.

The first main contribution of the thesis is the construction of smaller confidence sets for the least-squares estimate. To arrive at these confidence sets, we derive a novel tail inequality for vector-valued martingales. Based on this new confidence set, we improve the algorithms for the linear stochastic bandit problem.

The second main contribution is the introduction of a novel technique to construct confidence sets, which allows us to construct confidence sets given the predictions of any algorithm whose objective is to achieve low regret with respect to the quadratic loss while using linear predictors. As a demonstration of this new approach, we introduce the sparse variant of linear bandits and show that a recent online algorithm together with our conversion allows one to derive algorithms that can exploit if the unknown parameter vector is sparse.

In the second part of the thesis, we study the LQ control problem with unknown model parameters. We design an algorithm and prove a sublinear regret. We also show that similar techniques can be employed to design and analyze an algorithm for a more general problem with nonlinear dynamics but linear parametric uncertainty. To the best of our knowledge this is the the first time that regret bounds are derived for these classes of control problems.

Acknowledgements

I was extremely fortunate to work under the guidance of Csaba Szepesvári, who taught me invaluable lessons in research. I am also grateful for the support and the freedom that I enjoyed during my studies.

I thank our colleague David Pál with whom we obtained results of Chapters 3 and 4, while he was a post-doc here in Edmonton. I am thankful to my committee members, Richard Sutton, Dale Schuurmans, Russell Greiner, Biao Huang, and Vivek Borkar, for their feedback.

I also thank members of Reinforcement Learning and Artificial Intelligence (RLAI) lab for the friendly environment and all the great discussions and talks. I thank my friends in Department of Computing Science for making my time fun and memorable.

Last but not least, I would like to thank Kiana Q. Hajebi, for her endless love, support and encouragement! My special appreciation goes to my parents, for their never-ending support and dedication.

Table of Contents

[1] This chapter is based on the work by Abbasi-Yadkori, Pal, and Szepesvari (2011a) and Abbasi-Yadkori, Pal, and Szepesvari (2011b).

[2] This chapter is based on the work by Abbasi-Yadkori, Pal, and Szepesvari (2011a) and Abbasi-Yadkori, Pal, and Szepesvari (2011b).

[3] Results of Sections 5.1 and 5.2 have appeared in (Abbasi-Yadkori and Szepesvári, 2011).

List of Figures

Notations

Let $\mathcal{H}$ be a separable Hilbert space. Let $\mathcal{L}(\mathcal{H})$ be the space of all $\mathcal{H} \to \mathcal{H}$ linear operators. Let (e_i) be a countable orthonormal basis for $\mathcal{H}$. The inner product in $\mathcal{H}$ is denoted by $\langle ., . \rangle$. The outer product of vector x, denoted by $x \otimes x$, is defined as a linear operator such that, for any vector v, $(x \otimes x)v = \langle v, x \rangle x$. Let A^* denote the adjoint of operator A. We use $\|x\| = \sqrt{\langle x, x \rangle}$ to denote the norm of a vector $x \in \mathcal{H}$. For a positive definite self-adjoint operator A, the weighted norm of vector x is defined by $\|x\|_A = \sqrt{\langle x, Ax \rangle}$. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of the positive semidefinite matrix A, respectively. We use $A \succ 0$ to denote that A is positive definite, while we use $A \succeq 0$ to denote that it is positive semidefinite. We use $\mathbb{I}_{\{E\}}$ to denote the indicator function of event E.

We use O and Ω the "big-Oh" and "big-Omega" notations, respectively. That is, for $D \subset \mathbb{R}$, $f, g : D \to \mathbb{R}$, $a \in \mathbb{R} \cup \{-\infty, +\infty\}$, we say that $f = O(g)$ at a if $\limsup_{x \to a, x \in D} |f(x)/g(x)| < \infty$. Similarly, we say that $f = \Omega(g)$ at a if $\limsup_{x \to a, x \in D} |g(x)/f(x)| < \infty$. Usually, a is clear from the context and is suppressed. We use $\widetilde{O}$ to hide logarithmic factors in the big-O notation. We use $\wedge$ and $\vee$ to denote the minimum and the maximum, respectively, in addition to the more customary (but longer) min and max operators.

Throughout this thesis, we assume that the reader is familiar with basic concepts of calculus, Reproducing Kernel Hilbert Spaces (RKHS), and probability theory. Required background is summarized in Appendices A, B, and C.

Chapter 1

Introduction

Prediction problems can be formulated as a game between a learner and an environment, where the learner receives data from the environment and is asked to build a predictor with a small loss on future data. In the offline setting, the learner is given a dataset and the goal is to find a predictor that performs well on "future" data: Learning (i.e., finding the predictor) is one-shot, as is evaluation. In the online variant, on the other hand, learning and performance assessment are interleaved: Data arrives sequentially in discrete time steps. In each time step, the learner produces a predictor, which is evaluated on the next data point. Performance is measured by the total loss of predictions over time. Sequential prediction problems of this nature are called online learning problems (Cesa-Bianchi and Lugosi, 2006).

Oftentimes, the performance of an online learner is measured with respect to that of the best predictor in some comparison class. This gives rise to the concept of the learner's *regret*, which is defined as the difference between the total loss (up to some time) of the best competitor from the comparison class and the total loss of the learner. We say that an algorithm is learning with respect to a class of competitors if its regret with respect to the given class grows at most sublinearly with time, i.e. the average regret in the limit is nonpositive. This property is known as *Hannan consistency*.

1.1 Specific Problems

Let us now consider a few specific examples, a subset of which is the subject of this thesis.

Problem 1 Full Information Online Learning (Cesa-Bianchi and Lugosi, 2006) Consider the following game between the learner and the environment. At each round t, the learner chooses a vector a_t from a (known) set $D \subset \mathbb{R}^d$. Next, the environment reveals a loss function $\ell_t(.)$ (the sequence of loss functions is chosen ahead of the game) and the learner suffers the loss $\ell_t(a_t)$. The game is illustrated in Figure 1.1.

In the worst-case setting, the environment can choose any sequence of loss functions. A special case of this problem, which makes it easier from the point of view of the learner, when D is a convex, bounded region and ℓ_t is a convex loss function and the loss functions are bounded by some constant. The resulting problem is called *online convex optimization under full-information feedback* (Zinkevich, 2003, Shalev-Shwartz, 2011).

It might be tempting to ask for an algorithm that performs nearly as well as a competitor that chooses the best prediction at each round, i.e. to define the regret by

$$P_T = \sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \min_{a \in D} \ell_t(a)\,,$$

where T is the time horizon. It is easy to see that for any learner, P_T can grow linearly with T. Thus, no learner can be Hannan consistent under this criterion. A criterion that is much

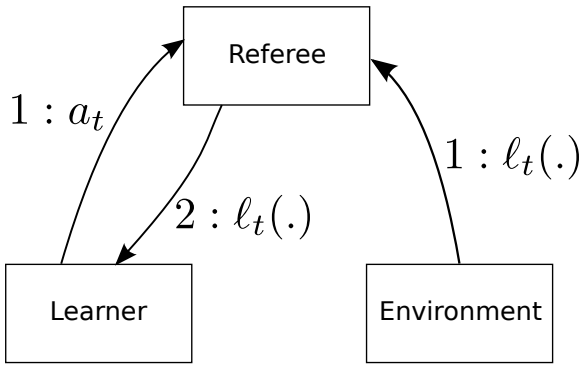


Figure 1.1: Full information online learning. The learner predicts a_t and then observes the loss function ℓ_t. The learner suffers the loss $\ell_t(a_t)$ which is recorded by the referee. Numbers show the ordering in which the interactions occur.

less demanding and that often makes Hannan consistent learning possible, is to restrict the set of competitors to ones that make the same decision in all timesteps, leading to the regret definition

$$R_T = \sum_{t=1}^{T} \ell_t(a_t) - \min_{a \in D} \sum_{t=1}^{T} \ell_t(a) \ .$$

Full-information online learning is well-suited to model supervised learning problems. For example, consider the problem of predicting the temperature in Edmonton. At time t, the learner predicts the temperature based on some historical information x_t. The prediction can be a simple linear function of the data x_t and the loss function can be quadratic: $\ell_t(a) = (y_t - \langle x_t, a \rangle)^2$, where y_t is the temperature for round t. The loss function satisfies the convexity assumption. In this case, competing with the best constant predictor means competing with the best linear predictor.

Problem 2 Bandit Information Online Learning (Abernethy et al., 2008) In a bandit problem, the loss function is revealed only at the point chosen by the learner. Bandit problem is more challenging than the full information online learning because the learner no longer has access to additional information about the loss function such as its derivatives. Following the notation of the previous example, this means that the learner only observes $\ell_t(a_t)$ at round t (as opposed to observing $\ell_t(\cdot)$, which was the case previously). The information flow between the learner and the environment is shown in Figure 1.2.

Many practical problems, such as web advertisement, online routing, recommendation systems, fit only the bandit framework (as opposed to the full information framework). In these applications the decision of the learner, a_t, is often viewed as an *action*. In what follows we will use the words action, prediction and decision interchangeably.

As an illustration of bandit problems, consider a web advertisement application. When a user visits the website, the learner shows an ad from a pool of ads. The loss could be zero if the user clicks on the ad, and it could be one otherwise. This gives rise to a function that assigns a binary value to every ad. However, only the loss of the ad shown will be available to the learner, thus making the problem an instance of bandit problems.

An important special case of the bandit problem is the *linear bandit problem*, where the action-set is a bounded, convex subset of some vector-space and the loss is a linear function of the action. We note in passing that if the action set is not convex, one can always define

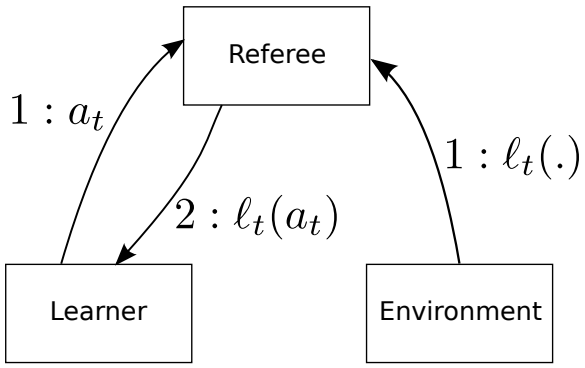


Figure 1.2: Bandit problem. As opposed to the full information setting, the loss function is never passed to the learner, only the loss suffered at the chosen prediction point a_t.

a new action set as the convex hull of the original action set and extend the loss function in a linear fashion. In practice, this means to use randomized actions (or distributions over the original action set as actions).

A special case of linear bandit problems is when the action set is the standard simplex of a finite-dimensional Euclidean space, giving rise to the so-called *multi-armed bandit (MAB)* problem (Auer et al., 2003).

Problem 3 Stochastic Linear Bandits (Auer et al., 2002a, Auer, 2002, Dani et al., 2008, Abbasi-Yadkori et al., 2011a) So far, no assumptions were made about how the loss functions arise, apart from that they are selected ahead of the game. A wide class of problems can be obtained by making stochasticity assumptions about the environment. For example, classical statistical learning theory problems can be obtained by making the assumption that $\ell_t(a) = \ell(a, \eta_t)$, for some (fixed) loss function ℓ and a sequence of independent, identically distributed (i.i.d.) random variables $(\eta_t)_t$. The stochastic setting often allows specific algorithms and tighter regret bounds. Its appeal is that despite this, it is still flexible enough to capture the essence of many real world problems.

As a specific example, consider *stochastic linear bandit* problems. In this case, $\ell_t(a) = \langle a, \theta_* \rangle + \eta_t(a)$, where $\theta_* \in \mathbb{R}^d$ is a fixed but unknown parameter vector and the elements of the *noise* sequence $(\eta_t(\cdot))_t$ for any fixed value of a are zero-mean random variables with bounded moments. We shall deal with model (in fact, a slight generalization of it) in Chapter 4. One of the contributions of this thesis is the design of algorithms that achieve better performance (both in theory and practice) than what was previously available.

Problem 4 Sparse Linear Stochastic Bandits (Abbasi-Yadkori et al., 2011b) We say that a vector is *sparse* when most of its elements are zero. It has been shown that more sample efficient algorithms can be designed for many machine learning problems when the parameter vector is sparse (Bühlmann and van de Geer, 2011). Sparse stochastic bandits is the sparse variant of the linear stochastic bandits. In particular, in a sparse linear stochastic bandit, the assumption is that most elements of θ_* are zero. The question then is whether there exist algorithms with regret that improves with increasing sparsity. This problem is particularly interesting because many applications have a large number of features, but only a few features are relevant.

A contribution of this thesis is to introduce the sparse stochastic bandit problem and to obtain tight regret bounds for it.

Problem 5 Online Reinforcement Learning (Szepesvári (2010), Section 3.2.4) Up until now, the environment was memoryless, in the sense that actions had no effect on future loss functions. The reinforcement learning (RL) problem is a more general problem where the *state* of the environment changes as a function of the current state and the action taken by the learner. Formally, such an environment can be described by a 6-tuple $(\mu, \mathcal{X}, D, \mathcal{Z}, p, \ell)$, giving rise a Markovian Decision Process. Here, μ is the distribution of the initial state; $\mathcal{X}$ is the state space; D is the action space; $\mathcal{Z}$ is the set of admissible state-action pairs defined by

$$\mathcal{Z} = \{(x, a) \,:\, x \in \mathcal{X},\, a \in D(x)\}\,,$$

where $D(x)$ is the set of available actions at state x; $p : \mathcal{X} \times D \times \mathcal{X} \to \mathbb{R}$ is a stochastic kernel[1] on $\mathcal{X}$ given $\mathcal{Z}$, also known as the transition law; and $\ell : \mathcal{X} \times D \to \mathbb{R}$ is the loss (a.k.a. cost) function. Notice that, now, the loss is a function of both the environment's state and the action chosen by the learner. The interaction between the learner and the environment is shown in Figure 1.3. There are a limited number of results when the stochasticity assumption is relaxed.

As before, the objective is to have low regret with respect to a class of competitors. In this setting, a standard competitor set is the set of (stationary) policies. Thus, each competitor is a mapping from the state space to the action space, choosing in each state one of the admissible actions. Thus, the regret has the form of

$$R_T = \sum_{t=1}^{T} \ell(x_t, a_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \ell(x_t, \pi(x_t))\,, \tag{1.1}$$

where Π is the class of policies. We refer the reader to standard RL textbooks (Sutton and Barto, 1998, Bertsekas and Tsitsiklis, 1996) for further reading on the popular approaches to the RL problem and how to choose a suitable class of policies Π.

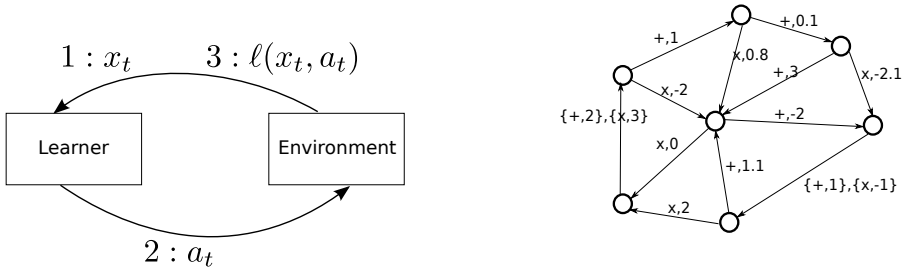
Regret bounds exist only for *finite* Markov Decision Processes (Burnetas and Katehakis, 1997, Borkar, 2000, Bartlett and Tewari, 2009, Jaksch et al., 2010) when both the state and action spaces are finite sets. A contribution of this thesis is to extend such results to MDPs with continuous state-action spaces under some structural assumptions (see Problems 6 and 7 below). We study the case when the loss is stochastic. In contrast, Neu et al. (2010a) obtain regret bounds for finite loop-free stochastic shortest path problems when the reward function is determined by an *oblivious* adversary, while Neu et al. (2010b) extend these results to finite MDPs, but make the additional assumption that the transition law is "uniformly mixing". These papers assume that the transition probabilities are known. An extension of these results to the case when the transition probabilities are unknown is presented in the paper by Neu et al. (2012).

Remark 1.1 Discounted Losses A large portion of RL literature studies discounted variant of the problem, where losses are discounted by a discount factor $0 < \gamma < 1$ and the total loss of policy π starting from state x is defined as $L_{x,\pi} = \sum_{t=1}^{\infty} \gamma^t \ell(x_t, \pi(x_t))$, where $x_1 = x$. When discounting, we are basically assuming that future losses are less important than more immediate ones. An optimal policy is defined by $\pi_*^x = \operatorname{argmin}_{\pi \in \Pi} \mathbb{E}\left[L_{x,\pi}\right]$. An ϵ-optimal policy starting from state x is one that satisfies $L_{x,\pi} < L_{x,\pi_*} + \epsilon$. Notice that when the loss function is bounded by $B > 0$, $L_{x,\pi}$ is bounded by $B/(1-\gamma)$. This makes the analysis of such problems somewhat easier.

A number of papers have studied the problem in a PAC-learning[2] framework (Kearns and Singh, 1998, Brafman and Tennenholtz, 2002, Kakade, 2003, Strehl et al., 2006, Szita and

[1] Let $\mathcal{B}(\mathcal{X})$ denote the Borel σ-algebra of $\mathcal{X}$. A stochastic kernel on $\mathcal{X}$ given $\mathcal{Z}$ is a function $p(.|.)$ such that for each $z \in \mathcal{Z}$, $p(.|z)$ is a probability measure on $\mathcal{X}$ and for each $B \in \mathcal{B}(\mathcal{X})$, $p(B|.)$ is a measurable function on $\mathcal{Z}$.

[2] PAC stands for "probably approximately correct".



(a) (b)

Figure 1.3: (a) Online reinforcement learning problem. At round t, the learner observes the environment's current state x_t and based on it, takes action a_t, which is sent to the environment. In response, the environment reveals the loss $\ell(x_t, a_t)$ and moves its state to x_{t+1}. (b) An example of an RL problem. The action space is $D = \{+, \times\}$ and each vertex is a state. The learner starts from one of the vertices and travels on the graph with the objective of incurring low losses. The losses associated with the actions are shown on edges, while the directions of the edges show the direction of state transitions (state transitions are deterministic). Notice that the learner's actions change what future loss functions it will get.

Szepesvári, 2010, Jain and Varaiya, 2010). To state such results, we first define the sample complexity of an algorithm. The definition is stated in the form given here as Definition 1 in (Szita and Szepesvári, 2010).

Definition 1.2 Let $\epsilon > 0$ be a prescribed accuracy and $\delta > 0$ be an allowed probability of failure. The expression $\zeta(\epsilon, \delta, N, K, \gamma, C_m)$ is a sample complexity bound for algorithm $\mathcal{A}$, if the following holds: Take any $\epsilon > 0$, $\delta \in (0, 1)$, $N > 0$, $K > 0$, $\gamma \in [0, 1)$, $C_m > 0$ and any MDP M with N states, K actions, discount factor γ, and losses bounded by C_m. Let π be the policy that algorithm $\mathcal{A}$ runs. Let $\mathcal{A}$ interact with M, resulting in the process $x_1, a_1, x_2, a_2, \ldots$. Then, independently of the choice of x_1, with probability at least $1 - \delta$, the number of timesteps such that $L_{x_t,\pi} > L_{x_t,\pi_*^{x_t}} + \epsilon$ is at most $\zeta(\epsilon, \delta, N, K, \gamma, C_m)$. An algorithm with sample complexity that is polynomial in $1/\epsilon$, $\log(1/\delta)$, N, K, $1/(1-\gamma)$, C_m is called PAC-MDP (probably correct in MDPs).

The following theorem is a recent result in PAC-learning framework.

Theorem 1.3 (Szita and Szepesvári, 2010) Fix some prescribed accuracy $\epsilon > 0$, failure probability $\delta \in (0, 1)$, and discount factor $\gamma \in [0, 1)$. Let $M = (\mu, \mathcal{X}, D, \mathcal{Z}, p, \ell)$ be an MDP with $|\mathcal{X}| = N$ states and $|D| = K$ actions, with non-negative losses, and a value $L \in \mathbb{R}$ that is an upper bound on all discounted cumulated losses. If the MOREMAX algorithm, defined in (Szita and Szepesvári, 2010), runs on MDP M, then with probability at least $1 - \delta$, the number of rounds t for which $L_{x_t,\text{MOREMAX}} > L_{x_t,\pi_*^{x_t}} + \epsilon$ is bounded by $\widetilde{O}\left(\frac{NKL^2}{(1-\gamma)^4\epsilon^2}\right)$.

Kakade et al. (2003) obtain sample complexity bounds for infinite state space MDPs under the assumptions that the state transition and reward functions are uniformly Lipschitz, while the state space is compact.

Remark 1.4 Convergence Results for Temporal-Difference (TD) Methods Asymptotic behavior of *temporal-difference methods* (Sutton, 1988) in large state and action spaces is studied both in *on-policy* (Tsitsiklis and Van Roy, 1997) and *off-policy* (Sutton et al., 2009b,a, Maei et al., 2009) settings. All these results concern the policy estimation problem, i.e., estimating the value of a fixed policy. The available results for the control problem,

i.e., estimating the value of the optimal policy, are more limited (Maei et al., 2010) and prove only convergence to local optimum of some objective function. It is not clear if and under what conditions these TD control methods converge to the optimal policy.

Problem 6 Linearly Parametrized Control The linearly parametrized control problem is a special case of the RL problem when the next state is a noisy linear function of some features of the previous state and the action taken,

$$x_{t+1} = \Theta_* \varphi(x_t, a_t) + w_{t+1} \; . \tag{1.2}$$

Here, $\mathcal{X} \subset \mathbb{R}^n$, $D \subset \mathbb{R}^d$, $\Theta_* \in \mathbb{R}^{n \times m}$ is an unknown matrix, $\varphi : \mathbb{R}^{n+d} \to \mathbb{R}^m$ is a feature mapping, and $w_t \in \mathbb{R}^n$ is a zero mean random variable that satisfies certain martingale and tail properties (to be specified later). The difference with the more general RL problem is that, here, the noise vector w_t has an additive effect in the model and is also independent of the state.

Strehl and Littman (2008) study this problem in the discounted setting and obtain sample complexity bounds. They assume that the loss function is known or otherwise has a linear form. Further, they assume that the ℓ^2 norm of feature mapping is less than 1, i.e., $\sup_{x \in \mathcal{X}, a \in D} \|\varphi(x, a)\| < 1$. In this thesis, we obtain regret bounds for the linearly parametrized control problem in a more general setting (see Chapter 5).

Problem 7 Linear Quadratic Problem (Aström and Wittenmark, 1973) The linear quadratic (LQ) problem is a special case of Problem 6 when the next state is a noisy linear function of the previous state and the action taken,

$$x_{t+1} = A_* x_t + B_* a_t + w_{t+1} \; , \tag{1.3}$$

and the loss is a quadratic function of the state and action:

$$\ell(x_t, a_t) = x_t^\top Q x_t + a_t^\top R a_t \; .$$

Here, $A_* \in \mathbb{R}^{n \times n}$ and $B_* \in \mathbb{R}^{n \times d}$ are unknown matrices, while $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{d \times d}$ are known matrices. The LQ problem plays a fundamental role in the control literature.

1.2 Optimism in the Face of Uncertainty

A main topic of the thesis is that a careful study of linear prediction problems leads to improved algorithms for the stochastic online learning problems. In the thesis, we will demonstrate this for Problems 3, 4, 6, and 7. The algorithms that we study are based on a common underlying idea—the *optimism-in-the-face-of-uncertainty* (OFU) principle.

Optimism in the face of uncertainty is a general principle that can be employed to design efficient algorithms in many stochastic online learning problems. To simplify the discussion of the principle, we restrict ourselves to the bandit problem in this section.

Consider the stochastic linear bandit problem. If the learner knew θ_*, he could simply take the action $a_* = \operatorname{argmin}_{a \in D} \langle a, \theta_* \rangle$ in every time step to minimize the (expected) loss (D is a convex, compact set, so the minimum is well-defined). When θ_* is unknown, the learner can rely only on an estimate. One simple idea, known as the *certainty equivalence principle* (Simon, 1956), is to estimate θ_* by some means such as using a least-squares method and behave as if the (least-squares) estimate $\widehat{\theta}_t$ was the true parameter vector. It is easy to show that an algorithm that relies on the certainty equivalence principle can get stuck with a sub-optimal choice, which leads to a linear regret. We demonstrate this by means of an example.

Example 1 Let the decision set of a stochastic linear bandit problem be $D = \{e_1, e_2\}$, where e_i is the unit vector along ith coordinate in $\mathbb{R}^2$. As noted beforehand, when D contains unit vectors, we get what is called a multi-armed bandit problem. In this case,

there are two "arms", or actions. Let the loss be such that if the first action (e_1) is chosen, then the loss is a random number in the $[0,1]$ interval with mean μ_1, while if the second action (e_2) is chosen, the loss is deterministic and takes on the fixed value $\mu_2 \in [0,1]$. To map this into our framework, let (ξ_t) be a sequence of i.i.d. random variables taking values in $[0,1]$ and whose mean is μ_1. Then, $\ell_t(a) = a_1\xi_t + a_2\mu_2 = \langle a, \theta_* \rangle + a_1(\xi_t - \mu_1)$, where $\theta_* = (\mu_1, \mu_2)^\top$. Thus, the loss indeed takes the desired form with $\eta_t(a) = a_1(\xi_t - \mu_1)$.

Let $y_t \in [0,1]$ be the loss observed at time t. Assume that $\mu_2 > \mu_1$. The least-squares estimate of μ_i at time t is

$$\widehat{\mu}_{i,t} = \frac{\sum_{s=1}^{t} \mathbb{I}_{\{x_s = e_i\}} y_s}{\sum_{s=1}^{t} \mathbb{I}_{\{x_s = e_i\}}},$$

where $\mathbb{I}_{\{.\}}$ is the indicator function. Assume that each action is taken once at the beginning, from which point on the certainty equivalence principle is followed. This means that

$$a_{t+1} = \operatorname*{argmin}_{e_1, e_2} \{\langle e_1, \widehat{\mu}_{1,t} \rangle, \langle e_2, \widehat{\mu}_{2,t} \rangle\}$$

for $t \geq 3$. Clearly, the event $\widehat{\mu}_{1,2} > \widehat{\mu}_{2,2} = \mu_2$ happens with positive probability. When this event happens, by induction we can see that action e_1 will never be chosen again. Indeed, when action two is chosen, only $\widehat{\mu}_{2,t}$ has a chance of being changed. However, since the payoff of action 2 is deterministic, the estimate will never be changed. Hence, the algorithm will keep using the second (suboptimal) action, leading to a linear lower bound on the regret.

As this example demonstrates, the certainty equivalence principle can get stuck with a sub-optimal action for long periods of time or even indefinitely. It is thus necessary to allocate a portion of time to try those actions with higher estimated losses: The algorithms need to "explore" actions that look "suboptimal". How often or rather, when to explore such actions is the main issue in designing efficient bandit algorithms. Clearly, exploration should be tapered off with time: If this does not happen, the algorithm would still pay a linear regret. The problem of balancing between exploring and exploiting is called the *exploration-exploitation dilemma*.

Perhaps the simplest exploration method is to take random actions with a certain rate to obtain more information about the parameter vector. For example, the learner in the previous game can take one of the two actions uniformly at random once in every few rounds. The exploration rate should be tuned to minimize regret. This method, known as the *forced-exploration method* or *ϵ-greedy method* (Lai and Wei, 1981, 1982, 1987, Chen and Guo, 1987), is simple to implement and with proper tuning it can often be made quite competitive in practice.

A major issue with this simple idea is that the optimal exploration rate will depend on the problem structure, which is often unknown to the learner. In Example 1, little deliberation shows that one would need to explore for a longer fraction of the time to discriminate between the two actions for smaller values of the (hidden) "gap" parameter $\Delta \doteq \mu_2 - \mu_1$. To see why this is the case, consider the following informal argument (for a precise argument with identical conclusions, see Dani and Hayes 2006). Assume the learner takes T^α exploratory actions at the *beginning* of the game, where T is the time horizon and $0 < \alpha < 1$ is a tuneable parameter that governs the amount of exploration. During this period, the learner suffers a regret of $\Omega(T^\alpha \Delta)$. With a simple application of Hoeffding's Inequality (see Appendix C), we get that the event $A = \{\widehat{\mu}_{1,T^\alpha} > \widehat{\mu}_{2,T^\alpha}\}$ happens with a probability that is bounded by $\exp(-2T^\alpha \Delta^2)$. Further, one can argue that this probability (in the worst-case) cannot be "much" smaller either (e.g., by an application of Stirling's formula). Under event A, the learner takes the sub-optimal action for the rest of the game. Thus, the expected regret up to time T can be written as

$$\begin{aligned} \mathbb{E}[R_T] &= \mathbb{E}[\text{Regret(Exploration Phase)}] + \mathbb{E}[\text{Regret(Exploitation Phase)}] \\ &\approx T^\alpha \Delta + (T - T^\alpha) e^{-2T^\alpha \Delta^2} \Delta \,. \end{aligned} \tag{1.4}$$

If the learner chooses $\alpha \leq 2/3$, the second term of (1.4) is bounded by $\Omega(T^{2/3})$ when $\Delta = T^{-1/3}$. Otherwise, the first term is bounded by $\Omega(T^{2/3})$ when $\Delta = 1$. In any case, for any *a priori fixed amount of exploration*, the learner can suffer a regret as large as $\Omega(T^{2/3})$. On the other hand, it is well known that there exist algorithms (some of which will be discussed below) that are able to achieve a regret of size at most $O(T^{1/2})$ *independently* of what problem they are used for (thus, we say that they achieve a uniform, worst-case bound of $O(T^{1/2})$) (Auer et al., 2002a, Bubeck and Audibert, 2010). In fact, it is also known that for the bandit problems considered here, $O(T^{1/2})$ is the best regret possible (Auer et al., 2002a). One way of expressing that forced-exploration schemes are unable to achieve this lower bound (i.e., the optimal growth rate of regret) uniformly over all problems is that they do not adapt to the difficulty of the individual bandit problems.

1.2.1 The OFU principle

The Optimism in the Face of Uncertainty (OFU) principle, proposed by Lai and Robbins (1985), elegantly addresses the exploration-exploitation dilemma (it appears that the principle has been rediscovered at least once by Campi (1997) who calls it the "bet on the best" principle). The basic idea is to maintain a confidence set for the parameter vector and then in every round choose an estimate from the confidence set together with an action so that the predicted expected loss is minimized, i.e., the estimate-action pair is chosen optimistically. The OFU principle is known to have better adaptivity properties than (e.g.) forced-exploration schemes in the sense that often it leads to algorithms that are able to achieve the minimax regret rate.

To see the OFU principle in practice, consider again the two-action bandit problem of Example 1. To implement the OFU principle, we need to come up with confidence set for $\theta_* = (\mu_1, \mu_2)$. In this example it makes sense to seek an appropriate confidence set in the form of a Cartesian product, $C_{1,t}(\delta/2) \times C_{2,t}(\delta/2) \subset \mathbb{R}^2$, where for a given confidence parameter $0 < \delta < 1$, $C_{i,t}(\delta) = \{\mu \ : \ |\mu - \widehat{\mu}_{i,t}| \leq c_{i,t}(\delta)\}$ is a (random) interval centered around the empirical estimate of the mean loss of action i with (half-)width $c_{i,t}(\delta)$ chosen such that the event $|\mu_i - \widehat{\mu}_{i,t}| \leq c_{i,t}(\delta)$ holds with probability at least $1 - \delta$. Then, by the union bound, $\mathbb{P}\left(\theta_* \in C_{1,t}(\delta/2) \times C_{2,t}(\delta/2)\right) \geq 1 - \delta$. The optimistic loss estimate of action i is then $\widetilde{y}_{i,t} = \widehat{\mu}_{i,t} - c_{i,t}(\delta/2)$ and the optimistic algorithm (that implements the OFU principle) plays action $a_{t+1} = \operatorname{argmin}_i \widetilde{y}_{i,t}$ in round $t + 1$.

The width, $c_{i,t}(\delta)$, typically scales like $O(\sqrt{\log(1/\delta)/N_{i,t}})$, where $N_{i,t}$ is the number of times action i was played up to time t. If an action is played only a small number of rounds, then its confidence interval will be large, which means its optimistic loss estimate will be small, which increases the chance of choosing such an action. To understand how the OFU principle leads to an algorithm that "adapts" to the size of the gap, Δ, consider the case when Δ is large. In this case, the algorithm can easily discriminate between the two actions and thus the exploration rate will be automatically small. Otherwise, if the gap is small, the discrimination will be difficult and, as expected, the algorithm will explore more often.

We should note that the OFU principle, by construction, is limited to stochastic problems. Thus, the principle does not apply when no stochasticity assumptions are made on the environment. It also appears that in the *stochastic partial monitoring* problems, the OFU principle is not sufficient (Bartók, 2012). The fundamental reason is that, in these problems, one has to explicitly reason about the "value of information"; actions differ in terms of how much information we can gain by using them about the environment. The optimism principle focuses too narrowly on the role of losses (or rewards), thus not leaving sufficient room to reason about the indirect, "information-value" of using the actions. However, when a stochastic model for the environment is available and we directly observe losses, the OFU principle has been shown to be quite successful for a wide range of problems, including bandit problems of various types (Lai and Robbins, 1985, Katehakis and Robbins, 1995, Burnetas and Katehakis, 1996, Auer et al., 2002a, 2007, Auer, 2002, Li et al., 2010, Filippi et al., 2010) or online reinforcement learning problems (Burnetas and Katehakis,

1997, Campi, 1997, Kakade, 2003, Bittanti and Campi, 2006, Strehl and Littman, 2008, Bartlett and Tewari, 2009, Jaksch et al., 2010, Szita and Szepesvári, 2010). Indeed, one of the contributions of this thesis is to add further evidence that the OFU principle can be applied to an even wider range of problems. In particular, we will show, both theoretically and empirically that the OFU principle can be effectively applied to even complicated online reinforcement learning problems such as the LQ problem. It remains to be seen whether alternative approaches, such as those by Abernethy et al. (2008) that were developed for non-stochastic bandit problems, can be applied to stochastic online reinforcement learning problems.

Chapter 2

Summary of Contributions

The thesis makes contributions to constructing confidence sets for linear prediction problems and demonstrates how the new confidence set construction techniques lead, through the OFU principle, to learning algorithms with improved learning speed.

2.1 Construction of Confidence Sets by Vector-Valued Self-Normalized Processes

The first confidence set construction technique builds on and extends techniques whose history goes back to Robbins and Siegmund (1970). The key idea is to construct faithful confidence sets whose shape is strongly dictated by the data, avoiding conservative upper bounds by relying on so-called self-normalized bounds (see Chapter 3, Theorem 3.11).

With our new technique, we vastly reduce the size of the confidence sets of Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010), and Srinivas et al. (2010). First, our confidence sets are valid uniformly over all time steps, which immediately saves $\log(T)$ factor by avoiding the otherwise needed union bound. Second, our confidence sets are "more empirical" in the sense that some worst-case quantities from the old bounds are replaced by empirical quantities that are always smaller, sometimes substantially. Further, the calculations are done for linear prediction over separable Hilbert spaces instead of finite-dimensional Euclidean spaces, thereby significantly extending the scope and applicability of the result. In particular, the result is applicable to popular nonparametric learning scenarios too, such as learning with Gaussian processes or with ridge regression over RKHS spaces. This difference is demonstrated through computer simulations in the further parts of the thesis, where the confidence sets are used in constructing online learning methods in various control learning problems.

2.2 A New Method to Construct Confidence Sets: Online-to-Confidence-Set-Conversion

The aforementioned confidence sets are constructed from predictions of the online least-squares method. Another contribution of this thesis is to show that, more generally, predictions of any online algorithm that predicts the responses of the chosen inputs in a sequential manner can be "converted" to a confidence set. The only assumption is that the online prediction algorithm comes with an upper bound on its regret[1] with respect to the best linear predictor using the quadratic prediction loss. The details of this conversion are explained in Section 3.5.

[1] This notion of regret, to be defined in Section 3.5, is different from the regret of the bandit problem.

One strength of our method is that it allows us to use any linear prediction algorithm as the underlying online algorithm, such as (online) least-squares (regularized or constrained) (Lai et al., 1979, Auer et al., 2002b, Vovk, 2001), online LASSO, the exponentiated gradient (EG) algorithm[2] (Kivinen and Warmuth, 1997), the p-norm algorithm (Grove et al., 2001, Gentile and Littlestone, 1999), the SeqSEW algorithm (Gerchinovitz, 2011), etc. These algorithms differ in terms of their biases towards different solutions. For example, some of these algorithms are biased towards sparse solutions, some of them are biased towards sparse inputs, etc. However, *all* the algorithms just mentioned satisfy the assumptions of the conversion, i.e., they work with quadratic prediction loss and for most of these algorithms a regret bound is known. Thanks to the generality of our solution, we can obtain a confidence set for each of these algorithms and, in fact, for any algorithm that might be developed in the future, too. An important consequence of our approach is that the confidence sets we derive from a regret bound for a given algorithm with a certain "bias" will inherit the "bias" from the algorithm.

Study of conversions and reductions between machine learning tasks has a long history (Blackwell, 1953, Birnbaum, 1961, Morse and Sacksteder, 1966, Conover and Iman, 1981, Littlestone, 1989, Bartlett et al., 1994, Kearns, 1998). Our online-to-confidence-set conversion can be compared with the online-to-batch conversions (Littlestone, 1989, Cesa-Bianchi et al., 2004, Dekel and Singer, 2006). However, there are two major differences between these two. First, online-to-batch conversions convert the predictions of a low-regret online algorithm into a single prediction with a low risk, whereas in our online-to-confidence-set conversion, we combine the predictions to construct a confidence set. Second, in online-to-batch conversions, one assumes that the data (i.e., the input-response pairs) are generated in an i.i.d. fashion (in fact, the risk is defined with respect to the underlying joint distribution), while in online-to-confidence-set conversion the inputs (a.k.a. covariates) can be chosen adversarially and only responses are stochastic. In summary, we are not aware of previous results on reductions of the type we consider.

2.3 Linear Bandit Problems

By applying the OFU principle to the linear stochastic bandit problem, the problem reduces to construction of confidence sets for the parameter vector.

When the decision set is a subset of $\mathbb{R}^d$, using our confidence sets, we improve regret of the ConfidenceBall algorithm of Dani et al. (2008). They showed that the regret of this algorithm is at most $O(d\log(T)\sqrt{T\log(T/\delta)})$ with probability at least $1-\delta$. We modify their algorithm so that it uses our new confidence sets and we show that its regret is at most $O(d\log(T)\sqrt{T}+\sqrt{dT\log(T/\delta)})$, which (roughly) removes a multiplicative factor $\sqrt{\log(T)}$ from their bound (See Theorem 4.1). Dani et al. (2008) also proved a problem-dependent regret bound. Namely, they showed that the regret of their algorithm is $O(\frac{d^2}{\Delta}\log(T/\delta)\log^2(T))$ where Δ is the "gap" as defined in (Dani et al., 2008). For our modified algorithm, we prove an improved $O(\frac{\log(1/\delta)}{\Delta}(\log(T)+d\log\log T)^2)$ bound (see Theorem 4.8). Apart from these theoretical improvements, we empirically demonstrate that the improved confidence bounds lead to significantly better performance.

Srinivas et al. (2010) obtain sublinear regret bounds when the decision set is a subset of a separable Hilbert space and the noise is Gaussian. We extend these results to sub-Gaussian noise and also obtain better dependence in terms of logarithmic and constant terms (see Theorem 4.1).

[2]EG is a variant of Winnow for linear prediction.

2.4 Sparse Stochastic Linear Bandits

Another contribution of this thesis is the introduction and study of a variant of the stochastic linear bandit problem, which we call *sparse stochastic linear bandits*. Sparsity, in recent years, became the line of attack for statistical problems that were previously thought unsolvable. The assumption that the underlying statistical model is sparse greatly decreases the sample size required to learn the model provided, of course, when that the model is indeed sparse. Several examples of algorithms that take advantage of sparsity are the Winnow algorithm (Littlestone, 1988), the LASSO (Tibshirani, 1996) and algorithms for compressed sensing (Candès, 2006).

With sparsity in mind, we investigate the sparse variant of the linear stochastic bandit problem. We focus on the situation when the underlying linear function is potentially sparse, i.e., many of its coefficients are zero, as can be expected to be the case in applications when the feature space is high-dimensional but only a few features are relevant (e.g., in web advertisement applications). We show that a recent online algorithm together with our online-to-confidence-set conversion allows us to derive algorithms that can exploit if the reward is a function of a sparse linear combination of the components of the chosen action. The details are given in Section 4.5.

Sparse linear bandit problem can be viewed as sequential decision making version of the feature selection problem. Its potential applications include problems where the dimensionality of the features describing the actions/contexts is typically high; such as the online optimization of contents of web pages, medical trials, web advertising management, etc.

2.5 Control Problems

The bandit setting relies on the assumption that the loss is a function of only the action taken by the learner. Although the bandit setting is a satisfactory model of many real-world decision making problems, it fails to fit many others. For example, consider a queueing problem where the server controls the service rate, which in turn determines the frequency at which the server processes the incoming jobs. The example is taken from (Lai and Yakowitz, 1995) and is shown in Figure 2.1. The bandit learner seeks to find a single best fixed service rate to have low cost. The cost, however, is a function of both learner's action (service rate) and the state of the system (number of jobs in the queue). The decisions made by the learner change the state. An optimal learner needs to change its action depending on the current state of system, a topic of reinforcement learning problem.

In the second part of the thesis (Chapter 5), we study two special cases of the RL problem. First, we apply our confidence sets to the linear quadratic (LQ) problem and derive the first finite-time regret bound for this problem. The LQ problem plays a central role in control theory thanks to its simplicity and elegance. Although, in practice, rarely does any control problem fit the LQ problem description, it is still one of the first choices of control engineers and it is also often used as a building block of other algorithms.

Our approach is to follow the OFU principle: We construct confidence sets from results of Chapter 3 and play optimistically with respect to them: At each round, the algorithm solves an OFU optimization problem to choose the next action. Unfortunately, the OFU optimization problem is not convex. We propose a gradient descent method for this optimization problem and show that it is effective in finding near-optimal decisions.

Finally, we show that similar techniques can be employed to design algorithms for the more general linearly parametrized control problems. Although we obtain $\widetilde{O}(\sqrt{T})$ regret bounds for this class of problems, the algorithm is computationally intractable. It remains an open problem to design efficient algorithm with nontrivial regret guarantees for this problem.

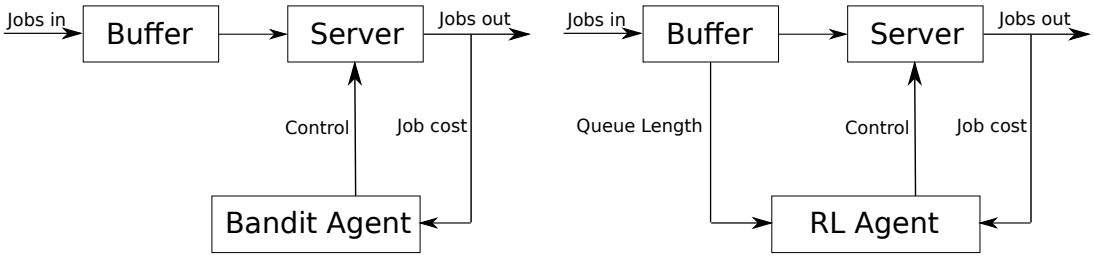


Figure 2.1: A queueing problem (Lai and Yakowitz, 1995). The loss for a job serviced is $l + Ca^2 s$, where l is the time spent in the queue, s is the service time spent by the server on this job, C is a parameter, and a is the service rate. The expected loss function is $\ell(x, a) = \mathbb{E}\left[(x + Ca^2)s(a)\right] = x/a + Ca$, where x is the number of jobs in the queue and $s(a)$ is the service time as a function of the service rate. (a) The queueing problem modelled as a bandit problem. The bandit agent is indifferent to the state of the system x. (b) The queueing problem as a reinforcement learning problem. The RL agent's action is a function of both the state and the loss observed.

Chapter 3

Online Least-Squares Prediction[1]

A large portion of machine learning is devoted to constructing point estimates of some unknown quantity given some "noisy data". A main issue with point estimates is that they lack a description of the remaining uncertainty about the unknown quantity. Confidence sets, on the other hand, allow one to characterize the remaining uncertainty. As Wasserman has written: "never give an estimator without giving a confidence set" (Wasserman, 1998, p. vii.). However useful confidence sets are on their own, in a number of sequential tasks they are in fact indispensable. Examples include stopping problems (Mnih et al., 2008), bandit problems (Auer et al., 2002a, Auer, 2002, Dani et al., 2008), variants of the pick-the-winner problem (Even-Dar et al., 2002, Mannor and Tsitsiklis, 2004, Mnih et al., 2008), reinforcement learning (Bartlett and Tewari, 2009, Jaksch et al., 2010), or active learning (Even-Dar et al., 2002).

In this chapter we investigate the problem of constructing confidence sets for the vector-coefficient of a linear function observed at a finite number of points in martingale noise (the exact conditions of our result will be stated in the next section). In other words, the uncertainty appears in a linear fashion. This is a popular and widely employed assumption thanks to its simplicity and mathematical elegance. The resulting linear prediction problem is widely studied in statistics, probability and statistical learning theory (Mardia et al., 1979, Weisberg, 1980, Seber, 1984).

We demonstrate two novel approaches to construct confidence sets for the unknown parameter vector. First, we employ tools from the theory of self-normalized processes to provide a simple and self-contained proof on the tail behaviour of a vector-valued martingale. We use the bound obtained to construct new confidence sets for the least-squares estimate that are tighter than those that were previously available.

Our second approach concerns a general method that allows one to construct high-probability confidence sets for linear prediction with correlated inputs given the predictions of *any* algorithm (e.g., online LASSO, exponentiated gradient algorithm, online least-squares, p-norm algorithm) targeting online learning with linear predictors and the quadratic loss. We call this technique an *online-to-confidence-set conversion*. In the next chapter, we will show how these new, tight confidence sets lead to improved performance (both theoretically, and empirically) for existing linear stochastic bandit algorithms.

3.1 Self-Normalized Processes

The study of self-normalized processes has a long history that goes back to William Sealy Gosset (a.k.a. Student) and is treated in detail in the recent book by de la Peña et al.

[1]This chapter is based on the work by Abbasi-Yadkori, Pal, and Szepesvari (2011a) and Abbasi-Yadkori, Pal, and Szepesvari (2011b).

(2009). As explained there, perhaps the most well-known result from these studies concerns the t-statistic and its limiting properties: Consider the problem of statistical inference on the mean μ of a normal distribution when the variance σ^2 is unknown. Let $\overline{m}_k = k^{-1}\sum_{i=1}^k m_i$ be the average of k i.i.d samples from the normal distribution and s_k be the sample standard deviation, $s_k^2 = (k-1)^{-1}\sum_{i=1}^k (m_i - \overline{m}_k)^2$. Gosset (Student) shows that the t-statistic $T_k = \sqrt{k}(\overline{m}_k - \mu)/s_k$ has the t-distribution with $k-1$ degrees of freedom. The t-distribution converges to the normal distribution as $k \to \infty$. Without loss of generality assume that $\mu = 0$. Then we can write

$$T_k = \sqrt{k}\frac{\overline{m}_k}{s_k} = \frac{S_k}{V_k^{1/2}}\left(\frac{k-1}{k - (S_k/V_k^{1/2})^2}\right)^{1/2}, \tag{3.1}$$

where $S_k = \sum_{i=1}^k m_i$ is a scalar-valued martingale and $V_k = \sum_{i=1}^k m_i^2$ is an increasing process. Let $N_k = S_k/V_k^{1/2}$ be a sum of random variables normalized by the square root of the cumulative variance, also known as a *self-normalized process*. Equation (3.1) implies that study of tail bounds for T_k reduces to that for N_k. It is also known that the limiting distributions of T_k and N_k coincide (Griffin, 2002).

In the next section, we study the vector-valued analog of N_k. Namely, we study self-normalized processes of the form $\|S_k\|^2_{\overline{V}_k^{-1}}$, where $S_k = \sum_{i=1}^k \eta_i m_{i-1}$ is a vector-valued martingale, η_i is a real-valued, R-sub-Gaussian random variable (to be defined in Section 3.2), m_i lies in a separable Hilbert space $\mathcal{H}$, and $\overline{V}_k = V + \sum_{i=1}^k m_{i-1} \otimes m_{i-1}$ is the corresponding *normalizing* operator with positive-definite, "regularizing" operator V. We employ the so-called method of mixtures of Robbins and Siegmund (1970) (see also de la Peña et al. 2009) to prove that for any $0 < \delta < 1$, with probability $1-\delta$,

$$\forall k \geq 0, \qquad \|S_k\|^2_{\overline{V}_k^{-1}} \leq 2R^2 \log\left(\frac{\det\left(I + M_{1:k}V^{-1}M_{1:k}^*\right)^{1/2}}{\delta}\right), \tag{3.2}$$

where $M_{1:k} : \mathcal{H} \to \mathbb{R}^k$ is the operator such that for any $v \in \mathcal{H}$, the ith element of $M_{1:k}v$ is $\langle m_i, v\rangle$. A less general version of the bound, applicable only to the finite-dimensional setting, can be derived from the works of de la Peña et al. (2004, 2009) by following some parts of our derivations. The merit of our new proof is thus that it is more generally applicable and it is also self-contained.

The above bound improves the previous bound of Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010) which were derived for the finite-dimensional setting (i.e., for Euclidean spaces), as well as improving on the bound of Srinivas et al. (2012) which was derived for separable Hilbert spaces. The bound is applicable to virtually any online least-squares problem. The bound that we derive, immediately gives rise to tight confidence sets and pointwise error bounds for the online least-squares estimate that can replace the confidence sets in existing linear stochastic bandit algorithms. In particular, if the true parameter vector is θ_*, we prove pointwise error bounds of the form

$$\forall \delta \in (0,1), \quad \mathbb{P}\left(\forall k, \forall m \in \mathcal{H}, \left|\langle m, \widehat{\theta}_k\rangle - \langle m, \theta_*\rangle\right| \leq \|m\|_{\overline{V}_k^{-1}} \beta_k(\delta)\right) \geq 1-\delta,$$

for the least-squares estimate $\widehat{\theta}_k$, where $\beta_k(\delta)$ denotes the square-root of the quantity on the RHS[2] of Equation 3.2 (cf. Theorem 3.11). Such bounds hold for *any* vector $m \in \mathcal{H}$ and give rise to tight confidence sets for θ_*.

In one-dimensional problems, the bound in (3.2) is comparable to a self-normalized form that can be obtained from Freedman's inequality by using a peeling/stratification argument (see, e.g., Theorem 1 in the paper by Audibert et al. (2009))[3]: let random variables

[2]The abbreviation RHS stands for "right-hand side"

[3]Audibert et al. (2009) show the derivation for i.i.d random variables. But, with a similar derivation, the same form can be proven for martingale difference sequences.

$m_1, m_2, \ldots$ be bounded in $[0, b]$. Let $G_s = \sum_{i=1}^{s}(m_i - \overline{m}_s)^2/s$ be the empirical variance. For any $k \in \mathbb{N}$ and $x > 0$, with probability at least $1 - 3\inf_{1<\alpha\le 3}\left(\frac{\log k}{\log \alpha} \wedge k\right) e^{-x/\alpha}$, for any $s \in \{1, \ldots, k\}$,

$$|\overline{m}_s - \mu| \le \sqrt{\frac{2G_s x}{s}} + \frac{3bx}{s} \; .$$

Compared to (3.2), the price is a $\log\log k$ factor.

3.2 Vector-Valued Martingale Tail Inequalities

Let $(\mathcal{F}_k; k \ge 1)$ be a filtration, $(m_k; k \ge 1)$ be a $\mathcal{H}$-valued stochastic process adapted to $(\mathcal{F}_k)$, $(\eta_k; k \ge 2)$ be a real-valued martingale difference process adapted to $(\mathcal{F}_k)$. Assume that η_k is conditionally sub-Gaussian in the sense that there exists some $R > 0$ such that for any $\gamma \in \mathbb{R}$, $k \ge 2$,

$$\mathbb{E}[\exp(\gamma\eta_k) \,|\, \mathcal{F}_{k-1}] \le \exp\left(\frac{\gamma^2 R^2}{2}\right) \quad \text{a.s.} \tag{3.3}$$

The sub-Gaussian condition automatically implies that $\mathbb{E}[\eta_k \,|\, \mathcal{F}_{k-1}] = 0$. Furthermore, it also implies that $\mathrm{Var}\,[\eta_t \,|\, F_{k-1}] \le R^2$ and thus we can think of R^2 as the (conditional) variance of the noise. An example of R-sub-Gaussian η_k is a zero-mean Gaussian noise with variance at most R^2, or a bounded zero-mean noise lying in an interval of length at most $2R$. Consider the martingale

$$S_t = \sum_{k=1}^{t-1} \eta_{k+1} m_k \tag{3.4}$$

and the processes

$$V_t = \sum_{k=1}^{t-1} m_k \otimes m_k, \qquad \overline{V}_t = V + V_t, \quad t \ge 1, \tag{3.5}$$

where V is a positive definite operator such that for any m, Vm is $\mathcal{F}_1$-measurable. Let $M_{1:t} : \mathcal{H} \to \mathbb{R}^{t-1}$ be the operator such that for any $v \in \mathcal{H}$, the kth element of $M_{1:t}v$ is $\langle m_k, v\rangle$. Then, V_t can be written as $V_t = M_{1:t}^* M_{1:t}$.

Let $\mathcal{N}(m, B)$ denote the Gaussian measure on $\mathcal{H}$ with mean $m \in \mathcal{H}$ and a positive definite self-adjoint trace class covariance operator B (Maniglia and Rhandi, 2004). Let $\det(I + C)$ denote the Fredholm determinant of $I + C$, where C is any trace class operator. The next two lemmas are stated as Lemma 1.2.7 and Proposition 1.2.8 in (Maniglia and Rhandi, 2004). We will use these lemmas to calculate certain expectations under certain Gaussian measures.

Lemma 3.1 (Invariance of Gaussian measures under affine transformations) Let $\mathcal{H}$ be a separable Hilbert spaces. Consider the affine transformation $F : \mathcal{H} \to \mathcal{H}$ defined by $F(x) = Qx + z$, where $Q \in \mathcal{L}(\mathcal{H})$ and $z \in \mathcal{H}$. If we set $\mu_k = \mathcal{N}(m, B) \circ F^{-1}$, the measure defined by $\mu_k(A) = \mathcal{N}(m, B)(F^{-1}(A))$, $A \in \mathcal{F}_k$, then

$$\mu_k = \mathcal{N}(Qm + z, QBQ^*) \; .$$

Lemma 3.2 (Integration of $e^{\frac{\alpha}{2}\|x\|^2}$ under Gaussian measures) Let $\mathcal{N}(m, B)$ be a Gaussian measure on $\mathcal{H}$. Then there is an orthonormal basis (e_n) of $\mathcal{H}$ such that $Be_n = \lambda_n e_n$, $\lambda_n \ge 0$, $n \in \mathbb{N}$. Moreover, for any $\alpha < \alpha_0 = \inf_n \frac{1}{\lambda_n}$, we have

$$\int_{\mathcal{H}} e^{\frac{\alpha}{2}\|x\|^2} \mathcal{N}(m, B)(dx) = (\det(I - \alpha B))^{-1/2} \exp\left(\frac{\alpha}{2}\langle (I - \alpha B)^{-1} m, m\rangle\right) \; .$$

Note that B, being a covariance operator of a Gaussian measure, is trace class. Thus, so is $-\alpha B$. As a consequence, $\det(I-\alpha B)$ is well-defined. In what follows, we use the equality with $\alpha = -1$.

The following inequality, which is a standard tool in proving tail inequalities, will play a crucial role in our proof:

Lemma 3.3 Consider (η_t), (m_t) as defined above and let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)$. Let $\lambda \in \mathcal{H}$ be arbitrary and consider

$$P_t^\lambda = \exp\left(\sum_{k=1}^{t-1}\left[\frac{\eta_{k+1}\langle\lambda, m_k\rangle}{R} - \frac{1}{2}\langle\lambda, m_k\rangle^2\right]\right).$$

Then P_τ^λ is almost surely well-defined and

$$\mathbb{E}\left[P_\tau^\lambda\right] \leq 1.$$

Proof. The proof is standard, at least until the stopping time τ is considered. We give the proof for the sake of completeness. First, we claim that $P_t = P_t^\lambda$ is a supermartingale. Let

$$D_k = \exp\left(\frac{\eta_{k+1}\langle\lambda, m_k\rangle}{R} - \frac{1}{2}\langle\lambda, m_k\rangle^2\right).$$

Observe that by the choice of $\gamma = \langle\lambda, m_k\rangle/R$ in (3.3), we have $\mathbb{E}\left[D_k\,|\mathcal{F}_k\right] \leq 1$. Clearly, D_{k-1} is $\mathcal{F}_k$-adapted, as is P_k. Further,

$$\mathbb{E}\left[P_t|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[D_1\cdots D_{t-2}D_{t-1}\,|\mathcal{F}_{t-1}\right] = D_1\cdots D_{t-2}\,\mathbb{E}\left[D_{t-1}\,|\mathcal{F}_{t-1}\right] \leq P_{t-1},$$

showing that (P_t) is indeed a supermartingale.

Now, this immediately leads to the desired result when $\tau = t$ for some deterministic time t. This is based on the fact that the mean of any supermartingale can be bounded by the mean of its first element. In the case of (P_t), for example, we have $\mathbb{E}\left[P_t\right] = \mathbb{E}\left[\mathbb{E}\left[P_t|\mathcal{F}_{t-1}\right]\right] \leq \mathbb{E}\left[P_{t-1}\right] \leq \ldots \leq \mathbb{E}\left[P_1\right] = \mathbb{E}\left[D_1\right] \leq 1$.

Now, in order to consider the general case, let $S_t = P_{\tau\wedge t}$. It is well known that (S_t) is still a supermartingale with $\mathbb{E}\left[S_t\right] \leq \mathbb{E}\left[S_1\right] = \mathbb{E}\left[P_1\right] = 1$. Further, since P_t was non-negative, so is S_t. Hence, by the convergence theorem for non-negative supermartingales, almost surely, $\lim_{t\to\infty} S_t$ exists, i.e., P_τ is almost surely well-defined. Further, $\mathbb{E}\left[P_\tau\right] = \mathbb{E}\left[\liminf_{t\to\infty} S_t\right] \leq \liminf_{t\to\infty}\mathbb{E}\left[S_t\right] \leq 1$ by Fatou's Lemma. □

We now show how to obtain a self-normalized bound for vector-valued martingales using the method of mixtures, originally used by Robbins and Siegmund (1970) to evaluate boundary crossing probabilities for Brownian motion.

Theorem 3.4 Let (η_t), (m_t), (S_t), (V_t), and $(\mathcal{F}_t)$ be as before and let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)$. Let $s > 0$ be an arbitrary integer and U_s be a positive definite, deterministic operator whose inverse is trace class. Then, for any $0 < \delta < 1$, with probability $1-\delta$,

$$\|S_\tau\|^2_{(V_\tau+U_s)^{-1}} \leq 2R^2\log\left(\frac{\det\left(I + M_{1:\tau}U_s^{-1}M_{1:\tau}^*\right)^{1/2}}{\delta}\right). \tag{3.6}$$

Proof. Without loss of generality, assume that $R = 1$ (by appropriately scaling S_t, this can always be achieved). Let

$$G_t(\lambda) = \exp\left(\,\langle\lambda, S_t\rangle - \tfrac{1}{2}\,\|\lambda\|^2_{V_t}\,\right).$$

Notice that $\langle\lambda, S_t\rangle = \sum_{k=1}^{t-1}\eta_{k+1}\langle\lambda, m_k\rangle$, $\|\lambda\|^2_{V_t} = \sum_{k=1}^{t-1}\langle\lambda, m_k\rangle^2$, and that by Lemma 3.3, the mean of $G_\tau(\lambda)$ is not larger than one.

Let Λ be a Gaussian random variable that is independent of all the other random variables and whose covariance operator is U_s^{-1}. Define

$$G_t = \mathbb{E}\left[G_t(\Lambda)|\mathcal{F}_\infty\right] .$$

Clearly, we still have $\mathbb{E}\left[G_\tau\right] = \mathbb{E}\left[\,\mathbb{E}\left[\,G_\tau(\Lambda)\,|\,\Lambda\,\right]\right] \leq 1$.

Let us calculate G_t: Let $f = \mathcal{N}(0, U_s^{-1})$ denote the Gaussian measure underlying Λ. We have that

$$\langle \lambda, S_t\rangle - \tfrac{1}{2}\,\|\lambda\|_{V_t}^2 = \tfrac{1}{2}\,\|S_t\|_{V_t^{-1}}^2 - \tfrac{1}{2}\,\left\|\lambda - V_t^{-1}S_t\right\|_{V_t}^2$$

Thus,

$$G_t = \exp\left(\tfrac{1}{2}\,\|S_t\|_{V_t^{-1}}^2\right)\int_{\mathcal{H}} \exp\left(-\tfrac{1}{2}\,\left\|\lambda - V_t^{-1}S_t\right\|_{V_t}^2\right) f(d\lambda) .$$

Let $\nu = M_{1:t}(\lambda - V_t^{-1}S_t)$ and $g = \mathcal{N}(-M_{1:t}V_t^{-1}S_t, M_{1:t}U_s^{-1}M_{1:t}^*)$. By Lemma 3.1 we have that

$$G_t = \exp\left(\tfrac{1}{2}\,\|S_t\|_{V_t^{-1}}^2\right)\int_{\mathcal{H}} \exp\left(-\tfrac{1}{2}\,\|\nu\|^2\right) g(d\nu) .$$

Now, by Lemma 3.2 we calculate that

$$G_t = \det\left(I + M_{1:t}U_s^{-1}M_{1:t}^*\right)^{-1/2}\exp\left(\tfrac{1}{2}\,\|S_t\|_{V_t^{-1}}^2 - \tfrac{1}{2}\,\|S_t\|_{V_t^{-1}M_{1:t}(I+M_{1:t}U_s^{-1}M_{1:t}^*)^{-1}M_{1:t}V_t^{-1}}^2\right) .$$

By elementary algebra, we obtain that

$$V_t^{-1} - V_t^{-1}M_{1:t}(I + M_{1:t}U_s^{-1}M_{1:t}^*)^{-1}M_{1:t}V_t^{-1} = (U_s + V_t)^{-1}$$

and thus

$$G_t = \det\left(I + M_{1:t}U_s^{-1}M_{1:t}^*\right)^{-1/2}\exp\left(\tfrac{1}{2}\,\|S_t\|_{(U_s+V_t)^{-1}}^2\right) .$$

Now, from $\mathbb{E}\left[G_\tau\right] \leq 1$, we get

$$\begin{aligned}
\mathbb{P}\left(\|S_\tau\|_{(U_s+V_\tau)^{-1}}^2 > 2\log\left(\det\left(I + M_{1:\tau}U_s^{-1}M_{1:\tau}^*\right)^{1/2}\;\frac{1}{\delta}\right)\right) & \\
&= \mathbb{P}\left(\frac{\exp\left(\tfrac{1}{2}\,\|S_\tau\|_{(U_s+V_\tau)^{-1}}^2\right)}{\delta^{-1}\det\left(I + M_{1:\tau}U_s^{-1}M_{1:\tau}^*\right)^{1/2}} > 1\right) \\
&\leq \mathbb{E}\left[\frac{\exp\left(\tfrac{1}{2}\,\|S_\tau\|_{(U_s+V_\tau)^{-1}}^2\right)}{\delta^{-1}\det\left(I + M_{1:\tau}U_s^{-1}M_{1:\tau}^*\right)^{1/2}}\right] \\
&= \mathbb{E}\left[G_\tau\right]\delta \leq \delta,
\end{aligned}$$

thus finishing the proof. □

Notice that the previous result does not apply to important special cases such as $U_s = \lambda I$, $\lambda > 0$. Next we extend Theorem 3.4 to positive-definite regularizers that are not necessarily trace class.

Corollary 3.5 (Self-Normalized Bound for Vector-Valued Martingales) Let (η_t), (m_t), (S_t), (V_t), and $(\mathcal{F}_t)$ be as before and let τ be a stopping time with respect to the filtration $(\mathcal{F}_t)$. Assume that V is a deterministic positive-definite operator with a bounded inverse. Then, for any $0 < \delta < 1$, with probability $1 - \delta$,

$$\|S_\tau\|_{\overline{V}_\tau^{-1}}^2 \leq 2R^2\log\left(\frac{\det\left(I + M_{1:\tau}V^{-1}M_{1:\tau}^*\right)^{1/2}}{\delta}\right) . \tag{3.7}$$

Proof. First, let us construct a sequence (U_s^{-1}) of trace-class operators with limit V^{-1} in the weak-operator topology.[4] Let (e_i) be an orthonormal basis of $\mathcal{H}$. For $x \in \mathcal{H}$, $x = \sum_i \lambda_i e_i$, define $W_s x = \sum_{i=1}^s \lambda_i V^{-1} e_i$. Note that W_s is well-defined and is positive definite, trace-class. Thus, a positive definite inverse of W_s exists; let's denote the inverse of W_s by U_s. It remains to be shown that V^{-1} is the limit of (W_s) in the weak-operator topology. To check this, take $x, y \in \mathcal{H}$ and let $\lambda_i = \langle x, e_i \rangle$. Then, $\langle y, W_s x \rangle = \sum_{i=1}^s \lambda_i \langle y, V^{-1} e_i \rangle = \langle (V^{-1})^* y, \sum_{i=1}^s \lambda_i e_i \rangle$. Thus, $|\langle y, W_s x \rangle - \langle (V^{-1})^* y, x \rangle| = |\langle (V^{-1})^* y, \sum_{i=1}^s \lambda_i e_i - x \rangle| \leq \|(V^{-1})^* y\| \| \sum_{i=1}^s \lambda_i e_i - x\| \to 0$ as $s \to \infty$, which finishes the proof.

Let

$$B_s = 2R^2 \log \left(\frac{\det \left(I + M_{1:\tau} U_s^{-1} M_{1:\tau}^* \right)^{1/2}}{\delta} \right), \quad B = 2R^2 \log \left(\frac{\det \left(I + M_{1:\tau} V^{-1} M_{1:\tau}^* \right)^{1/2}}{\delta} \right).$$

Note that $M_{1:\tau} V^{-1} M_{1:\tau}^*$ is a $\tau \times \tau$ matrix and hence B is well-defined. Now, since U_s^{-1} converges to V^{-1} in the weak-operator topology, each element of the matrix $M_{1:\tau} V^{-1} M_{1:\tau}^*$ is the almost sure limit of the corresponding element of the sequence $(M_{1:\tau} U_s^{-1} M_{1:\tau}^*)_s$. Let $Z_s = \mathbf{1}\left\{ \|S_\tau\|^2_{(V_\tau + U_s)^{-1}} \leq B_s \right\}$, and $Z = \mathbf{1}\left\{ \|S_\tau\|^2_{V_\tau^{-1}} \leq B \right\}$. We have that $\|S_\tau\|^2_{(V_\tau + U_s)^{-1}} \to \|S_\tau\|^2_{V_\tau^{-1}}$ and $B_s \to B$ almost surely. Thus, $Z_s \to Z$ almost surely. Because $Z_s \leq 1$ for all s, by Lebesgue's dominated convergence theorem we get that $\mathbb{E}[Z_s] \to \mathbb{E}[Z]$. By Theorem 3.4, we have that $\mathbb{E}[Z_s] \geq 1 - \delta$ for any s. Thus, $\limsup \mathbb{E}[Z_s] \geq 1 - \delta$, which implies that $\mathbb{E}[Z] \geq 1 - \delta$, finishing the proof. □

Corollary 3.6 (Uniform Bound) Under the same assumptions as in the previous corollary, for any $0 < \delta < 1$, with probability $1 - \delta$,

$$\forall t \geq 1, \qquad \|S_t\|^2_{V_t^{-1}} \leq 2R^2 \log \left(\frac{\det \left(I + M_{1:t} V^{-1} M_{1:t}^* \right)^{1/2}}{\delta} \right). \tag{3.8}$$

Proof. We will use a stopping time construction, which goes back at least to Freedman (1975). Define the bad event

$$B_t(\delta) = \left\{ \omega \in \Omega : \|S_t\|^2_{V_t^{-1}} > 2R^2 \log \left(\frac{\det \left(I + M_{1:t} V^{-1} M_{1:t}^* \right)^{1/2}}{\delta} \right) \right\} \tag{3.9}$$

We are interested in bounding the probability that $\bigcup_{t \geq 1} B_t(\delta)$ happens. Define $\tau(\omega) = \min\{t \geq 1 : \omega \in B_t(\delta)\}$, with the convention that $\min \emptyset = \infty$. Then, τ is a stopping time. Further,

$$\bigcup_{t \geq 1} B_t(\delta) = \{\omega : \tau(\omega) < \infty\}.$$

Thus, by Theorem 3.4

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t \geq 1} B_t(\delta) \right) &= \mathbb{P}(\tau < \infty) \\
&= \mathbb{P}\left(\|S_\tau\|^2_{V_\tau^{-1}} > 2R^2 \log \left(\frac{\det \left(I + M_{1:t} V^{-1} M_{1:t}^* \right)^{1/2}}{\delta} \right), \tau < \infty \right) \\
&\leq \mathbb{P}\left(\|S_\tau\|^2_{V_\tau^{-1}} > 2R^2 \log \left(\frac{\det \left(I + M_{1:t} V^{-1} M_{1:t}^* \right)^{1/2}}{\delta} \right) \right) \\
&\leq \delta .
\end{aligned}$$

□

[4] That is, for any $x, y \in \mathcal{H}$, $\langle x, U_s^{-1} y \rangle \to \langle x, V^{-1} y \rangle$.

The quantity $\log\det\left(I + M_{1:\tau}V^{-1}M_{1:\tau}^*\right)$ is bounded for several kernels in (Srinivas et al., 2012). Here, we demonstrate a simple bound for two special cases.

Corollary 3.7 Assume that the vectors (m_k) come from a finite, K-element set, $\{v_1, \ldots, v_K\} \subset \mathcal{H}$. Then

$$\log\det\left(I + M_{1:\tau}V^{-1}M_{1:\tau}^*\right) \le K\log\left(1 + \frac{\tau}{K}\right) .$$

Proof. Choose regularizer V such that for $i \le K$, $V^{-1/2}v_i = e_i$. Thus, each row of $M_{1:\tau}V^{-1/2}$ is an identity vector. Let C be the $K \times K$ diagonal operator with each diagonal element $C_{s,s} = N_s + 1$, where N_s is the number of times we have observed v_s. Then, by the matrix determinant lemma, it is easy to see that

$$\det\left(I + M_{1:\tau}V^{-1}M_{1:\tau}^*\right) = \det(C) .$$

This means

$$\det\left(I + M_{1:\tau}V^{-1}M_{1:\tau}^*\right) = \prod_{s=1}^{K}(1 + N_s) \le \left(1 + \frac{\tau}{K}\right)^K .$$

□

By the same matrix-determinant lemma, we also get a bound for the case when $\mathcal{H}$ is finite-dimensional:

$$\det(\overline{V}_t) = \det(V)\det(I + M_{1:t}V^{-1}M_{1:t}^\top) ,$$

resulting in the finite-dimensional version of Corollary 3.6:

Corollary 3.8 (Finite-Dimensional Case) Assume the same as in Corollary 3.6 and assume that $\mathcal{H} = \mathbb{R}^d$ for some positive integer d. Then, for any $0 < \delta < 1$, with probability $1 - \delta$,

$$\forall t \ge 1, \qquad \|S_t\|_{\overline{V}_t^{-1}} \le R\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} .$$

This result can be compared with a recent result that can be extracted from the paper by Rusmevichientong and Tsitsiklis (2010).

Theorem 3.9 (Rusmevichientong and Tsitsiklis (2010)) Consider the finite-dimensional version of the processes (S_t), $(\overline{V}_t)$ as defined above and let

$$\kappa = \sqrt{3 + 2\log((L^2 + \mathrm{trace}(V))/\lambda_0)} .$$

Then, for any $0 < \delta < 1$, $t \ge 2$, with probability at least $1 - \delta$,

$$\|S_t\|_{\overline{V}_t^{-1}} \le 2\kappa^2 R\sqrt{\log t}\sqrt{d\log(t) + \log(1/\delta)} .$$

Remark 3.10 By combining Corollary 3.8 and Lemma E.3 in Appendix E, we get a simple worst case bound that holds with probability $1 - \delta$:

$$\forall t \ge 1, \quad \|S_t\|^2_{\overline{V}_t^{-1}} \le d\,R^2\log\left(\frac{\mathrm{trace}(V) + tL^2}{d\,\delta}\right) .$$

Despite the use of the crude upper bound in Lemma E.3, we see that the new bound is still considerably better than that of Theorem 3.9. Note that the $\log(t)$ factor cannot be removed from this new bound, as shown by Problem 3, page 203 in the book by de la Peña et al. (2009).

3.3 Optional Skipping

Consider the case when $d = 1$, $m_k = \varepsilon_k \in \{0, 1\}$, i.e., the case of an optional skipping process (see Appendix C for definitions). Then, using again $V = I = 1$, $\overline{V}_t = 1 + \sum_{k=1}^{t-1} \varepsilon_k \doteq 1 + N_t$ and thus the expression studied becomes

$$\|S_t\|_{\overline{V}_t^{-1}} = \frac{|\sum_{k=1}^{t-1} \varepsilon_k \eta_{k+1}|}{\sqrt{1 + N_t}} \; .$$

We also have $\log\det(\overline{V}_t) = \log(1 + N_t)$. Thus, we get, with probability $1 - \delta$

$$\forall s \geq 1, \quad \left|\sum_{k=1}^{s-1} \varepsilon_k \eta_{k+1}\right| \leq \sqrt{2(1 + N_s)\, \log\left(\frac{(1 + N_s)^{1/2}}{\delta}\right)}. \tag{3.10}$$

If we apply Doob's optional skipping and Hoeffding-Azuma (see Appendix C), with a union bound (see, e.g., the paper of Bubeck et al. (2008)), we would get, for any $0 < \delta < 1$, $t \geq 3$, with probability $1 - \delta$,

$$\forall s \in \{1, \ldots, t\}, \qquad \left|\sum_{k=1}^{s-1} \varepsilon_k \eta_{k+1}\right| \leq \sqrt{2N_s\, \log\left(\frac{2t}{\delta}\right)}. \tag{3.11}$$

The major difference between these bounds is that (3.11) depends explicitly on t, while (3.10) does not. This has the positive effect that one need not recompute the bound if N_t does not grow, which helps e.g. in the paper of Bubeck et al. (2008) to improve the computational complexity of the HOO algorithm.

Instead of a union bound, it is possible to use a "peeling device" to replace the conservative $\log t$ factor in the above bound by essentially $\log\log t$. This is done e.g. in Garivier and Moulines (2008) in their Theorem 22.[5] From their derivations, the following one sided, uniform bound can be extracted (see Remark 24, page 19): For any $0 < \delta < 1$, $t \geq 3$, with probability $1 - \delta$,

$$\forall s \in \{1, \ldots, t\}, \qquad \sum_{k=1}^{s-1} \varepsilon_k \eta_{k+1} \leq \sqrt{\frac{4\, N_s}{1.99}\, \log\left(\frac{6 \log t}{\delta}\right)}. \tag{3.12}$$

As noted by Garivier and Moulines (2008), due to the law of iterated logarithm (see Appendix C), the scaling of the RHS as a function of t cannot be improved in the worst-case. However, this leaves open the possibility of deriving a maximal inequality that depends on t only through N_t.

3.4 Application to Least-Squares Estimation

In this section we first apply Theorem 3.4 to derive confidence intervals for least-squares estimation, where the covariate process is an arbitrary process. In particular, our assumption on the data is as follows:

Assumption A1 Linear Response Assumption Let $(\mathcal{F}_k)$ be a filtration, (m_1, y_1), …, (m_t, y_t) be a sequence of random variables over $\mathcal{H} \times \mathbb{R}$ such that m_k is $\mathcal{F}_k$-measurable, and y_k is $\mathcal{F}_{k+1}$-measurable ($k = 1, 2, \ldots$). Assume that there exists $\theta_* \in \mathcal{H}$ such that $\mathbb{E}[y_k|\mathcal{F}_k] = \langle m_k, \theta_* \rangle$, i.e., $\eta_{k+1} = y_k - \langle m_k, \theta_* \rangle$ is a martingale difference sequence ($\mathbb{E}[\eta_{k+1}|\mathcal{F}_k] = 0$, $k = 1, 2, \ldots$) and that η_k is R-sub-Gaussian.

[5] They give their theorem as ratios, which they should not, since their inequality then fails to hold for $N_t = 0$. However, this is easy to remedy by reformulating their result as we do it here.

We shall call the random variables m_k covariates and the random variables y_k the responses. Note that the assumption allows any sequential generation of the covariates.

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

$$\widehat{\theta}_t = (M^* M + \lambda I)^{-1} M^* Y, \qquad \widehat{\theta}_1 = 0, \tag{3.13}$$

where $M = M_{1:t}$ and $M^* = M^*_{1:t}$ and $Y = (y_1, \dots, y_{t-1})^\top$. We further let $\eta = (\eta_2, \dots, \eta_t)^\top$.

We are interested in deriving a confidence bound on the error of predicting the mean response $\langle m, \theta_* \rangle$ at an arbitrarily chosen random covariate m using the least-squares predictor $\langle m, \widehat{\theta}_t \rangle$. Using

$$\begin{aligned}
\widehat{\theta}_t &= (M^* M + \lambda I)^{-1} M^* (M\theta_* + \eta) \\
&= (M^* M + \lambda I)^{-1} M^* \eta + (M^* M + \lambda I)^{-1} (M^* M + \lambda I)\theta_* - \lambda (M^* M + \lambda I)^{-1} \theta_* \\
&= (M^* M + \lambda I)^{-1} M^* \eta + \theta_* - \lambda (M^* M + \lambda I)^{-1} \theta_* \,,
\end{aligned}$$

we get

$$\begin{aligned}
\langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle &= \langle m, M^* \eta \rangle_{(M^* M + \lambda I)^{-1}} - \lambda \langle m, \theta_* \rangle_{(M^* M + \lambda I)^{-1}} \\
&= \langle m, M^* \eta \rangle_{\overline{V}_t^{-1}} - \lambda \langle m, \theta_* \rangle_{\overline{V}_t^{-1}},
\end{aligned}$$

where $\overline{V}_t = M^* M + \lambda I$. Note that $\overline{V}_t$ is positive definite (thanks to $\lambda > 0$) and hence so is $\overline{V}_t^{-1}$, so the above inner product is well-defined. Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
|\langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle| &\le \|m\|_{\overline{V}_t^{-1}} \left(\|M^* \eta\|_{\overline{V}_t^{-1}} + \lambda \, \|\theta_*\|_{\overline{V}_t^{-1}} \right) \\
&\le \|m\|_{\overline{V}_t^{-1}} \left(\|M^* \eta\|_{\overline{V}_t^{-1}} + \lambda^{1/2} \, \|\theta_*\| \right),
\end{aligned}$$

where we used that $\|\theta_*\|^2_{\overline{V}_t^{-1}} \le 1/\lambda_{\min}(\overline{V}_t) \, \|\theta_*\|^2 \le 1/\lambda \, \|\theta_*\|^2$. Fix any $0 < \delta < 1$. By Corollary 3.6, with probability at least $1 - \delta$,

$$\forall t \ge 1, \quad \|M^* \eta\|_{\overline{V}_t^{-1}} \le R \sqrt{2 \log \left(\frac{\det(I + M_{1:t} M^*_{1:t} / \lambda)^{1/2}}{\delta} \right)}.$$

Therefore, on the event where this inequality holds, one also has

$$|\langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle| \le \|m\|_{\overline{V}_t^{-1}} \left(R \sqrt{2 \log \left(\frac{\det(I + M_{1:t} M^*_{1:t} / \lambda)^{1/2}}{\delta} \right)} + \lambda^{1/2} \, \|\theta_*\| \right).$$

Similarly, we can derive a worst-case bound. The result is summarized in the following statement:

Theorem 3.11 Let $(m_1, y_1), \dots, (m_{t-1}, y_{t-1})$, $m_k \in \mathcal{H}$, $y_k \in \mathbb{R}$ satisfy the linear model Assumption A1 with some $R > 0$, $\theta_* \in \mathcal{H}$ and let $(\mathcal{F}_t)$ be the associated filtration. Assume that $\|\theta_*\| \le S$. Consider the ℓ^2-regularized least-squares parameter estimate $\widehat{\theta}_t$ with regularization coefficient $\lambda > 0$ (cf. (3.13)). Let m be an arbitary, $\mathcal{H}$-valued random variable. Let $\overline{V}_t = \lambda I + \sum_{k=1}^{t-1} m_k \otimes m_k$ be the regularized design matrix underlying the covariates. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, for any $m \in \mathcal{H}$,

$$\forall t \ge 1, \quad |\langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle| \le \|m\|_{\overline{V}_t^{-1}} \left(R \sqrt{2 \log \left(\frac{\det(I + M_{1:t} M^*_{1:t} / \lambda)^{1/2}}{\delta} \right)} + \lambda^{1/2} \, S \right). \tag{3.14}$$

Further, if the covariates satisfy $\|m_k\| \le L$, $k = 1, \dots, t-1$, then with probability $1 - \delta$, for any $m \in \mathcal{H}$,

$$\forall t \ge 1, \quad |\langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle| \le \|m\|_{\overline{V}_t^{-1}} \left(R \sqrt{\frac{(t-1)L^2}{\lambda} + 2 \log \left(\frac{1}{\delta} \right)} + \lambda^{1/2} \, S \right). \tag{3.15}$$

Remark 3.12 We see that $\lambda \to \infty$ increases the second term (the "bias term") in the parenthesis of the estimate. In fact, $\lambda \to \infty$ for n fixed gives $\lambda^{1/2} \|m\|_{V_t^{-1}} \to$ const (as it should be). Decreasing λ, on the other hand increases $\|m\|_{V_t^{-1}}$ and the log term, while it decreases the bias term $\lambda^{1/2} S$.

Remark 3.13 Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with underlying kernel function k, and K be the Gramian matrix. It is easy to see that

$$\lambda(M^*M + \lambda I)^{-1} = I - M^*(MM^* + \lambda I)M.$$

Thus,

$$\begin{aligned} \lambda \|m\|^2_{\overline{V}_t^{-1}} &= \langle m, m\rangle - \|m\|^2_{M^*(MM^*+\lambda I)^{-1}M} \\ &= k(m,m) - k(m,.)^\top (K + \lambda I)^{-1} k(m,.) \end{aligned} \tag{3.16}$$

Further, by matrix-determinant lemma, we have that

$$\det(I + MM^*/\lambda) = \det(I + K/\lambda). \tag{3.17}$$

Equations (3.16) and (3.17) let us compute the RHS of (3.14) when $\mathcal{H}$ is a RKHS.

Theorem 3.11 can be compared with Theorem 6 of Srinivas et al. (2010).

Theorem 3.14 (Srinivas et al. (2010)) Let $\delta \in (0,1)$. Assume that the noise variables have a Gaussian distribution with variance σ^2. Define

$$\beta_t = 2\|\theta_*\|^2 + 150 \log^3(t/\delta) \log\det(I + \sigma^{-2} M_{1:t} M_{1:t}^*) \ .$$

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

$$\forall t \geq 1, \quad \left|\langle m, \widehat{\theta}_t\rangle - \langle m, \theta_*\rangle\right| \leq \beta_{t+1}^{1/2} \|m\|_{(\sigma^{-2}V_t + I)^{-1}} \ .$$

Apart from improving in logarithmic terms and constants, our bound applies to the substantially more general case when the noise is sub-Gaussian, while the result of Srinivas et al. (2010), although they claim otherwise, applies only to Gaussian noise. In particular, the proof of Lemma 7.2 of Srinivas et al. (2010) uses the explicit form of Gaussian probability distribution function and it is unclear how this could be avoided with their proof technique.

From Theorem 3.11, we immediately obtain confidence bounds for θ_*:

Corollary 3.15 Under the conditions of Theorem 3.11, with probability at least $1-\delta$,

$$\forall t \geq 1, \quad \left\|\widehat{\theta}_t - \theta_*\right\|_{\overline{V}_t} \leq R\sqrt{2\log\left(\frac{\det(I + M_{1:t}M_{1:t}^*/\lambda)^{1/2}}{\delta}\right)} + \lambda^{1/2} S.$$

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

$$\forall t \geq 1, \quad \left\|\widehat{\theta}_t - \theta_*\right\|_{\overline{V}_t} \leq R\sqrt{\frac{(t-1)L^2}{\lambda} + 2\log\left(\frac{1}{\delta}\right)} + \lambda^{1/2} S.$$

Proof. Plugging in $m = \overline{V}_t(\widehat{\theta}_t - \theta_*)$ into (3.14), we get

$$\left\|\widehat{\theta}_t - \theta_*\right\|^2_{\overline{V}_t} \leq \left\|\overline{V}_t(\widehat{\theta}_t - \theta_*)\right\|_{\overline{V}_t^{-1}} \left(R\sqrt{2\log\left(\frac{\det(I + M_{1:t}M_{1:t}^*/\lambda)^{1/2}}{\delta}\right)} + \lambda^{1/2} S\right). \tag{3.18}$$

Now, $\left\|\overline{V}_t(\widehat{\theta}_t - \theta_*)\right\|^2_{\overline{V}_t^{-1}} = \left\|\widehat{\theta}_t - \theta_*\right\|^2_{\overline{V}_t}$ and therefore either $\left\|\widehat{\theta}_t - \theta_*\right\|_{\overline{V}_t} = 0$, in which case the conclusion holds, or we can divide both sides of (3.18) by $\left\|\widehat{\theta}_t - \theta_*\right\|_{\overline{V}_t}$ to obtain the desired result. □

In what follows, we denote the "radius" of the confidence set at time t by $\beta_t(\delta)$:

$$\beta_t(\delta) = \left(R\sqrt{2\log\left(\frac{\det(I + M_{1:t}M_{1:t}^*/\lambda)^{1/2}}{\delta}\right)} + \lambda^{1/2}\, S \right)^2 .$$

Remark 3.16 In fact, the theorem and the corollary are equivalent. To see this note that $\langle m, \widehat{\theta}_t - \theta_* \rangle = \langle \overline{V}_t^{1/2}(\widehat{\theta}_t - \theta_*), \overline{V}_t^{-1/2} m \rangle$, thus

$$\sup_{m \neq 0} \frac{|\langle m, \widehat{\theta}_t - \theta_* \rangle|}{\|m\|_{\overline{V}_t^{-1}}} = \left\| \widehat{\theta}_t - \theta_* \right\|_{\overline{V}_t} .$$

Corollary 3.17 Assume the same as in Theorem 3.11 except that inputs $m_i \in \mathbb{R}^d$ lie in an Euclidean space. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 1$, θ_* lies in the set

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

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

$$C_t' = \left\{ \theta \in \mathbb{R}^d \,:\, \left\| \widehat{\theta}_t - \theta \right\|_{\overline{V}_t} \leq R\sqrt{d\log\left(\frac{1 + (t-1)L^2/\lambda}{\delta}\right)} + \lambda^{1/2}\, S \right\} .$$

The above bound could be compared with a similar bound of Dani et al. (2008) whose bound, under identical conditions, states that (with appropriate initialization) with probability $1 - \delta$,

$$\text{for all } t \text{ large enough} \quad \left\| \widehat{\theta}_t - \theta_* \right\|_{\overline{V}_t} \leq R \max\left\{ \sqrt{128\, d \log(t)\, \log\left(\frac{t^2}{\delta}\right)}, \frac{8}{3}\, \log\left(\frac{t^2}{\delta}\right) \right\} , \tag{3.19}$$

where large enough means that t satisfies $0 < \delta < t^2 e^{-1/16}$. Denote by $\sqrt{\tilde{\beta}_t(\delta)}$ the RHS in the last bound. The restriction on t comes from the fact that $\tilde{\beta}_t(\delta) \geq 2d(1 + 2\log(t))$ is needed in the proof of the last inequality of their Theorem 5.

On the other hand, Rusmevichientong and Tsitsiklis (2010) proved that for any *fixed* $t \geq 2$, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\left\| \widehat{\theta}_t - \theta_* \right\|_{\overline{V}_t} \leq 2\,\kappa^2 R\sqrt{\log t}\, \sqrt{d\, \log(t) + \log(1/\delta)} + \lambda^{1/2} S\,,$$

where $\kappa = \sqrt{3 + 2\log((L^2 + \mathrm{trace}(V))/\lambda}$. To get a uniform bound one can use a union bound with $\delta_t = \delta/t^2$. Then $\sum_{t=2}^{\infty} \delta_t = \delta(\frac{\pi^2}{6} - 1) \leq \delta$. This thus gives that for any $0 < \delta < 1$, with probability at least $1 - \delta$,

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

This is tighter than (3.19), but is still lagging behind the result of Corollary 3.17. Note that the new confidence set seems to require the computation of a determinant of a matrix, a potentially expensive step. However, one can speed up the computation by using the matrix determinant lemma, exploiting that the matrix whose determinant is needed is obtained via a rank-one update (cf. the proof of Lemma E.1 in the Appendix). This way, the determinant can be kept up-to-date with linear time computation.

3.5 Online-to-Confidence-Conversion

The confidence set of Section 3.4 is essentially constructed from the predictions of the online least-squares method. In this section we show that, more generally, the predictions of any online algorithm that predicts the responses of the chosen inputs in a sequential manner can be "converted" to a confidence set.

In online linear prediction, we assume that in round t an online algorithm receives $m_t \in \mathcal{H}$, predicts $\widehat{y}_t \in \mathbb{R}$, receives $y_t \in \mathbb{R}$ and suffers a loss $\ell_t(\widehat{y}_t)$ where $\ell_t(y) = (y - y_t)^2$ is the *quadratic prediction loss*. In online linear prediction, one makes no assumptions on the sequence $\{(m_t, y_t)\}_{t=1}^{\infty}$, perhaps except for bounds on the norm of m_t and magnitude of y_t. In fact, the sequence $\{(m_t, y_t)\}_{t=1}^{\infty}$ can be chosen in an adversarial fashion.

The task of the online algorithm is to keep its T-step cumulative loss $\sum_{t=1}^{T} \ell_t(\widehat{y}_t)$ as low as possible. We compare the loss of the algorithm with the loss of the strategy that uses a fixed weight vector $\theta \in \mathcal{H}$ and in round t predicts $\langle \theta, m_t \rangle$ – this is why the problem is called linear prediction. The difference of the losses is called the *regret with respect to* θ and formally we write it as

$$\rho_T(\theta) = \sum_{t=1}^{T} \ell_t(\widehat{y}_t) - \sum_{t=1}^{T} \ell_t(\langle \theta, m_t \rangle) \; .$$

The construction of algorithms with "small" regret $\rho_T(\theta)$ is an important topic in the online learning literature. Examples of algorithms designed to achieve this include variants of the least-squares method (projected or regularized), the exponentiated gradient algorithm, the p-norm regularized algorithm, online LASSO, SeqSEW, etc.

Suppose now that we feed an online algorithm for linear prediction with a stochastic sequence $\{(m_t, y_t)\}_{t=1}^{\infty}$ generated according to the model described above. Let the sequence of predictions produced by the algorithm be $\{\widehat{y}_t\}_{t=1}^{\infty}$. The following theorem states that from the sequence $\{\widehat{y}_t\}_{t=1}^{\infty}$ of predictions we can construct high-probability confidence sets C_t for θ_*. Moreover, as we will see the volume of the set C_t will be related to the regret of the algorithm; the smaller the regret of the algorithm, the smaller the volume of C_t is. The theorem below states the precise result.

Theorem 3.18 (Online-to-Confidence-Set Conversion) Assume that $\{F_t\}_{t=1}^{\infty}$ is a filtration and for any $t \geq 1$, m_t is an $\mathcal{H}$-valued, F_t-measurable random variable and η_t is a real-valued, F_t-measurable random variable that is conditionally R-sub-Gaussian. Define $y_t = \langle \theta_*, m_t \rangle + \eta_{t+1}$, where $\theta^* \in \mathbb{R}^d$ is the true parameter. Suppose that we feed $\{(m_t, y_t)\}_{t=1}^{\infty}$ into an online prediction algorithm that, for all $t \geq 1$, admits a regret bound

$$\rho_t(\theta_*) \leq B_t, \qquad \text{(almost surely)}$$

where $\{B_t\}_{t=1}^{\infty}$ is some sequence of $\{F_{t+1}\}_{t=1}^{\infty}$-adapted non-negative random variables. Then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, the true parameter θ_* lies in the intersection of the sets

$$C_T = \left\{ \theta \in \mathcal{H} \, : \, \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta, m_t \rangle)^2 \leq 1 + 2B_T + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1 + B_T}}{\delta} \right) \right\} ,$$

where $T \geq 1$.

The proof of the theorem can be found in Section 3.5.1.

Notice that, as expected, the confidence sets C_T in the theorem can be constructed from observable quantities: the data $m_1, m_2, \ldots, m_T$, $y_1, y_2, \ldots, y_T$, the predictions $\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_T$ of the linear prediction algorithm, the regret bound B_T, the "variance" of the noise R^2 and the confidence parameter δ. Finally, it is not hard to see that since C_T is a sub-level set of a non-negative quadratic function in θ, it is an ellipsoid, possibly, with some of the axes infinitely long.

An important feature of the confidence sets constructed in Theorem 3.18 is that they are based on regret bounds B_T, which can themselves be *data-dependent bounds* on the regret. Although we will not exploit this in the later sections of the paper, in practice, the use of such data dependent bounds (which exists for a large number of the algorithms mentioned) is highly recommended.

Another important feature of the bound is that the unknown parameter vector belongs to the intersection of all the the confidence sets constructed, i.e., the confidence sets hold the true parameter vector *uniformly in time*. This property is useful both because it leads to simpler algorithm designs and also to simpler analysis. Note that usually this property is achieved by taking a union bound, where the failure probability δ at time step T would be divided by a diverging function of T in the definition of the confidence set. With our techniques, we were able to avoid this union bound, which is expected to give better results in practice. In particular, if the online algorithm is "lucky" in that its regret B_T does not grow, or grows very slowly, our confidence set shrink faster than if a union bound was used to ensure uniformity in time.

It turns out that the fact that confidence sets constructed in Theorem 3.18 can be unbounded, might potentially lead to trouble (this happens when the vectors (m_t) do not span the full space $\mathcal{H}$). To deal with this issue, we slightly modify the confidence sets: If we know a priori that $\|\theta_*\| \le E$ we can add $\|\theta\|^2 \le E^2$ to the inequality defining C_T in the theorem. (Other *a priori* information can also be added; though using the Hilbert-space norm leads to computational advantages as we will see.) This leads to the following obvious corollary.

Corollary 3.19 (Regularized Confidence Sets) Assume the same as in Theorem 3.18 and additionally assume that $\|\theta_*\| \le E$. Then, for any $\delta \in (0, 1/4]$, with probability at least $1-\delta$, the true parameter θ_* lies in the intersections of the sets

$$C_T = \left\{ \theta \in \mathcal{H} \,:\, \|\theta\|^2 + \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta, m_t \rangle)^2 \le E^2 + 1 + 2B_T + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_T}}{\delta} \right) \right\},$$

where $T \ge 1$.

Of course, it would be better to take intersection of the confidence sets from Theorem 3.18 and the set $\{\theta \,:\, \|\theta\| \le E\}$ instead, since the resulting confidence set would be smaller than the confidence set constructed in the corollary. However, the resulting confidence set would no longer be an ellipsoid and this might complicate matters later. The confidence set constructed in the corollary is always a bounded non-degenerate ellipsoid and this allows a relatively simple analysis.

Corollary 3.20 Assume the same as in Corollary 3.19. The confidence sets are contained in larger ellipsoids

$$C_{t-1} \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta - \widehat{\theta}_t\|_{\overline{V}_t}^2 \le \beta_t(\delta) \right\},$$

where

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\|\theta\|^2 + \sum_{s=1}^{t-1} (\widehat{y}_s - \langle \theta, m_s \rangle)^2 \right).$$

Proof. Consider the event A when $\theta_* \in \bigcap_{t=1}^{\infty} C_t$. By Corollary 3.19, the event A occurs with probability at least $1-\delta$.

The set C_{t-1} is an ellipsoid underlying the covariance matrix $\overline{V}_t = I + \sum_{s=1}^{t-1} m_s m_s^\top$ and center

$$\widehat{\theta}_t = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\|\theta\|^2 + \sum_{s=1}^{t-1} (\widehat{y}_s - \langle \theta, m_s \rangle)^2 \right).$$

The ellipsoid C_{t-1} is non-empty since θ_* lies in it (on the event A). Therefore $\widehat{\theta}_t \in C_{t-1}$. We can thus express the ellipsoid as

$$C_{t-1} = \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta - \widehat{\theta}_t\|^2_{\overline{V}_t} + \|\widehat{\theta}_t\|^2 + \sum_{s=1}^{t-1} \left(\widehat{y}_s - \langle \widehat{\theta}_t, m_s \rangle \right)^2 \le \beta_t(\delta) \right\} .$$

The ellipsoid is contained in a larger ellipsoid

$$C_{t-1} \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta - \widehat{\theta}_t\|^2_{\overline{V}_t} \le \beta_t(\delta) \right\} = \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta - \widehat{\theta}_t\|_{\overline{V}_t} \le \sqrt{\beta_t(\delta)} \right\} .$$

□

Remark 3.21 As mentioned in Remark 3.16, we can also obtain a pointwise error bound: with probability at least $1-\delta$, for any $m \in \mathcal{H}$,

$$\forall t \ge 1, \quad \left| \langle m, \widehat{\theta}_t \rangle - \langle m, \theta_* \rangle \right| \le \|m\|_{\overline{V}_t^{-1}} \sqrt{\beta_t(\delta)} .$$

3.5.1 Proof of Theorem 3.18

To prove Theorem 3.18, we will need Corollary 3.6 from Section 3.2 and Propositions D.1 and D.2 from Appendix D.

Proof of Theorem 3.18. With probability one,

$$\begin{aligned}
B_T &\ge \rho_T(\theta_*) \\
&= \sum_{t=1}^{T} \ell_t(\widehat{y}_t) - \ell_t(\langle \theta_*, m_t \rangle) \\
&= \sum_{t=1}^{T} (\widehat{y}_t - y_t)^2 - (\langle \theta_*, m_t \rangle - y_t)^2 \\
&= \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle - \eta_{t+1})^2 - \eta_{t+1}^2 \\
&= \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 - 2\eta_{t+1}(\widehat{y}_t - \langle \theta_*, m_t \rangle) .
\end{aligned}$$

Thus, with probability one,

$$\sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 \le B_T + 2 \sum_{t=1}^{T} \eta_{t+1}(\widehat{y}_t - \langle \theta_*, m_t \rangle) . \tag{3.20}$$

The sequence $\{\sum_{t=1}^{T} \eta_{t+1}(\widehat{y}_t - \langle \theta_*, m_t \rangle)\}_{T=1}^{\infty}$ is a martingale adapted to $\{F_{T+1}\}_{T=1}^{\infty}$. We upper bound its tail using Corollary 3.6 with $V = 1$.

Corollary 3.6 gives that with probability at least $1-\delta$, for all $T \ge 1$

$$\left| \sum_{t=1}^{T} \eta_{t+1}(\widehat{y}_t - \langle \theta_*, m_t \rangle) \right| \le R \sqrt{2 \left(1 + \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 \right) \ln \left(\frac{\sqrt{1 + \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2}}{\delta} \right)} .$$

Combining with (3.20), we get

$$\begin{aligned}
\sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 \le B_T + 2R &\sqrt{2 \left(1 + \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 \right)} \\
&\times \sqrt{\ln \left(\frac{\sqrt{1 + \sum_{t=1}^{T} (\widehat{y}_t - \langle \theta_*, m_t \rangle)^2}}{\delta} \right)} .
\end{aligned} \tag{3.21}$$

From this point on, we just need to "solve" this inequality. More precisely, our goal is to isolate a simple function of θ_*. We proceed as follows. We first add 1 to the both sides of the inequality and introduce the notation $z = \sqrt{1 + \sum_{t=1}^{T}(\widehat{y}_t - \langle \theta_*, m_t \rangle)^2}$, $a = B_T + 1$ and $b = 2R\sqrt{2\ln(z/\delta)}$. With this notation, we can write the last equation equivalently in the form

$$z^2 \leq a + bz \; .$$

Since $a \geq 0$ and $b \geq 0$ (since $z \geq 1$ and $\delta \in (0, 1/4]$) we can apply Proposition D.1 and obtain that

$$z \leq b + \sqrt{a} \; .$$

Substituting for b we have

$$z \leq R\sqrt{8\ln(z/\delta)} + \sqrt{a} \; .$$

Introducing the notation $c = \sqrt{a}$ and $f = R\sqrt{8}$ we can write the last inequality equivalently as

$$z \leq c + f\sqrt{\ln(z/\delta)} \; .$$

Therefore, by Proposition D.2,

$$z \leq c + f\sqrt{2\ln\left(\frac{f+c}{\delta}\right)} \; .$$

Substituting for c, a and f we get

$$z \leq \sqrt{B_T + 1} + 4R\sqrt{\ln\left(\frac{R\sqrt{8} + \sqrt{1+B_T}}{\delta}\right)} \; .$$

Squaring both sides and using the inequality $(u+v)^2 \leq 2u^2 + 2v^2$ valid for any $u, v \in \mathbb{R}$, we have

$$z^2 \leq 2B_T + 2 + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_T}}{\delta}\right) \; .$$

Substituting for z^2 and subtracting 1 from both sides we get

$$\sum_{t=1}^{T}(\widehat{y}_t - \langle \theta_*, m_t \rangle)^2 \leq 1 + 2B_T + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1+B_T}}{\delta}\right) \; .$$

This means that $\theta_* \in C_T$ and the proof is finished. □

Chapter 4

Stochastic Linear Bandits[1]

Stochastic linear bandit problem in a separable Hilbert space $\mathcal{H}$ is a sequential decision-making problem where in each round t, the learner is given a decision set $D_t \subseteq \mathcal{H}$ from which he has to choose an action a_t. Subsequently he observes loss $\ell_t(a_t) = \langle a_t, \theta_* \rangle + \eta_t$ where $\theta_* \in \mathcal{H}$ is an unknown parameter and η_t is a random noise satisfying conditions of Assumption A1 of the previous chapter. In what follows, y_t denotes the loss at time t, $\ell_t(a_t)$.

The goal of the learner is to minimize his total loss $\sum_{t=1}^{T} \langle a_t, \theta_* \rangle$ accumulated over the course of T rounds. Clearly, with the knowledge of θ_*, the optimal strategy is to choose in round t the point $a_{*,t} = \operatorname{argmin}_{a \in D_t} \langle a, \theta_* \rangle$ that minimizes the loss. This strategy would accumulate total loss $\sum_{t=1}^{T} \langle a_{*,t}, \theta_* \rangle$. It is thus natural to evaluate the learner relative to this optimal strategy. The difference of the learner's total loss and the total loss of the optimal strategy is called the *pseudo-regret* (Audibert et al., 2009) of the algorithm, which can be formally written as

$$R_T = \left(\sum_{t=1}^{T} \langle a_t, \theta_* \rangle \right) - \left(\sum_{t=1}^{T} \langle a_{*,t}, \theta_* \rangle \right) = \sum_{t=1}^{T} \langle a_t - a_{*,t}, \theta_* \rangle \ .$$

As compared to the regret, the pseudo-regret has the same expected value, but lower variance because the additive noise η_t is removed. However, the omitted quantity is uncontrollable, hence we have no interest in including it in our results (the omitted quantity would also cancel, if η_t was a sequence which is independently selected of $a_1, \ldots, a_t$.) In what follows, for simplicity we use the word *regret* instead of the more precise pseudo-regret in connection to R_T.

The goal of the algorithm is to keep the regret R_T as low as possible. As a bare minimum, we require that the algorithm is Hannan consistent, i.e., $R_T/T \to 0$ with probability one.

Several variants and special cases of the problem exist, differing on what the set of available actions is in each round. For example, the standard stochastic multi-armed bandit (MAB) problem is a special case of the linear stochastic bandit problem where the set of available actions in each round is the standard orthonormal basis of $\mathbb{R}^d$. The MAB problem is introduced by Thompson (1933) and Robbins (1952) and is extensively studied in the literature. Gittins (1979) studied the *discounted* problem in a Bayesian framework. The Bayes rule and the dynamic programming can, in principle, give us the optimal action in a Bayesian framework; however, the procedure can be computationally intractable. Gittins' contribution was to show that the Bayesian computations can be efficiently performed in this problem. Although this is a significant result, Gittins' setting is different than ours, which is not Bayesian and not discounted.

Lai and Robbins (1985) study the problem in a non-Bayesian parametric setting and design an asymptotically optimal algorithm. The algorithm provably achieves a regret

[1] This chapter is based on the work by Abbasi-Yadkori, Pal, and Szepesvari (2011a) and Abbasi-Yadkori, Pal, and Szepesvari (2011b).

that is within a constant factor of the lower bound for this problem. The algorithm of Lai and Robbins (1985) was simplified and further developed in (Katehakis and Robbins, 1995, Burnetas and Katehakis, 1996). These results have two limitations: their asymptotic nature, and the assumption that the loss is a member of a known parametric family.[2] These two limitations were later removed by Auer et al. (2002a) who studied the non-parametric problem and proposed the UPPER CONFIDENCE BOUND (UCB) algorithm with finite-time regret bounds.

There is a parallel line of research on bandit problems that place no stochasticity assumptions on the environment. In such problems, the environment can be adversarial and is free to choose any loss function as long as it satisfies certain assumptions such as linearity, convexity, boundedness, etc. Auer et al. (2003) proposed the EXP3 algorithm for the adversarial version of the MAB problem and proved a $O(\sqrt{KT \log T})$ regret bound, where K is the number of actions. This bound has an extra $\log(T)$ factor compared to the lower bound proved by Lai and Robbins (1985). The gap was later filled in by Bubeck and Audibert (2010) who proposed an algorithm with a matching $O(\sqrt{KT})$ regret bound.

The linear bandit problem was first introduced by Auer (2002) under the name "linear reinforcement learning". In this version of the problem, the set of available actions changes from timestep to timestep, but has the same finite cardinality in each step. Auer (2002) proposed two algorithms for this problem: the simpler LINREL algorithm that was proposed without analysis, and the more complicated SUPLINREL algorithm with a $\widetilde{O}(\log^{3/2} K\sqrt{dT})$ regret bound, where d is the dimensionality of the unknown parameter vector and K is the number of actions. Later, Li et al. (2010), Chu et al. (2011) studied the problem in the context of web advertisement.

Another variant of the linear bandit problem, studied by Dani et al. (2008), Abbasi-Yadkori (2009), Rusmevichientong and Tsitsiklis (2010), Abbasi-Yadkori et al. (2011a), is the case when the set of available actions does not change between timesteps but the set can be an almost arbitrary, even infinite, bounded subset of a finite-dimensional vector space. Dani et al. (2008) proved a $\widetilde{O}(d\sqrt{T})$ regret bound for the LINREL algorithm. They also showed that the upper bound is tight by proving a $\widetilde{O}(d\sqrt{T})$ lower bound for the linear bandit problem. This seems to contradict the $\widetilde{O}(\sqrt{dT})$ upper bound of Auer (2002), but notice that the settings are slightly different because the action set of (Auer, 2002) is finite (though changing) and the "data" is subject to different constaints (ℓ^∞ vs. ℓ^2). Abbasi-Yadkori (2009) and Rusmevichientong and Tsitsiklis (2010) independently analyzed forced-exploration schemes in the linear bandit problem and derived problem-dependent $\widetilde{O}(d\sqrt{T})$ regret bounds. In this chapter, we show that the result of Dani et al. (2008) can be improved in terms of the regret bound, while both the result of Auer (2002) and Dani et al. (2008) can be improved in terms of the algorithms' computational complexity. We have also introduced the regularization to the linear bandit problem to have algorithms that are adaptive to the sparsity or other regularities of the problem.

Another variant of the linear bandit problem, dubbed "sleeping bandits" and studied by Kleinberg et al. (2008a), is the case when the set of available actions changes from timestep to timestep, but it is always a subset of the standard orthonormal basis of $\mathbb{R}^d$. Related problems were also studied by Abe et al. (2003), Walsh et al. (2009), Dekel et al. (2010).

More generally, the set of available actions might lie in a separable Hilbert space. Srinivas et al. (2010) study this problem under the additional assumption that the noise is Gaussian.[3] We extend their results to the sub-Gaussian noise and also improve the regret bounds in terms of logarithmic terms and constants.

Abernethy et al. (2008) studied the linear bandit problem in a finite-dimensional non-stochastic framework and obtained a $\widetilde{O}(d^{3/2}\sqrt{T})$ regret bound, which is not scaling optimally

[2] We note that the results of (Burnetas and Katehakis, 1996) applies to non-parametric discrete distributions with finite support.

[3] Srinivas et al. (2010) claim that the noise can be any arbitrary bounded random variable, but in their proof, they in fact use the specific probability distribution function of the Gaussian random variables.

for $t := 1, 2, \ldots$ **do**
  $(a_t, \widetilde{\theta}_t) = \operatorname{argmin}_{(a,\theta)\in D_t \times C_{t-1}} \langle a, \theta \rangle$
  Play a_t and observe loss y_t
  Update C_t
end for

Figure 4.1: OFUL ALGORITHM

in dimensionality for this class of problems. More recently, Bubeck et al. (2012) proposed an algorithm with a $O(\sqrt{dT \log N})$ regret bound for any finite action set with N actions. From this algorithm, they obtain an algorithm with a $\widetilde{O}(d\sqrt{T})$ regret bound for the more general setting of compact action sets and a $\widetilde{O}(\sqrt{dT})$ regret bound for the case when the action set is the Euclidean ball. Unlike the algorithm of Abernethy et al. (2008), this algorithm is computationally intractable. Bubeck et al. (2012), however, show that the algorithm can be efficiently implemented for two special cases: the hypercube and the Euclidean ball.

4.1 Optimism in the Face of Uncertainty

A natural and successful way to design an algorithm in many stochastic online learning problems is the optimism in the face of uncertainty principle (OFU). The basic idea, as explained in Chapter 1, is that the algorithm maintains a confidence set $C_{t-1} \subseteq \mathcal{H}$ for the unknown parameters. It is required that C_{t-1} can be calculated from past history and "with high probability" the unknown parameters lies in C_{t-1}.

In summary, the algorithm chooses an optimistic estimate

$$\widetilde{\theta}_t = \operatorname*{argmin}_{\theta \in C_{t-1}} (\min_{a \in D_t} \langle a, \theta \rangle)$$

and then chooses action $a_t = \operatorname{argmax}_{a \in D_t} \langle a, \widetilde{\theta}_t \rangle$ that minimizes the loss according to the estimate $\widetilde{\theta}_t$. Equivalently, and more compactly, the algorithm chooses the pair

$$(a_t, \widetilde{\theta}_t) = \operatorname*{argmin}_{(a,\theta)\in D_t \times C_{t-1}} \langle a, \theta \rangle \,,$$

which *jointly* minimizes the loss. We call the resulting algorithm the OFUL ALGORITHM for "optimism in the face of uncertainty linear bandit algorithm". Pseudo-code of the algorithm is given in Figure 4.1.

4.2 Regret Analysis of the OFUL algorithm

We now give a bound on the regret of the OFUL algorithm when run with confidence sets C_t constructed in Corollary 3.15 in the previous chapter. We will need to assume that expected losses are bounded. We can view this as a bound on θ_* and the bound on the decision sets D_t. The next theorem states a bound on the regret of the algorithm. The proofs, which are largely based on the work of Dani et al. (2008) and are included for completeness, can be found in Appendix E.1.

Theorem 4.1 (Regret of OFUL) Let $(a_1, y_1), (a_2, y_2), \ldots$ satisfy the Linear Response Assumption A1 of the previous chapter. Let $A_{1:t} : \mathcal{H} \to \mathbb{R}^{t-1}$ be an operator such that for any $v \in \mathcal{H}$, the kth element of $A_{1:t}v$ is $\langle a_k, v \rangle$. Assume that for all t and all $a \in D_t$, $\langle a, \theta_* \rangle \in [-1, 1]$. Then, with probability at least $1 - \delta$, the regret of the OFUL algorithm satisfies

$$\forall T \geq 1, \quad R_T \leq 4\sqrt{\beta_T(\delta) T \log \det(I + A_{1:T+1} A^*_{1:T+1}/\lambda)} \,,$$

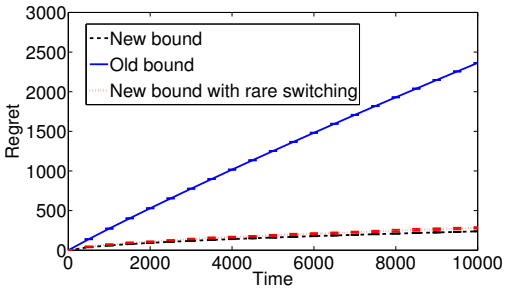


Figure 4.2: The application of the new confidence sets (constructed in Corollary 3.15) to a linear bandit problem. A 2-dimensional linear bandit, where the parameter vector and the actions are from the unit ball. The regret of OFUL is significantly better compared to the regret of ConfidenceBall of Dani et al. (2008). The noise is a zero mean Gaussian with standard deviation $\sigma = 0.1$. The probability that confidence sets fail is $\delta = 0.0001$. The experiments are repeated 10 times and the average and the standard deviation over these 10 runs are shown.

where

$$\beta_T(\delta) = \left(R\sqrt{2\log\left(\frac{\det(I + A_{1:T}A_{1:T}^*/\lambda)^{1/2}}{\delta}\right)} + \lambda^{1/2}\, S \right)^2 .$$

Remark 4.2 In a d-dimensional space, the above bound with the choice of $V = \lambda I$ reduces to

$$R_T \le 4\sqrt{Td\log(\lambda + (T-1)L/d)}\left(\lambda^{1/2}S + R\sqrt{2\log(1/\delta) + d\log(1 + (T-1)L/(\lambda d))}\right) .$$

Remark 4.3 When the action set is finite and has only k members, with an appropriate choice of V, we get that

$$\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) \le k\log\left(1 + \frac{t}{k}\right),$$

independently of the dimensionality of the space that the actions are embedded into (see Corollary 3.7). Thus, the bound in the above theorem is in the order of $k\sqrt{T}$. Although the optimal rate for the MAB problem scales as $\sqrt{kT}$, it is still interesting to obtain a dimensionality-independent bound for a linear bandit algorithm. We note that such a dimensionality-independent bound can not be obtained for the algorithm of Dani et al. (2008), because the dimensionality appears in the size of their confidence ellipsoid. This shows another advantage of constructing tight data-driven confidence sets.

Figure 4.2 shows the experiments with the new confidence set (constructed in Corollary 3.15). The regret of OFUL is significantly better compared to the regret of ConfidenceBall of Dani et al. (2008). The figure also shows a version of the algorithm that has a similar regret to the OFUL algorithm, but spends about 350 times less computation in this experiment. Next, we explain how we can achieve this computation saving.

4.2.1 Saving Computation

In this section, we show that we essentially need to recompute $\widetilde{\theta}_t$ only $O(\log T)$ times up to time T and hence saving computations.[4] The idea is to recompute $\widetilde{\theta}_t$ whenever

[4]Note this is very different than the common "doubling trick" in online learning literature. The doubling is used to cope with a different problem. Namely, the problem when the time horizon T is unknown ahead of time.

Input: Constant $C > 0$
$\tau = 1$ {This is the last time step that we changed $\widetilde{\theta}_t$}
for $t := 1, 2, \ldots$ **do**
  if $\det(I + A_{\tau:t} V_\tau^{-1} A_{\tau:t}^*) > 1 + C$ **then**
    $(a_t, \widetilde{\theta}_t) = \operatorname{argmin}_{(a,\theta) \in D_t \times C_{t-1}} \langle a, \theta \rangle$.
    $\tau = t$.
  end if
  $a_t = \operatorname{argmin}_{a \in D_t} \langle a, \widetilde{\theta}_\tau \rangle$.
  Play a_t and observe loss y_t.
end for

Figure 4.3: The RARELY SWITCHING OFUL ALGORITHM

$\det(V_t)$ increases by a constant factor $(1 + C)$. We call the resulting algorithm the RARELY SWITCHING OFUL algorithm and its pseudo-code is given in Figure 4.3. As the next theorem shows its regret bound is essentially the same as the regret for OFUL.

Theorem 4.4 Under the same assumptions as in Theorem 4.1, with probability at least $1 - \delta$, for all $T \geq 1$, the regret of the RARELY SWITCHING OFUL ALGORITHM satisfies

$$R_T \leq 4\sqrt{(1 + C)\beta_T(\delta) T \log \det(I + A_{1:T+1} V^{-1} A_{1:T+1}^*)} \, .$$

Remark 4.5 In the finite-dimensional case, we get that

$$R_T \leq 4\sqrt{(1 + C) T d \log(\lambda + (T - 1)L/d)} \left(\lambda^{1/2} S + R\sqrt{2 \log 1/\delta + d \log(1 + (T - 1)L/(\lambda d))} \right) \, .$$

The proof of the theorem is given in Appendix E.2. The proof is based on the following lemma whose proof can also be found in Appendix E.2.

Lemma 4.6 Let A, B and C be positive semi-definite operators such that $A = B + C$. Let $C = D^* D$ be a decomposition of C. Assume that $D B^{-1} D^*$ is a trace-class operator. Then, we have that

$$\sup_{a \neq 0} \frac{\langle a, Aa \rangle}{\langle a, Ba \rangle} \leq \det(I + D B^{-1} D^*) \, .$$

Remark 4.7 In the finite-dimensional case, the lemma can also be stated as follows: Let A, B and C be positive semi-definite matrices such that $A = B + C$. Then, we have that

$$\sup_{a \neq 0} \frac{a^\top A a}{a^\top B a} \leq \frac{\det(A)}{\det(B)} \, .$$

Figure 4.4 shows a simple experiment with the RARELY SWITCHING OFUL ALGORITHM.

4.2.2 Problem Dependent Bound

For simplicity, we restrict ourselves to the finite-dimensional fixed decision sets in this section. First we explain the notion of "gap" as defined in (Dani et al., 2008). An extremal point of the decision set D is a point that is not a proper convex combination of points in D. The set of all extremal points in D is denoted by Γ. It can be shown that a linear function on D is minimized in a point in Γ. Define the set of sub-optimal extremal points

$$\Gamma_- = \{a \in \Gamma \, : \, \langle a, \theta_* \rangle > \langle a_*, \theta_* \rangle\} \, .$$

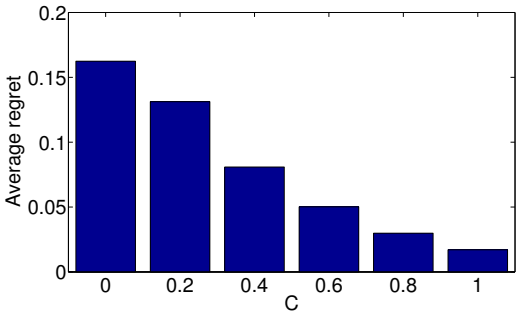


Figure 4.4: Regret against computation. We fixed the number of times the algorithm is allowed to update its action in OFUL. For larger values of C, the algorithm changes action less frequently, hence, will play for a longer time period. The figure shows the average regret obtained during the given time periods for the different values of C. Thus, we see that by increasing C, one can actually lower the average regret per time step for a given fixed computation budget.

Then define the gap, Δ, as

$$\Delta = \inf_{a \in \Gamma_-} \langle a, \theta_* \rangle - \langle a_*, \theta_* \rangle \ .$$

Intuitively, Δ is the difference between the losses of the best and the "second best" action in the decision set D. Note that when D is a ball, $\Delta = 0$. When D is a polytope, $\Delta > 0$.

The regret of OFUL can be upper bounded in terms of Δ as follows.

Theorem 4.8 Assume that the action set is contained in a Euclidean ball of radius L in $\mathbb{R}^d$. Assume that $\lambda \geq 1$ and $\|\theta_*\|_2 \leq S$ where $S \geq 1$. With probability at least $1 - \delta$, for all $T \geq 1$, the regret of the OFUL algorithm satisfies

$$R_T \leq \frac{16 R^2 \lambda S^2}{\Delta} \Big(\log(L(T-1)) + (d-1) \log \frac{64 R^2 \lambda S^2 L}{\Delta^2} \\ + 2(d-1) \log \left(d \log \frac{d\lambda + (T-1)L^2}{d} + 2 \log(1/\delta) \right) + 2 \log(1/\delta) \Big)^2 \ .$$

The proof of the theorem can be found in the Appendix E.3.

The problem dependent regret of (Dani et al., 2008) scales like $O(\frac{d^2}{\Delta} \log^3 T)$, while our bound scales like $O(\frac{1}{\Delta}(\log^2 T + d \log T + d^2 \log \log T))$.

4.2.3 Multi-Armed Bandits

In this section we show that a modified version of UCB has with high probability constant regret.

Let μ_i be the expected loss of action $i = 1, 2, \ldots, d$. Let $\mu_* = \min_{1 \leq i \leq d} \mu_i$ be the expected loss of the best action, and let $\Delta_i = \mu_i - \mu_*$, $i = 1, 2, \ldots, d$, be the "gaps" with respect to the best action. We assume that if we choose action I_t in round t we obtain loss $\mu_{I_t} + \eta_{t+1}$. Let $N_{i,t}$ denote the number of times we have played action i up to time t, and $\overline{X}_{i,t}$ denote the average of the losses received by action i up to time t. We construct confidence intervals for the expected losses μ_i in the following lemma. (The proof can be obtained from Equation 3.10 and an additional union bound over actions.)

Lemma 4.9 (Confidence Intervals) Assuming that the noise η_t is conditionally 1-sub-Gaussian, with probability at least $1 - \delta$,

$$\forall i \in \{1, 2, \ldots, d\}, \ \forall t \geq 1, \qquad |\overline{X}_{i,t} - \mu_i| \leq c_{i,t} \ ,$$

where

$$c_{i,t} = \sqrt{2\frac{(1+N_{i,t})}{N_{i,t}^2}\log\left(\frac{d(1+N_{i,t})^{1/2}}{\delta}\right)}\,. \tag{4.1}$$

Using these confidence intervals, we modify the UCB algorithm of Auer et al. (2002a) and change the action selection rule accordingly. Hence, at time t, we choose the action

$$I_t = \operatorname*{argmin}_i \overline{X}_{i,t} - c_{i,t}\,. \tag{4.2}$$

We call this algorithm UCB(δ).

The main difference between UCB(δ) and UCB is that the length of the confidence interval $c_{i,t}$ depends neither on T, nor on t. This allows us to prove the following result that the regret of UCB(δ) is constant. The proof can be found in Appendix E.4.

Theorem 4.10 (Regret of UCB(δ)) Assume that the noise η_t is conditionally 1-sub-Gaussian. Then with probability at least $1-\delta$, the total regret of UCB(δ) is bounded as

$$R_T \le \sum_{i:\Delta_i>0}\left(3\Delta_i + \frac{16}{\Delta_i}\log\frac{2d}{\Delta_i\delta}\right)\,.$$

Figure 4.5 compares two versions of the UCB(δ) algorithm: one that uses a Hoeffding-based confidence interval, and the other with confidence interval (4.1). As we can see, the regret of UCB(δ) is improved with the new bound.

Remark 4.11 Lai and Robbins (1985) prove that for any suboptimal arm j,

$$\mathbb{E}\left[N_{i,t}\right] \ge \frac{\log t}{D(p_j, p_*)}\,,$$

where D is the KL-divergence, and p_* and p_j are the probability density functions of the optimal arm and arm j, respectively. This lower bound does not contradict Theorem 4.10, as Theorem 4.10 only states a high probability upper bound on the regret. Note that UCB(δ) takes δ as its input. Because with probability δ, the regret in time t can be $O(t)$, on expectation, the algorithm might have a regret of $O(t\delta)$. If we select $\delta = 1/t$, then we get $O(\log t)$ upper bound on the expected regret.

Remark 4.12 If we are interested in an average regret result, then, with a slight modification of the proof technique, we can obtain a result similar to what Auer et al. (2002a) prove.

Remark 4.13 Coquelin and Munos (2007) and Audibert et al. (2009) prove similar high-probability constant regret bounds for variations of the UCB algorithm. Compared to their bounds, our bound is tighter thanks to that with the new self-normalized tail inequality we can avoid one union bound. The improvement can also be seen in experiment as the curve that we get for the performance of the algorithm of Coquelin and Munos (2007) is almost exactly the same as the curve that is labelled Old Bound in Figure 4.5.

4.3 Alternative Methods for Stochastic Linear Bandit Problems

In this section, we briefly explain a number of other approaches to the linear stochastic bandit problem. In the next section, we will compare the performance of these algorithms and ours on a real-world dataset. We start by a slight generalization of the linear stochastic bandit problem.

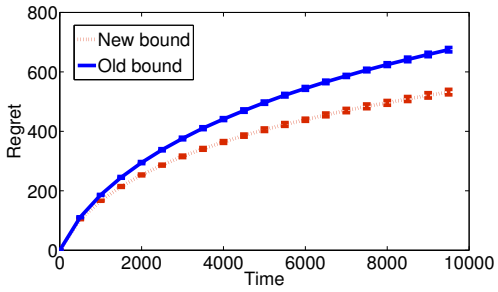


Figure 4.5: The regret against time for two versions of the UCB(δ) algorithm: one that uses a Hoeffding-based confidence interval (referred to as OLD BOUND), and the other with confidence interval (4.1) (referred to as NEW BOUND). The results are shown for a 10-armed bandit problem, where the mean value of each arm is fixed to some value in $[0, 1]$. The regret of UCB(δ) is improved with the new bound. The noise is a zero-mean Gaussian with standard deviation $\sigma = 0.1$. The value of δ is set to 0.0001. The experiments are repeated 10 times and the average together with the standard deviation are shown.

Changing Action Sets

In the rest of this chapter we will consider a slight generalization of the stochastic linear bandit problem. The generalization is important from the point of view applications, but it does not present any added difficulties for the algorithms discussed so far, or the algorithms that we will subsequently discuss. So far, we have assumed that the actions available in each time step belong to the same set $D \subset \mathbb{R}^d$. In many applications, however, the set of actions may change in each time step (some actions may expire, some other actions may become available). Thus, in what follows, we will allow this set to change. The set of admissible actions for time step t will be denoted by $D_t \subset \mathbb{R}^d$. Naturally, the set D_t will be announced at the beginning of the round, before the algorithm has to choose an action.

Thompson Sampling

Thompson (1933) introduced the multi-armed bandit (MAB) problem and proposed a simple Bayesian mechanism for action selection in such problems. The algorithm takes a prior distribution and a probability model on the losses of the arms/actions as input. Given the observations, the posterior distribution is obtained using the Bayes' rule, from which the algorithm samples the next loss estimate that is the basis of choosing the next action.

Recently, Chapelle and Li (2011) showed experimentally that the multi-armed version of the algorithm is competitive in a number of experiments. The first finite-time analysis of the method has appeared in Agrawal and Goyal (2012) and Kaufmann et al. (2012). These results, although promising, do not match the minimax optimal regret rates of the MAB problem.

In this section, we show an implementation of Thompson sampling in the linear bandit setting. Let $\mathcal{N}_n(m, \Sigma)$ denote the n-dimensional normal distribution with mean vector m and covariance matrix Σ. Let χ^2_ν denote the chi-square distribution with ν degrees of freedom. We assume that, given action a_t, the loss is distributed according to

$$y_t \mid a_t, \theta, \sigma^2 \;\sim\; \mathcal{N}_1(\langle a_t, \theta \rangle, \sigma^2)\,. \tag{4.3}$$

Further, we assume that the parameter vector θ and the variance σ^2 together have a normal-(inverse chi-square) distribution:

$$\begin{aligned} \theta \mid \sigma^2 \;&\sim\; \mathcal{N}_d(\mu_t, \sigma^2(\nu_t \Psi_t)^{-1})\,, \\ \sigma^{-2} \;&\sim\; (\tau_t^2 \nu_t)^{-1} \chi^2_{\nu_t}\,. \end{aligned} \tag{4.4}$$

for $t := 1, 2, \ldots$ **do**
  Draw (θ, σ^2) from normal-(inverse chi-square) distribution (4.4)
  For each action $a \in D_t$, draw a loss from (4.3)
  Let a_t be the action with the lowest sample loss
  Play action a_t and observe loss y_t
  Update hyper-parameters by (4.5)
end for

Figure 4.6: THOMPSON SAMPLING for linear bandits.

The variables $\mu_t, \nu_t, \Psi_t, \tau_t$ are called hyper-parameters as they specify the distribution of parameters θ and σ^2. The prior is conjugate: the posterior has the same form given new data. The hyper-parameters are updated as follows:

$$\begin{aligned}
\nu_{t+1} &= \nu_t + 1\,, \\
\nu_{t+1}\Psi_{t+1} &= \nu_t \Psi_t + a_t^\top a_t\,, \\
\nu_{t+1}\Psi_{t+1}\mu_{t+1} &= \nu_t \Psi_t \mu_t + a_t^\top y_t\,, \\
\nu_{t+1}(\tau_{t+1}^2 + \mu_{t+1}^\top \Psi_{t+1}\mu_{t+1}) &= \nu_t(\tau_t^2 + \mu_t^\top \Psi_t \mu_t) + y_t^2\,. \qquad (4.5)
\end{aligned}$$

The pseudo-code of the algorithm is given in Figure 4.6. We currently have no theoretical guarantees for this algorithm.

Exponentially weighted Stochastic (EwS) Algorithm

Maillard (2011) proposes the *exponentially weighted stochastic* (EWS) algorithm for MAB problems. Let $N_{a,t}$ be the number of times that action a is played up to time t, and $\widehat{\mu}_{a,t}$ be the empirical mean loss of action a at time t. At each round, the algorithm computes the empirical gaps

$$\forall a \in D_t, \quad \widehat{\Delta}_{a,t} = \max_{b \in D_t} \widehat{\mu}_{a,t} - \widehat{\mu}_{b,t}\,,$$

and then defines a distribution from which the next action is drawn:

$$p_t(a) = \frac{\exp(-2N_{a,t-1}\widehat{\Delta}_{a,t-1}^2)}{\sum_{b \in D_t} \exp(-2N_{b,t-1}\widehat{\Delta}_{b,t-1}^2)}\,.$$

The EWS algorithm can be generalized to the linear bandit setting. Using the same notations as before, we define the empirical gaps

$$\forall a \in D_t, \quad \widehat{\Delta}_{a,t} = \max_{b \in D_t} \langle a, \widehat{\theta}_t \rangle - \langle b, \widehat{\theta}_t \rangle\,, \qquad (4.6)$$

and the distribution from which the next action is drawn from the density

$$p_t(a) = \frac{\exp\left(-2\,\|a\|_{\overline{V}_t^{-1}}\,\widehat{\Delta}_{a,t-1}^2\right)}{\sum_{b \in D_t} \exp\left(-2\,\|b\|_{\overline{V}_t^{-1}}\,\widehat{\Delta}_{b,t-1}^2\right)}\,. \qquad (4.7)$$

The Pseudo-code of the algorithm is shown in Figure 4.7. We currently have no theoretical guarantees for this algorithm.

for $t := 1, 2, \ldots$ **do**
  Compute empirical gaps by (4.6) for each action $a \in D_t$
  Compute distribution (4.7)
  Draw an action a_t from the distribution
  Play action a_t and observe loss y_t
  Update $\widehat{\theta}_t$ by the least-squares method
end for

Figure 4.7: EwS for linear bandits

Bandits based on Generalized Linear Models (GLM)

Modelling a bounded loss, such as that in the web advertisement, by an unconstrained linear function can lead to suboptimal performance. One approach to constrain the loss is to use a generalized linear model (GLM); given action a and parameter vector θ_*, the expected loss y is assumed to have the form of

$$\mathbb{E}\left[\, y \,|\, a, \theta_* \,\right] = \rho(\langle a, \theta_* \rangle)\,,$$

where the *inverse link function* $\rho : \mathbb{R} \to [0, 1]$ is a strictly increasing, continuously differentiable function.

An inverse link function ρ with the said properties gives rise to a so-called *canonical exponential model* that determines the distribution of the losses as a function of the action and the parameter. It can be shown that the maximum likelihood estimator (MLE) underlying this exponential model satisfies

$$F(\theta) \doteq \sum_{k=1}^{t-1} (y_k - \rho(\langle a_k, \theta \rangle)) a_k \;=\; 0\,.$$

Filippi et al. (2010) propose to find the MLE estimator by using Newton's method to find the root of this equation and then play optimistically. Newton's update rule gives the recursion

$$\widehat{\theta}_t^{\,(p+1)} = \widehat{\theta}_t^{\,(p)} - \left(\sum_{k=1}^{t-1} \rho'\left(\langle a_k, \widehat{\theta}_t^{\,(p)} \rangle \right) \left(1 - \rho\left(\langle a_k, \widehat{\theta}_t^{\,(p)} \rangle \right) \right) a_k a_k^\top \right)^{-1} F(\widehat{\theta}_t^{\,(p)})\,, \qquad (4.8)$$

where ρ' denotes the derivative of ρ. This update has to be iterated until convergence, giving rise to $\widehat{\theta}_t$. Note that Newton's method is guaranteed to converge due to the properties of ρ. Further, in an efficient implementation the inverse matrix can be computed incrementally using the Sherman-Morrison formula. Finally, given the most recent estimate, an optimistic action is chosen by solving

$$a_t = \operatorname*{argmin}_{a \in D_t} \left\{ \rho(\langle a, \theta_t \rangle) - \beta_t \left\| a \right\|_{V_t^{-1}} \right\}\,, \qquad (4.9)$$

where β_t is an appropriate increasing function. The pseudo-code of the algorithm is given in Figure 4.8.

4.4 Experiments

We tested the linear bandit algorithms presented in the previous section, along with OFUL in the recent "Exploration/Exploitation" challenge[5]. We present the results in this section.

[5]The challenge was part of an ICML 2012 workshop on Exploration/Exploitation dilemma. See https://explochallenge.inria.fr

for $t := 1, 2, \ldots$ **do**
  Find the optimistic action a_t by solving (4.9)
  Play action a_t and observe loss y_t
  Update $\widehat{\theta}_t$ by Newton's method (4.8)
end for

Figure 4.8: GLM for linear bandits

Problem Definition and Dataset

For evaluation, the organizers of the challenge used the Yahoo! front page article recommendation dataset.[6] This dataset has been generated by Yahoo! and saved for the evaluation of bandit methods. How the actual evaluation of bandit algorithms (which assume the availability of an environment to interact with instead of some "static" dataset) is done will be explained in the next section (the idea is to use importance weighting), while in this section we focus on the description of the underlying problem and how the data was collected and what it contains.

The contextual bandit problem underlying this dataset is to pick an article at any time t from a pool of available articles at that time, given some information about the user visiting the webpage. The headline of the article picked is then displayed at a prominent part of Yahoo's webpage. The goal is to recommend articles that the users will find appealing. Whether the user found an article appealing is measured based on whether the user clicked on the article. Accordingly, the reward (which is the negated loss) given to the bandit algorithm is 1, if the user clicks on the chosen article, and is 0 otherwise. This can be modelled by introducing the reward function

$$r_t(a) = \begin{cases} 1, & \text{if the user who visits the website at time } t \text{ clicks on article } a; \\ 0, & \text{otherwise.} \end{cases}$$

The *clickthrough rate* (CTR) achieved by an algorithm is the average reward achieved by the algorithm over a given period of time.

The dataset is collected as follows: at each timestep, a user visits the Yahoo! front page. The behaviour policy observes the pool of articles and it then chooses one article uniformly at random to show to the user. The user's feedback and other information (timestamp, user vector, and available articles) are then recorded.

Each entry in the dataset contains a timestamp, a 136-dimensional binary vector representing the user visiting the front page at the given time step, the identifiers of the available articles (around 30 of them), the identifier of the displayed article, and whether the user clicked on the article. The identifier shows the age of an article; articles with larger identifiers are more recent. The first element in the user vector is constant and always equals to 1. The remaining elements encode other user information such as age, gender, etc, but at the time of the contest, the meaning of these bits was not revealed to the contestants.

The dataset contains 30×10^6 entries, from which $9,200,000$ entries are used in the training phase. The split is based on time (older data is used for the first, training phase). The total number of articles during the training phase is 246, while the total number of articles in the full dataset is 652. At any timestep, the number of available articles is less than 30. The pool of available articles, however, changes over time (see Figure 4.9).

Given that there are around 30 articles at each timestep, the probability of the event that a policy's recommendation matches the choice of the behaviour policy (the displayed article) is 1/30. Naturally, only when this happens does the algorithm get useful feedback (see also the next section for further explanation). Thus, there are around $300,000$ evaluations for

[6] http://webscope.sandbox.yahoo.com/catalog.php?datatype=r%20%28%E2%80%9CR6%E2%80%B3%29

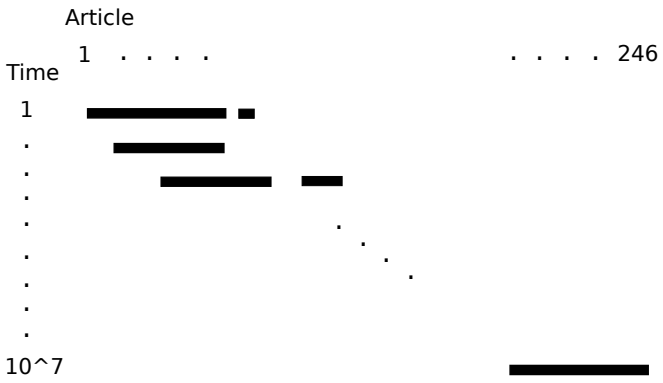


Figure 4.9: The pool of available articles changes over time. The total number of articles during the training phase is 246. Black bars show the subset of the articles that are available at any given time.

any algorithm. Submissions are constrained to run in less than 36 hours. To fulfil this time constraint, an algorithm should choose an article on average within 5 ms and update its policy within 50 ms when receiving feedback.

Off-Policy Evaluation

In a bandit problem, the loss function is revealed only for the action chosen; we do not know what would have happened if another action was chosen at a given time. Thus, when recording the dataset, the dataset will miss this information and if a bandit algorithm to be evaluated will deviate from the action used to collect the data, the algorithm cannot be given feedback. This is in contrast to the supervised learning setting where no matter what decision one makes, the decision can be evaluated. One approach to this problem is to repeat the data collection process for each bandit algorithm. This approach, however, is inpractical as testing on a real-world system, such as the Yahoo! front page article recommendation, can be expensive.

Another approach to this problem is to use a single dataset to evaluate different bandit policies. The problem of evaluating policies (algorithms) other than the one that was used to generate the data is known as the *off-policy evaluation* problem in reinforcement learning (Sutton and Barto, 1998). In this literature, the data gathering policy is called the behaviour policy, while the policy that we want to evaluate is called the target policy.

How can we evaluate a target policy based on some data generated by a behavior policy that may be different from the target policy? Assume that contexts-reward function pairs $(x_t, r_t(\cdot))_t$ are sampled in an i.i.d. fashion. Let a_t be the choice of the behavior policy b at time t, and let a'_t be the choice of the target policy π at time t. The goal is to estimate the expected average reward that is achieved by the target policy. The idea is simply to reject timesteps where a_t and a'_t differ – a form of "rejection sampling", giving rise to

$$\widehat{r}_{\pi,T} = \frac{\sum_{t=1}^{T} \mathbb{I}_{\{a_t = a'_t\}} r_t(a_t)}{\sum_{t=1}^{T} \mathbb{I}_{\{a_t = a'_t\}}}, \tag{4.10}$$

where T is the size of the dataset. Assuming that the behavior policy selects actions uniformly at random and that the target policy is a "stationary policy" (i.e., it chooses its

action based on the current context only), one can show that this estimator is unbiased, i.e, $\mathbb{E}\left[\widehat{r}_{\pi,T}\right] = \mathbb{E}\left[r(\pi(x_1))\right]$ (Li et al., 2011). This suggests using this estimator in evaluating the performance of bandit methods (the only issue is that bandit algorithms "learn", i.e., the policy underlying a bandit algorithm is not stationary).

Submission Rules

Submissions are not allowed to log any part of the data; therefore, no offline processing is possible. The only feedback that participants receive is a file that shows the CTR of the submitted algorithm at every $200,000$ timesteps.

Article Recommendation as a Linear Stochastic Bandit Problem

A simple approach to the article recommendation problem is to treat articles independently and use a MAB algorithm, such as UCB, for action selection. The CTR of UCB is shown Figure 4.10.

The problem can be more naturally modelled as a linear bandit problem by assuming that for each article l and user x_t, the probability of click is linear in some features of the article and the user, $\mathbb{P}\left(r_t(x_t, l) = 1\right) = \langle a(x_t, l), \theta_* \rangle$. The action set in this linear bandit problem has the form of

$$D_t = \left\{ a(x_t, l) \,:\, l \in L_t \right\},$$

where L_t is the set of articles at time t.

Action Representations

As mentioned earlier, we can treat articles as if there was no information shared between them and use a MAB algorithm for action selection. In that case, the action set has the form of

$$D_t = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\}, \quad \theta_* = \begin{pmatrix} \theta_{1,*} \\ \theta_{2,*} \\ \vdots \\ \theta_{k,*} \end{pmatrix} \in \mathbb{R}^k,$$

where $k = 256$ is the number of articles.

This simple baseline can be improved by considering additional information, so that we can generalize over the set of articles. First notice that more recent articles are more likely to receive clicks. To see this, notice the CTR difference between AlwaysLast and AlwaysFirst algorithms in Figure 4.10, which always pick the most recent or oldest article, respectively. Another source of information that can be exploited is the user vector. It is reasonable to assume that similar users have similar click behaviour. We explain three methods to encode this information in the action representation.

The first action representation, which we call UserInfo, represents an action by a vector of length $246 \times 136 = 33,456$. For article $i \in \{1, \ldots, 246\}$, all elements of the corresponding action vector are zero except 136 of them in positions $136(i-1)$ to $136i$ that contain the user vector. Thus, the action set has the form of

$$D_t = \left\{ \begin{pmatrix} x_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ x_t \\ \vdots \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ x_t \end{pmatrix} \right\}, \quad \theta_* = \begin{pmatrix} \theta_{1,*} \\ \theta_{2,*} \\ \vdots \\ \theta_{k,*} \end{pmatrix} \in \mathbb{R}^{k \times n},$$

where $n = 136$ is the dimensionality of the user vector and each $\theta_{i,*}$ is an n-dimensional vector.

The second action representation, that we call AgeInfo, represents an action by a vector of length 248. Let i_t be the most recent article available at time t. Let $f_{i,t}$ be the age

difference between article i and article i_t. We represent the action associated with article i by a vector a that has non-zero values only at $a_i = a_{247} = 1$ and $a_{248} = f_{i,t}$. Thus, the action set has the form of

$$D_t = \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \\ f_{1,t} \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 1 \\ f_{2,t} \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 1 \\ f_{k,t} \end{pmatrix} \right\}, \quad \theta_* = \begin{pmatrix} \theta_{1,*} \\ \theta_{2,*} \\ \vdots \\ \theta_{k,*} \end{pmatrix} \in \mathbb{R}^{k+2} .$$

The third action representation, that we call HYBRID, represents an action by a vector of length $246 \times 136 + 2 = 33,458$. The first $33,456$ elements of action i are identical to USERINFO, the $33,457$th element is 1, and the last element is $f_{i,t}$. Thus, the action set has the form of

$$D_t = \left\{ \begin{pmatrix} x_t \\ 0 \\ \vdots \\ 0 \\ 1 \\ f_{1,t} \end{pmatrix}, \begin{pmatrix} 0 \\ x_t \\ \vdots \\ 0 \\ 1 \\ f_{2,t} \end{pmatrix}, \ldots, \begin{pmatrix} 0 \\ 0 \\ \vdots \\ x_t \\ 1 \\ f_{k,t} \end{pmatrix} \right\}, \quad \theta_* = \begin{pmatrix} \theta_{1,*} \\ \theta_{2,*} \\ \vdots \\ \theta_{k,*} \end{pmatrix} \in \mathbb{R}^{k \times n+2} .$$

Next we explain the computational complexity of different algorithms with the above action representations.

Computational Issues

Consider the OFUL algorithm. The optimistic value of action a is

$$\langle a, \widehat{\theta}_t \rangle + \|a\|_{\overline{V}_t^{-1}} \beta_t(\delta) .$$

Because action vectors are sparse in these experiments, both terms can be computed efficiently even in large dimensions. It only remains to update $\widehat{\theta}_t$ and $\overline{V}_t^{-1}$ efficiently. Computing $\widehat{\theta}_t$ by the least-squares method also requires $\overline{V}_t^{-1}$. So we first discuss the complexity of updating $\overline{V}_t^{-1}$.

Computing $\overline{V}_t^{-1}$ can be computationally expensive in high-dimensional problems. Because $\overline{V}_t$ is updated by a rank-one matrice, we can use the Sherman-Morrison formula to improve the computational cost of the update of the inverse from $O(d^3)$ to $O(d^2)$:

$$(\overline{V}_t + a_t a_t^\top)^{-1} = \overline{V}_t^{-1} - \frac{\overline{V}_t^{-1} a_t a_t^\top \overline{V}_t^{-1}}{1 + a_t^\top \overline{V}_t^{-1} a_t} .$$

The formula is fast enough for the AGEINFO representation. The combination of the OFUL algorithm and the AGEINFO action representation is denoted by OFUL-AGEINFO in Figure 4.10. In the USERINFO representation, the $\overline{V}_t$ matrix has a block-diagonal structure, which again allows fast updates. This combination is denoted by OFUL-USERINFO in Figure 4.10.

We lose the block-diagonal structure with the HYBRID representation and the updates can no longer be done within the time limits. The OFUL algorithm estimates θ_* by the least-squares method. Instead of this, we can update the estimate by the gradient descent method, which is computationally efficient even with the high-dimensional HYBRID representation. Let $\alpha > 0$ be a learning rate, $\widehat{\theta}_t$ be the estimate at time t, $a_{\text{HYBRID},t}$ be the action vector in the

HYBRID representation, and $a_{\text{AGEINFO},t}$ be the action vector in the AGEINFO representation. The gradient update rule is given by

$$\widehat{\theta}_t = \widehat{\theta}_{t-1} + 2\alpha(r_t - \langle a_{\text{HYBRID},t}, \widehat{\theta}_{t-1}\rangle)a_{\text{HYBRID},t} \ .$$

Computing the optimistic bonus requires $\overline{V}_t^{-1}$, which, due to the quadratic complexity of the Sherman-Morrison formula, can be updated within the time limits only in the AGEINFO representation. Adding the estimate and the optimistic bonus together, we get the optimistic estimate:

$$\langle a_{\text{HYBRID},t}, \widehat{\theta}_t\rangle + \|a_{\text{AGEINFO},t}\|_{\overline{V}_t^{-1}} \beta_t(\delta) \ .$$

This method is denoted by OFUL-GRADIENT in Figure 4.10.

The GLM algorithm has all the computational complexities of the OFUL algorithm with the additional cost of Newton's update (4.8). The update rule (4.8) is computationally expensive as it iterates over all past timesteps. Instead, we estimate Newton's update using a subset S_t that is sampled from past timesteps:

$$\widehat{\theta}_t^{(p+1)} = \widehat{\theta}_t^{(p)} - \left(\sum_{k\in S_t} \rho'\left(\langle a_k, \widehat{\theta}_t^{(p)}\rangle\right)\left(1-\rho\left(\langle a_k, \widehat{\theta}_t^{(p)}\rangle\right)\right)a_k a_k^\top\right)^{-1} F(\widehat{\theta}_t^{(p)}) \ .$$

Further, in each time step, again for increased speed, only one update of Newton's iteration is executed, so the update used in the experiments actually takes the above form.

Results

Results for the algorithms we have implemented are shown on Figure 4.10. The figure shows the scores that were achieved in Phase 1 of the competition when we have experimented with the various methods. As described earlier, in every time step AlwaysFirst chooses the oldest articles available in the pool, while AlwaysLast chooses the most recent ones. The difference between the performance of these two methods shows that the "age" of an article is indeed important. Random simply chooses one article uniformly at random. That its performance is better than that of AlwaysFirst is another indication that choosing recent articles is a good idea. That Thompson sampling did not perform well is surprising. It is possible that tweaking with its parameters may lead to much better results. That OFUL-Grad, which is a true "hybrid" did not perform well is less surprising. However, it is more surprising that GLM-AgeInfo did not perform well. It may be because the original GLM algorithm was replaced with a computationally cheaper alternative and as a result the performance got sacrificed. That UCB performed well is reassuring, as is that two variants of OFUL (OFUL-AgeInfo and OFUL-UserInfo) performed significantly better than UCB. In the final submission we chose to submit OFUL-AgeInfo, which was our submission that performed the best in Phase 1. We were worried about overfitting, but at the time there was a possibility that on the top X contestants will proceed to the second phase, hence we felt it is better to submit our best performing method.

In the actual competition, there were 38 participants, but only 25 of them obtained a score substantially higher than the UCB algorithm. Our best submission (OFUL-AGEINFO) ranked 7th in the Phase 1, while it ranked 8th in Phase 2. The scores achieved by the top 8 participants in the two phases are shown on Figure 4.11. As can be seen from the figure, all scores were quite close to each other (and much better than the score that was achieved by UCB in Phase 1, which was below 600).[7] It is remarkable, that the score of the top participants of Phase 1 all dropped in Phase 2. The actual scores are shown on Figures 4.12 and 4.13, for Phase 1 and Phase 2, respectively. The winner of Phase 2 did not use any user or article information, but used a variant of UCB tuned for normally

[7]The score is 1000 times the CTR as estimated with the off-policy evaluation method described earlier.

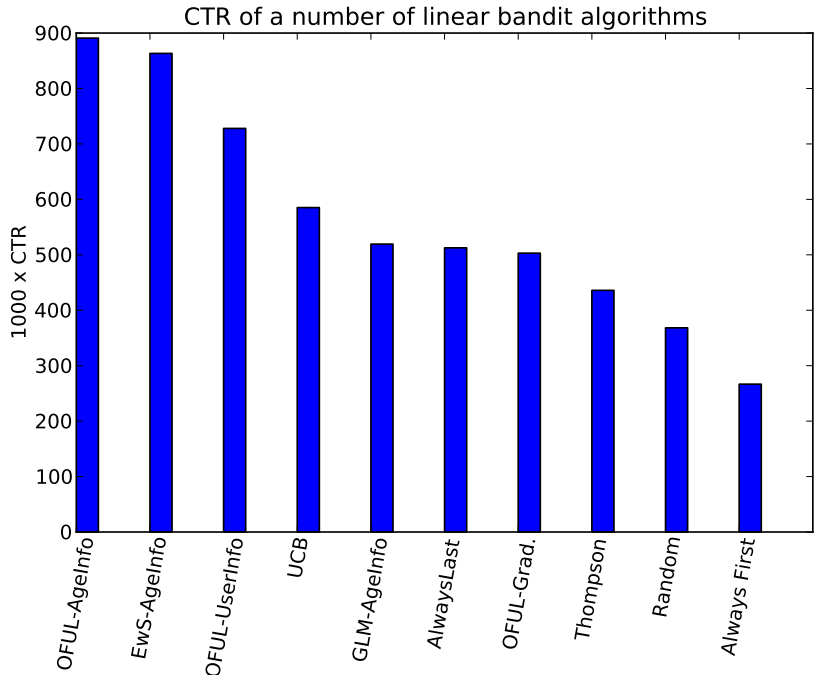


Figure 4.10: Clickthrough rate (CTR) of a number of linear bandit algorithms on Yahoo! front page article recommendation dataset.

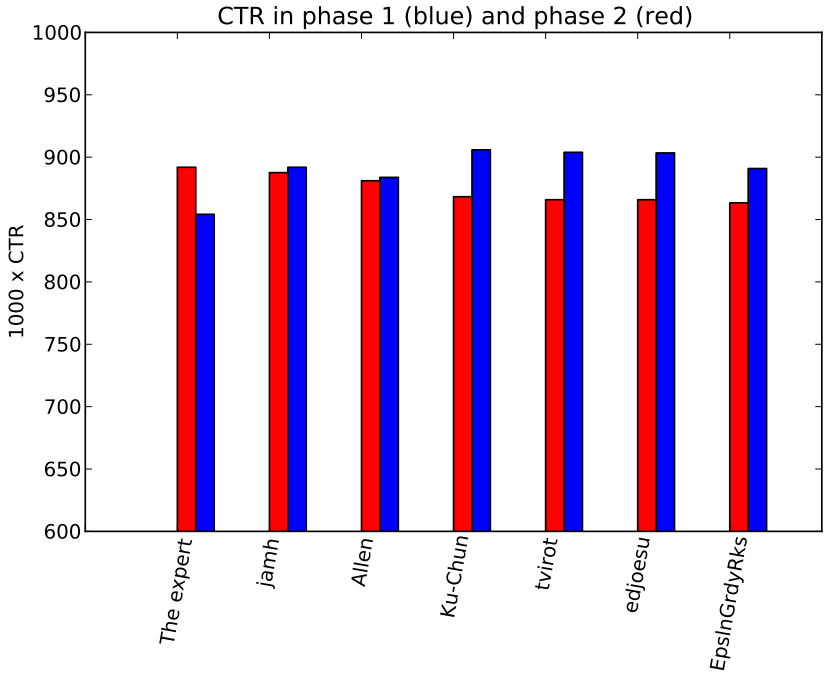


Figure 4.11: Scaled CTR of our algorithm compared to the scaled CTR of top three participants of the training and test phases. Our username is EpsilonGreedyRocks!

distributed random variables (the so-called UCB-Normal algorithm, Auer et al. 2002a). The main difference to UCB is the use of variance information. It is remarkable, though due to the small CTR values somewhat expected, that in this problem the use of variance improves the performance of UCB significantly. The winner of Phase 1 combined heuristic methods with algorithms based on the OFU principle (Chou and Lin, 2012). Based on the data available, it appears that the top participants of Phase 1, despite their limited access to the data, overfitted the data from Phase 1. However, the data also suggest that the final outcome can potentially be attributed to luck: the score of the winner of Phase 2 *increased* by approximately 50 points from Phase 1 to Phase 2. Given that the scores have to divided by 1000, the difference is 4×10^{-2}. A difference of this size, given that the size of the test data is in the range of 10^6, is on the edge of being detectable at the range of scores of 900: The variance of a binomial random variable with $p = 9 \times 10^{-4}$ and $n = 10^6$ trials is about 30. Therefore, scores within $30 - 60$ points of each other are expected to be statistically indistinguishable.

4.5 Sparse Bandits

In this section, we first define the sparse variant of the linear stochastic bandits, and then show how the so-called "optimism in the face of uncertainty" principle can be applied to this problem.

Our goal will be to design algorithms for which the regret is low if θ_* is sparse, that is, if most coordinates of θ_* are zero. This is what we call the *sparse* variant of the linear stochastic bandit problem.

4.5.1 Regret Analysis of OFUL

Consider the OFUL ALGORITHM defined in Figure 4.1 that uses the confidence set C_t constructed in Corollary 3.19 from an online linear prediction algorithm. To keep the analysis general, we leave the underlying linear prediction algorithm unspecified and we only assume that for all $T \geq 1$ it satisfies the regret bound $\rho_T(\theta_*) \leq B_T$.

We introduce a shorthand notation for the RHS of the inequality in Corollary 3.19 specifying the confidence set C_t:

$$\beta_t(\delta) = E^2 + 1 + 2B_t + 32R^2 \ln\left(\frac{R\sqrt{8} + \sqrt{1 + B_t}}{\delta}\right) .$$

The next two theorems upper bound the regret R_T of the resulting OFUL ALGORITHM[8]. The proofs are similar to the proofs of Theorems 4.1 and 4.8.

Theorem 4.14 (Regret of OFUL) Assume that $\|\theta_*\|_2 \leq E$ and assume that for all $t \geq 1$ and for all $a \in D_t$, $\|a\|_2 \leq A$ and $|\langle a, \theta_* \rangle| \leq G$. Then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, for any $T \geq 1$, the regret of the OFUL algorithm is bounded as

$$R_T \leq 2 \max\{1, G\} \sqrt{2T \log \det(1 + A_{1:T+1} A_{1:T+1}^* / \lambda) \max_{0 \leq t < T} \beta_t(\delta)} .$$

Theorem 4.15 (Problem Dependent Regret Bound of OFUL) Assume the same as in Theorem 4.14 and additionally assume that the gap Δ, as defined in Section 4.2.2, is positive. Then, for any $\delta \in (0, 1/4]$, with probability at least $1 - \delta$, for any $T \geq 1$, the regret of the OFUL algorithm is bounded as

$$R_T \leq \frac{8}{\Delta} \max\{1, G^2\} \log \det(1 + A_{1:T+1} A_{1:T+1}^* / \lambda) \max_{0 \leq t < T} \beta_t(\delta) .$$

[8]Note that R_T has nothing to do with $\rho_T(\theta)$ of Section 3.5, except for sharing the same name.

Results of phase 1 :

NAME	AFFILIATION	LAST SCORE (CTR * 10 000)	BEST SCORE (CTR * 10 000)	RANK
Ku-Chun	NTU	882.9	**905.9**	1
tvirot	MIT	903.9	**903.9**	2
edjoesu	MIT	889.9	**903.4**	3
Francis	ULg	865.9	**895.4**	4
jamh	UCM	888	**891.9**	5
exploreit	untitled	0 444.1 (incomplete)	**891.4**	6
EpsilonGreedyRocks	U of A	887.8	**890.9**	7
Exp	LUM	890.9	**890.9**	8
bigwhite	NCU of TW	869.4	**885.8**	9
Allen	NCU	735	**883.8**	10
hl	unknown	881.3	**881.3**	11
tianhuil	Princeton	862.6	**881.1**	12
Yildiz	MUL	872	**881.1**	13
tmann	Texas A&M University	875.3	**876.7**	14
ludc	UP	454.2	**876**	15
bhy	NUS	0 808.2 (incomplete)	**875.2**	16
Occam's	Razor	0 635.3 (incomplete)	**872.8**	17
Collegeboy	MUL	776.5	**872.1**	18
lucati	MUL	860.8	**869.2**	19
krk	cse@buffalo	382.1	**863.6**	20
the_expert	Montanuniversitaet Leoben	854.2	**854.2**	21
at	DIT	385.4	**790.1**	22
lzw	NUS	780.2	**780.2**	23
bandit_guy	unknown	439.9	**774.7**	24
szatymaz	LAL/CNRS	572.2	**672.7**	25

Figure 4.12: Results of the training phase.

Results of phase 2 (final stage):

NAME	AFFILIATION	SCORE (CTR * 10 000)	RANK
The expert	Montanuniversitaet Leoben	**891.9**	1
jamh	Universitad de Madrid	**887.6**	2
Allen	NCU	**881.1**	3
Yildiz	Montanuniversitaet Leoben	**874.3**	4
lucati	Montanuniversitaet Leoben	**873.4**	5
Ku-Chun	University of Taiwan	**868.3**	6
tvirot & edjoesu	MIT	**865.9**	7
tianhuil	Princeton	**863.4**	8
EpsilonGreedyRocks	U of A	**863.4**	9

Figure 4.13: Results of the test phase. Our username is EpsilonGreedyRocks.

To simplify, here and in the rest of the chapter we view E, A, G, R as constants. Then, when the action set is a subset of $\mathbb{R}^d$, the problem dependent and independent regrets of OFUL are $\widetilde{O}(dB_T \ln T/\Delta)$ and $\widetilde{O}(\sqrt{TdB_T})$, respectively. Consequently, *smaller regret bound for the online prediction algorithm translates (via Theorems 4.14 and 4.15) into a smaller regret bound for* OFUL.

As the theorems show, the regret of OFUL depends on the regret of the online learning algorithm that we use as a sub-routine to construct the confidence set. In particular, in order to achieve $O(\text{polylog}(T)\sqrt{T})$ uniform regret for OFUL, one needs an online learning algorithm with $O(\text{polylog}(T))$ regret bound.

Unfortunately, for some of the popular algorithms, such as the exponentiated gradient, the p-norm algorithms, and also for online LASSO, the best known regret bounds are of the order $O(\sqrt{T})$; see (Kivinen and Warmuth, 1997) and (Cesa-Bianchi and Lugosi, 2006, Chapter 11). The main reason for the mediocre $O(\sqrt{T})$ regret bounds seems to be that these algorithm use only gradient information about the quadratic prediction loss function $\ell_t(\langle \cdot, a_t \rangle)$.

Better bounds are available for, e.g., the online regularized least-squares algorithm (i.e., ridge regression) that also uses Hessian information:

Theorem 4.16 (Regret of Ridge Regression; Cesa-Bianchi and Lugosi 2006, Theorem 11.7) Let $\{\theta_t\}_{t=1}^{T+1}$ be the sequence generated by the Follow the Regularized Leader algorithm on the quadratic loss with the quadratic regularizer $R(\theta) = \|\theta\|_2^2/2$. The FTRL algorithm with learning rate $\eta > 0$ satisfies the following bound that holds for all $T \geq 1$ and all $(a_1, y_1), \ldots, (a_T, y_T)$

$$\sum_{t=1}^{T} \ell_t(\widehat{y}_t) \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \ell_t(\langle a_t, \theta \rangle) + \frac{\|\theta\|_2^2}{2\eta} \right\} + \frac{L_T d}{2} \log\left(1 + \frac{\eta A^2 T}{d}\right) ,$$

where $L_T = \max_{1 \leq t \leq T} \ell_t(\langle \theta_t, a_t \rangle)$.

By combining Theorems 4.14 and 4.16, we get that the regret of OFUL with ridge regression is $\widetilde{O}(d\sqrt{T})$. Note that this latter bound essentially matches the bound obtained by Dani et al. (2008) and is similar to the bound in Theorem 4.1.

In online linear prediction, one approach to exploit sparsity (when present) is to use an online ℓ^1-regularized least-squares method. To be able to demonstrate that sparsity can indeed be exploited in stochastic linear bandits, one then needs results similar to Theorem 4.16 for this algorithm, under sparsity assumption. This was an open problem until recently, when Gerchinovitz (2011) proposed the SeqSEW algorithm ("Sequential Sparse Exponential Weights" algorithm), which is based on the sparse exponential weighting algorithm introduced by Dalalyan and Tsybakov (2007), and proved the following logarithmic regret bound for it.

Theorem 4.17 (Regret of SeqSEW$^*_*$, Theorem 8 of Gerchinovitz 2011) The SeqSEW$^*_*$ algorithm introduced by Gerchinovitz (2011) satisfies the following bound that holds for all $T \geq 1$ and all sequences $(a_1, y_1), \ldots, (a_T, y_T)$, $(a_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$,

$$\sum_{t=1}^{T} \ell_t(\widehat{y}_t) \leq \inf_{\theta \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \ell_t(\langle a_t, \theta \rangle) + H_T(\theta) \right\} + (1 + 38 \max_{1 \leq t \leq n} y_t^2) G_T \,,$$

where $\ell_t(y) = (y_t - y)^2$,

$$H_T(\theta) = 256 \left(\max_{1 \leq t \leq T} y_t^2 \right) \|\theta\|_0 \log \left(e + \sqrt{\sum_{t=1}^{T} \|a_t\|^2} \right) + 64 \left(\max_{1 \leq t \leq T} y_t^2 \right) G_T \|\theta\|_0 \log \left(1 + \frac{\|\theta\|_1}{\|\theta\|_0} \right) \tag{4.11}$$

and

$$G_T = 2 + \log_2 \log \left(e + \sqrt{\sum_{t=1}^{T} \|a_t\|^2} \right) \,.$$

The pseudo-code of the SeqSEW$^{B,\eta}_{\tau}$ algorithm, which SeqSEW$^*_*$ extends, is given in Figure 4.14. The SeqSEW$^*_*$ algorithm differs from this algorithm in that it sets the values of τ, B and η on a data-dependent fashion. In particular SeqSEW$^*_*$ runs fresh copies SeqSEW$^{B,\eta}_{\tau}$ for non-overlapping time-periods of increasing lengths with values of τ, B, η set based on empirical quantities measured during previous rounds (details can be found in Section 3.3 of the paper by Gerchinovitz 2011). Theorem 4.17 motivates the OFUL algorithm presented in Figure 4.15 that uses the SeqSEW$^*_*$ algorithm of Gerchinovitz (2011) as an online learning sub-routine. By combining Theorems 4.14 and 4.17, we get that the regret of the OFUL with SeqSEW$^*_*$ algorithm (shown on Figure 4.15) is bounded, with probability at least $1 - \delta$, as

$$R_T \leq 2 \max\{1, G\} \sqrt{2Td \log \left(1 + \frac{TA^2}{d} \right) \max_{1 \leq t < T} \beta_t(\delta)} \,, \tag{4.12}$$

where

$$\beta_t(\delta) = E^2 + 1 + 2B_t(\theta_*) + 32R^2 \log \left(\frac{R\sqrt{8} + \sqrt{1 + B_t(\theta_*)}}{\delta} \right) \,,$$

$B_t(\theta_*) = H_t(\theta_*) + (1 + 38 \max_{1 \leq s \leq t} y_s^2) G_t$ and H_t, G_t are defined as in Theorem 4.17.

From Theorem 4.17 we obtain a confidence set that scales with the sparsity of θ_*. This confidence set is not computable unless if we assume that a prior bound is known on the

Input: threshold $B > 0$, learning rate $\eta > 0$, prior scale $\tau > 0$.
Prior distribution

$$p_0(du) = \prod_{j=1}^{d} \frac{(3/\tau)du_j}{2(1 + |u_j|/\tau)^4} \ .$$

for $t := 1, 2, \ldots$ **do**
  Observe input a_t
  Predict $\widehat{y}_t = \int_{\mathbb{R}^d} [\langle a_t, u\rangle]_B \, p_t(du)$
  Observe y_t
  Compute the posterior distribution

$$p_{t+1}(du) = \frac{1}{Z_{t+1}} \exp\left(-\eta \sum_{s=1}^{t} (y_s - [\langle a_s, u\rangle]_B)^2\right) p_0(du)\,,$$

  where

$$Z_{t+1} = \int_{\mathbb{R}^d} \exp\left(-\eta \sum_{s=1}^{t} (y_s - [\langle a_s, v\rangle]_B)^2\right) p_0(dv)$$

  is the normalizing factor.
end for

Figure 4.14: SeqSEW$^{B,\eta}_{\tau}$ algorithm. In the prediction step, the algorithm makes use of the truncation operator, $[y]_B = \max(\min(y, B), -B)$, where B is an *a priori* bound on the range of prediction values.

for $t := 1, 2, \ldots$ **do**
  Construct confidence set C_{t-1} by Corollary 3.19
  $(a_t, \widetilde{\theta}_t) = \mathrm{argmax}_{(a,\theta)\in D_t \times C_{t-1}} \ \langle a, \theta\rangle$
  Predict $\widehat{y}_t$ by SeqSEW$^*_*$
  Play action a_t and observe loss y_t
  Update C_t
end for

Figure 4.15: OFUL with SeqSEW$^*_*$

sparsity of θ_*. To relax this assumption, we would need data-driven regret bounds for the algorithm of Gerchinovitz (2011), a problem that is still open. Another open problem is to extend Theorem 4.17 to the separable Hilbert spaces, as it applies only to finite-dimensional spaces in its current form.

From the sub-Gaussianity assumption (3.3), we have that with probability $1-\delta$, for any time $t \le T$,

$$|y_t| \le G + R\sqrt{2\log(T/\delta)}\ .$$

Thus the regret (4.12) can be compactly written as $\widetilde{O}(\sqrt{d\|\theta_*\|_0 T})$. Compared to the $\widetilde{O}(d\sqrt{T})$ bound of Dani et al. (2008), the regret bound of OFUL with SeqSEW$^*_*$ is lower when $\|\theta_*\|_0 < d$, which is the case for sparse vectors. Similarly, by application of Theorem 4.17 to the problem dependent regret bound of OFUL in Theorem 4.15, the $\widetilde{O}(d^2 \log^3 T/\Delta)$ problem dependent bound of Dani et al. (2008) can be improved to $\widetilde{O}(d\|\theta_*\|_0 \log^2 T/\Delta)$.

Notice that the regret bound of OFUL with SeqSEW$^*_*$ still depends on d. A slight modification of the usual lower bound for d-armed bandit (Cesa-Bianchi and Lugosi, 2006, Chapter 6) will give us that even if sparsity is $p = 1$ then the regret must be $O(\sqrt{dT})$. More specifically, we choose d distributions for the d arms, so that a random arm has a small reward and all the others have reward zero. This is equivalent to having a sparse θ_* with one non-zero component. Antos and Szepesvári (2009) provide another lower bound of the same order when the action set is the unit ball. This shows that the $\sqrt{d}$ term in the regret is unavoidable, which is in contrast to sparsity regret bounds for the full information online learning problems.

4.5.2 Compressed Sensing and Bandits

Carpentier and Munos (2012) employ compressed sensing techniques to estimate the support of θ_* and achieve sparsity regret bounds of order of $\widetilde{O}(p\sqrt{T})$. Their setting is different than ours in two aspects. First, they consider the case when the action set is the unit ball, which makes it possible to satisfy the isotropic conditions that are required for compressed sensing. In contrast, our results hold for any bounded action set. The second difference, which also explains why they can avoid the $\sqrt{d}$ in their upper bound, is that they assume noise "in the parameters" in the sense that their loss function takes the form of $\ell_t = \langle a_t, \theta_* \rangle + \langle a_t, \eta_t \rangle$.

4.5.3 Experiments with Sparse Bandits

The SeqSEW$^*_*$ algorithm integrates over a multi-dimensional distribution to make a prediction. As integration in general can be computationally expensive, we should resort to sampling techniques such as particle filtering, Gibbs sampling, MCMC, etc. (as suggested by Gerchinovitz (2011), for example, the Langevin Monte-Carlo method of Dalalyan and Tsybakov (2009) could be adapted for this purpose). Instead, for ease of implementation, we use another sparsity online algorithm, called the Exponentiated Gradient (EG) algorithm, in the experiments.

We compare two versions of the OFUL algorithm: the basic algorithm, called OFUL-LS, with confidence sets that are constructed in Corollary 3.15; and another version, called OFUL-EG, whose confidence sets are constructed from predictions of the EG algorithm.

The Exponentiated Gradient (EG) Algorithm

The EG algorithm is a member of the family of the Linearized Proximal-Point algorithms (Rockafellar, 1976). The Linearized Proximal-Point algorithms predict

$$\theta_{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left[\eta \widetilde{\ell}_t(\theta) + D_R(\theta, \theta_t) \right] \tag{4.13}$$

at time t, where Θ is a convex set, $\widetilde{\ell}_t(\theta) = \langle \nabla \ell_t, \theta - \theta_t \rangle$ is the linearized loss, $\eta > 0$ is a learning rate, and D_R is the *Bregman divergence* corresponding to the *Legendre function* R,

$$D_R(u, v) = R(u) - R(v) - \langle \nabla R(v), u - v \rangle \ .$$

The function R is also known as the regularizer as it biases the predictor to specific solutions. We obtain different algorithms by choosing different regularizers. For example, the EG algorithm corresponds to the unnormalized negative entropy regularizer, $R(\theta) = \sum_{i=1}^{d} (\theta_i \log \theta_i - \theta_i)$, whose domain is the positive quadrant of $\mathbb{R}^d$.

In order to solve (4.13), we first solve the unconstrained minimization problem,

$$\overline{\theta}_{t+1} = \operatorname{argmin} \left[\eta \widetilde{\ell}_t(\theta) + D_R(\theta, \theta_t) \right] , \tag{4.14}$$

and then project the solution on Θ,

$$\theta_{t+1} = \operatorname*{argmin}_{\theta \in \Theta} D_R(\theta, \overline{\theta}_{t+1}) \ . \tag{4.15}$$

The unconstrained minimization (4.14) can be solved by finding the root of the gradient vector,

$$\eta \nabla \ell_t + \nabla D_R(\theta, \theta_t) = 0 \ .$$

It can be shown that $\nabla D_R(\theta, \theta_t) = \nabla R(\theta) - \nabla R(\theta_t)$, and thus, if R^* denotes the Legendre dual of R, we get

$$\begin{aligned} \overline{\theta}_{t+1} &= (\nabla R)^{-1}(\nabla R(\theta_t) - \eta \nabla \ell_t) \\ &= \nabla R^*(\nabla R(\theta_t) - \eta \nabla \ell_t), \end{aligned}$$

which is the "mirror-descent" form of the linearized proximal point algorithm (Nemirovski and Yudin, 1998, Beck and Teboulle, 2003). When the constraint set Θ is the d-dimensional simplex,

$$\Delta_d = \left\{ \theta \in \mathbb{R}^d \ : \ \sum_{i=1}^{d} \theta_i = 1 \text{ and } \theta_i \geq 0, \quad 1 \leq i \leq d \right\} \ ,$$

i.e., when $\Theta = \Delta_d$, the update rule for the EG algorithm can be compactly written as

$$\begin{aligned} \overline{\theta}_{t+1,i} &= \theta_{t,i} \exp\left(-\eta \nabla_i \ell_t(\theta_t)\right), \qquad \text{for } 1 \leq i \leq d \,, \\ \theta_{t+1} &= \frac{\overline{\theta}_{t+1}}{\left\| \overline{\theta}_{t+1} \right\|_1} \ . \end{aligned}$$

Let us consider now the regret of this algorithm. Let $A_\infty = \max_{1 \leq t \leq T} \|a_t\|_\infty^2$. Let $L_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta)$ be the total loss of a fixed predictor $\theta \in \mathbb{R}^d$, and $\widehat{L}_T = \sum_{t=1}^{T} \ell_t(\theta_t)$ be the total loss of the learner. Kivinen and Warmuth (1997) show that if the parameter space Θ is the d-dimensional simplex and the loss function is the quadratic loss, then, for all $\theta \in \Delta_d$, the Exponentiated Gradient algorithm with learning rate $\eta = \sqrt{\frac{2 \log d}{A_\infty \widehat{L}_T}}$ satisfies

$$\rho_T(\theta) = \widehat{L}_T - L_T(\theta) \leq 2 A_\infty \log(d) + 2 \sqrt{A_\infty \log(d) L_T(\theta)} \ . \tag{4.16}$$

In the current setting with the linear observation model, $y_t = \langle \theta_*, a_t \rangle + \eta_t$, and quadratic loss functions, $\ell_t(\theta) = (y_t - \langle \theta, a_t \rangle)^2$, under the additional assumption that the noise is bounded by σ, we have that

$$L_T(\theta_*) = \sum_{t=1}^{T} (y_t - \langle \theta_*, a_t \rangle)^2 \leq \sum_{t=1}^{T} \eta_t^2 \leq \sigma^2 T \ .$$

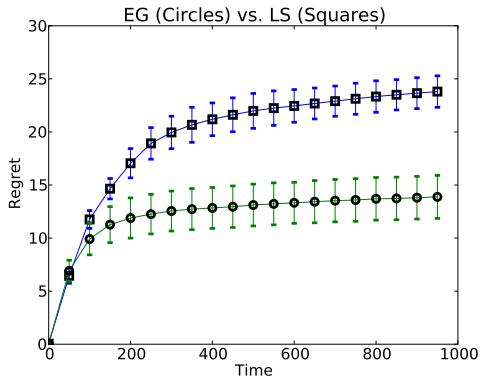


Figure 4.16: The EG algorithm achieves a smaller regret compared to the least-squares method on a prediction problem when the unknown parameter vector is sparse. At each round, we generate a random input vector a_t in $\{-1,+1\}^{200}$. The parameter vector θ_* has only 10 non-zero elements, each being equal to 0.1. The algorithm observes $\langle \theta_*, a_t \rangle$ corrupted by a Gaussian noise drawn from $\mathcal{N}(0, 0.1^2)$. The time horizon is $T = 1000$. We set the least-squares regularizer to $\lambda = 1$, and the EG time-varying learning rate to $\sqrt{2\log(d)/t}$.

Thus, we have

$$\rho_T(\theta_*) = \widehat{L}_T - L_T(\theta_*) \le 2A_\infty \log(d) + 2\sigma\sqrt{A_\infty \log(d) T}\,. \tag{4.17}$$

Compared to this regret bound, which scales with the root of the length of the horizon T, the SeqSEW algorithm enjoys a stronger $O(\log(T))$ regret bound. However, as we explained at the beginning of this section, it is much easier to implement EG. Further, the regret bound presented here might be too conservative. Nevertheless, in contrast to the least-squares method whose regret scales linearly with dimensionality, the regret bound (4.16) displays only a logarithmic dependence on the dimension d.

The assumption that $\theta \in \Theta = \Delta_d$ may seem overly restrictive. However, the algorithm and the analysis are not hard to extend to the case when Θ is the ℓ^1 unit-ball. For this, just note that for $\theta \in \mathbb{R}^d$, $\|\theta\|_1 \le 1$, we can write $\theta = \theta^+ - \theta^-$, where $\|\theta^+\|_1, \|\theta^-\|_1 \le 1$ and $\theta^+, \theta^- \ge 0$ ($\ge$ is meant to be componentwise, i.e., we just decompose θ into its positive and negative parts). Then, keeping two sets of weights, one for the positive part, one for the negative part of the weight, both updated using the EG algorithm (no projection is necessary when the updated parameter vector has a 1-norm below one), while predicting with $\theta_t = \theta_t^+ - \theta_t^-$, we get that essentially the same result holds for the resulting EG$\pm$ algorithm Grove et al. (2001).

Experimental Results

We compare the two bandit methods derived from ridge regression and EG using a synthetic problem. However, first we compare the two underlying prediction methods on some artificial problem, which is constructed to favour EG to verify whether EG indeed has some advantage over ridge regression.

Ridge Regression vs. EG in a Prediction Setting At each round, we generate a random input vector a_t in $\{-1,+1\}^{200}$. The parameter vector θ_* has only 10 non-zero elements, each being equal to 0.1 (thus, $\theta_* \in \Delta_d$, $d = 200$). The algorithm observes $\langle \theta_*, a_t \rangle$ corrupted by a Gaussian noise drawn from $\mathcal{N}(0, 0.1^2)$. The time horizon is $T = 1000$. We set the least-squares regularizer to $\lambda = 1$, and the EG time-varying learning rate to $\sqrt{2\log(d)/t}$.

for $t := 1, 2, \dots$ **do**
  $a_t = \operatorname{argmax}_{a \in D} \langle \widehat{\theta}_t, a \rangle + \|a\|_{\overline{V}_t^{-1}} \beta_t(\delta)$, where $\widehat{\theta}_t$ and $\beta_t(\delta)$ are defined in (4.18) and (4.19), respectively
  Given $\{(a_s, y_s)\}_{s=1}^{t-1}$, EG predicts $\widehat{y}_t$
  Play action a_t and observe loss y_t
end for

Figure 4.17: The OFUL-EG algorithm.

Figure 4.16 shows that the EG algorithm achieves a smaller regret compared to the least-squares method. This implies that the OFUL-EG algorithm, whose confidence sets are constructed from the predictions of the EG algorithm, might achieve a better performance compared to the OFUL-LS algorithm. Next, we compare these two algorithms.

Comparison of OFUL-EG and OFUL-LS Recall from Section 3.5 that, given an online learning algorithm (such as EG) that produces predictions $\{\widehat{y}_1, \widehat{y}_2, \dots\}$ for inputs $\{a_1, a_2, \dots\}$, and admits a regret bound $\rho_t(\theta_*) \le B_t$, we can construct a confidence ellipsoid with centre

$$\widehat{\theta}_{t+1} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left(\|\theta\|_2^2 + \sum_{s=1}^{t} (\widehat{y}_s - \langle \theta, a_s \rangle)^2 \right) \tag{4.18}$$

and width

$$\beta_t(\delta) = 1 + 2B_t + 32R^2 \ln \left(\frac{R\sqrt{8} + \sqrt{1 + B_t}}{\delta} \right) . \tag{4.19}$$

For the EG algorithm, Inequality (4.17) implies that we can use $B_t = 2A_\infty \log(d) + 2\sigma\sqrt{A_\infty \log(d) t}$. The definition of the OFUL-EG algorithm is shown in Figure 4.17. The OFUL-LS algorithm is identical to the one studied in Section 4.1 (see Figure 4.1).

We compare these two linear bandit algorithms on synthetic data. The action set is $k = 200$ randomly generated vectors in $\{-1, +1\}^{200}$. The parameter vector θ_* has only 10 non-zero elements, each being equal to 0.1. The algorithm observes $\langle \theta_*, a_t \rangle$ corrupted by a Gaussian noise drawn from $\mathcal{N}(0, 0.1^2)$. The time horizon is $T = 1000$. We set the least-squares regularizer to $\lambda = 1$, and the EG learning rate to $\eta = 1$. Figure 4.18-(a) shows that the OFUL-LS algorithm outperforms the OFUL-EG algorithm in this experiment. This better performance can be attributed to the fact that the OFUL-LS algorithm uses very tight confidence sets (constructed in Chapter 3). However, there is still room for improving OFUL-EG.

In particular, if we study the proof of Corollary 3.20 more carefully, we realize that the confidence width (4.19) can be tightened to

$$\beta_t(\delta) = 1 + 2B_t + 32R^2 \ln \left(\frac{R\sqrt{8} + \sqrt{1 + B_t}}{\delta} \right) - \sum_{s=1}^{t} \left(\widehat{y}_s - \langle \widehat{\theta}_{t+1}, x_s \rangle \right)^2 . \tag{4.20}$$

A close inspection of the proof of Theorem 3.18 also reveals that the confidence width can be further reduced to

$$\beta_t(\delta) = \left(\sqrt{B_t + 1} + 4R \sqrt{\ln \left(\frac{R\sqrt{8} + \sqrt{1 + B_t}}{\delta} \right)} \right)^2 - 1 - \sum_{s=1}^{t} \left(\widehat{y}_s - \langle \widehat{\theta}_{t+1}, x_s \rangle \right)^2 . \tag{4.21}$$

These two modifications greatly improve the performance of the OFUL-EG algorithm. Figures 4.18-(b,c) show the performance of the OFUL-EG algorithm using the improved confidence widths (4.20) and (4.21).

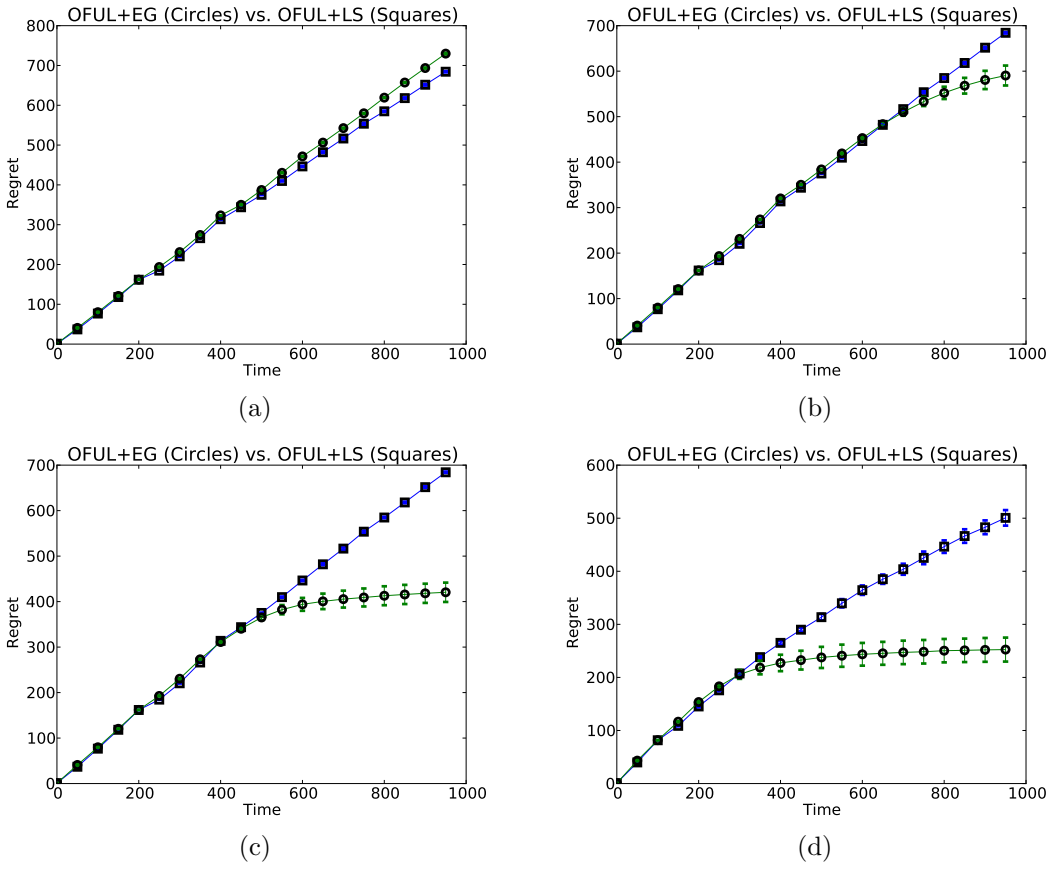


Figure 4.18: Comparing the OFUL-EG and the OFUL-LS algorithms on synthetic data. The action set is $k = 200$ randomly generated vectors in $\{-1, +1\}^{200}$. The parameter vector θ_* has only 10 non-zero elements, each being equal to 0.1. The algorithm observes $\langle \theta_*, a_t \rangle$ corrupted by a Gaussian noise drawn from $\mathcal{N}(0, 0.1^2)$. The time horizon is $T = 1000$. We set the least-squares regularizer to $\lambda = 1$, and the EG learning rate to $\eta = 1$. (a) The OFUL-LS algorithm outperforms the OFUL-EG algorithm (b) The OFUL-EG algorithm with the improved confidence width (4.20) outperforms the OFUL-LS algorithm (c) Improving the regret of the OFUL-EG algorithm with confidence width (4.21) (d) Experimenting with a problem with a smaller dimensionality and action set, $k = 100$, $d = 100$.

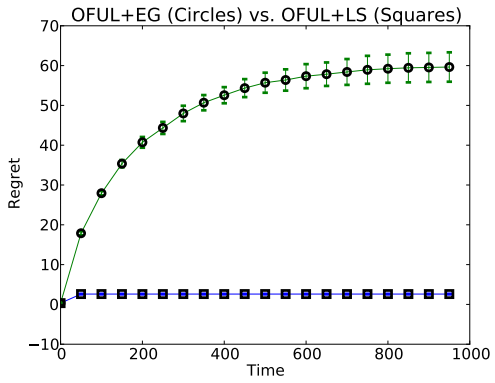


Figure 4.19: Comparing the OFUL-EG and the OFUL-LS algorithms on synthetic data. The action set is $k = 5$ randomly generated vectors in $\{-1, +1\}^{200}$.

Interestingly, the OFUL-LS is the winning algorithm when the number of actions is small. Figure 4.19 shows an experiment with only $k = 5$ actions. This phenomena can be explained by noting that the regret of the OFUL-LS algorithm depends on $\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right)$ (see Theorem 4.1), which, in a finite-action setting, can be bounded as

$$\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) \le k\log\left(1 + \frac{t}{k}\right),$$

independently of the dimensionality of the space that the actions are embedded into (see Corollary 3.7).

Chapter 5

Linearly Parametrized Control Problems[1]

Up until now, the environment was memoryless, in the sense that actions had no effect on future loss functions. This chapter studies a more general problem where the *state* of the environment changes as a function of the current state and the action taken by the learner.

In the first part of the chapter, we study the average loss linear quadratic (LQ) control problem with unknown model parameters, also known as the adaptive control problem in the control community. The problem is to minimize the average loss of a controller that operates in an environment whose dynamics is linear, while the loss is a quadratic function of the state and the control. The optimal solution is a linear feedback controller that can be computed in a closed form from the matrices underlying the dynamics and the loss. In the learning problem, the topic of this chapter, the dynamics of the environment is unknown. This problem is challenging since the control actions influence both the loss and the rate at which the dynamics is learned, a topic of adaptive control. The objective in this case is to minimize the regret of the controller, i.e. to minimize the difference between the average loss incurred by the learning controller and that of the optimal controller. In this dissertation, for the first time, we show an adaptive controller and prove that, under some assumptions, its expected regret is bounded by $\widetilde{O}(\sqrt{T})$. We build on the results of Chapter 3 on online linear estimation and the results in adaptive control design, the latter of which we survey next.

When the model parameters are known and the state is fully observed, we can derive the optimal controller from the principles of dynamic programming. The version of the problem that deals with the unknown model parameters is called the adaptive control problem. The early attempts to solve this problem relied on the certainty equivalence principle (Simon, 1956). The idea, as explained in Chapter 1, was to estimate the unknown parameters from observations and then design a controller as if the estimated parameters are the true parameters. Later, it was realized that the certainty equivalence principle did not necessarily provide enough information to reliably estimate the parameters and the estimated parameters could converge to incorrect values with positive probability (Becker et al., 1985). This in turn might lead to suboptimal performance.

To avoid failure to identify the system dynamics, methods that actively explore to gather information have been developed (Lai and Wei, 1982, 1987, Chen and Guo, 1987, Chen and Zhang, 1990, Fiechter, 1997, Bradtke et al., 1994, Lai and Ying, 2006, Campi and Kumar, 1998, Bittanti and Campi, 2006, Al-Tamimi et al., 2007). However, up to our best knowledge, so far no finite-time regret bounds, but only asymptotic results are known for these methods. One exception is the work of Fiechter (1997). However, the main result of this work is a PAC-type result and in a discounted total expected cost framework.

Most of the aforementioned methods use forced-exploration schemes to provide the suf-

[1] Results of Sections 5.1 and 5.2 have appeared in (Abbasi-Yadkori and Szepesvári, 2011).

ficient exploratory information. The idea, as explained in Chapter 1, is to take exploratory actions according to a fixed and appropriately designed schedule. However, the forced-exploration schemes lack strong worst-case regret bounds, even in the simplest problems (see e.g. Dani and Hayes (2006), Section 6 and the explanation in Chapter 1). Unlike the preceding methods, Campi and Kumar (1998) proposes an algorithm based on the OFU principle, which they call the Bet On the Best (BOB) principle. However, Campi and Kumar (1998) only show asymptotic optimality, i.e., the average loss of their algorithm converges to that of the optimal policy in the limit. In this chapter, we modify the algorithm and the proof technique of Campi and Kumar (1998) and extend their work to derive a finite time regret bound. The modification of the algorithm was necessary to make our proof go through. The results presented here, just like of the previous chapters, build upon the results of Chapter 3 that provide confidence sets for linear estimation with dependent covariates.

The OFU principle has also been applied to learning in *finite* Markov Decision Processes, both in a regret minimization (e.g., Bartlett and Tewari 2009, Jaksch et al. 2010) and a PAC-learning setting (e.g., Kearns and Singh 1998, Brafman and Tennenholtz 2002, Kakade 2003, Strehl et al. 2006, Szita and Szepesvári 2010). In the PAC-MDP framework there has been some work to extend the OFU principle to infinite Markov Decision Problems under various assumptions. For example, Lipschitz assumptions and finiteness of the action set have been used by Kakade et al. (2003), while Strehl and Littman (2008) explored linearly parametrized models with stable feature mappings. (A stable matrix has a ℓ^2 norm less than 1.) However, none of these works obtain regret bounds for MDPs with continuous state and action spaces, a setting we study in this chapter. Continuous action spaces in the context of bandits have been explored in a number of works, such as the works of Kleinberg (2005), Auer et al. (2007), Kleinberg et al. (2008b) and in a linear setting by Auer (2002), Dani et al. (2008), Rusmevichientong and Tsitsiklis (2010) and Abbasi-Yadkori et al. (2011a).

One potential problem with the proposed algorithm is its computational requirements; the algorithm needs to solve a computationally expensive optimization problem at each round. In Section 5.4, we derive a gradient algorithm for this optimization problem (with no guarantees) and investigate the behavior of this algorithm empirically. The experiments attest that the method is indeed successful in achieving sublinear regret, while keeping the cost of computations at a manageable level.

In the second part of the chapter, we study the adaptive control problem with linearly parametrized dynamics. More specifically, the expected value of the next state is assumed to be linear in some *features* of the current state/action pair. We propose a similar OFU-based method and show that, under some assumptions, its expected regret is bounded by $\widetilde{O}(\sqrt{T})$. Whether this (or a similar) method can be implemented efficiently remains to be seen.

5.1 The Linear Quadratic (LQ) Control Problem

We consider the discrete-time infinite-horizon linear quadratic (LQ) control problem:

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* a_t + w_{t+1}\,, \\ \ell(x_t, a_t) &= x_t^\top Q x_t + a_t^\top R a_t\,, \end{aligned} \tag{5.1}$$

where $t = 1, \dots$, $a_t \in \mathbb{R}^d$ is the action at time t, $x_t \in \mathbb{R}^n$ is the state at time t, $\ell(x_t, a_t) \in \mathbb{R}$ is the loss at time t, w_{t+1} is the "noise", $A_* \in \mathbb{R}^{n\times n}$ and $B_* \in \mathbb{R}^{n\times d}$ are unknown matrices while $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{d\times d}$ are known (positive definite) matrices. We will denote $\ell(x_t, a_t)$ by ℓ_t. For simplicity, $x_1 = 0$. The problem is to design a controller based on past observations to minimize the average expected loss

$$J(a_1, a_2, \dots) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\ell_t\right]\,. \tag{5.2}$$

Let J_* be the optimal (lowest) average loss. The regret up to time T of a controller incurring a loss of ℓ_t at time t is defined by

$$R_T = \sum_{t=1}^{T} (\ell_t - J_*) \,,$$

i.e., the regret is the difference between the performance of the controller and the (expected average) performance of the optimal controller that has full information about the system dynamics. As usual, regret can be interpreted as a measure of the loss due to not knowing the system dynamics.

5.1.1 Assumptions

In this section, we state our assumptions on the noise and the system dynamics. The role of assumptions on the system dynamics is to ensure that the optimal control problem has a well-defined solution.
Define

$$\Theta_*^\top = \begin{pmatrix} A_* \,, & B_* \end{pmatrix} , \qquad m = n + d \,, \qquad \text{and} \qquad z_t = \begin{pmatrix} x_t \\ a_t \end{pmatrix} .$$

Thus, the state transition can be written as

$$x_{t+1} = \Theta_*^\top z_t + w_{t+1} \,.$$

Assumption A2 (Linear Model Assumption) Let $(\mathcal{F}_t; t \geq 1)$ be a filtration, (z_1, x_2), …, (z_t, x_{t+1}) be a sequence of random variables over $\mathbb{R}^m \times \mathbb{R}^n$ such that:

(i) z_t, x_t are $\mathcal{F}_t$-measurable;

(ii) For any $t \geq 1$,

$$\mathbb{E}\left[x_{t+1} | \mathcal{F}_t\right] = z_t^\top \Theta_* \,,$$

i.e., $w_{t+1} = x_{t+1} - z_t^\top \Theta_*$ is a martingale difference sequence ($\mathbb{E}\left[w_{t+1} | \mathcal{F}_t\right] = 0$, $t = 1, 2, \ldots$);

(iii) $\mathbb{E}\left[w_{t+1} w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$;

(iv) The random variables w_t are component-wise sub-Gaussian in the sense that there exists a constant $L > 0$ such that for any $\gamma \in \mathbb{R}$, and index $j \in \{1, \ldots, n\}$,

$$\mathbb{E}\left[\exp(\gamma w_{t+1,j}) \,|\, \mathcal{F}_t\right] \leq \exp(\gamma^2 L^2 / 2) \,.$$

The assumption $\mathbb{E}\left[w_{t+1} w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$ makes the analysis more readable. In Section 5.2, we will sketch how this assumption could be removed.

Our next assumption concerns the system dynamics. This assumption will make sure that that optimal control problem for the true system is well-posed. Coincidentally, it also ensures that an optimal controller (knowing the system parameters) can be found efficiently. To introduce this assumption, we need some more background about optimal LQ control. First, we introduce the concepts of reachability and observability.

Reachability means that for any initial state x_0 and final state x_f, there exists a sequence of control vectors that after n timesteps, bring the state of the system $x_{t+1} = A_* x_t + B_* a_t$ to x_f. When $x_{t+1} = A_* x_t + B_* a_t$ is reachable, we also say that the pair (A_*, B_*) is reachable. For linear systems it holds that the definition will not change if the restriction that the state has to be brought to x_f in at most n steps was removed. (cf. Section 3.8.4 of (Hendricks et al., 2008)). *Observability* means that, given at least n consecutive measurements of the

form $z_t = Cx_t$, we can infer the initial state x_0 of the system $x_{t+1} = A_* x_t$ (Hendricks et al., 2008). When $x_{t+1} = A_* x_t$, $z_t = Cx_t$ is observable, we also say that the pair (A_*, C) is observable.

The following result gives a complete algebraic characterization of the reachability and observability (Hendricks et al., 2008, p. 140, Theorem RD2 and p. 149, Theorem OD2):

Proposition 5.1 A pair (A, B), where A is an $n \times n$ matrix and B is an $n \times d$ matrix, is reachable if and only if the $n \times nd$ matrix

$$[B \; AB \; \ldots \; A^{n-1}B]$$

has full rank. A pair (A, C), where A is an $n \times n$ matrix and C is an $d \times n$ matrix, is observable if and only if the pair $(A^\top, C^\top)$ is reachable.

Let $\Theta^\top = (A, B)$. Assume that (A, B) is reachable and $(A, Q^{1/2})$ is observable. (Here, $Q^{1/2}$ is any matrix M satisfying $Q = M^\top M$. Note that it does not matter which of the possible roots of Q one uses, $(A, Q^{1/2})$ is either observable for all of the roots, or it is not observable for any of them as it follows from the definition of observability. Given our assumptions, there is a unique solution $P(\Theta)$ in the class of positive-semidefinite symmetric matrices to the so-called *Riccati equation*

$$P(\Theta) = Q + A^\top P(\Theta)A - A^\top P(\Theta)B(B^\top P(\Theta)B + R)^{-1}B^\top P(\Theta)A \,. \tag{5.3}$$

The *optimal control law* for a LQ system with parameters Θ is

$$a_t = K(\Theta)x_t \,, \tag{5.4}$$

where

$$K(\Theta) = -(B^\top P(\Theta)B + R)^{-1}B^\top P(\Theta)A$$

denotes the so-called optimal *gain matrix* (Bertsekas, 2001, V. 2, p. 273). The average loss of this controller, which is equal to the optimal average loss for the system, satisfies

$$J(\Theta) = \mathrm{trace}(P(\Theta))$$

(Bertsekas, 2001, V. 2, p. 273) (in particular, $J_* = J(\Theta_*) = \mathrm{trace}(P(\Theta_*))$).

Under the feedback law (5.4), the closed-loop behavior is

$$x_{t+1} = (A + BK(\Theta))x_t + w_{t+1} \,.$$

Thus, the stability of the closed-loop system is controlled by the matrix $A + BK(\Theta)$. As it is well-known, under the same assumptions, the matrix $A + BK(\Theta)$ is stable, i.e., its ℓ^2-norm is less than one.

We are ready to state our assumptions on the system:

Assumption A3 (Reachability and Observability Assumption) Fix the constants $S, C > 0$, $\Lambda \in [0, 1)$ and define the set

$$\mathcal{S} = \mathcal{S}_0 \cap \mathcal{S}_1 \cap \mathcal{S}_2 \cap \left\{\Theta \in \mathbb{R}^{n \times m} \,:\, \mathrm{trace}(\Theta^\top \Theta) \le S^2\right\} \,,$$

where

$$\begin{aligned}
\mathcal{S}_0 &= \left\{\Theta = (A, B) \in \mathbb{R}^{n \times m} \,:\, (A, B) \text{ is reachable, } (A, Q^{1/2}) \text{ is observable } \right\} , \\
\mathcal{S}_1 &= \left\{\Theta = (A, B) \in \mathbb{R}^{n \times m} \,:\, \|A + BK(A, B)\| \le \Lambda\right\} , \\
\mathcal{S}_2 &= \left\{\Theta = (A, B) \in \mathbb{R}^{n \times m} \,:\, \|K(\Theta)\| \le C\right\} .
\end{aligned}$$

We assume that $\Theta_* \in \mathcal{S}$ with known S.

By the boundedness of $\mathcal{S}$, we also obtain the boundedness of $P(\Theta)$ (Anderson and Moore, 1971). The corresponding constant will be denoted by D:

$$D \doteq \sup \{\|P(\Theta)\| \,:\, \Theta \in \mathcal{S}\} \ . \tag{5.5}$$

The assumption that $\Theta_* \in \mathcal{S}_0$ can be relaxed somewhat; it turns out that the reachability and observability assumptions can be replaced by weaker *stabilizability* and *detectability* conditions, which are both necessary and sufficient for the optimal LQ problem to have a nontrivial solution (Bertsekas, 2001, V. 1, P. 141).[2] We have decided to use the stronger conditions to simplify the presentation. The assumptions $\Theta_* \in \mathcal{S}_1$, $\Theta_* \in \mathcal{S}_2$ for some values of Λ and C are not restrictive given $\Theta_* \in \mathcal{S}_0$ (since the values of Λ and C need not be known). These assumptions allow us to state our regret bounds for any $\Theta_* \in \mathcal{S}$ (i.e., the bound will depend on Λ and C; as Λ approaches 1 or C approaches infinity, the bound will grow unbounded). The assumption that $\mathrm{trace}(\Theta_*^\top \Theta_*) \le S^2$ for a *known* value of S *is* restrictive. We note that the previous works by Campi and Kumar (1998) and Bittanti and Campi (2006) also make the same assumption. We leave it for future work to remove this assumption.

5.1.2 Parameter estimation

We need high-probability confidence sets to implement the OFU principle. The derivation of the confidence set is based on the results of Chapter 3. Define

$$e_t(\Theta) = \lambda \,\mathrm{trace}(\Theta^\top \Theta) + \sum_{s=1}^{t-1} \mathrm{trace}((x_{s+1} - \Theta^\top z_s)(x_{s+1} - \Theta^\top z_s)^\top) \ .$$

Let $\widehat{\Theta}_t$ be the ℓ^2-regularized least-squares estimate of Θ_* with regularization parameter $\lambda > 0$:

$$\widehat{\Theta}_t = \operatorname*{argmin}_{\Theta} \ e_t(\Theta) = (Z^\top Z + \lambda I)^{-1} Z^\top X \,, \tag{5.6}$$

where Z and X are the matrices whose rows are $z_1^\top, \dots, z_{t-1}^\top$ and $x_2^\top, \dots, x_t^\top$, respectively. The next theorem, whose proof following along the lines of Corollary 3.15 and is hence omitted, constructs high probability confidence sets for the unknown parameter matrix Θ_*.

Theorem 5.2 Let $(z_1, x_2), \dots, (z_t, x_{t+1})$, $z_s \in \mathbb{R}^m$, $x_s \in \mathbb{R}^n$ satisfy the Linear Model Assumption A2 with some $L > 0$, $\Theta_* \in \mathbb{R}^{m \times n}$, $\mathrm{trace}(\Theta_*^\top \Theta_*) \le S^2$ and let $\mathcal{F} = (\mathcal{F}_t)$ be the associated filtration. Consider the ℓ^2-regularized least-squares parameter estimate $\widehat{\Theta}_t$ with regularization coefficient $\lambda > 0$ (cf. (5.6)). Let

$$\overline{V}_t = \lambda I + \sum_{s=1}^{t-1} z_s z_s^\top$$

be the regularized design matrix underlying the covariates. Define

$$\beta_t(\delta) = \left(nL \sqrt{2 \log \left(\frac{\det(\overline{V}_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \lambda^{1/2} S \right)^2 \ . \tag{5.7}$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\mathrm{trace}((\widehat{\Theta}_t - \Theta_*)^\top \overline{V}_t (\widehat{\Theta}_t - \Theta_*)) \le \beta_t(\delta) \ .$$

In particular, $\mathbb{P}\left(\Theta_* \in C_t(\delta),\, t = 1, 2, \dots\right) \ge 1 - \delta$, where

$$C_t(\delta) = \left\{ \Theta \in \mathbb{R}^{n \times m} \,:\, \mathrm{trace}\left\{ (\Theta - \widehat{\Theta}_t)^\top \overline{V}_t (\Theta - \widehat{\Theta}_t) \right\} \le \beta_t(\delta) \right\} \ .$$

Notice that the construction of the confidence set uses the knowledge of both L and S.

[2] Stabilizability differs from reachability in that it requires that the "unreachable" part of the state is stable. Similarly, detectability differs from observability in that it requires that the "unobservable" part of the state is stable.

Inputs: $S > 0, L > 0, \delta \in (0,1), \lambda > 0, Q \in \mathbb{R}^{n\times n}, R \in \mathbb{R}^{d\times d}$.
Set $\overline{V}_0 = \lambda I$ and $\widehat{\Theta}_0 = 0$.
$(\widetilde{A}_0, \widetilde{B}_0) = \widetilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in C_0(\delta)\cap \mathcal{S}} J(\Theta)$.
for $t := 0, 1, 2, \ldots$ **do**
  if $\det(\overline{V}_t) > 2\det(\overline{V}_0)$ **then**
    Calculate $\widehat{\Theta}_t$ using (5.6).
    Find $\widetilde{\Theta}_t$ such that $J(\widetilde{\Theta}_t) \le \inf_{\Theta \in C_t(\delta)\cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}$.
    Let $\overline{V}_0 = \overline{V}_t$.
  else
    $\widetilde{\Theta}_t = \widetilde{\Theta}_{t-1}$.
  end if
  Calculate a_t based on the current parameters, $a_t = K(\widetilde{\Theta}_t) x_t$.
  Execute control, observe new state x_{t+1}.
  Save (z_t, x_{t+1}) into the dataset, where $z_t^\top = (x_t^\top, a_t^\top)$.
  $\overline{V}_{t+1} := \overline{V}_t + z_t z_t^\top$.
end for

Figure 5.1: The OFULQ ALGORITHM for the LQ problem.

5.1.3 The OFULQ Algorithm

The adaptive controller uses the OFU principle. The idea is to construct the confidence set for the unknown parameter Θ_* as in Theorem 5.2 with an appropriately selected value of $\delta = \delta_t$ and then select the parameter that gives the smallest average loss over all parameters within the confidence set. Given the parameter selected, use the controls as specified by the optimal control law (5.4) using the parameter found.

However, there are two issues with the method described so far. First, solving for the minimizer of $J(\Theta)$ over $C_t(\delta) \cap \mathcal{S}$ may be too demanding. Therefore, we relax this requirement to finding a parameter vector $\widetilde{\Theta}_t$ from $C_t(\delta) \cap \mathcal{S}$ such that

$$J(\widetilde{\Theta}_t) \le \inf_{\Theta \in C_t(\delta)\cap \mathcal{S}} J(\Theta) + \frac{1}{\sqrt{t}}\,. \tag{5.8}$$

We will later argue that the relaxed requirement will not cause a significant increase of the regret.

The second issue is that performance may also get harmed if the controller is changed too frequently (the initial state distribution has to converge to the steady distribution underlying the controller for the controller's actual average reward to be close to the expected long-term average reward of the controller). Thus, to prevent the loss increase resulting from too frequent controller changes, we limit the frequency with which the controller can be switched. In particular, the idea is not to switch to a new controller before a significant amount of new information is collected about the current parameter estimates (this idea was used earlier, e.g., in the paper by Jaksch et al. (2010)). More specifically, the algorithm changes controllers only when the determinant of $\overline{V}_t$ is increased by a constant factor (chosen to be 2 in the algorithm). We call the resulting algorithm the OFULQ ALGORITHM for "optimism in the face of uncertainty linear quadratic algorithm". The details of the algorithm are given in Figure 5.1.

5.2 Analysis

In this section we give our main result together with its proof. Our main result states that with high probability the regret is of order $O(\sqrt{T})$.

Theorem 5.3 Fix $Q, R \succ 0$ and assume that A3 holds for some values of $S, \Lambda, C > 0$. Consider the OFULQ ALGORITHM with parameters $L, S > 0$, $\delta \in (0,1)$, $\lambda > 0$ and Q, R. Assume that $((z_t, x_{t+1}))_{t \geq 1}$ satisfies A2 with constant $L > 0$. Then, it holds that for any $0 < \delta < 1$, for any time T, with probability at least $1 - \delta$, the regret of OFULQ satisfies

$$R_T = \widetilde{O}\left(\sqrt{T \log(1/\delta)}\right) ,$$

where the constant hidden is a problem dependent constant.

Remark 5.4 We see that the major assumption is that the algorithm needs to know a bound on $\mathrm{trace}(\Theta_*^\top \Theta_*)$ and on the "sub-Gaussianity" constant $L > 0$. These bounds are used in the construction of the confidence set.

Remark 5.5 The assumption $\mathbb{E}\left[w_{t+1} w_{t+1}^\top \,|\, \mathcal{F}_t\right] = I_n$ in the Linear Model Assumption A2 makes the analysis more readable. Alternatively, we could assume that $\mathbb{E}\left[w_{t+1} w_{t+1}^\top \mid \mathcal{F}_t\right] = G_*$, which is unknown. Then the optimal average loss becomes $J(\Theta_*, G_*) = \mathrm{trace}(P(\Theta_*) G_*)$. The only change in the OFULQ ALGORITHM is in the computation of $\widetilde{\Theta}_t$, which will have the following form:

$$(\widetilde{\Theta}_t, \widetilde{G}) = \operatorname*{argmin}_{(\Theta, G) \in C_t(\delta)} J(\Theta),$$

where $C_t(\delta)$ is now a confidence set over Θ_* and G_*. The rest of the analysis remains identical, provided that an appropriate confidence set is constructed.

Remark 5.6 In online learning problems (see Problem 1 in Chapter 1), obtaining a high probability regret bound is usually considered a stronger result than obtaining a bound on the expected regret. This is because in such problems the loss function is assumed to be bounded so that a simple argument obtains a bound on the expected regret: assume we have a high probability bound of the form,

$$\mathbb{P}\left(R_T \leq B_T \log(1/\delta)\right) \geq 1 - \delta \, .$$

Assume that the loss function is bounded by L, which implies that $R_T \leq LT$. Then we have that for any $\delta \in (0,1)$,

$$\mathbb{E}\left[R_T\right] \leq B_T \log(1/\delta) + \delta L T \, .$$

Choose $\delta = 1/T$ and obtain $\mathbb{E}\left[R_T\right] \leq B_T \log(T) + L$.

This argument is not applicable in the current LQ setting with unbounded loss functions. Notice that, in the worst case, the state vector can grow exponentially. Thus, Theorem 5.3 provides only a high probability regret bound and cannot be used to obtain a bound on the expected regret. We leave it for future work to derive a bound on the expected regret.

The least-squares estimation error in Theorem 5.2 scales with the size of the state and action vectors. First we show that with high probability the norm of the state vector grows slowly. Given the well-behavedness of the state, we decompose the regret and analyze each term using appropriate concentration inequalities to prove Theorem 5.3. We might expect that in order to have a sublinear regret, the state should converge to zero as time goes to infinity. But notice that the state is perturbed by a sub-Gaussian noise, that, with high probability, can take values as large as $\log(t)$ up to time t.

5.2.1 Bounding $\|x_t\|$

We choose an error probability, $\delta > 0$. Given this, we define two "good events" in the probability space Ω.

Definition 5.7 We define the event that the confidence sets hold for $s = 1, \dots, t$,

$$E_t = \{\omega \in \Omega : \forall s \le t, \quad \Theta_* \in C_s(\delta/4)\} ,$$

and the event that the state vector stays "small":

$$F_t = \{\omega \in \Omega : \forall s \le t, \quad \|x_s\| \le \Upsilon_t\} ,$$

where

$$\Upsilon_t = \frac{1}{1-\Lambda}\left(\frac{\Psi}{\Lambda}\right)^m \left(G\left(Z_t^m \beta_t(\delta/4)^{1/2}\right)^{1/m+1} + 2L\sqrt{n\log(4nt(t+1)/\delta)} \right) ,$$

and

$$\begin{aligned}
\Psi &= 1 \vee \sup_{\Theta\in\mathcal{S}} \|A_* + B_* K(\Theta)\| , & Z_t &= \max_{0\le s\le t} \|z_s\| , \\
G &= 2\left(2Sm^m\sqrt{mH_1H_3}\right)^{1/m+1} , & H_1 &> 16 \vee \tfrac{4S^2H_2^2H_3}{m} , \\
H_2 &= \sup_{Y\ge 1} \tfrac{1}{Y}\left(nL\sqrt{m\log\left(\tfrac{1+TY/\lambda}{\delta}\right)} + \lambda^{1/2}S \right) , & H_3 &= 16^{m-2}(1 \vee S^{2(m-2)}) .
\end{aligned}$$

In what follows, we let $E = E_T$ and $F = F_T$.

We show that $E \cap F$ holds with high probability and on $E \cap F$, the state vector grows slowly.

Lemma 5.8 $\mathbb{P}(E \cap F) \ge 1 - \delta/2$.

The proof is in Appendix F.2. The proof considers only the case when $Z_t > 1$, as otherwise, the state vector is obviously bounded. The proof first shows that $\left\|(\Theta_* - \widetilde{\Theta}_t)^\top z_t\right\|$ is well-controlled except for a small number of occasions. Given this and a proper decomposition of the state update equation, we prove that the state vector stays smaller than Υ_t. Notice that Υ_t itself depends on $\beta_t(\delta)$ and Z_t, which in turn depend on x_t. The next lemma takes the additional step to bound $\|x_t\|$. The proof is in Appendix F.2.

Lemma 5.9 For appropriate problem dependent constants $C_1 > 0, C_2 > 0$ (which are independent of t, δ, T), for any $t \ge 1$, it holds that $\mathbb{I}_{\{F_t\}} \max_{1\le s\le t} \|x_s\| \le X_t$, where

$$X_t = Y_t^{m+1}$$

and

$$Y_t \doteq (e \vee \lambda m(e-1) \vee 4(C_1\log(1/\delta) + C_2\log(t/\delta))\log^2(4(C_1\log(1/\delta) + C_2\log(t/\delta))) .$$

5.2.2 Regret Decomposition

Given the previous bound on the state vector, we decompose the regret and analyze each term using appropriate concentration inequalities. From the Bellman optimality equations[3] for the LQ problem, we get that (Bertsekas, 2001, V. 2, p. 228–229)

$$\begin{aligned}
J(\widetilde{\Theta}_t) + x_t^\top P(\widetilde{\Theta}_t) x_t &= \min_a \left\{ x_t^\top Q x_t + a^\top R a + \mathbb{E}\left[\widetilde{x}_{t+1}^{a\top} P(\widetilde{\Theta}_t) \widetilde{x}_{t+1}^a \;\middle|\; \mathcal{F}_t \right] \right\} \qquad (5.9) \\
&= x_t^\top Q x_t + a_t^\top R a_t + \mathbb{E}\left[\widetilde{x}_{t+1}^{a_t\top} P(\widetilde{\Theta}_t) \widetilde{x}_{t+1}^{a_t} \;\middle|\; \mathcal{F}_t \right] ,
\end{aligned}$$

[3]The existence of a solution for (5.9) follows from the existence of a solution for the Riccati equation.

where $\widetilde{x}^a_{t+1} = \widetilde{A}_t x_t + \widetilde{B}_t a + w_{t+1}$ and $(\widetilde{A}_t, \widetilde{B}_t) = \widetilde{\Theta}_t$. Thus,

$$\begin{aligned}
J(\widetilde{\Theta}_t) + x_t^\top P(\widetilde{\Theta}_t) x_t &= x_t^\top Q x_t + a_t^\top R a_t \\
&\quad + \mathbb{E}\left[(\widetilde{A}_t x_t + \widetilde{B}_t a_t + w_{t+1})^\top P(\widetilde{\Theta}_t)(\widetilde{A}_t x_t + \widetilde{B}_t a_t + w_{t+1}) \;\middle|\; \mathcal{F}_t \right] \\
&= x_t^\top Q x_t + a_t^\top R a_t + \mathbb{E}\left[(\widetilde{A}_t x_t + \widetilde{B}_t a_t)^\top P(\widetilde{\Theta}_t)(\widetilde{A}_t x_t + \widetilde{B}_t a_t) \;\middle|\; \mathcal{F}_t \right] \\
&\quad + \mathbb{E}\left[w_{t+1}^\top P(\widetilde{\Theta}_t) w_{t+1} \;\middle|\; \mathcal{F}_t \right] \\
&= x_t^\top Q x_t + a_t^\top R a_t + \mathbb{E}\left[(\widetilde{A}_t x_t + \widetilde{B}_t a_t)^\top P(\widetilde{\Theta}_t)(\widetilde{A}_t x_t + \widetilde{B}_t a_t) \;\middle|\; \mathcal{F}_t \right] \\
&\quad + \mathbb{E}\left[x_{t+1}^\top P(\widetilde{\Theta}_t) x_{t+1} \;\middle|\; \mathcal{F}_t \right] \\
&\quad - \mathbb{E}\left[(A_* x_t + B_* a_t)^\top P(\widetilde{\Theta}_t)(A_* x_t + B_* a_t) \;\middle|\; \mathcal{F}_t \right] \\
&= x_t^\top Q x_t + a_t^\top R a_t + \mathbb{E}\left[x_{t+1}^\top P(\widetilde{\Theta}_t) x_{t+1} \;\middle|\; \mathcal{F}_t \right] \\
&\quad + (\widetilde{A}_t x_t + \widetilde{B}_t a_t)^\top P(\widetilde{\Theta}_t)(\widetilde{A}_t x_t + \widetilde{B}_t a_t) \\
&\quad - (A_* x_t + B_* a_t)^\top P(\widetilde{\Theta}_t)(A_* x_t + B_* a_t),
\end{aligned}$$

where in the third equality we have used $x_{t+1} = A_* x_t + B_* a_t + w_{t+1}$ and the martingale property of the noise. Thus,

$$\sum_{t=1}^{T} J(\widetilde{\Theta}_t) + R_1 = \sum_{t=1}^{T} \left(x_t^\top Q x_t + a_t^\top R a_t \right) + R_2 + R_3 \,,$$

where

$$R_1 = \sum_{t=1}^{T} \left\{ x_t^\top P(\widetilde{\Theta}_t) x_t - \mathbb{E}\left[x_{t+1}^\top P(\widetilde{\Theta}_{t+1}) x_{t+1} \middle| \mathcal{F}_t \right] \right\} , \tag{5.10}$$

$$R_2 = \sum_{t=1}^{T} \mathbb{E}\left[x_{t+1}^\top (P(\widetilde{\Theta}_t) - P(\widetilde{\Theta}_{t+1})) x_{t+1} \middle| \mathcal{F}_t \right] , \tag{5.11}$$

$$R_3 = \sum_{t=1}^{T} \left((\widetilde{A}_t x_t + \widetilde{B}_t a_t)^\top P(\widetilde{\Theta}_t)(\widetilde{A}_t x_t + \widetilde{B}_t a_t) - (A_* x_t + B_* a_t)^\top P(\widetilde{\Theta}_t)(A_* x_t + B_* a_t) \right) . \tag{5.12}$$

Thus, on $E \cap F$,

$$\begin{aligned}
\sum_{t=1}^{T} (x_t^\top Q x_t + a_t^\top R a_t) &= \sum_{t=1}^{T} J(\widetilde{\Theta}_t) + R_1 - R_2 - R_3 \\
&\le T J(\Theta_*) + R_1 - R_2 - R_3 + 2\sqrt{T} \,,
\end{aligned}$$

where the inequality follows from the choice of $\widetilde{\Theta}_t$ and the fact that on E, $\Theta_* \in C_t(\delta)$. Thus, on $E \cap F$,

$$R(T) \le R_1 - R_2 - R_3 + 2\sqrt{T} \,. \tag{5.13}$$

In the following subsections, we bound R_1, R_2, and R_3.

5.2.3 Bounding $\mathbb{I}_{\{E \cap F\}} R_1$

We start by showing that with high probability all noise terms are small.

Lemma 5.10 With probability $1-\delta/8$, for any $t \le T$, $\|w_t\| \le Ln\sqrt{2n\log(8nT/\delta)}$.

Proof. From the Linear Model Assumption A2, we have that for any index $1 \le i \le n$ and any time t,

$$\mathbb{P}\left(|w_{t,i}| \le L\sqrt{2\log(8/\delta)}\right) \ge 1-\delta/8 \ .$$

With an union bound on time and dimension, we get that, with probability $1-\delta/8$, for any $t \le T$, $\|w_t\| \le Ln\sqrt{2n\log(8nT/\delta)}$. □

Lemma 5.11 Let R_1 be as defined by (5.10). Let $W = Ln\sqrt{2n\log(8nT/\delta)}$, $\nu > 0$ be an arbitrary positive constant, and

$$B'_\delta = \left(\nu + TD^2S^2X^2(1+C^2)\right)\log\left(\frac{4n\nu^{-1/2}}{\delta}\left(\nu + TD^2S^2X^2(1+C^2)\right)^{1/2}\right) \ .$$

With probability at least $1-\delta/2$,

$$\mathbb{I}_{\{E\cap F\}}R_1 \le 2DW^2\sqrt{2T\log(8/\delta)} + n\sqrt{B'_\delta} \ .$$

Proof. Let $f_{t-1} = A_*x_{t-1} + B_*a_{t-1}$ and $P_t = P(\widetilde{\Theta}_t)$. Write

$$R_1 = x_1^\top P(\widetilde{\Theta}_1)x_1 - x_{T+1}^\top P(\widetilde{\Theta}_{T+1})x_{T+1} + \sum_{t=2}^{T}\left(x_t^\top P(\widetilde{\Theta}_t)x_t - \mathbb{E}\left[x_t^\top P(\widetilde{\Theta}_t)x_t|\mathcal{F}_{t-1}\right]\right) .$$

Because P is positive semi-definite and $x_1 = 0$, the first term, is bounded by zero. The second term can be decomposed as follows:

$$\sum_{t=2}^{T}\left(x_t^\top P_tx_t - \mathbb{E}\left[x_t^\top P_tx_t|\mathcal{F}_{t-1}\right]\right) = \sum_{t=2}^{T} f_{t-1}^\top P_tw_t + \sum_{t=2}^{T}\left(w_t^\top P_tw_t - \mathbb{E}\left[w_t^\top P_tw_t|\mathcal{F}_{t-1}\right]\right) .$$

We bound each term separately. Let $v_t^\top = f_{t-1}^\top P_t$ and

$$G_1 = \mathbb{I}_{\{E\cap F\}}\sum_{t=2}^{T} v_t^\top w_t = \mathbb{I}_{\{E\cap F\}}\sum_{t=2}^{T}\sum_{i=1}^{n} v_{t,i}w_{t,i} = \sum_{i=1}^{n}\mathbb{I}_{\{E\cap F\}}\sum_{2=1}^{T} v_{t,i}w_{t,i} \ .$$

Let $M_{T,i} = \sum_{t=1}^{T} v_{k,i}w_{k,i}$. By Corollary 3.6, on some event $G_{\delta,i}$ that holds with probability at least $1-\delta/(4n)$, for any $T \ge 0$,

$$M_{T,i}^2 \le 2R^2\left(\nu + \sum_{2=1}^{T} v_{t,i}^2\right)\log\left(\frac{4n\nu^{-1/2}}{\delta}\left(\nu + \sum_{2=1}^{T} v_{t,i}^2\right)^{1/2}\right) \doteq B_{\delta,i} \ .$$

On $E\cap F$, $\|v_t\| \le DSX\sqrt{1+C^2}$ and thus, $v_{t,i} \le DSX\sqrt{1+C^2}$. Thus, on $G_{\delta,i}$, $\mathbb{I}_{\{E\cap F\}}M_{t,i}^2 \le B'_\delta$. Thus, we have

$$G_1 \le \sum_{i=1}^{n}\sqrt{B'_{\delta,i}}$$

on $\cap_{i=1}^{n}G_{\delta,i}$, that holds w.p. $1-\delta/4$.

Define $U_t = w_t^\top P_tw_t - \mathbb{E}\left[w_t^\top P_tw_t|\mathcal{F}_{t-1}\right]$ and its truncated version $\widetilde{U}_t = U_t\mathbb{I}_{\{U_t\le 2DW^2\}}$. Define $G_2 = \sum_{t=2}^{T} U_t$ and $\widetilde{G}_2 = \sum_{t=1}^{T}\widetilde{U}_t$. By Lemma C.7,

$$\mathbb{P}\left(G_2 > 2DW^2\sqrt{2T\log\frac{8}{\delta}}\right) \le \mathbb{P}\left(\max_{2\le t\le T} U_t \ge 2DW^2\right) + \mathbb{P}\left(\widetilde{G}_2 > 2DW^2\sqrt{2T\log\frac{8}{\delta}}\right) \ .$$

By Lemma 5.10 and Azuma's inequality, each term on the right hand side is bounded by $\delta/8$. Thus, w.p. $1-\delta/4$,

$$G_2 \leq 2DW^2\sqrt{2T\log\frac{8}{\delta}}\,.$$

Summing up the bounds on G_1 and G_2 gives the result that holds w.p. at least $1-\delta/2$,

$$\mathbb{I}_{\{E\cap F\}}R_1 \leq 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B'_\delta}\,.$$

□

5.2.4 Bounding $\mathbb{I}_{\{E\cap F\}}|R_2|$

We can bound $\mathbb{I}_{\{E\cap F\}}|R_2|$ by simply showing that the OFULQ ALGORITHM rarely changes the policy, and hence most terms in (5.11) are zero.

Lemma 5.12 On the event $E\cap F$, the OFULQ ALGORITHM changes the policy at most

$$m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)$$

times up to time T.

Proof. If the OFULQ ALGORITHM has changed the policy K times up to time T, then we should have that $\det(\overline{V}_T)\geq\lambda^m 2^K$. On the other hand, we have

$$\lambda_{\max}(\overline{V}_T)\leq\lambda+\sum_{t=1}^{T-1}\|z_t\|^2\leq\lambda+TX_T^2(1+C^2)\,,$$

where C is the upper bound on the norm of $K(.)$ (see Assumption A3). Thus, it holds that

$$\lambda^m 2^K\leq(\lambda+TX_T^2(1+C^2))^m\,.$$

Solving for K, we get

$$K\leq m\log_2\left(1+\frac{TX_T^2(1+C^2)}{\lambda}\right)\,.$$

□

Lemma 5.13 Let R_2 be as defined by (5.11). Then we have

$$\mathbb{I}_{\{E\cap F\}}|R_2|\leq 2DX_T^2 m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)\,.$$

Proof. On event $E\cap F$, we have at most $K=m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)$ policy changes up to time T. So at most K terms in the summation (5.11) are non-zero. Each term in the summation is bounded by $2DX_T^2$. Thus,

$$\mathbb{I}_{\{E\cap F\}}|R_2|\leq 2DX_T^2 m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)\,.$$

□

5.2.5 Bounding $\mathbb{I}_{\{E\cap F\}}|R_3|$

The summation $\sum_{t=1}^{T}\left\|(\Theta_*-\widetilde{\Theta}_t)^\top z_t\right\|^2$ will appear while bounding $|R_3|$. So we first upper bound this summation, whose analysis requires Lemma E.2 from Appendix E and the following lemma whose proof can be found in Appendix F. Notice that Lemma 5.14 is the matrix counter-part of Lemma 4.6.

Lemma 5.14 Let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times m}$ be positive semi-definite matrices such that $A \succeq B$. Then, we have

$$\sup_{X\neq 0} \frac{\left\|X^\top A X\right\|}{\|X^\top B X\|} \leq \frac{\det(A)}{\det(B)} .$$

Lemma 5.15 On $E \cap F$, it holds that

$$\sum_{t=1}^{T} \left\|(\Theta_* - \widetilde{\Theta}_t)^\top z_t\right\|^2 \leq \frac{16}{\lambda}(1+C^2)X_T^2\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)} .$$

Proof. Consider round t. Let $s_t = (\Theta_* - \widetilde{\Theta}_t)^\top z_t$. Let $\tau \leq t$ be the last round that the policy is changed. So $\widetilde{\Theta}_t = \widetilde{\Theta}_\tau$ and $s_t = (\Theta_* - \widetilde{\Theta}_\tau)^\top z_t$. By Triangle inequality we have

$$\|s_t\| \leq \left\|(\Theta_* - \widehat{\Theta}_\tau)^\top z_t\right\| + \left\|(\widehat{\Theta}_\tau - \widetilde{\Theta}_\tau)^\top z_t\right\| . \tag{5.14}$$

Next, we bound each term on the RHS. For all $\Theta \in C_\tau$,

$$\begin{aligned}
\left\|(\Theta - \widehat{\Theta}_\tau)^\top z_t\right\| &\leq \left\|\overline{V}_t^{1/2}(\Theta - \widehat{\Theta}_\tau)\right\| \|z_t\|_{\overline{V}_t^{-1}} && \text{(Cauchy-Schwarz inequality)}\\
&\leq \left\|\overline{V}_\tau^{1/2}(\Theta - \widehat{\Theta}_\tau)\right\| \sqrt{\tfrac{\det(\overline{V}_t)}{\det(\overline{V}_\tau)}}\, \|z_t\|_{\overline{V}_t^{-1}} && \text{(Lemma 5.14)}\\
&\leq \sqrt{2}\left\|\overline{V}_\tau^{1/2}(\Theta - \widehat{\Theta}_\tau)\right\| \|z_t\|_{\overline{V}_t^{-1}} && \text{(Choice of } \tau)\\
&\leq \sqrt{2\beta_\tau(\delta/4)}\, \|z_t\|_{\overline{V}_t^{-1}} , && (\lambda_{\max}(M) \leq \mathrm{trace}(M) \text{ for } M \succeq 0)
\end{aligned}$$

Applying the inequality to Θ_* and $\widetilde{\Theta}_\tau$, together with (5.14) gives

$$\|s_t\|^2 \leq 8\beta_\tau(\delta/4)\, \|z_t\|^2_{\overline{V}_t^{-1}} .$$

By the fact that $\widetilde{\Theta}_t \in \mathcal{S}$ we have that

$$\|z_t\|^2_{\overline{V}_t^{-1}} \leq \frac{\|z_t\|^2}{\lambda} \leq \frac{(1+C^2)X_T^2}{\lambda} .$$

It follows then that

$$\begin{aligned}
\sum_{t=1}^{T} \|s_t\|^2 &\leq \frac{8}{\lambda}(1+C^2)X_T^2\beta_T(\delta/4)\sum_{t=0}^{T}(\|z_t\|^2_{\overline{V}_t^{-1}} \wedge 1)\\
&\leq \frac{16}{\lambda}(1+C^2)X_T^2\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)} . && \text{(Lemma E.2).}
\end{aligned}$$

□

Now, we are ready to bound R_3.

Lemma 5.16 Let R_3 be as defined by (5.12). Then we have

$$\mathbb{I}_{\{E\cap F\}}\,|R_3| \leq \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2 SD\left(\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T} .$$

Proof. We have that

$$\begin{aligned}
\mathbb{I}_{\{E\cap F\}}|R_3| &\le \mathbb{I}_{\{E\cap F\}}\sum_{t=1}^{T}\left|\left\|P(\widetilde{\Theta}_t)^{1/2}\widetilde{\Theta}_t^\top z_t\right\|^2-\left\|P(\widetilde{\Theta}_t)^{1/2}\Theta_*^\top z_t\right\|^2\right| && \text{(Tri. ineq.)}\\
&\le \mathbb{I}_{\{E\cap F\}}\left(\sum_{t=1}^{T}\left(\left\|P(\widetilde{\Theta}_t)^{1/2}\widetilde{\Theta}_t^\top z_t\right\|-\left\|P(\widetilde{\Theta}_t)^{1/2}\Theta_*^\top z_t\right\|\right)^2\right)^{1/2} && \text{(C.-S. ineq.)}\\
&\quad\times\left(\sum_{t=1}^{T}\left(\left\|P(\widetilde{\Theta}_t)^{1/2}\widetilde{\Theta}_t^\top z_t\right\|+\left\|P(\widetilde{\Theta}_t)^{1/2}\Theta_*^\top z_t\right\|\right)^2\right)^{1/2}\\
&\le \mathbb{I}_{\{E\cap F\}}\left(\sum_{t=1}^{T}\left\|P(\widetilde{\Theta}_t)^{1/2}(\widetilde{\Theta}_t-\Theta_*)^\top z_t\right\|^2\right)^{1/2} && \text{(Tri. ineq.)}\\
&\quad\times\left(\sum_{t=1}^{T}\left(\left\|P(\widetilde{\Theta}_t)^{1/2}\widetilde{\Theta}_t^\top z_t\right\|+\left\|P(\widetilde{\Theta}_t)^{1/2}\Theta_*^\top z_t\right\|\right)^2\right)^{1/2}\\
&\le \frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2SD\left(\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T}\,. && \text{((5.5), L. 5.15)}
\end{aligned}$$

□

5.2.6 Putting Everything Together

Proof of Theorem 5.3. By (5.13) and Lemmas 5.11, 5.13, 5.16 we have that with probability at least $1-\delta/2$,

$$\begin{aligned}
\mathbb{I}_{\{E\cap F\}}(R_1-R_2-R_3) &\le 2DX_T^2m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)+2DW^2\sqrt{2T\log\frac{8}{\delta}}+n\sqrt{B'_\delta}\\
&\quad+\frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2SD\left(\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T}\,.
\end{aligned}$$

Thus, on $E\cap F$, with probability at least $1-\delta/2$,

$$\begin{aligned}
R(T) &\le 2DX_T^2m\log_2\left(1+TX_T^2(1+C^2)/\lambda\right)+2DW^2\sqrt{2T\log\frac{8}{\delta}}+n\sqrt{B'_\delta}\\
&\quad+\frac{8}{\sqrt{\lambda}}(1+C^2)X_T^2SD\left(\beta_T(\delta/4)\log\frac{\det(\overline{V}_T)}{\det(\lambda I)}\right)^{1/2}\sqrt{T}\,.
\end{aligned}$$

Further, on $E\cap F$, by Lemmas 5.9 and E.2,

$$\log\det\overline{V}_T\le m\log\left(\frac{\lambda m+T(1+C^2)X_T^2}{\lambda m}\right)+\log\det\lambda I\,.$$

Plugging in the above inequality gives the final bound, which, by Lemma 5.8, holds with probability $1-\delta$. □

5.3 Extension to Non-Linear Dynamics

The optimization problem (5.8) is defined over reachable and observable matrices. The reachability and observability assumptions guarantee that the Riccati equation has a solution, which in turn guarantees a solution to the average cost optimality equation (ACOE). Thus, a solution to the ACOE is guaranteed for a LQ problem whose model is specified by

the solution of (5.8). Thus, we could write the optimality equation (5.9), from which we obtained the regret decomposition.

In this section, we study a more general control problem with a non-linear transition law. It turns out that the same proof technique still can be employed under the weaker assumption that the average cost optimality inequality (ACOI) has a solution. It only remains to find conditions analogous to reachability and observability, under which, the ACOI has a solution in a non-linear control problem. We take a general approach and make the following assumption:

Assumption A4 Let $\mathcal{X} \subset \mathbb{R}^n$ be the state space. Let $p(. \,|\, ., ., \Theta) : \mathcal{X} \times D \to \mathcal{X}$ be the transition law parametrized by Θ. Let Θ_* be the true parameter. There exists $J(\Theta_*)$ and $h(., \Theta_*) : \mathcal{X} \to \mathbb{R}$ that satisfy the following average cost optimality inequality (ACOI) for any $x \in \mathcal{X}$:

$$J(\Theta_*) + h(x, \Theta_*) \geq \min_{a \in D(x)} \left\{ \ell(x, a) + \int h(y, \Theta_*) p(dy \,|\, x, a, \Theta_*) \right\} .$$

Further, we assume that there exists an oracle that, for any possible value of Θ, can determine if this ACOI has a solution and if a solution exists, provide one.

We further make the following assumption:

Assumption A5 For $\Theta = \Theta_*$, the optimal average loss, J_* is well-defined. Further, it holds that $J(\Theta_*) \leq J_*$.

With a slight abuse of the concepts, we will call the quantity $J(\Theta)$ the average loss of the optimal policy, while function $h(., \Theta)$ will be called the value function. In what follows, we denote $h(., \Theta_*)$ by $h_*(.)$. We denote the set of parameters for which the ACOI has a solution by $\mathcal{K}$. By Assumption A4, $\Theta_* \in \mathcal{K}$.

The following examples, taken from (Arapostathis et al., 1993, Hernández-Lerma and Lasserre, 1996), show two cases when Assumption A4 is satisfied.

Example 2 Let $\mathcal{F}(\mathcal{X})$ be the space of bounded functions on state space $\mathcal{X}$, and $\mathcal{F}_y(\mathcal{X})$ be the subspace of functions f in $\mathcal{F}(\mathcal{X})$ such that $f(y) = 0$ for some given point $y \in \mathcal{X}$. Define the span of a function f as

$$s(f) \doteq \sup_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} f(x) .$$

Recall the definition of a Markov Decision Process from Chapter 1. Assume that the loss is bounded, lower semi-continuous[4], non-negative, and inf-compact[5] on the set of admissible state-action pairs. Assume that the transition kernel is strongly continuous. Define the mapping $T_y : \mathcal{F}_y(\mathcal{X}) \to \mathcal{F}_y(\mathcal{X})$ by

$$T_y f(x) = Tf(x) - Tf(y), \qquad x \in \mathcal{X} , \tag{5.15}$$

where

$$Tf(x) = \min_{a \in D(x)} \left[\ell(x, a) + \int f(y) p(dy|x, a) \right] .$$

Define the total variation norm of a finite signed measure λ on $\mathcal{X}$ by

$$\|\lambda\|_V \doteq \sup_{B \in \mathcal{B}(\mathcal{X})} \lambda(B) - \inf_{B \in \mathcal{B}(\mathcal{X})} \lambda(B) .$$

In particular, we have

$$\|p - q\|_V = 2 \sup_{B \in \mathcal{B}(\mathcal{X})} |p(B) - q(B)|$$

[4] A function f is lower semi-continuous at x_0 if $\liminf_{x \to x_0} f(x) \geq f(x_0)$.

[5] The loss function ℓ is inf-compact if the set $\{a \in D(x) : \ell(x, a) \leq r\}$ is compact for any x in $\mathcal{X}$ and r in $\mathbb{R}$.

for probability measures p and q. Suppose the transition kernel satisfies

$$\forall (x,a),\,(x',a') \in \mathcal{Z}, \qquad \|p(.|x,a) - p(.|x',a')\|_V \leq 2\beta\,, \tag{5.16}$$

where $0 < \beta < 1$ and $\mathcal{Z}$ is the set of admissible state-action pairs. Under this condition, it can be shown that T_y is a contraction mapping. Thus, by the Banach Fixed-Point Theorem, we get that there is a unique function $h \in \mathcal{F}_y(\mathcal{X})$ that satisfies $T_y h = h$. By substituting h in (5.15) we get the ACOE

$$\forall x \in \mathcal{X}, \qquad J + h(x) = Th(x)\,,$$

where $J = Th(y)$.

To see an example when condition (5.16) holds, consider the case when the state transition is given by

$$x_{t+1} = f(x_t, a_t) + g(x_t) w_{t+1}\,,$$

the state and action spaces are compact sets; $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ are bounded, continuous, and $g(.) > 0$; and (w_t) is a sequence of independent $\mathcal{N}(0, I)$ random vectors.

Example 3 Assume that the loss is lower semi-continuous, non-negative, and inf-compact on the set of admissible state-action pairs. Assume that the transition kernel is strongly continuous. Define the total discounted loss by

$$V_\gamma(\pi, x) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) \,|\, x_1 = x\right]$$

and the optimal total discounted loss by

$$V_\gamma^*(x) = \inf_{\pi \in \Pi} V_\gamma(x, \pi)\,,$$

where Π is the class of all policies. Suppose there exists a state $y \in \mathcal{X}$ and constants $\beta \in (0,1)$ and $M \geq 0$ such that for all $\gamma \in [\beta, 1)$,

$$(1-\gamma) V_\gamma^*(y) \leq M\,. \tag{5.17}$$

Further, assume that there is a constant $N \geq 0$ and a non-negative function $b(.)$ on $\mathcal{X}$ such that for any $x \in \mathcal{X}$ and any $\gamma \in [\beta, 1)$,

$$-N \leq V_\gamma^*(x) - V_\gamma^*(y) \leq b(x)\,. \tag{5.18}$$

Under these assumptions, it can be shown that there exist $0 < \rho^* \leq M$ and a sequence $\gamma_n \uparrow 1$ such that for any $x \in \mathcal{X}$,

$$\lim_{n\to\infty} (1-\gamma_n) V_{\gamma_n}^*(x) = \rho^*\,. \tag{5.19}$$

Further, under the same assumptions, it can be shown that there exists a constant J and a function $h : \mathcal{X} \to \mathbb{R}$ such that $h(y) = 0$, for any $x \in \mathcal{X}$

$$-N \leq h(x) \leq b(x)\,,$$

and (J, h) satisfy the ACOI

$$J + h(x) \geq \min_{a \in D(x)} \left[\ell(x,a) + \int h(y) p(dy \,|\, x, a)\right]\,, \quad x \in \mathcal{X}\,.$$

Condition (5.17) is satisfied, for example, for bounded losses.

We can show that the ACOE also has a solution by making the following additional assumptions:

(i) The function $b(.)$ in (5.18) is measurable and that for any $x \in \mathcal{X}$ and $a \in D(x)$, it satisfies $\int_{\mathcal{X}} b(y)p(dy|x,a) < \infty$

(ii) The sequence $(V^*_{\gamma_n}(x) - V^*_{\gamma_n}(y))$ is equicontinuous,[6] where (γ_n) is the sequence that satisfies (5.19).

In this section, we assume that the model has the form of

$$x_{t+1} = \Theta_*^\top \varphi(x_t, a_t) + w_{t+1}\,,$$

where $\Theta_* \in \mathbb{R}^{m\times n}$ is an unknown matrix, $\varphi : \mathcal{Z} \to \mathbb{R}^m$ is a feature mapping, and the noise w_{t+1} has the same properties as before (see the Linear Model Assumption A2). We consider any loss function $\ell(x_t, a_t)$ as long as a smoothness assumption on the *value function* is satisfied (see Assumption A7).

We apply techniques similar to those in Section 5.2 to obtain sublinear regret bounds for this family of control problems. Before stating our main result, we make a number of assumptions on the loss and the transition law. We make the following assumption on the boundedness of the parameter matrix:

Assumption A6 The unknown parameter Θ_* is a member of set $\mathcal{S}$ defined by

$$\mathcal{S} = \left\{\Theta \in \mathbb{R}^{n\times(n+d)} \,:\, \mathrm{trace}(\Theta^\top\Theta) \le S^2\right\}\,.$$

We also make an assumption on the smoothness of the value function:

Assumption A7 Lipschitz Continuity There exists $B > 0$ such that for all $\Theta \in \mathcal{S}$, $h(0,\Theta) = 0$ and for all $x, x' \in \mathcal{X}$, $|h(x,\Theta) - h(x',\Theta)| \le B\,\|x - x'\|$.

Assumption A8 Optimal Policies We assume that for any $\Theta \in \mathcal{K}$, an optimal deterministic stationary policy exists and the algorithm has access to an oracle that returns such a policy. We denote the output of this method by $a(.,\Theta)$.

This condition could again be relaxed, since we do not need the actual optimal policy, only an approximately optimal policy.

Finally, we make the following assumption on the feature mapping:

Assumption A9 There exist $0 < \Lambda < 1$, $K > 0$, $k > 0$ such that for all $\Theta \in \mathcal{S}$ and for all $x \in \mathcal{X}$,

$$\left\|\Theta^\top\varphi(x, a(x,\Theta))\right\| \le \Lambda\,\|x\|\,, \tag{5.20}$$

$$\|\varphi(x, a(x,\Theta))\| \le K\,\|x\|^k\,. \tag{5.21}$$

Condition 5.20 will let us show that the state vector is "well-behaved". The condition can be relaxed in a bounded state space. We can obtain similar results by replacing (5.20) and (5.21) with the less restrictive conditions that $\left\|\Theta^\top\varphi(x, a(x,\Theta))\right\| \le \Lambda\,\|x\| + \Lambda'\,, \Lambda' > 0$ and $\|\varphi(x, a(x,\Theta))\| \le K\,\|x\|^k + K'\,, K' > 0$. We have decided to use the stronger conditions to simplify the presentation.

Let $\widehat{\Theta}_t$ be the ℓ^2-regularized least-squares estimate of Θ_* with regularization parameter $\lambda > 0$. Theorem 5.2 gives a high-probability confidence set around Θ_*:

$$C_t(\delta) = \Bigg\{\Theta \,:\, \mathrm{trace}((\widehat{\Theta}_t - \Theta)^\top \overline{V}_t(\widehat{\Theta}_t - \Theta)) \le \left(nL\sqrt{2\log\left(\frac{\det(\overline{V}_t)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} + \lambda^{1/2}\,S\right)^2\Bigg\},$$

[6] A class of functions $\mathcal{F}$ is equicontinuous if for any $x \in \mathcal{X}$, for any $\epsilon > 0$, there exists a set G such that $x \in G$ and for any $f \in \mathcal{F}$ and $y \in G$,

$$|f(y) - f(x)| < \epsilon\,.$$

Inputs: $S > 0, \delta \geq 0, L, \lambda > 0$.
Set $\overline{V}_0 = \lambda I$ and $\widehat{\Theta}_0 = 0$.
$\widetilde{\Theta}_0 = \operatorname{argmin}_{\Theta \in C_0(\delta) \cap \mathcal{S} \cap \mathcal{K}} J(\Theta)$.
for $t := 0, 1, 2, \ldots$ **do**
  if $\det(\overline{V}_t) > 2 \det(V_0)$ **then**
    Calculate the ridge-regression estimate $\widehat{\Theta}_t$.
    Find $\widetilde{\Theta}_t$ such that $J(\widetilde{\Theta}_t) \leq \inf_{\Theta \in C_t(\delta) \cap \mathcal{S} \cap \mathcal{K}} J(\Theta) + \frac{1}{\sqrt{t}}$.
    Let $\overline{V}_0 = \overline{V}_t$.
  else
    $\widetilde{\Theta}_t = \widetilde{\Theta}_{t-1}$.
  end if
  Calculate control $a_t = a(x_t, \widetilde{\Theta}_t)$.
  Execute control, observe new state x_{t+1}.
  Save $(\varphi(x_t, a_t), x_{t+1})$ into the dataset.
  $\overline{V}_{t+1} := \overline{V}_t + \varphi(x_t, a_t)\varphi(x_t, a_t)^\top$.
end for

Figure 5.2: The OFUNLQ ALGORITHM: The implementation of the OFU principle for non-linear control problems.

where $\overline{V}_t = \lambda I + \sum_{s=1}^{t-1} \varphi(x_s, a_s)\varphi(x_s, a_s)^\top$, L is defined in Assumption A2, S is defined in Assumption A6, and δ is the confidence level.

As before, the objective is to have low average loss. We use the OFU principle as follows: at time t, we pick a parameter $\widetilde{\Theta}_t \in C_t(\delta) \cap \mathcal{S} \cap \mathcal{K}$ that has a small average loss $J(\widetilde{\Theta}_t)$:

$$J(\widetilde{\Theta}_t) \leq \inf_{\Theta \in C_t(\delta) \cap \mathcal{S} \cap \mathcal{K}} J(\Theta) + \frac{1}{\sqrt{t}} \,. \tag{5.22}$$

We assume that there exists an optimization method that, given the model, returns the optimal policy. We denote the output of this method by $a(., \Theta)$. At time t, we take action $a_t = a(x_t, \widetilde{\Theta}_t)$. Because Θ_* belongs to the confidence set, we have that $J(\widetilde{\Theta}_t) \leq J_* + 1/\sqrt{t}$. The details of the algorithm are given in Figure 5.2.

5.3.1 Analysis

The following theorem states a high probability bound on the regret of the OFUNLQ algorithm:

Theorem 5.17 Assume that A4–A9 hold for some values of $0 < \Lambda < 1, S, B, k, K > 0$. Assume that $((\varphi(x_t, a_t), x_{t+1}))_{t \geq 1}$ satisfies A2 with constant $L > 0$. Consider the OFUNLQ algorithm with parameters $L, S > 0$, $\delta \in (0, 1)$, and $\lambda > 0$. Then, it holds that for any $0 < \delta < 1$, for any time T, with probability at least $1 - \delta$, the regret of OFUNLQ satisfies

$$R_T = \widetilde{O}\left(\sqrt{T \log(1/\delta)}\right) \,,$$

where the hidden constant depends on the problem.

The proof is similar to that of Theorem 5.3. The least-squares estimation error in Theorem 3.15 scales with the size of the state vectors. First we show that with high probability the norm of the state vector grows slowly. Given the well-behavedness of the state, we decompose the regret and analyze each term using appropriate concentration inequalities to finish the proof.

Bounding $\|x_t\|$

We choose an error probability $\delta > 0$. Given this, we define two "good events" in the probability space Ω.

Definition 5.18 We define the event that the confidence sets hold for $s = 1, \dots, t$,

$$E_t = \{\omega \in \Omega : \forall s \le t, \quad \Theta_* \in C_s(\delta/4)\} ,$$

and the event that the state vector stays "small":

$$F_t = \{\omega \in \Omega : \forall s \le t, \quad \|x_s\| \le \Upsilon_t\} ,$$

where

$$\Upsilon_t = \frac{1}{1-\Lambda}\left(\frac{\Psi}{\Lambda}\right)^m \left(G\left(Z_t^m \beta_t (\delta/4)^{1/2}\right)^{1/m+1} + 2L\sqrt{n\log(4nt(t+1)/\delta)}\right) ,$$

and

$$\begin{aligned}
\Psi &= 1 \vee \max_{t\le T} \left\|\Theta_*^\top \varphi(x_t, a_t)\right\| / \|x_t\| , & Z_t &= \max_{1\le s\le t} \|\varphi(x_s, a_s)\| , \\
G &= 2\left(2Sm^m\sqrt{mH_1H_3}\right)^{1/m+1} , & H_1 &> 16 \vee \tfrac{4S^2H_2^2H_3}{m} , \\
H_2 &= \sup_{Y\ge 0} \tfrac{1}{Y}\left(nL\sqrt{m\log\left(\tfrac{1+TY/\lambda}{\delta}\right)} + \lambda^{1/2}S\right) , & H_3 &= 16^{m-2}(1 \vee S^{2(m-2)}) .
\end{aligned}$$

In what follows, we let $E = E_T$ and $F = F_T$.

We show that $E \cap F$ holds with high probability and on $E \cap F$, the state vector grows slowly.

Lemma 5.19 $\mathbb{P}(E \cap F) \ge 1 - \delta/2$.

Lemma 5.20 For appropriate problem dependent constants $C_1 > 0, C_2 > 0$ (which are independent of t, δ, T), for any $t \ge 0$, it holds that $\mathbb{I}_{\{F_t\}} \max_{1\le s\le t} \|x_s\| \le X_t$, where

$$X_t = Y_t^{m+1}$$

and

$$Y_t \doteq (e \vee \lambda m(e-1) \vee 4(C_1\log(1/\delta) + C_2\log(t/\delta))\log^2(4(C_1\log(1/\delta) + C_2\log(t/\delta))) .$$

The proofs of these lemmas are very similar to those for Lemmas 5.8 and 5.9. The only difference is in the proof of Lemma 5.8, which now has the following form: let $M_t = \Theta_* - \widetilde{\Theta}_t$ and

$$g_t = \begin{cases} \Lambda & t \notin \mathcal{T}_T \\ \Psi & t \in \mathcal{T}_T \end{cases}$$

We can write the state update as

$$x_{t+1} = \Gamma_t + r_{t+1} ,$$

where

$$\Gamma_t = \begin{cases} \widetilde{\Theta}_t^\top \varphi(x_t, u_t) & t \notin \mathcal{T}_T \\ \Theta_*^\top \varphi(x_t, u_t) & t \in \mathcal{T}_T \end{cases}$$

and

$$r_{t+1} = \begin{cases} M_t^\top \varphi(x_t, u_t) + w_{t+1} & t \notin \mathcal{T}_T \\ w_{t+1} & t \in \mathcal{T}_T \end{cases}$$

Thus, we can write

$$\begin{aligned}\|x_t\| &\le \|\Gamma_{t-1}\| + \|r_t\| \le g_{t-1}\|x_{t-1}\| + \|r_t\| = g_{t-1}g_{t-2}\|x_{t-2}\| + \|r_t\| + g_{t-1}\|r_{t-1}\| \\ &= g_{t-1}g_{t-2}g_{t-3}\|x_{t-3}\| + \|r_t\| + g_{t-1}\|r_{t-1}\| + g_{t-1}g_{t-2}\|r_{t-2}\| = \cdots = g_{t-1}\dots g_{t-t}\|x_{t-t}\| \\ &+ \|r_t\| + g_{t-1}\|r_{t-1}\| + g_{t-1}g_{t-2}\|r_{t-2}\| + \cdots + g_{t-1}g_{t-2}\dots g_{t-(t-1)}\left\|r_{t-(t-1)}\right\| \\ &= \sum_{k=1}^{t}\left(\prod_{s=k}^{t-1} g_s\right)\|r_k\| \ .\end{aligned}$$

Thus, we have that

$$\prod_{s=k}^{t-1} g_s \le \Psi^{n+d}\Lambda^{t-k-(n+d)} \ .$$

Thus, we have that

$$\begin{aligned}\|x_t\| &\le \left(\frac{\Psi}{\Lambda}\right)^{n+d}\sum_{k=0}^{t-1}\Lambda^{t-k-1}\|r_{k+1}\| \\ &\le \frac{1}{1-\Lambda}\left(\frac{\Psi}{\Lambda}\right)^{n+d}\max_{0\le k\le t-1}\|r_{k+1}\| \ .\end{aligned}$$

The rest of the proof is identical to the proof of Lemma 5.8.

Regret Decomposition

Define $\widetilde{x}_{t+1}^{a_t} = \widetilde{\Theta}_t^\top\varphi(x_t,a_t) + w_{t+1}$. From Assumption A4, provided that the confidence set does not fail at time t, we get that

$$\begin{aligned}J(\widetilde{\Theta}_t) + h_t(x_t) &\ge \min_{a\in D(x_t)}\left\{\ell(x_t,a) + \mathbb{E}\left[h_t(\widetilde{x}_{t+1}^a)\,|\,\mathcal{F}_t\right]\right\} \\ &= \ell(x_t,a_t) + \mathbb{E}\left[h_t(\widetilde{x}_{t+1}^{a_t})\,|\,\mathcal{F}_t\right] \\ &= \ell(x_t,a_t) + \mathbb{E}\left[h_t(\widetilde{\Theta}_t^\top\varphi(x_t,a_t) + w_{t+1})\,|\,\mathcal{F}_t\right] \\ &= \ell(x_t,a_t) + \mathbb{E}\left[h_t((\widetilde{\Theta}_t - \Theta_*)^\top\varphi(x_t,a_t) + \Theta_*^\top\varphi(x_t,a_t) + w_{t+1})\,|\,\mathcal{F}_t\right] \\ &= \ell(x_t,a_t) + \mathbb{E}\left[h_t(\epsilon_t + x_{t+1})\,|\,\mathcal{F}_t\right] \ ,\end{aligned}$$

where $\epsilon_t = (\widetilde{\Theta}_t - \Theta_*)^\top\varphi(x_t,a_t)$. Thus, on $E\cap F$,

$$\begin{aligned}R(T) &= \sum_{t=1}^{T}(\ell(x_t,a_t) - J_*) \\ &\le \sum_{t=1}^{T}(\ell(x_t,a_t) - J(\widetilde{\Theta}_t)) \qquad\qquad \text{(by A5)} \\ &\le \sum_{t=1}^{T}(h_t(x_t) - \mathbb{E}\left[h_t(x_{t+1}+\epsilon_t)\,|\,\mathcal{F}_t\right]) \ .\end{aligned}$$

Thus, we can bound the regret on $E\cap F$,

$$\begin{aligned}R(T) &\le h_1(x_1) - h_{T+1}(x_{T+1}) + \sum_{t=1}^{T}(h_{t+1}(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}+\epsilon_t)\,|\,\mathcal{F}_t\right]) \\ &\le \sum_{t=1}^{T}(h_{t+1}(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}+\epsilon_t)\,|\,\mathcal{F}_t\right]) \ .\end{aligned}$$

The OFU algorithm changes the policy only when the confidence set is halved. (so, $h_{t+1} = h_t$ most of time.) Let A_t denote the event that the algorithm has changed its policy at time t. Then, on $E \cap F$,

$$\begin{aligned}
R(T) &\leq \sum_{t=1}^{T} \left(h_{t+1}(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t\right]\right) \\
&= \sum_{t=1}^{T} (h_{t+1}(x_{t+1}) - h_t(x_{t+1})) + \sum_{t=1}^{T} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t\right]) \\
&\leq 2BX_T \sum_{t=1}^{T} \mathbb{I}_{\{A_t\}} + \sum_{t=1}^{T} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1} + \epsilon_t) \,|\, \mathcal{F}_t\right]) \\
&\leq 2BX_T \sum_{t=1}^{T} \mathbb{I}_{\{A_t\}} + B \sum_{t=1}^{T} \|\epsilon_t\| + \sum_{t=1}^{T} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}) \,|\, \mathcal{F}_t\right]) \,.
\end{aligned}$$

Define

$$R_1 = 2BX_T \sum_{t=1}^{T} \mathbb{I}_{\{A_t\}} \,, \tag{5.23}$$

$$R_2 = B \sum_{t=1}^{T} \|\epsilon_t\| \,, \tag{5.24}$$

$$R_3 = \sum_{t=1}^{T} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}) \,|\, \mathcal{F}_t\right]) \,. \tag{5.25}$$

We bound these terms in a number of lemmas.

Lemma 5.21 On the event $E \cap F$, the OFU algorithm changes its policy at most

$$m \log_2 \left(1 + TK^2 X_T^{2k} / \lambda\right)$$

times up to time T.

Proof. If the algorithm has changed the policy M times up to time T, then we should have that $\det(\overline{V}_T) \geq \lambda^m 2^M$. On the other hand, we have

$$\lambda_{\max}(\overline{V}_T) \leq \lambda + \sum_{t=1}^{T-1} \|\varphi(x_t, a_t)\|^2 \leq \lambda + TK^2 X_T^{2k} \,.$$

Thus, it holds that

$$\lambda^m 2^M \leq (\lambda + TK^2 X_T^{2k})^m \,.$$

As a result, we have

$$M \leq m \log_2 \left(1 + \frac{TK^2 X_T^{2k}}{\lambda}\right) \,.$$

□

This lemma implies that R_1 can be bounded as follows:

Lemma 5.22 Let R_1 be as defined by Equation (5.23). Then we have that

$$\mathbb{I}_{\{E \cap F\}} R_1 \leq 2BX_T m \log_2 \left(1 + TK^2 X_T^{2k} / \lambda\right) \,.$$

Next, we bound R_2 by an argument identical to the one used to prove Lemma 5.15.

Lemma 5.23 Assume that $E \cap F$ holds. Then we have that

$$\sum_{t=1}^{T} \left\| (\Theta_* - \widetilde{\Theta}_t)^\top \varphi(x_t, a_t) \right\|^2 \leq \frac{8}{\lambda} K^2 X_T^{2k} \beta_T(\delta/4) \log\det(\overline{V}_T) \ .$$

Lemma 5.24 Let R_2 be as defined by Equation (5.24). Then we have that

$$\mathbb{I}_{\{E \cap F\}} R_2 \leq B \sqrt{\frac{8}{\lambda} T K^2 X_T^{2k} \beta_T(\delta/4) \log\det(\overline{V}_T)} \ .$$

Similarly, we bound R_3 by an argument similar to the one used to prove Lemma 5.16.

Lemma 5.25 Let R_3 be as defined by Equation (5.25). Then with probability at least $1 - \delta/2$,

$$\mathbb{I}_{\{E \cap F\}} R_3 \leq H \sqrt{8T \log 4/\delta} \, ,$$

where $H = 2B(\Psi + W)$ and $W = Ln\sqrt{2n \log(4nT/\delta)}$.

Proof. Note that $E_{t+1} \subset E_t$ and $F_{t+1} \subset F_t$, and so $\mathbb{I}_{\{E_{t+1} \cap F_{t+1}\}} \leq \mathbb{I}_{\{E_t \cap F_t\}}$, and in particular, since $E = E_T$, $F = F_T$, $\mathbb{I}_{\{E \cap F\}} \leq \mathbb{I}_{\{E_t \cap F_s\}}$ holds for any $t, s \leq T$. We have that

$$\mathbb{I}_{\{E \cap F\}} R_3 \leq \sum_{t=1}^{T} \mathbb{I}_{\{E_{t+1} \cap F_t\}} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}) | \mathcal{F}_t\right]).$$

Define

$$D_t = \mathbb{I}_{\{E_{t+1} \cap F_t\}} (h_t(x_{t+1}) - \mathbb{E}\left[h_t(x_{t+1}) | \mathcal{F}_t\right])$$

and its truncated version $D_t^c = D_t \wedge H$. Define the supermartingale

$$M_\tau = \sum_{t=1}^{\tau} D_t^c, \ M_0 = 0 \ .$$

This is a supermartingale, since E_{t+1} and F_t are $\mathcal{F}_t$ measurable. Let A be the event that for $t \leq T$, $w_t \leq W$. If A holds then,

$$\begin{aligned} D_t &= \mathbb{I}_{\{E_{t+1} \cap F_t\}} \Big(h_t\left(\Theta_*^\top \varphi(x_t, a_t) + w_{t+1}\right) - \mathbb{E}\left[h_t\left(\Theta_*^\top \varphi(x_t, a_t) + w_{t+1}\right) | \mathcal{F}_t \right] \Big) \\ &\leq 2B \left\| \Theta_*^\top \varphi(x_t, a_t) \right\| + B \Big(\|w_{t+1}\| + \mathbb{E}\left[\|w_{t+1}\| \, | \mathcal{F}_t \right] \Big) \\ &\leq 2B(\Psi + W) = H \ . \end{aligned}$$

Thus, under A, $D_t = D_t^c$. Then, by Azuma's inequality, we get that

$$\mathbb{P}\left(M_T > H\sqrt{2T \log(4/\delta)} \right) \leq \delta/4 \ .$$

Thus,

$$\mathbb{P}\left(A, \sum_{t=1}^{T} D_t > H\sqrt{2T \log(4/\delta)} \right) \leq \delta/4 \ .$$

Thus

$$\mathbb{P}\left(\sum_{t=1}^{T} D_t \leq H\sqrt{2T \log(4/\delta)} \right) \geq 1 - \delta/4 - \mathbb{P}\left(A^c\right) \geq 1 - \delta/2 \ .$$

Then, by Lemma 5.10, we get that with probability at least $1 - \delta/2$,

$$\mathbb{I}_{\{E \cap F\}} R_3 \leq H\sqrt{8T \log 4/\delta} \ .$$

□

Gradient
Input: Ellipsoid covariance matrix V, ellipsoid center $\widehat{\Theta}$, step-size α.
$\Theta_0 = \widehat{\Theta}$.
for $i := 0, 1, \ldots, C$ **do**
  Compute $\nabla_\Theta \operatorname{trace}(P(\Theta))$ by (5.26),(5.27).
  $\Theta_{i+1} = \Theta_i - \alpha \nabla_\Theta \operatorname{trace}(P(\Theta))$.
  $\Theta_{i+1} = \text{Project}(\Theta_{i+1}, V, I)$.
end for
Return Θ_C.

Figure 5.3: The projected gradient descent method for solving the OFU optimization.

Putting Everything Together and Proving Theorem 5.17

Proof of Theorem 5.17. The proof follows in a straightforward manner from Lemmas 5.22, 5.24, and 5.25. □

5.4 Computational Issues and Experiments

In this section, we derive two incremental gradient descent methods to find an approximate solution to problem (5.8). We study experimentally their behavior on a simple example. Finally, we study the behavior of the OFULQ algorithm that uses these optimization methods on a simple idealized web server control problem.

5.4.1 Incremental Methods for Finding an Optimistic Parameter

Recall that $J(\Theta) = \operatorname{trace}(P(\Theta))$, where $P(\Theta)$ is the Riccati solution. At round t, we solve the optimization problem

$$\inf_{\Theta \in C_t(\delta) \cap \mathcal{S}} \operatorname{trace}(P(\Theta))$$

approximately to find an optimistic estimate $\widetilde{\Theta}_t$. Perhaps the simplest approach is to use an iterative method such as the projected gradient descent method,

$$\widetilde{\Theta}_t \leftarrow \Pi_{C_t(\delta)} \left(\widetilde{\Theta}_t - \alpha \nabla_\Theta \operatorname{trace}(P(\Theta)) \right) ,$$

where $\nabla_\Theta f$ is the gradient of f with respect to Θ, $\alpha > 0$ is a step-size, $C_t(\delta)$ is the confidence ellipsoid at time t, and Π_R is the Euclidean projection on R. The resulting algorithm is shown in Figure 5.3.

Alternatively, we can use Newton's method with iterations

$$\widetilde{\Theta}_t \leftarrow \Pi_{C_t(\delta)} \left(\widetilde{\Theta}_t - \alpha H_\Theta(\operatorname{trace}(P(\Theta)))^{-1} \nabla_\Theta \operatorname{trace}(P(\Theta)) \right) ,$$

where $H_\Theta(f)$ is the Hessian of f with respect to Θ. Next we derive the gradient and the Hessian of $\operatorname{trace}(P(\Theta))$ and the projection rules. In general, Newton's method is expected to converge with a fewer iterations but at the price of a higher per-iteration computational effort. Whether a gradient method, or Newton's method should be used in a specific application is expected to depends on the specific of the problem.

Notice that the above optimization algorithms project onto $C_t(\delta)$ instead of $C_t(\delta) \cap \mathcal{S}$. This is done for ease of implementation. Later, we will show experimentally that omitting $\mathcal{S}$ does not cause major problems, at least in the problem we tested the algorithm on. In the next two subsections we show how the gradient and Hessian of the objective function J can be computed. This is followed by the description of how the projection step can be implemented.

Gradient Computation

To simplify the presentation, we use P to denote $P(\Theta)$. Let $[M_{jk}]_{j=1\ldots r,\, k=1\ldots c}$ be a $r \times c$ matrix whose (j,k)th element is M_{jk}. If the dimensionality of the matrix can be understood from the context, we omit r and c and just write $[M_{jk}]_{j,k}$.

The derivation of the gradient and the Hessian of trace(P) in general multidimensional problems is as follows. We have that

$$\nabla_A \operatorname{trace}(P) = \sum_{i=1}^{n} \nabla_A P_{ii} = \sum_{i=1}^{n} \left[\frac{\partial P_{ii}}{\partial A_{jk}}\right]_{j,k}, \tag{5.26}$$

$$\nabla_B \operatorname{trace}(P) = \sum_{i=1}^{n} \nabla_B P_{ii} = \sum_{i=1}^{n} \left[\frac{\partial P_{ii}}{\partial B_{jk}}\right]_{j,k}. \tag{5.27}$$

Define

$$G = Q + A^\top PA - A^\top PB(B^\top PB + R)^{-1}B^\top PA - P$$

and

$$g = \operatorname{trace}(G).$$

If A, B, P satisfy the Riccati equation, then $g = 0$. By the Implicit Function Theorem (Theorem A.2 in Appendix A), we get that for any $1 \le i, j \le n\,, 1 \le k \le d$,

$$\frac{\partial P_{ii}}{\partial A_{jk}} = -\frac{\partial g/\partial A_{jk}}{\partial g/\partial P_{ii}}, \tag{5.28}$$

$$\frac{\partial P_{ii}}{\partial B_{jk}} = -\frac{\partial g/\partial B_{jk}}{\partial g/\partial P_{ii}}. \tag{5.29}$$

Thus, we only need to compute

$$\frac{\partial g}{\partial A_{jk}} = \operatorname{trace}\left(\frac{\partial G}{\partial A_{jk}}\right), \tag{5.30}$$

$$\frac{\partial g}{\partial B_{jk}} = \operatorname{trace}\left(\frac{\partial G}{\partial B_{jk}}\right), \tag{5.31}$$

$$\frac{\partial g}{\partial P_{ii}} = \operatorname{trace}\left(\frac{\partial G}{\partial P_{ii}}\right). \tag{5.32}$$

It can be shown that

$$\begin{aligned}
\frac{\partial G}{\partial A_{jk}} &= A^\top C 1_{jk} + 1_{kj} CA\,,\\
\frac{\partial G}{\partial B_{jk}} &= -H 1_{kj} PA + H(B^\top P 1_{jk} + 1_{kj} PB)H^\top - A^\top P 1_{jk} H^\top\,,\\
\frac{\partial G}{\partial P_{ii}} &= A^\top 1_{ii} A - 1_{ii} - HB^\top 1_{ii} A - A^\top 1_{ii} BH^\top + HB^\top 1_{ii} BH^\top\,,
\end{aligned}$$

where $H = A^\top PB(B^\top PB + R)^{-1}$. Then, given (5.28)–(5.32), we can compute derivatives via the help of (5.26) and (5.27).

Hessian Computation

Next, we show how the Hessian can be computed. The second-order derivatives of trace$(P(\Theta))$ can be obtained from (5.28) and (5.29), which, in turn, requires the second-order derivatives of g with respect to A, B, P. These derivatives can be obtained from the following equations:

$$\begin{aligned}
\frac{\partial^2 G}{\partial A_{jk}\partial B_{j'k'}} &= PB(B^\top PB + R)^{-1} 1_{k'j'} P - P 1_{j'k'} (B^\top PB + R)^{-1} B^\top P\\
&\quad + PB(B^\top PB + R)^{-1}(B^\top P 1_{j'k'} + 1_{k'j'} PB)(B^\top PB + R)^{-1} B^\top P\,,
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 G}{\partial P_{ii}\partial B_{j'k'}} &= -\frac{\partial H}{\partial P_{ii}} 1_{k'j'} + \frac{\partial H}{\partial P_{ii}}(B^\top P 1_{j'k'} + 1_{k'j'} P B) H^\top \\
&\quad + H(B^\top 1_{ii} 1_{j'k'} + 1_{k'j'} 1_{ii} B) H^\top + H(B^\top P 1_{j'k'} + 1_{k'j'} P B)\frac{\partial H^\top}{\partial P_{ii}}\,, \\
\frac{\partial^2 G}{\partial P_{ii}\partial A_{j'k'}} &= 1_{kj} 1_{ii} A + A^\top 1_{ii} 1_{jk} - H B^\top 1_{ii} 1_{jk} - 1_{kj} 1_{ii} B H^\top \\
&\quad - 1_{kj} P B (B^\top P B + R)^{-1} B^\top 1_{ii} A - A^\top 1_{ii} B (B^\top P B + R)^{-1} B^\top P 1_{jk} \\
&\quad + 1_{kj} P B (B^\top P B + R)^{-1} B^\top 1_{ii} B H^\top + H B^\top 1_{ii} B (B^\top P B + R)^{-1} B^\top P 1_{jk}\,, \\
\frac{\partial^2 G}{\partial A_{jk}\partial A_{j'k'}} &= 1_{kj} C 1_{j'k'} + 1_{j'k'} C 1_{jk}\,, \\
\frac{\partial^2 G}{\partial B_{jk}\partial B_{j'k'}} &= -\frac{\partial H}{\partial B_{j'k'}} 1_{kj} P A - A^\top P 1_{jk}\frac{\partial H^\top}{\partial B_{j'k'}} + \frac{\partial H}{\partial B_{j'k'}}(B^\top P 1_{jk} + 1_{kj} P B) H^\top \\
&\quad + H(B^\top P 1_{jk} + 1_{kj} P B)\frac{\partial H^\top}{\partial B_{j'k'}} + H(1_{k'j'}^\top P 1_{jk} + 1_{kj} P 1_{j'k'}) H^\top\,, \\
\frac{\partial H}{\partial B_{jk}} &= A^\top P 1_{jk} (B^\top P B + R)^{-1} \\
&\quad - A^\top P B (B^\top P B + R)^{-1}(1_{kj} P B + B^\top P 1_{jk})(B^\top P B + R)^{-1}\,, \\
\frac{\partial H}{\partial P_{ii}} &= A^\top 1_{ii} B (B^\top P B + R)^{-1} \\
&\quad - A^\top P B (B^\top P B + R)^{-1}(B^\top 1_{ii} B + R)(B^\top P B + R)^{-1}\,.
\end{aligned}
$$

From these, the Hessian can be computed using simple algebra.

Projection Step

Because the confidence set (the optimization space) is an ellipsoid, we discuss how to project a point on an ellipsoid. The derivation of how this can be done can be found in, e.g., the work of Kiseliov (1994) and is essentially an application of Lagrange multipliers and Newton's method. Consider the ellipsoid

$$
U = \{u : \mathbb{R}^k \;:\; \mathrm{trace}(u^\top M u) \le 1\}\,.
$$

Imagine (abstractly) that the goal is to project $u_0 \in \mathbb{R}^k$ on U according to a weighted Euclidean norm, i.e., the problem is to compute

$$
\widehat{u} = \operatorname*{argmin}_{u \in U} (u - u_0)^\top N (u - u_0)\,.
$$

Define the Lagrangian

$$
L(u, \mu) = (u - u_0)^\top N (u - u_0) + \mu(\mathrm{trace}(u^\top M u) - 1)\,.
$$

From $\partial L/\partial u = 0$ we get the projection

$$
\widehat{u}(\mu) = (N + \mu M)^{-1} N u_0\,. \tag{5.33}
$$

From $\partial L/\partial \mu = 0$ we get that

$$
\mathrm{trace}(u(\mu)^\top M u(\mu)) - 1 = 0\,. \tag{5.34}
$$

We solve for μ and then obtain the projection from (5.33). Define

$$
G(\mu) = \mathrm{trace}(\widehat{u}(\mu)^\top M \widehat{u}(\mu)) - 1\,. \tag{5.35}
$$

Project
Input: Point u_0, ellipsoid covariance matrix M, projection weight matrix N.
$\mu_0 = 0$.
for $i := 0, 1, \ldots, C$ **do**
  Compute $G(\mu_i)$ and $G'(\mu_i)$ by (5.35),(5.36).
  $\mu_{i+1} = \mu_i - \frac{G(\mu_i)}{G'(\mu_i)}$.
end for
Return $(N + \mu_{C+1}M)^{-1}Nu_0$.

Figure 5.4: Projection of a point on an ellipsoid.

We solve for the zero of $G(\mu) = 0$ iteratively using Newton's method:

$$\mu_{s+1} = \mu_s - \frac{G(\mu_s)}{G'(\mu_s)} \, .$$

The derivative of $G(\mu)$ can be computed as follows:

$$G'(\mu) = \operatorname{trace}\left(\frac{\partial}{\partial \mu}\widehat{u}(\mu)^\top M \widehat{u}(\mu)\right) = 2\operatorname{trace}\left(\frac{\partial \widehat{u}(\mu)^\top}{\partial \mu} M \widehat{u}(\mu)\right) , \tag{5.36}$$

$$\frac{\partial \widehat{u}(\mu)}{\partial \mu} = \left\{\frac{\partial}{\partial \mu}(N + \mu M)^{-1}\right\} Nu_0 = -(N + \mu M)^{-1} M (N + \mu M)^{-1} Nu_0 \, .$$

Putting together the pieces, we arrive at the iterative projection algorithm whose pseudocode is shown in Figure 5.4.

5.4.2 Finding the Optimistic Parameters: Numerical Illustration

Figure 5.5 shows the objective function $J(\Theta) = \operatorname{trace}(P(\Theta))$ when the state and action spaces are 1-dimensional ($n, d = 1$). Based on this figure, we conjecture that $\operatorname{trace}(P(\Theta))$ is convex in A. In fact, the hypothesis can be shown to be true in the 1-dimensional case. However, it is not known if the hypothesis is true in higher dimensions. From Figure 5.5, it also appears that if the confidence set is away from $B = 0$, the gradient descent method might be effective in solving the OFU optimization problem. That the critical line is $B = 0$ is not a coincidence: Knowing the "sign" of the control clearly plays a crucial role. The practical consequence of this for multidimensional systems is that the gradient method will only be reliable when "signs" in the some kind of "eigenstructure" of matrix B is identified with high probability. Determining the exact condition remains for future work. Nevertheless, we speculate that in this initial phase, whose length should probably be determining using a stopping rule, it is better to use randomized controls.

Let us now illustrate the behavior of Newton's method, the earlier described projected gradient descent method, and a simple discretization method for minimizing $\operatorname{trace}(P(\Theta))$ over a fixed confidence set when $n, d = 1$. Figure 5.6-(a) shows that Newton's method finds near optimal solutions in just two steps. The behavior of the other two methods are shown in Figure 5.6-(b,c). Figure 5.7 shows the fixed confidence ellipsoid and a sample trajectory of the gradient procedure that converges to the minimum of $\operatorname{trace}(P(\Theta))$ over the confidence set. The step-size of the gradient method is chosen to be $1/\sqrt{\lambda_1 + \lambda_2}$, where λ_1 and λ_2 are the eigenvalues of the covariance matrix underlying the ellipsoid. We observed that the gradient method with this step-size performs well on many problems.

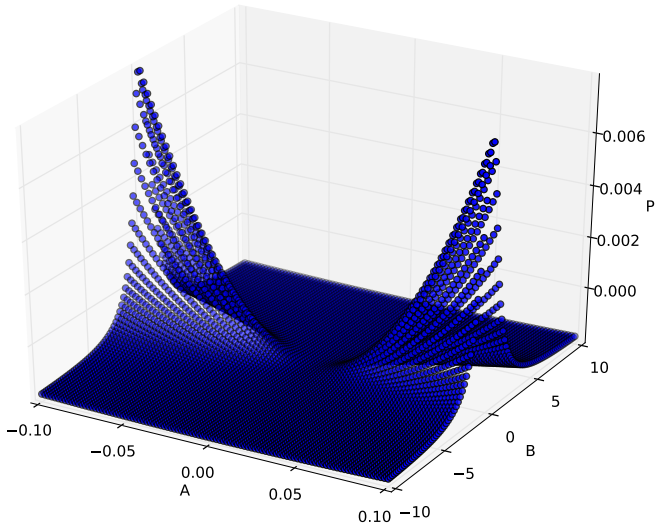
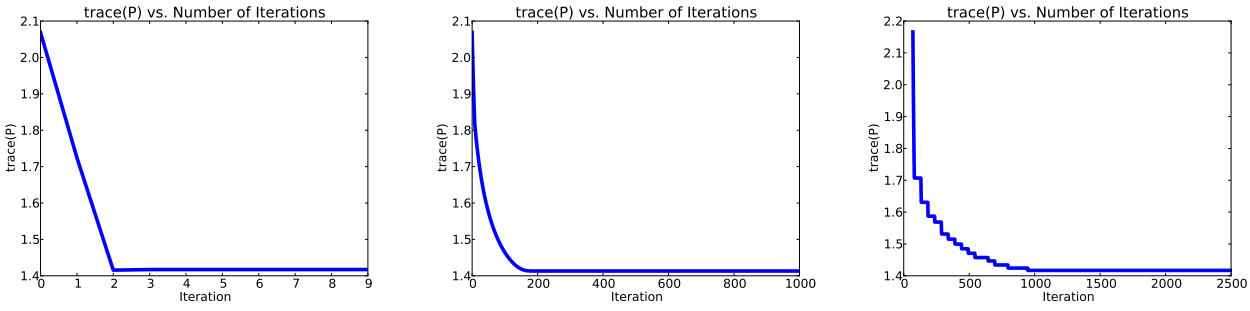


Figure 5.5: Values of the objective function $J(\Theta) = \text{trace}(P(\Theta))$ as a function of $\Theta = (A, B)$, where $A, B \in \mathbb{R}$. Here, $P(\Theta)$ is the solution of the Riccati Equation (5.3).



(a) (b) (c)

Figure 5.6: (a) Newton's method. (b) Gradient descent method. (c) Uniform discretization.

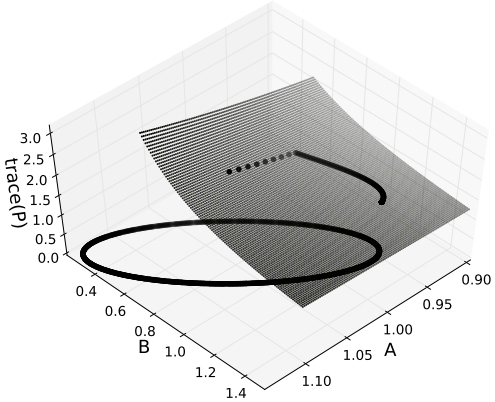


Figure 5.7: A sample trajectory of the projected gradient descent method.

5.4.3 Illustration of OFULQ

The purpose of this section is to illustrate the behavior of OFULQ on a simple control problem. As the control problem, we choose a web server control problem. This control problem is described first, which will be followed by the description of our results.

Web Server Control Application

Next, we illustrate the behavior of OFULQ on a web server control problem. The problem is taken from Section 7.8.1 of the book by Hellerstein et al. (2004) (this example is also used in Section 3.4 of the book by Aström and Murray (2008)). An Apache HTTP web server processes the incoming connections that arrive on a queue. Each connection is assigned to an available process. A process drops the connection if no requests have been received in the last KeepAlive seconds. At any given time, there are at most MaxClients active processes. The values of the KeepAlive and MaxClients parameters, denoted by a_{ka} and a_{mc} respectively, are chosen by a control algorithm. Increasing a_{mc} and a_{ka} results in faster and longer services to the connections, but also increases the CPU and memory usage of the server. MaxClients is bounded in $[1, 20]$, while KeepAlive is bounded in $[1, 1024]$. The state of the server is determined by the average processor load $x_{cpu} \in [0, 1]$ and the relative memory usage $x_{mem} \in [0, 1]$. A *operating point of interest* of the system is given by

$$x_{cpu} = 0.58\,, \qquad a_{ka} = 11s\,, \qquad x_{mem} = 0.55\,, \qquad a_{mc} = 600\,.$$

A linear model around the operating point is assumed, resulting in a model of the form

$$\begin{pmatrix} x_{cpu}(t+1) \\ x_{mem}(t+1) \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{21} \end{pmatrix} \begin{pmatrix} x_{cpu}(t) \\ x_{mem}(t) \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{21} \end{pmatrix} \begin{pmatrix} a_{ka}(t) \\ a_{mc}(t) \end{pmatrix} + \begin{pmatrix} w_1(t+1) \\ w_2(t+1) \end{pmatrix}\,,$$

where $(w_1(t+1), w_2(t+1))_t$ is an i.i.d. sequence of Gaussian random variables, with a diagonal covariance matrix. Note that these state and action variables are in fact the deviations from the operating point. Hellerstein et al. (2004) fitted this model to an Apache HTTP server and obtained the parameters

$$A = \begin{pmatrix} 0.54 & -0.11 \\ -0.026 & 0.63 \end{pmatrix}, \qquad B = \begin{pmatrix} -85 & 4.4 \\ -2.5 & 2.8 \end{pmatrix} \times 10^{-4}\,,$$

while the noise standard deviation was measured to be 0.1. Hellerstein et al. (2004) found that these parameters provided a reasonable fit to their data.

For control purpose, the following cost matrices were chosen (cf. Example 6.9 of Aström and Murray (2008)):

$$Q = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}, \qquad R = \begin{pmatrix} 1/50^2 & 0 \\ 0 & 0.1^6 \end{pmatrix}\,.$$

Numerical Results

We compare a forced-exploration method, a Q-learning method, and the OFULQ algorithm on this problem. We run these experiments in NumPy on a dual-core Linux machine with a 2.16 gigahertz CPU and 3 gigabytes of RAM. The time horizon in these experiments is $T = 10,000$. We repeat each experiment 50 times and report the mean and the standard deviation of the observations. In these experiments, a single run takes, on average, 4.83 seconds with the forced-exploration method and 11.54 seconds with the OFULQ algorithm.

The forced-exploration algorithm takes $\sqrt{T} = 100$ random exploratory actions at the beginning, and then takes the greedy action with respect to the least-squares estimate for the rest of the episode. During the exploration phase, we sample the random action according to $a_{ka} \sim U(1, 100)$ and $a_{mc} \sim U(1, 10)$, where $U(a, b)$ is the uniform distribution over $[a, b]$. The top-left subfigure of Figure 5.8 shows the regret of the forced-exploration method.

Bradtke et al. (1994) prove convergence of a Q-learning method on discounted deterministic LQ problems. The algorithm maintains a current policy $K_t \in \mathbb{R}^{d\times n}$, which is a linear mapping from state to action. For every N_e steps of policy estimation, one step of policy improvement is performed. At each round, the output of the current policy is perturbed by a noise vector to obtain the action, $a_t = K_t x_t + u_t$. We choose $N_e = 30$ and discount factor $\gamma = 0.9$. Each element of the noise vector is drawn from $U(0, 1)$. The top-right subfigure of Figure 5.8 shows the regret of the Q-learning method, which looks almost linear. We also experimented with an R-learning method (Sutton and Barto, 1998), which is the average loss version of the Q-learning algorithm. However, the value function estimate diverges in our experiments.

The inputs to the OFULQ algorithm are $\delta = 1/T, \lambda = 1, R = 0.1, S = 1$. The gradient module takes 50 steps to solve each OFU optimization problem. The learning rate is

$$\alpha = \sqrt{\frac{0.001}{\mathrm{trace}(\overline{V}_t/\beta_t(\delta))}}\ .$$

The reason for dividing by the trace of $\overline{V}_t$ is that if some eigenvalue of $\overline{V}_t$ is big then in some direction the ellipsoid will have a small diameter. Hence, the inverse of the trace is an indicator of how big the confidence set is, note we want to take smaller steps when the ellipsoid is smaller. Similarly, $\beta_t(\delta)$, is the radius of the ellipsoid and then it makes sense to increase the step-size as a function of $\beta_t(\delta)$. The square root is needed to match the dimensions of the step-size to that of these two other quantities. The constant 0.001 is chosen in an ad hoc fashion (based on prior experience with other problems).

In the projected gradient method, we apply Newton's update for computing the projection of the unconstrained parameter at most $1,000$ times or until the projected point lies inside the confidence ellipsoid (on average, the Newton update is used only for about 15 steps, which takes about 0.0028 seconds). Solving the OFU problem requires 0.286 seconds on average. Empirically, we have found that the optimistic estimates always lied in the set $\mathcal{S}$. We also note that the algorithm always makes less than 30 switches.

The bottom-left subfigure of Figure 5.8 shows the regret of the OFULQ algorithm, which is slightly worse than what we get for the forced-exploration method. We explain this observation by noting that, in this problem, we need large inputs to reliably estimate matrix B, which has small elements. The forced-exploration method takes large random actions, whose scale is manually set by the programmer, while the OFULQ algorithm has no way to find the right scaling at the beginning. However, it can also be seen from the figure, that the rate at which the regret of OFULQ increases starts to decrease, while the forced-exploration method, once exploration is turned off, does not adapt and hence the rate of increase of its regret stays constant (illustrating the benefit of algorithms that adapt online).

To make the comparison with the forced exploration method fair(er), we add an initial exploration period to the OFULQ algorithm. A reasonable-looking suggestion is that this initial phase should last $(n + d) \times 10$ time steps; in other words, by the end, we expect to see 10 samples for each dimension.

The bottom-right subfigure of Figure 5.8 shows the regret of this algorithm. Now, OFULQ clearly outperforms the forced-exploration method – both in terms of its mean regret and variance. It is also interesting to know that this new algorithm makes about 15 switches. For illustration purposes, average trajectories of the state and action vectors are shown in Figure 5.9 for this last version of OFULQ. The evolution of the least-squares estimates for the elements of the parameter matrix for the same case is shown in Figures 5.10 and 5.11.

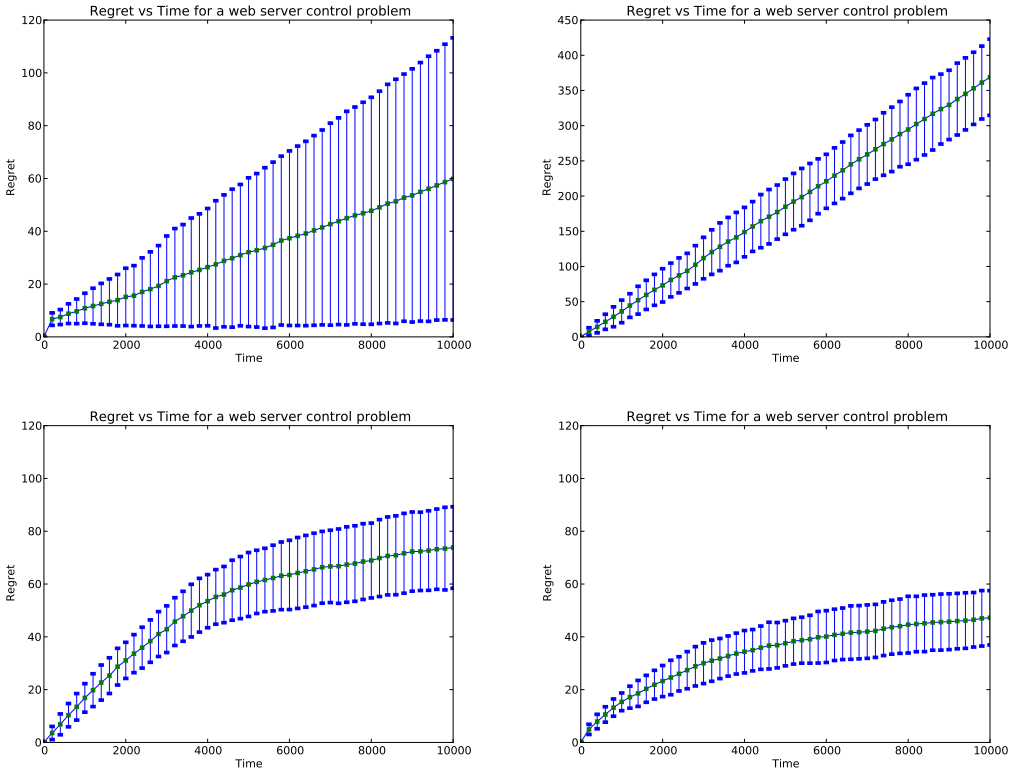


Figure 5.8: Regret vs time for a web server control problem. (Top-left): regret of the forced-exploration method. (Top-right): regret of a Q-learning method. (Bottom-left) regret of the OFULQ algorithm. (Bottom-right): regret of the OFULQ algorithm with the initial exploration.

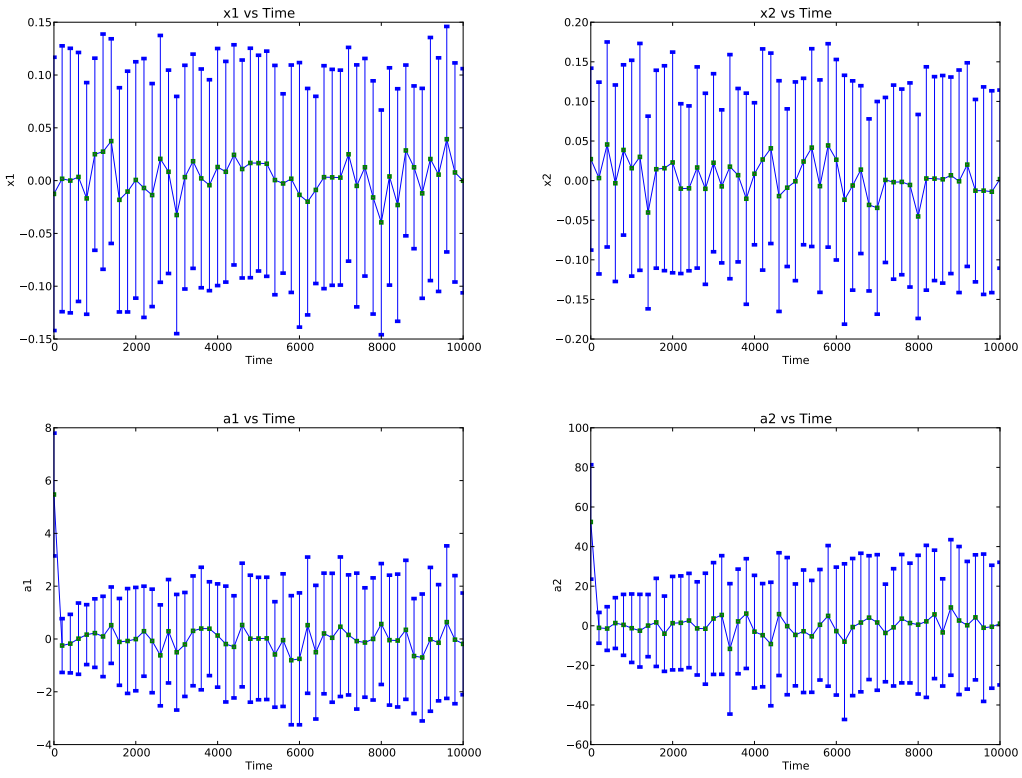


Figure 5.9: The trajectory of the state and action vectors. (Top left): x_{cpu} vs. time. (Top right): x_{mem} vs. time. (Bottom left): a_{ka} vs. time. (Bottom right): a_{mc} vs. time.

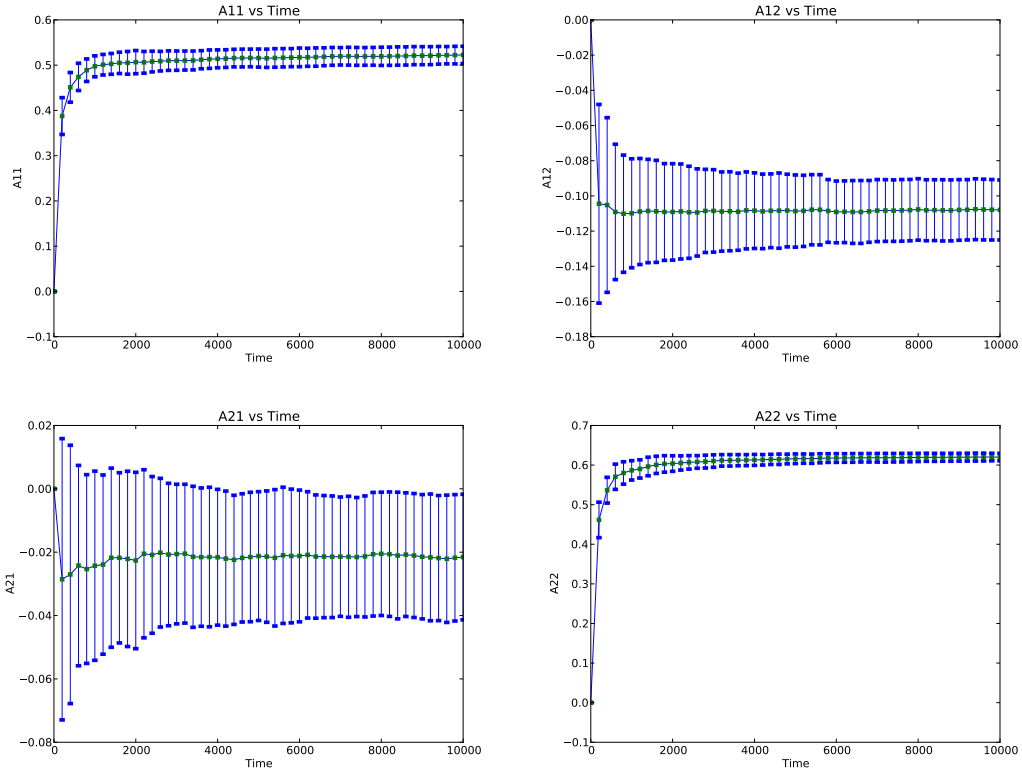


Figure 5.10: The least-squares estimate for matrix A vs time.

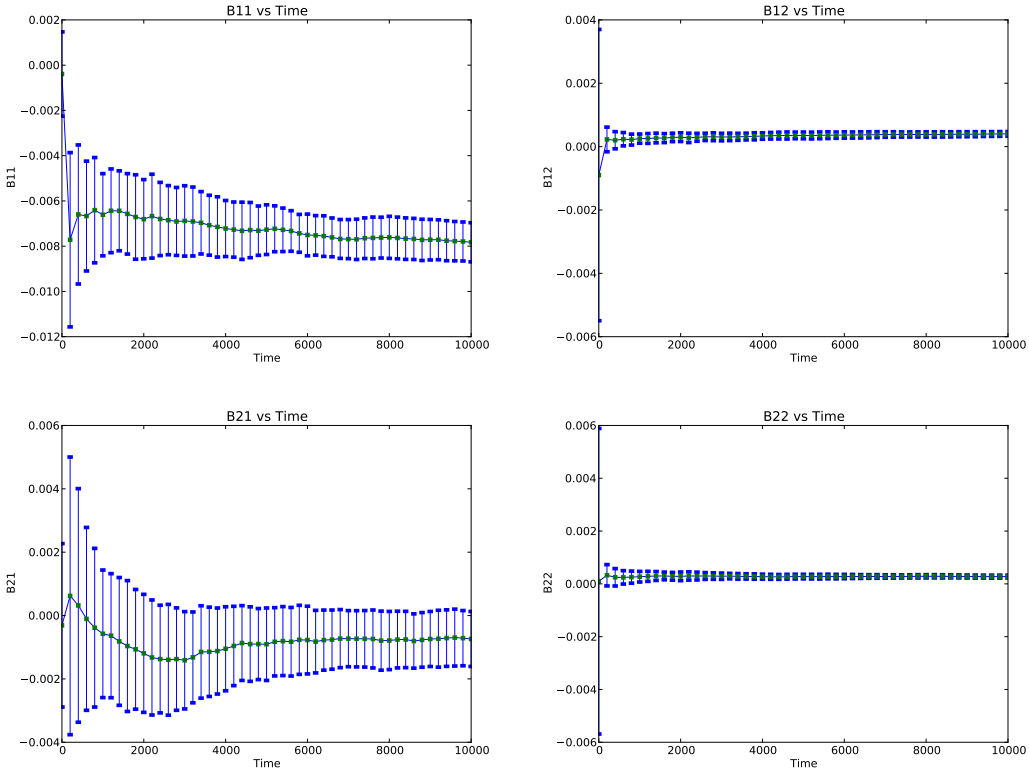


Figure 5.11: Least-squares estimate for matrix B vs time.

Chapter 6

Conclusions

We employed the "optimism in the face of uncertainty" principle to design efficient algorithms for a number of important decision making problems, such as the linear stochastic bandit and the linear quadratic control problems. The main requirement to the application of the OFU principle is that a pointwise error bound for the online prediction is available. Tighter error bounds often translate into better online performance.

The first contribution of this thesis was to derive new tighter error bounds for the online least-squares method. Our approach was to relate the estimation error to a self-normalized quantity for vector-valued martingales. We also showed that the predictions of any online algorithm with quadratic losses can be combined in a certain fashion to give a confidence set whose size scales only with the average regret of the online algorithm. This general result allowed us to construct confidence sets smaller than what was previously available when the unknown parameter vector is assumed to be sparse.

Equipped with tight confidence sets, we obtained linear bandit algorithms that achieve regret bounds with both improved theoretical and empirical performance. In particular, the effectiveness of the proposed methods was demonstrated on the Yahoo! article recommendation dataset, whereas we compared algorithms tuned for the sparse and nonsparse situations on synthetic data. These experiments fully confirmed the theoretical predictions: the algorithm specialized to the sparse setting performs favourably in comparison to other linear bandit methods when the parameter vector is sparse. We believe further progress can be made on this front by deriving confidence sets from tighter data-driven regret bounds. Obtaining such regret bounds remain an open problem.

The second main topic of the thesis is the application of the OFU principle to control problems. We first studied the basic linear quadratic control problem that plays a fundamental role in the control literature. We designed an algorithm and proved the first finite-time regret bound for this algorithm. We also proposed a gradient-based method to approximately solve the OFU optimization problem; thereby making it possible to implement the algorithm in practical way for high-dimensional problems. Our experiments show that the proposed gradient-based technique indeed achieves a sublinear regret. These results are encouraging as they show that the OFU principle can be successfully applied to a class of control problems with continuous state and action spaces.

Finally, we showed that a similar technique can be employed to design and analyze algorithms for more general control problems with non-linear transition laws, but linear uncertainty. For such problems, we also obtained sublinear regret bounds. However, at the moment, these results remain of purely theoretical interest only, as there are no known efficient ways to implement the required OFU optimization problem.

Bibliography

Yasin Abbasi-Yadkori. Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, Department of Computing Science, University of Alberta, 2009.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the Twenty Fifth Annual Conference on Neural Information Processing Systems*, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011b.

Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37:263–293, 2003.

Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21th Annual Conference on Learning Theory*, 2008.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

A. Al-Tamimi, D. Vrabie, M. Abu-Khalaf, and F.L. Lewis. Model-free approximate dynamic programming schemes for linear systems. In *International Joint Conference on Neural Networks*, 2007.

Brian D. O. Anderson and John B. Moore. *Linear Optimal Control*. Prentice-Hall, 1971.

András Antos and Csaba Szepesvári. Lower bounds for linear stochastic bandits. Personal Communication, 2009.

András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

Aristotle Arapostathis, Vivek S. Borkar, Emmanuel Fernández-Gaucherand, Mrinal K. Ghosh, and Steven I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, 1993.

Karl J. Aström and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.

Karl Johan Aström and Björn Wittenmark. On self tuning regulators. *Automatica*, 9(2): 185–199, 1973.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19): 1876–1902, 2009.

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

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

Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002b.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003.

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *Proceedings of the Twentieth Annual Conference on Learning Theory*, 2007.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Conference on Learning Theory*, 1994.

Gábor Bartók. *The role of information in online learning.* PhD thesis, Department of Computing Science, University of Alberta, 2012.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

Arthur Becker, P. R. Kumar, and Ching-Zong Wei. Adaptive control with the stochastic approximation algorithm: Geometry and convergence. *IEEE Trans. on Automatic Control*, 30(4):330–338, 1985.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 2nd edition, 2001.

Dimitri P. Bertsekas and John Tsitsiklis. *Neuro-Dynamic Programming.* Athena scientific optimization and computation series. Athena Scientific, 1996.

Allan Birnbaum. On the foundations of statistical inference: Binary experiments. *The Annals of Mathematical Statistics*, 32(2):414–435, 1961.

Sergio Bittanti and Marco C. Campi. Adaptive control of linear time invariant systems: the "Bet On the Best" principle. *Communications in Information and Systems*, 6(4):299–320, 2006.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Vivek S. Borkar. Sample complexity for markov chain self-tuner. *Systems & Control Letters*, 41(2):95–104, 2000.

S. J. Bradtke, B. E. Ydstie, and A. G. Barto. Adaptive linear quadratic control using policy iteration. Technical report, 1994.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Sébastien Bubeck and Jean-Yves Audibert. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836, 2010.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in X-armed bandits. In *Proceedings of the Twenty Second Annual Conference on Neural Information Processing Systems*, 2008.

Sébastien Bubeck, Nicolò Cesa-Bianchi, and Sham M. Kakade. Towards minimax policies for online linear optimization with bandit feedback. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Marco C. Campi. Achieving optimality in adaptiv control: the "Bet On the Best" approach. In *Proceedings of the 36th IEEE Conference on Decision and Control*, 1997.

Marco C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Emmanuel J. Candès. Compressive sampling. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 1433–1452, 2006.

Alexandra Carpentier and Rémi Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Proceedings of Fifteenth International Conference on Artificial Intelligence and Statistics*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Proceedings of the Twenty Fifth Annual Conference on Neural Information Processing Systems*, 2011.

Han-Fu Chen and Lei Guo. Optimal adaptive control and consistent parameter estimates for ARMAX model with quadratic cost. *SIAM Journal on Control and Optimization*, 25 (4):845–867, 1987.

Han-Fu Chen and Ji-Feng Zhang. Identification and adaptive control for systems with unknown orders, delay, and coefficients. *Automatic Control, IEEE Transactions on*, 35 (8):866–877, August 1990.

Ku-Chun Chou and Hsuan-Tien Lin. Balancing between estimated reward and uncertainty during news article recommendation for ICML 2012 exploration and exploitation challenge. presented at the ICML 2012 Workshop on Exploration and Exploitation Challenge, 2012.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

W. J. Conover and Ronald L. Iman. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3):124–129, 1981.

Pierre-Arnaud Coquelin and Rémi Munos. Bandit algorithms for tree search. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

Arnak S. Dalalyan and Alexandre B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 83—92, 2009.

Varsha Dani and Thomas P. Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2006.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, pages 355–366, 2008.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, 32(3):1902–1933, 2004.

Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

Ofer Dekel and Yoram Singer. Data-driven online to batch conversions. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, 2006.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Robust selective sampling from single and multiple teachers. In *Proceedings of the 23rd Conference on Learning Theory*, 2010.

Joseph L. Doob. *Stochastic Processes*. John Wiley and Sons, 1953.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Learning Theory*, 2002.

Claude-Nicolas Fiechter. PAC adaptive control of linear systems. In *Proceedings of the 10th Annual Conference on Computational Learning Theory*, 1997.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the Twenty Fourth Annual Conference on Neural Information Processing Systems*, 2010.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. Technical report, LTCI, 2008.

Claudio Gentile and Nicolas Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the Twelfth Annual Conference on Learning Theory*, 1999.

Sèbastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

William S. Gosset (Student). On the probable error of a mean. *Biometrika*, 6:1–25, 1908.

Philip S. Griffin. Tightness of the Student t-statistic. *Electronic Communications in Probability*, 7:181–190, 2002.

Adam J. Grove, Nicolas Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. *Machine Learning Journal*, 43(3):179–210, 2001.

Robert C. Gunning and Hugo Rossi. *Analytic Functions of Several Complex Variables*. Prentice-Hall, New York, 1965.

Joseph L. Hellerstein, Yixin Diao, Sujay Parekh, and Dawn M. Tilbury. *Feedback Control of Computing Systems*. John Wiley & Sons, Inc., 2004.

Elbert Hendricks, Ole Jannerup, and Paul Haase Sørensen. *Linear systems control: deterministic and stochastic methods*. Springer, 2008.

Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, 1996.

Raul Jain and Pravin Varaiya. Simulation-based optimization of markov decision processes: An empirical process theory approach. *Automatica*, 46(8):1297–1304, 2010.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Sham Kakade, Michael Kearns, and John Langford. Exploration in metric state spaces. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

Sham M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 2nd edition, 2002.

Michael N. Katehakis and Herbert E. Robbins. Sequential choice from several populations. *Proceedings of National Academy of Sciences*, 92:8584–8565, 1995.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An optimal finite time analysis. Arxiv preprint http://arxiv.org/abs/1205.4217, 2012.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

Michael Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. In *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998.

Yu. N. Kiseliov. Algorithms of projection of a point onto an ellipsoid. *Lithuanian Mathematical Journal*, 34(2):141–159, 1994.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, January 1997.

Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, 2005.

Robert D. Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, pages 1–28, 2008a.

Robert D. Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, 2008b.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Tze Leung Lai and Ching-Zong Wei. Adaptive control of linear dynamic systems. Technical report, Columbia University, 1981.

Tze Leung Lai and Ching-Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25:466–481, March 1987.

Tze Leung Lai and Sidney Yakowitz. Machine learning and nonparametric bandit theory. *IEEE Transactions on Automatic Control*, 40:1199–1209, 1995.

Tze Leung Lai and Zhiliang Ying. Efficient recursive estimation and adaptive control in stochastic regression and ARMAX models. *Statistica Sinica*, 16:741–772, 2006.

Tze Leung Lai, Herbert Robbins, and Ching-Zong Wei. Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences*, 75(7): 3034–3036, 1979.

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*, 2010.

Lihong Li, Wei Chu, John Langford, , and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011.

Nicolas Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

Nicolas Littlestone. From online to batch learning. In *Annual Workshop on Computational Learning Theory: Proceedings of the second annual workshop on Computational learning theory*, 1989.

H. R. Maei, Cs. Szepesvári, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, 2009.

H. R. Maei, Cs. Szepesvári, S. Bhatnagar, and R. S. Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Odalric-Ambrym Maillard. *Apprentissage séquentiel: bandits, statistique et renforcement.* PhD thesis, INRIA, 2011.

Stefania Maniglia and Abdelaziz Rhandi. *Gaussian Measures on Separable Hilbert Spaces and Applications.* Lecture Notes of the University of Lecce, Italy, Quaderno, 2004.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Kanti V. Mardia, John T. Kent, and John M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

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

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

Norman Morse and Richard Sacksteder. Statistical isomorphism. *The Annals of Mathematical Statistics*, 37(1):203–214, 1966.

Arkadi Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1998.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest path problem. In *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010a.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online Markov decision processes under bandit feedback. In *Proceedings of the Twenty-Fourth Annual Conference on Neural Information Processing Systems*, 2010b.

Gergely Neu, András György, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 805–813, 2012.

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

Herbert Robbins and David Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *Annals of Math. Statistics*, 41:1410–1429, 1970.

R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(1):877–898, 1976.

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

Bryan P. Rynne and Martin A. Youngson. *Linear Functional Analysis*. Springer, 2nd edition, 2008.

George A. F. Seber. *Multivariate Observations*. Wiley, 1984.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Herbert A. Simon. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica*, 24(1):74–81, 1956.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *To appear in the IEEE Transactions on Information Theory*, 2012.

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

Alexander L. Strehl and Michael L. Littman. Online linear regression and its application to model-based reinforcement learning. In *Proceedings of the Twenty Second Annual Conference on Neural Information Processing Systems*, 2008.

Alexander L. Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L. Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, Cs. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009a.

R. S. Sutton, Cs. Szepesvári, and H. R. Maei. A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2009b.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Bradford Book. MIT Press, 1998.

Csaba Szepesvári. Multi-task learning. Notes, 2009.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodolical)*, 58(1):267–288, 1996.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 42(5): 674–690, 1997.

Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Thomas J. Walsh, István Szita, Carlos Diuk, and Michael L. Littman. Exploring compact reinforcement-learning representations with linear regression. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 1998.

Sanford Weisberg. *Applied Linear Regression*. Wiley, 1980.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

Appendix A

Background in Calculus and Linear Algebra

Definition A.1 (Lipschitz Continuity) Let $D \subset \mathbb{R}^d$, $h : D \to \mathbb{R}$. If there exists $L \geq 0$ such that for all $a_1, a_2 \in D$, $|h(a_1) - h(a_2)| \leq L \|a_1 - a_2\|$, then we say that function h is Lipschitz continuous with constant L.

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 A.2 (Implicit Function Theorem) Let $G : \mathbb{R}^{s+k} \to \mathbb{R}^s$ 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 $\hat{g} = g$.

Lemma A.3 (Sherman-Morrison Formula) Assume that $V \in \mathbb{R}^{d \times d}$ is an invertible matrix and $a, b \in \mathbb{R}^d$ are vectors. If $1 + b^\top A^{-1} a \neq 0$, then we have that

$$(V + ab^\top)^{-1} = V^{-1} - \frac{A^{-1} a b^\top A^{-1}}{1 + b^\top A^{-1} a} .$$

Lemma A.4 (Matrix-Determinant Lemma) Let $V \in \mathbb{R}^{n \times n}$ be an invertible matrix and A, B be $n \times m$ matrices. Then it holds that

$$\det(V + AB^\top) = \det(V) \det(I + B^\top V^{-1} A) .$$

Appendix B

Reproducing Kernel Hilbert Spaces

The following definitions and theorems are taken from (Rynne and Youngson, 2008).

Definition B.1 (Inner Product Space) A real or complex vector space $\mathcal{H}$ with an inner product $\langle .,. \rangle$ is called an inner product space.

Definition B.2 (Hilbert Space) An inner product space that is complete with respect to the metric associated with the norm induced by the inner product is called a Hilbert space.

Definition B.3 (Orthonormal Sequence) Let $\mathcal{H}$ be an inner product space. A sequence $(e_i) \subset \mathcal{H}$ is said to be an orthonormal sequence if $\|e_i\| = 1$ for all $i \in \mathbb{N}$, and $\langle e_i, e_j \rangle = 0$ for all $i, j \in \mathbb{N}$ with $i \neq j$.

Theorem B.4 Any infinite-dimensional inner product space $\mathcal{H}$ contains an orthonormal sequence.

Definition B.5 (Orthonormal Basis) Let $\mathcal{H}$ be a Hilbert space and let (e_i) be an orthonormal sequence in $\mathcal{H}$. Then (e_i) is called an orthonormal basis for $\mathcal{H}$ if and only if for all $m \in \mathcal{H}$, $m = \sum_{i=1}^{\infty} \langle m, e_i \rangle e_i$.

Definition B.6 (Separable Space) A space is separable if it contains a countable dense subset.

Theorem B.7 (Separable Hilbert Space) A Hilbert space $\mathcal{H}$ is separable if and only if it has an orthonormal basis.

Definition B.8 (Trace Class Operator) Let $\mathcal{H}$ be a separable Hilbert space with orthonormal basis (e_i). A bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ is trace class if $\sum_{i=1}^{\infty} \langle (A^*A)^{1/2} e_i, e_i \rangle$ is finite.

Note that in this definition the choice of (e_i) is arbitrary. However, it can be shown that if $\sum_{i=1}^{\infty} \langle (A^*A)^{1/2} e_i, e_i \rangle$ is finite for some orthonormal basis, then it is also finite for any other basis.

The next definitions and propositions are stated in the form given in (Szepesvári, 2009).

Definition B.9 (Reproducing Kernel Hilbert Space) Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel function. Consider the linear space

$$\mathcal{H}_0 = \{ \sum_{k=1}^{t} \lambda_k k(x_k, .) \ : \ t \in \mathbb{N}, \lambda_k \in \mathbb{R}, x_k \in \mathcal{X} \} .$$

Pick $f, g \in \mathcal{H}_0$. Without loss of generality, we may assume that $f = \sum_{k=1}^{t} \lambda_k k(x_k, .)$, $g = \sum_{k=1}^{t} \lambda'_k k(x_k, .)$, i.e., f and g are defined using the same set of points in $\mathcal{X}$. Define the inner product of f and g by

$$\langle f, g \rangle = \sum_{k,j} \lambda_k \lambda_j k(x_k, x_j) \; .$$

The Reproducing Kernel Hilbert Space (RKHS) underlying k is defined as the closure of $\mathcal{H}_0$ with respect to the norm induced by so-defined inner product. This will be denoted by $\mathcal{H}$.

Proposition B.10 The following statements are true:

- The inner product in the above definition is well-defined.
- For any $x \in \mathcal{X}$, $k(x, .) \in \mathcal{H}_0$ and for any $x, x' \in \mathcal{X}$,

  $$\langle k(x, .), k(x', .) \rangle = k(x, x') \; .$$

  This is the so-called "reproducing property" of k.
- $\mathcal{H}$ is a Hilbert space.
- If $f = \sum_i \alpha_i k(x_i, .) \in \mathcal{H}$ then

  $$\|f\|_{\mathcal{H}}^2 \doteq \langle f, f \rangle = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \; .$$

Definition B.11 (Feature Map) Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel. We say that $\Phi : \mathcal{X} \to \mathbb{R}^{\mathbb{N}}$ is a feature map underlying k if

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle$$

holds for any $x, x' \in \mathcal{X}$, where in the RHS the inner product is the usual ℓ^2-inner product.

Proposition B.12 Let Φ be a feature-map underlying k and let $\mathcal{H}$ be the RKHS corresponding to k. For any $f \in \mathcal{H}$,

$$\|f\|_{\mathcal{H}}^2 = \inf\{\|\theta\|_2^2 \; : \; f = \theta^\top \Phi\} \, .$$

In particular, $\mathcal{H} = \{\theta^\top \Phi \;\; : \;\; \|\theta\|_2^2 < +\infty\}$.

Appendix C

Tools from Probability Theory

The following definitions and lemmas are stated in the form given here in (Kallenberg, 2002).

Lemma C.1 (Fatou's Lemma) For any measurable functions $f_1, f_2, \cdots \geq 0$ on $(\Omega, \mathcal{F}, \mu)$, we have

$$\liminf_{t\to\infty} \mu f_t \geq \mu \liminf_{t\to\infty} f_t \ .$$

Theorem C.2 (Lebesgue Dominated Convergence) Let $f, f_1, f_2, \ldots$ and $g, g_1, g_2, \ldots$ be measurable functions on $(\Omega, \mathcal{F}, \mu)$ with $|f_t| \leq g_t$ for all t, and such that $f_t \to f$, $g_t \to g$, and $\mu g_t \to \mu g < \infty$. Then $\mu f_t \to \mu f$.

Definition C.3 (Stopping Time) Let $\mathcal{F} = (\mathcal{F}_t)$ be a filtration. We say that random time τ is $\mathcal{F}$-stopping time if $\{\tau \leq t\} \in \mathcal{F}_t$ for every t.

Definition C.4 (Martingale) Given any filtration $\mathcal{F} = (\mathcal{F}_k)$, we say that a sequence $x = (x_k)$ forms a martingale with respect to $\mathcal{F}$ if $\mathbb{E}\,[x_k|\mathcal{F}_{k-1}] = x_{k-1}$ a.s. for all k.

An example of a martingale is the fortune of a gambler at round t, denoted by x_t. The game is considered fair if $\mathbb{E}\,[x_t|\mathcal{F}_{t-1}] = x_{t-1}$, which is the martingale property. If the gambler skips some rounds of a fair game, we expect that the resulting game is still fair. The following definition and theorem are stated in the form given here in (Doob, 1953).

Definition C.5 (Optional Skipping) Let (y_k) be a stochastic process, and let $\mathcal{F} = (\mathcal{F}_t)$ be a filtration with the following properties:

(i) $\mathbb{E}\,[\,|y_k|\,] < \infty\,, \quad k \geq 1.$

(ii) Random variable y_k is either measurable with respect to $\mathcal{F}_k$ or is equal almost everywhere to a function which is.

(iii) For all $k \geq 1$ either $\mathbb{E}\,[y_{k+1}|\mathcal{F}_k] = 0$ with probability 1, or the process is real and

(iv) $\mathbb{E}\,[y_{k+1}] \geq 0$ with probability 1.

Let $i_1, i_2, \ldots$ be random variables taking on integral values, and having the following properties: $1 < i_1 < i_2 < \cdots < \infty$, and $\{\omega : i_j(\omega) = k\} \in \mathcal{F}_{k-1}$ for $k \geq j$, neglecting sets of measure 0. Define $\widetilde{y}_j$ by

$$\widetilde{y}_j = y_{i_j}\,, \quad j \geq 1\ .$$

The process $(\widetilde{y}_k)$ will be said to be obtained from the process (y_k) by optional skipping.

Theorem C.6 Suppose that a process $(y_k; \mathcal{F}_k)$ with the properties (i), (ii), (iv) of the preceding definition is transformed into the process $(\widetilde{y}_k)$ by optional skipping, and suppose that

$$\mathbb{E}\,[|\widetilde{y}_k|] < \infty\,, \quad k \geq 1\ . \tag{C.1}$$

Then the process $(\widetilde{y}_k)$ satisfies

$$\mathbb{E}\left[\widetilde{y}_{k+1} \mid \widetilde{y}_1, \ldots, \widetilde{y}_k\right] \geq 0\,, \quad k \geq 1\,, \tag{C.2}$$

with probability 1. If (iv) is replaced by (iii), there is equality in (C.2) with probability 1. The condition (C.1) is satisfied if either of the following conditions is satisfied.

(i) Each i_j defining the skipping is a bounded random variable, with probability 1.

(ii) There is a number K such that, for each j,

$$\mathbb{E}\left[|y_{k+1}| \mid \mathcal{F}_k\right] \leq K\,, \quad k \leq i_j(\omega)\,,$$

with probability 1.

Lemma C.7 Let $x_1, \ldots, x_t$ be random variables. Let $a \in \mathbb{R}$. Let $S_t = \sum_{k=1}^t x_k$ and $\widetilde{S}_t = \sum_{k=1}^t x_k \mathbb{I}_{\{x_k \leq a\}}$. Then it holds that

$$\mathbb{P}\left(S_t > x\right) \leq \mathbb{P}\left(\max_{1\leq k\leq t} x_k \geq a\right) + \mathbb{P}\left(\widetilde{S}_t > x\right)\,.$$

Proof.

$$\begin{aligned}\mathbb{P}\left(S_t \geq x\right) &\leq \mathbb{P}\left(\max_{1\leq k\leq t} x_k \geq a\right) + \mathbb{P}\left(S_t \geq x,\ \max_{1\leq k\leq t} x_k \leq a\right)\\ &\leq \mathbb{P}\left(\max_{1\leq k\leq t} x_k \geq a\right) + \mathbb{P}\left(\widetilde{S}_t \geq x\right)\,.\end{aligned}$$

□

Theorem C.8 (Hoeffding's Inequality) Let $x_1, \ldots, x_t$ be independent real-valued random variables such that for each $k = 1, \ldots, t$ there exist some $a_k \leq b_k$ such that $\mathbb{P}\left(a_k \leq x_k \leq b_k\right) = 1$. Then for every $\epsilon > 0$,

$$\mathbb{P}\left(\sum_{k=1}^t x_k - \mathbb{E}\left[\sum_{k=1}^t x_k\right] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^t (b_k - a_k)^2}\right)$$

and

$$\mathbb{P}\left(\sum_{k=1}^t x_k - \mathbb{E}\left[\sum_{k=1}^t x_k\right] \leq -\epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{k=1}^t (b_k - a_k)^2}\right)\,.$$

Theorem C.9 (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 $v_k \in [a_k, a_k + c_k]$ for some random variable a_k, measurable with respect to $x_1, \ldots, x_{k-1}$ and a positive constant c_k. If $S_t = \sum_{k=1}^t v_k$, then for any $\epsilon > 0$,

$$\mathbb{P}\left(S_t > \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^t c_k^2}\right)$$

and

$$\mathbb{P}\left(S_t < -\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{k=1}^t c_k^2}\right)\,.$$

Theorem C.10 (Freedman's Inequality) Let $x_1, \ldots, x_t$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_k)_{1\leq k\leq t}$ and with $|x_k| \leq K$. Let

$$S_k = \sum_{s=1}^k x_s$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_t^2 = \sum_{k=1}^{t} \mathbb{E}\left[x_k^2 | \mathcal{F}_{k-1}\right] \ .$$

Then for all constants $\epsilon, v > 0$,

$$\mathbb{P}\left(\max_{k=1\ldots t} S_k > \epsilon \text{ and } \Sigma_t^2 \leq v\right) \leq \exp\left(-\frac{\epsilon^2}{2(v + K\epsilon/3)}\right) ,$$

and therefore,

$$\mathbb{P}\left(\max_{k=1\ldots t} S_k > \sqrt{2v\epsilon} + (\sqrt{2}/3)K\epsilon \text{ and } \Sigma_t^2 \leq v\right) \leq e^{-\epsilon} .$$

The following theorem is stated in the form given here in (de la Peña et al., 2009).

Theorem C.11 (Law of the Iterated Logarithm) Let $x_1, x_2, \ldots$ be independent and identically distributed random variables and let $S_t = \sum_{k=1}^{t} x_k$. If $\mathbb{E}\left[x_1^2\right] < \infty$ and $\mathbb{E}\left[x_1\right] = \mu$, $\mathrm{Var}\left[x_1\right] = \sigma^2$, then

$$\begin{aligned}
\limsup_{t\to\infty} \frac{S_t - t\mu}{\sqrt{2t \log\log t}} &= \sigma \quad \text{a.s.,} \\
\liminf_{t\to\infty} \frac{S_t - t\mu}{\sqrt{2t \log\log t}} &= -\sigma \quad \text{a.s.,} \\
\limsup_{t\to\infty} \frac{\max_{1\leq k\leq t} |S_k - k\mu|}{\sqrt{2t \log\log t}} &= \sigma \quad \text{a.s.}
\end{aligned}$$

Conversely, if there exist finite constants a and τ such that

$$\limsup_{t\to\infty} \frac{S_t - ta}{\sqrt{2t \log\log t}} = \tau \quad \text{a.s.,}$$

then $a = \mathbb{E}\left[x_1\right]$ and $\tau^2 = \mathrm{Var}\left[x_1\right]$.

Appendix D

Some Useful Tricks

Proposition D.1 (Square-Root Trick) Let $a, b \geq 0$. If $z^2 \leq a + bz$ then $z \leq b + \sqrt{a}$.

Proof. Let $q(x) = x^2 - bx - a$. The condition $z^2 \leq a + bz$ can be expressed as $q(z) \leq 0$. The quadratic polynomial $q(x)$ has two roots

$$x_{1,2} = \frac{b \pm \sqrt{b^2 + 4a}}{2} \, .$$

The condition $q(z) \leq 0$ implies that $z \leq \max\{x_1, x_2\}$. Therefore,

$$z \leq \max\{x_1, x_2\} = \frac{b + \sqrt{b^2 + 4a}}{2} \leq b + \sqrt{a} \, ,$$

where we have used that $\sqrt{u + v} \leq \sqrt{u} + \sqrt{v}$ holds for any $u, v \geq 0$. □

Proposition D.2 (Logarithmic Trick) Let $c \geq 1$, $f > 0$, $\delta \in (0, 1/4]$. If $z \geq 1$ and $z \leq c + f\sqrt{\ln(z/\delta)}$ then $z \leq c + f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)}$.

Proof. Let $g(x) = x - c - f\sqrt{\ln(x/\delta)}$ for any $x \geq 1$. The condition $z \leq c + f\sqrt{\ln(z/\delta)}$ can be expressed as $g(z) \leq 0$. For large enough x, the function $g(x)$ is increasing. This is easy to see, since $g'(x) = 1 - \frac{f}{2x\sqrt{\ln(x/\delta)}}$. Namely, it is not hard see $g(x)$ is increasing for $x \geq \max\{1, f/2\}$ since for any such x, $g'(x)$ is positive.

Clearly, $c + f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)} \geq \max\{1, f/2\}$ since $c \geq 1$ and $\delta \in (0, 1/4]$. Therefore, it suffices to show that

$$g\left(c + f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)}\right) \geq 0 \, .$$

This is verified by the following calculation

$$
\begin{aligned}
g\left(c+f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)}\right) &= c+f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)}-c-f\sqrt{\ln\left(\frac{c+f\sqrt{2\ln\left((c+f)/\delta\right)}}{\delta}\right)} \\
&= f\sqrt{2\ln\left(\frac{c+f}{\delta}\right)}-f\sqrt{\ln\left(\frac{c+f\sqrt{2\ln\left((c+f)/\delta\right)}}{\delta}\right)} \\
&= f\sqrt{\ln\left(\frac{c+f}{\delta}\right)^2}-f\sqrt{\ln\left(\frac{c+f\sqrt{2\ln\left((c+f)/\delta\right)}}{\delta}\right)} \\
&\geq f\sqrt{\ln\left(\frac{c+f}{\delta}\right)^2}-f\sqrt{\ln\left(\frac{(c+f)\sqrt{2\ln\left((c+f)/\delta\right)}}{\delta}\right)} \\
&= f\sqrt{\ln\left(A^2\right)}-f\sqrt{\ln\left(A\sqrt{2\ln A}\right)} \\
&\geq 0,
\end{aligned}
$$

where have defined $A=(c+f)/\delta$ and the last inequality follows from that $A^2 \geq A\sqrt{2\ln A}$ for any $A>0$. □

Appendix E

Proofs of theorems of Chapter 4

E.1 Proof of Theorem 4.1

Lemma E.1 We have that

$$\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) = \prod_{k=1}^{t-1}(1 + \|a_k\|^2_{\overline{V}_k^{-1}}) \,.$$

Proof. First consider the finite dimensional case. Elementary algebra gives

$$\begin{aligned}\det(\overline{V}_t) &= \det(\overline{V}_{t-1} + a_{t-1}a_{t-1}^\top) = \det(\overline{V}_{t-1})\det(I + \overline{V}_{t-1}^{-1/2}a_{t-1}(\overline{V}_{t-1}^{-1/2}a_{t-1})^\top) \\ &= \det(\overline{V}_{t-1})\,(1 + \|a_{t-1}\|^2_{\overline{V}_{t-1}^{-1}}) = \det(V)\prod_{k=1}^{t-1}\left(1 + \|a_k\|^2_{\overline{V}_k^{-1}}\right) \,, \end{aligned} \tag{E.1}$$

where we used that all the eigenvalues of a matrix of the form $I + mm^\top$ are one except one eigenvalue, which is $1 + \|m\|^2$ and which corresponds to the eigenvector m.

Let $A_{1:t,n}$ be the first n columns of matrix $A_{1:t}$, W_n be the $n \times n$ version of V, and $a_{k,n}$ be the first n elements of a_k. Similarly, define $\overline{W}_{t,n}$. Define

$$S_\infty \;=\; \frac{\det(I + A_{1:t}V^{-1}A_{1:t}^*)}{\prod_{k=1}^{t-1}(1 + \|a_k\|^2_{\overline{V}_k^{-1}})}$$

and

$$S_n \;=\; \frac{\det(I + A_{1:t,n}W_n^{-1}A_{1:t,n}^*)}{\prod_{k=1}^{t-1}(1 + \|a_{k,n}\|^2_{\overline{W}_{k,n}^{-1}})} \,.$$

Both these quantities are well-defined. We know that $S_n = 1$ for all n and so $\lim_{n\to\infty} S_n = 1$. It remains to show that $\lim_{n\to\infty} S_n = S_\infty$. For simplicity, assume that V is diagonal. We know that if H and H_n are $t \times t$ matrices and $H_n \to H$, then, $\det(H_n) \to \det(H)$. This holds because determinant is a polynomial function of elements of a matrix. We show that each element of $I + A_{1:t,n}W_n^{-1}A_{1:t,n}^*$ converges to the corresponding element in $I + A_{1:t}V^{-1}A_{1:t}^*$. Consider the (i,j)th element of the first matrix. It is equal to $\sum_{l=1}^{n} a_{i,l}a_{j,l}/V_{l,l}$, which by definition converges to $\sum_{l=1}^{\infty} a_{i,l}a_{j,l}/V_{l,l}$, which is the (i,j)th element of the second matrix. This shows that the numerator of S_n is converging to the numerator of S_∞. Similarly, we can show that the denominator of S_n is also converging to the denominator of S_∞. Thus, $\lim_{n\to\infty} S_n = S_\infty$. Thus $S_\infty = 1$. Thus,

$$\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) = \prod_{k=1}^{t-1}(1 + \|a_k\|^2_{\overline{V}_k^{-1}}) \,,$$

finishing the proof. □

Lemma E.2 We have that

$$\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) \le \sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}}\,,$$
$$\sum_{k=1}^{t-1}(\|a_k\|^2_{\overline{V}_k^{-1}} \wedge 1) \le 2\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right)\,.$$

Proof. Using $\log(1+u) \le u$, we can bound

$$\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right) \le \sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}}\,.$$

Further, by $u \le 2\log(1+u)$, which holds when $u \in [0,1]$, we get that

$$\sum_{k=1}^{t-1}(\|a_k\|^2_{\overline{V}_k^{-1}} \wedge 1) \le 2\log\det\left(I + A_{1:t}V^{-1}A_{1:t}^*\right)\,.$$

□

In the finite-dimensional case, we get the following result.

Lemma E.3 Let $(a_k)_{k=1}^{\infty}$ be a sequence in $\mathbb{R}^d$, V be a $d \times d$ positive definite matrix and define $\overline{V}_t = V + \sum_{k=1}^{t-1} a_s a_s^\top$. Then, we have that

$$\log\left(\frac{\det(\overline{V}_t)}{\det(V)}\right) \le \sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}}\,.$$

Further, if $\|a_k\|_2 \le L$ for all k, then

$$\sum_{k=1}^{t-1} \min\left\{1, \|a_k\|^2_{\overline{V}_k^{-1}}\right\} \le 2(\log\det(\overline{V}_t) - \log\det V)$$
$$\le 2(d\log((\mathrm{trace}(V) + tL^2)/d) - \log\det V)\,,$$

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

$$\sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}} \le 2\log\frac{\det(\overline{V}_t)}{\det(V)}\,.$$

Proof. The trace of $\overline{V}_t$ is bounded by $\mathrm{trace}(V) + tL^2$ if $\|a_k\|_2 \le L$. Hence,

$$\det(\overline{V}_t) = \prod_{i=1}^{d} \lambda_i \le \left(\frac{\mathrm{trace}(V) + tL^2}{d}\right)^d$$

and therefore,

$$\log\det(\overline{V}_t) \le d\log((\mathrm{trace}(V) + tL^2)/d)\,,$$

finishing the proof of the second inequality. The sum $\sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}}$ can itself be upper bounded as a function of $\log\det(\overline{V}_t)$ provided that $\lambda_{\min}(V)$ is large enough. Notice $\|a_k\|^2_{\overline{V}_k^{-1}} \le \lambda^{-1}_{\min}(\overline{V}_k)\,\|a_k\|^2 \le L^2/\lambda_{\min}(V)$. Hence, we get that if $\lambda_{\min}(V) \ge \max(1, L^2)$,

$$\log\frac{\det(\overline{V}_t)}{\det V} \le \sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}} \le 2\log\frac{\det(\overline{V}_t)}{\det(V)}\,.$$

□

Most of this argument can be extracted from the paper of Dani et al. (2008). However, the idea goes back at least to Lai et al. (1979), Lai and Wei (1982). (a similar argument is used around Theorem 11.7 in the book by Cesa-Bianchi and Lugosi (2006).) Note that Lemmas B.9–B.11 of Rusmevichientong and Tsitsiklis (2010) also give a bound on $\sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}}$, with an essentially identical argument. Alternatively, we can use the bounding technique of Auer (2002) (see the proof of Lemma 13 there on pages 412–413) to derive a bound like $\sum_{k=1}^{t-1} \|a_k\|^2_{\overline{V}_k^{-1}} \le Cd \log t$ for a suitable chosen constant $C > 0$.

Remark E.4 By combining Corollary 3.6 and Lemma E.2, and assuming that $V = \lambda I$ for some $\lambda > 0$, we get a simple worst case bound that holds with probability $1 - \delta$:

$$\forall t \ge 0, \quad \|S_t\|^2_{\overline{V}_t^{-1}} \le R^2 \left(\frac{tL^2}{\lambda} + 2 \log\left(\frac{1}{\delta}\right) \right) . \tag{E.2}$$

Proof of Theorem 4.1. Lets decompose the instantaneous regret as follows:

$$\begin{aligned}
r_t &= \langle \theta_*, a_t \rangle - \langle \theta_*, a_{*,t} \rangle \\
&\le \langle \theta_*, a_t \rangle - \langle \widetilde{\theta}_t, a_t \rangle && \text{(because } (a_t, \widetilde{\theta}_t) \text{ is optimistic)} \\
&= \langle \theta_* - \widetilde{\theta}_t, a_t \rangle \\
&= \langle \theta_* - \widehat{\theta}_t, a_t \rangle + \langle \widehat{\theta}_t - \widetilde{\theta}_t, a_t \rangle \\
&= \left\| \theta_* - \widehat{\theta}_t \right\|_{\overline{V}_t^{-1}} \|a_t\|_{\overline{V}_t^{-1}} + \left\| \widehat{\theta}_t - \widetilde{\theta}_t \right\|_{\overline{V}_t^{-1}} \|a_t\|_{\overline{V}_t^{-1}} && \text{(Cauchy-Schwarz Inequality)} \\
&\le 2\sqrt{\beta_t(\delta)} \, \|a_t\|_{V_t^{-1}} . && \text{(E.3)}
\end{aligned}$$

By (E.3) and the fact that $r_t \le 2$, we get that

$$\begin{aligned}
r_t &\le 2 \min\left(\sqrt{\beta_t(\delta)} \, \|a_t\|^2_{\overline{V}_t^{-1}} , \, 1 \right) \\
&\le 2\sqrt{\beta_t(\delta)} \min\left(\|a_t\|^2_{\overline{V}_t^{-1}} , \, 1 \right) .
\end{aligned}$$

Thus, with probability at least $1 - \delta$, for any $T \ge 0$,

$$\begin{aligned}
R_T = \sum_{t=1}^{T} r_t &\le \sqrt{T \sum_{t=1}^{T} r_t^2} \le \sqrt{8\beta_T(\delta) T \sum_{t=1}^{T} \left(\|a_t\|^2_{\overline{V}_t^{-1}} \wedge 1 \right)} \\
&\le 4\sqrt{\beta_T(\delta) T \log\det(I + A_{1:T+1} A^*_{1:T+1}/\lambda)} ,
\end{aligned}$$

where the last step follows from Lemma E.2 with the choice of $V = \lambda I$.

□

E.2 Proof of Theorem 4.4

First, we prove Lemma 4.6.

Proof of Lemma 4.6. We first consider a simple case. Assume that $C = a \otimes a$ where $a \in \mathcal{H}$ and B is positive definite. Let $x \ne 0$ be an arbitrary vector. Using the Cauchy-Schwarz inequality, we get

$$\langle x, a \rangle^2 = \langle B^{1/2} x, B^{-1/2} a \rangle^2 \le \left\| B^{1/2} x \right\|^2 \left\| B^{-1/2} a \right\|^2 = \|x\|^2_B \, \|m\|^2_{B^{-1}} .$$

Thus,

$$\langle x, (B + a \otimes a) x \rangle \le \langle x, Bx \rangle + \|x\|^2_B \, \|a\|^2_{B^{-1}} = (1 + \|a\|^2_{B^{-1}}) \, \|x\|^2_B$$

and so

$$\frac{\langle x, Ax\rangle}{\langle x, Bx\rangle} \le 1 + \|a\|^2_{B^{-1}} = \det(I + aB^{-1}a^*)\,,$$

thus finishing the proof of this case.

If $C = D^*D = a_1 \otimes a_1 + \cdots + a_{t-1} \otimes a_{t-1}$, then define $V_s = B + a_1 \otimes a_1 + \cdots + a_{s-1} \otimes a_{s-1}$ and use

$$\begin{aligned}\frac{\langle x, Ax\rangle}{\langle x, Bx\rangle} &= \frac{\langle x, V_t x\rangle}{\langle x, V_{t-1}x\rangle}\frac{\langle x, V_{t-1} x\rangle}{\langle x, V_{t-2}x\rangle}\cdots\frac{\langle x, V_2 x\rangle}{\langle x, Bx\rangle}\\ &\le \prod_{k=1}^{t-1}(1 + \|a_k\|^2_{V_k^{-1}})\\ &= \det(I + DB^{-1}D^*)\,,\end{aligned}$$

where the last step follows from Lemma E.1. This finishes the proof of this case.

If C is a positive definite matrix, then the eigendecomposition of C gives $C = U^\top \Lambda U$, where U is orthonormal and Λ is positive diagonal matrix. This, in fact gives that C can be written as the sum of countably many rank-one matrices:

$$C = D^*D = \sum_{k=1}^{\infty} a_k \otimes a_k\,.$$

We finish the proof for the general case by noting that $\det(I + DB^{-1}D^*)$ is well-defined by the assumption that operator $DB^{-1}D^*$ is trace-class, and applying a simple limiting argument to get that

$$\prod_{k=1}^{\infty}(1 + \|a_k\|^2_{V_k^{-1}}) = \det(I + DB^{-1}D^*)\,.$$

□

Proof of Theorem 4.4. Let τ_t be the smallest time step less than or equal to t such that $\widetilde{\theta}_t = \widetilde{\theta}_{\tau_t}$. By an argument similar to the one used in Theorem 4.1, we have

$$r_t \le \langle \theta_* - \widehat{\theta}_{\tau_t}, a_t\rangle + \langle \widehat{\theta}_{\tau_t} - \widetilde{\theta}_{\tau_t}, a_t\rangle\,.$$

We also have that for all $\theta \in C_{\tau_t - 1}$ and any $a \in \mathcal{H}$,

$$\begin{aligned}|\langle \theta - \widehat{\theta}_{\tau_t}, a\rangle| &\le \left\|\overline{V}_t^{1/2}(\theta - \widehat{\theta}_{\tau_t})\right\| \|a\|_{\overline{V}_t^{-1}}\\ &\le \left\|\overline{V}_{\tau_t}^{1/2}(\theta - \widehat{\theta}_{\tau_t})\right\| \sqrt{\det(I + A_{\tau_t:t}V_{\tau_t}^{-1}A_{\tau_t:t})}\,\|a\|_{\overline{V}_t^{-1}}\\ &\le \sqrt{1+C}\left\|\overline{V}_{\tau_t}^{1/2}(\theta - \widehat{\theta}_{\tau_t})\right\| \|a\|_{\overline{V}_t^{-1}}\\ &\le \sqrt{(1+C)\beta_{\tau_t}}\,\|a\|_{\overline{V}_t^{-1}}\\ &\le \sqrt{(1+C)\beta_t}\,\|a\|_{\overline{V}_t^{-1}}\,,\end{aligned}$$

where the second step follows from Lemma 4.6, and the third step follows from the fact that at time t, $\det(I + A_{\tau_t:t}V_{\tau_t}^{-1}A_{\tau_t:t}) \le 1 + C$. The rest of the argument is identical to that of Theorem 4.1. We conclude that with probability at least $1 - \delta$, for all $T \ge 0$,

$$R_T \le 4\sqrt{(1+C)\beta_T T \log\det(I + A_{1:T+1}V^{-1}A^*_{1:T+1})}\,.$$

□

E.3 Proof of Theorem 4.8

First we state a matrix perturbation theorem from Stewart and Sun (1990).

Theorem E.5 (Stewart and Sun (1990), Corollary 4.9) Let A be a $d \times d$ symmetric matrix with eigenvalues $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_d$, E be a symmetric $d \times d$ matrix with eigenvalues $e_1 \geq e_2 \geq \ldots \geq e_d$, and $V = A + E$ denote a symmetric perturbation of A such that the eigenvalues of V are $\widetilde{\nu}_1 \geq \widetilde{\nu}_2 \geq \ldots \geq \widetilde{\nu}_d$. Then, for $i = 1, 2, \ldots, d$,

$$\widetilde{\nu}_i \in [\nu_i + e_d, \nu_i + e_1] \,.$$

Proof of Theorem 4.8. We start by bounding the regret in terms of $\log\det(\overline{V}_{T+1})$. We have that

$$\begin{aligned} R_T &= \sum_{t=1}^{T} r_t \leq \sum_{t=1}^{T} \frac{r_t^2}{\Delta} \\ &\leq \frac{16\beta_T(\delta)}{\Delta} \log\det(\overline{V}_{T+1}) \,, \end{aligned} \tag{E.4}$$

where the first inequality follows from the fact that either $r_t = 0$ or $\Delta < r_t$, and the second inequality can be extracted from the proof of Theorem 4.1. Let b_t be the number of times that we have played a sub-optimal action (an action a_s for which $\langle \theta_*, a_s \rangle - \langle \theta_*, a_* \rangle \geq \Delta$) up to time t. Next, we bound $\log\det(\overline{V}_t)$ in terms of b_t.

We apply Theorem E.5 to bound the eigenvalues of $\overline{V}_t$. Let $E_t = \sum_{s:a_s \neq a_*}^{t} a_s a_s^\top$ and $A_t = \overline{V}_t - E_t = (t - b_t) a_* a_*^\top$. Let the eigenvalues of $\overline{V}_t$ and E_t be $\lambda_1 \geq \cdots \geq \lambda_d$ and $e_1 \geq \cdots \geq e_d$, respectively. The only non-zero eigenvalue of A_t is $(t - b_t)L_*$, where $L_* = a_*^\top a_* \leq L$. By Theorem E.5, we have that

$$\lambda_1 \in [(t - b_t)L_* + e_d,\, (t - b_t)L_* + e_1]$$

and

$$\forall i \in \{2, \ldots, d\} \,, \quad \lambda_i \in [e_d, e_1] \,.$$

Thus,

$$\begin{aligned} \det(\overline{V}_t) = \prod_{i=1}^{d} \lambda_i &\leq ((t - b_t)L_* + e_1) e_1^{d-1} \\ &\leq ((t - b_t)L + e_1) e_1^{d-1} \,. \end{aligned}$$

Therefore,

$$\log\det(\overline{V}_t) \leq \log((t - b_t)L + e_1) + (d - 1)\log e_1 \,.$$

Because $\mathrm{trace}(E) = \sum_{s:a_s \neq a_*}^{t} \mathrm{trace}(a_s a_s^\top) \leq L b_t$, we conclude that $e_1 \leq L b_t$. Thus,

$$\begin{aligned} \log\det(\overline{V}_t) &\leq \log((t - b_t)L + L b_t) + (d - 1)\log(L b_t) \\ &= \log(Lt) + (d - 1)\log(L b_t) \,. \end{aligned} \tag{E.5}$$

After some calculations, we can show that

$$\beta_t \log\det \overline{V}_t \leq 4R^2\lambda S^2 \left(\log\det \overline{V}_t + 2\log\frac{1}{\delta} \right)^2 \tag{E.6}$$

$$\leq 4R^2\lambda S^2 \left(d\log\frac{d\lambda + tL^2}{d} + 2\log\frac{1}{\delta} \right)^2 \,, \tag{E.7}$$

where the second inequality follows from Lemma E.3. Thus, by (E.4) and noting that $R_t \geq b_t \Delta$,

$$b_t \leq \frac{16\beta_t}{\Delta^2} \log\det(\overline{V}_t) \tag{E.8}$$

$$\leq \frac{64R^2\lambda S^2}{\Delta^2}\left(d\log\frac{d\lambda + tL^2}{d} + 2\log\frac{1}{\delta}\right)^2 . \tag{E.9}$$

Thus, with probability $1-\delta$, for all $T \geq 0$,

$$R_T \leq \frac{16\beta_T}{\Delta} \log\det(\overline{V}_{T+1}) \tag{E.4}$$

$$\leq \frac{64R^2\lambda S^2}{\Delta}(\log\det(\overline{V}_{T+1}) + 2\log(1/\delta))^2 \tag{E.6}$$

$$\leq \frac{16R^2\lambda S^2}{\Delta}(\log(L(T+1)) + (d-1)\log(Lb_{T+1}) + 2\log(1/\delta))^2 \tag{E.5}$$

$$\leq \frac{16R^2\lambda S^2}{\Delta}\Big(\log(L(T+1)) + (d-1)\log\frac{64R^2\lambda S^2 L}{\Delta^2} \tag{E.9}$$

$$+ 2(d-1)\log\left(d\log\frac{d\lambda + (T+1)L^2}{d} + 2\log(1/\delta)\right) + 2\log(1/\delta)\Big)^2 ,$$

finishing the proof. □

E.4 Proof of Theorem 4.10

Proof. Suppose that the confidence intervals do not fail. When we play action i, the lower estimate of the action is below μ^*. Thus,

$$c_{i,s} \geq \frac{\Delta_i}{2} .$$

Substituting $c_{i,s}$ and squaring gives

$$\frac{N_{i,s}^2 - 1}{N_{i,s}+1} \leq \frac{N_{i,s}^2}{N_{i,s}+1} \leq \frac{4}{\Delta_i^2}\left(2\log\frac{d(1+N_{i,s})^{1/2}}{\delta}\right) .$$

By applying Lemma 8 of Antos et al. (2010), we get that for all $s \geq 0$,

$$N_{i,s} \leq 3 + \frac{16}{\Delta_i^2}\log\frac{2d}{\Delta_i\delta} .$$

By substituting the above inequality in $R_T = \sum_{i\neq i_*}\Delta_i N_{i,n}$, we get that with probability at least $1-\delta$, the total regret is bounded by

$$R_T \leq \sum_{i:\Delta_i>0}\left(3\Delta_i + \frac{16}{\Delta_i}\log\frac{2d}{\Delta_i\delta}\right) .$$

□

Appendix F

Proofs of theorems of Chapter 5

F.1 Proof of Lemma 5.14

Proof of Lemma 5.14. We first consider a simple case. Let $A = B + vv^\top$ and B be a positive definite matrix. Let $X \neq 0$ be an arbitrary matrix. Using the Cauchy-Schwarz inequality and the fact that for any matrix M, $\left\|M^\top M\right\| = \|M\|^2$, we get

$$\left\|X^\top vv^\top X\right\| = \left\|v^\top X\right\|^2 = \left\|v^\top B^{-1/2} B^{1/2} X\right\|^2 \leq \left\|v^\top B^{-1/2}\right\|^2 \left\|B^{1/2} X\right\|^2 .$$

Thus,

$$\begin{aligned}\left\|X^\top (B + vv^\top) X\right\| &\leq \left\|X^\top BX\right\| + \left\|v^\top B^{-1/2}\right\|^2 \left\|B^{1/2} X\right\|^2 \\ &= \left(1 + \left\|v^\top B^{-1/2}\right\|^2\right) \left\|B^{1/2} X\right\|^2 ,\end{aligned}$$

and so

$$\frac{\left\|X^\top AX\right\|}{\left\|X^\top BX\right\|} \leq 1 + \left\|v^\top B^{-1/2}\right\|^2 .$$

We also have that

$$\det(A) = \det(B + vv^\top) = \det(B) \det(I + B^{-1/2} v (B^{-1/2} v)^\top) = \det(B)(1 + \|v\|_{B^{-1}}^2),$$

thus finishing the proof of this case.

More generally, if $A = B + v_1 v_1^\top + \cdots + v_{t-1} v_{t-1}^\top$ then define $V_s = B + v_1 v_1^\top + \cdots + v_{s-1} v_{s-1}^\top$ and use

$$\frac{\left\|X^\top AX\right\|}{\left\|X^\top BX\right\|} = \frac{\left\|X^\top V_t X\right\|}{\left\|X^\top V_{t-1} X\right\|} \frac{\left\|X^\top V_{t-1} X\right\|}{\left\|X^\top V_{t-2} X\right\|} \cdots \frac{\left\|X^\top V_2 X\right\|}{\left\|X^\top BX\right\|} .$$

By the above argument, since all the terms are positive, we get

$$\frac{\left\|X^\top AX\right\|}{\left\|X^\top BX\right\|} \leq \frac{\det(V_t)}{\det(V_{t-1})} \frac{\det(V_{t-1})}{\det(V_{t-2})} \cdots \frac{\det(V_2)}{\det(B)} = \frac{\det(V_t)}{\det(B)} = \frac{\det(A)}{\det(B)} ,$$

the desired inequality.

Finally, by SVD, if $C \succ 0$, C can be written as the sum of at most m rank-one matrices, finishing the proof for the general case. □

$\mathcal{B}_{T+1} = \emptyset$
for $t := T, T-1, \ldots$ **do**
  Initialize $\mathcal{B}_t = \mathcal{B}_{t+1}$
  while $\left\|\pi(M_t, \mathcal{B}_t^\perp)\right\|_F > m\epsilon$ **do**
    Choose a column of M_t, v, such that $\left\|\pi(v, \mathcal{B}_t^\perp)\right\|_F > \epsilon$
    Update $\mathcal{B}_t = \mathcal{B}_t \oplus \{v\}$
  end while
end for

Figure F.1: Obtaining subspaces $\mathcal{B}_t$ for $t \le T$.

F.2 Bounding $\|x_t\|$ - Proof of Lemmas 5.8 and 5.9

Define

$$
\begin{aligned}
H_1 &> 16 \vee \frac{4S^2 H_2^2 H_3}{m}, \\
H_3 &= 16^{m-2}(1 \vee S^{2(m-2)}), \\
H_2 &= \sup_{Y \ge 1} \frac{1}{Y}\left(nL\sqrt{m \log\left(\frac{1+TY/\lambda}{\delta}\right)} + \lambda^{1/2} S\right) .
\end{aligned}
$$

Recall that $z_t = (x_t^\top, a_t^\top)^\top$, $\widetilde{\Theta}_t$ is the optimistic estimate of Θ_* at round t, and E is the event that all confidence sets hold up to round T. First, we show that $\left\|(\Theta_* - \widetilde{\Theta}_t)^\top z_t\right\|$ is well-controlled except for a small number of rounds. Given this and a proper decomposition of the state update equation, we prove that $\|x_t\|$ stays smaller than α_t.

Let $\pi(v, \mathcal{B})$ and $\pi(M, \mathcal{B})$ be projections of vector v and matrix M on subspace $\mathcal{B} \subset \mathbb{R}^m$, where the projection of matrix M is done column-wise. Let $\mathcal{B} \oplus \{v\}$ be the span of $\mathcal{B}$ and v. Let $\mathcal{B}^\perp$ be the subspace orthogonal to $\mathcal{B}$ such that $\mathcal{B} \oplus \mathcal{B}^\perp = \mathbb{R}^m$.

Let $M_t = \Theta_* - \widetilde{\Theta}_t$. Fix a real number $0 \le \epsilon \le 1$. Define a sequence of subspaces $\mathcal{B}_t$ as follows: Set $\mathcal{B}_{T+1} = \emptyset$. For $t = T, \ldots, 1$, initialize $\mathcal{B}_t = \mathcal{B}_{t+1}$. Then while $\left\|\pi(M_t, \mathcal{B}_t^\perp)\right\|_F > m\epsilon$, choose a column of M_t, v, such that $\left\|\pi(v, \mathcal{B}_t^\perp)\right\|_F > \epsilon$ and update $\mathcal{B}_t = \mathcal{B}_t \oplus \{v\}$. The process is shown in Figure F.1. The sequence is defined such that the column space of $M_T, \ldots, M_t$ has *little* outside subspace $\mathcal{B}_t$; after finishing with round t, we will have

$$
\left\|\pi(M_t, \mathcal{B}_t^\perp)\right\| \le \left\|\pi(M_t, \mathcal{B}_t^\perp)\right\|_F \le m\epsilon . \tag{F.1}
$$

Let $\mathcal{T}_T$ be the set of timesteps (including repetitions) at which subspace $\mathcal{B}_t$ expands. The cardinality of this set, $c(t)$, is at most m. Denote these timesteps by $t_1 \ge t_2 \ge \cdots \ge t_{c(t)}$. Let $i(t) = \max\{1 \le i \le c(t) : t_i \ge t\}$. Let $N_t = \{v_1, \ldots, v_{c(t)}\}$ be the set of vectors that are added to $\mathcal{B}_t$ during the expansion rounds. Let $B_j = \text{span}(v_1, \ldots, v_j)$. Notice that $B_{c(t)} = \mathcal{B}_t$. We can write $v_k = w_k + u_k$, where $w_k \in B_{k-1}$, $u_k \perp B_{k-1}$, $\|u_k\| \ge \epsilon$, and $\|v_k\| \le 2S$.

The following lemma shows that any vector can be represented approximately by members of N_t.

Lemma F.1 We have that

$$
\forall x,\, \forall j \in \{1, \ldots, c(t)\}, \quad \sum_{k=1}^{j} \langle v_k, x\rangle^2 \ge \frac{\epsilon^4}{H_1} \frac{\epsilon^{2(j-2)}}{16^{j-2}(1 \vee S^{2(j-2)})} \left\|\pi(x, B_j)\right\|^2 , \tag{F.2}
$$

and thus,

$$
H_1 H_3 \sum_{i=1}^{i(t)} \left\|M_{t_i}^\top x\right\|^2 \ge \epsilon^{2c(t)} \left\|\pi(x, \mathcal{B}_t)\right\|^2 . \tag{F.3}
$$

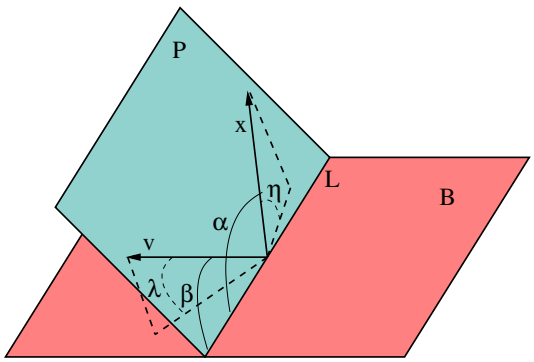


Figure F.2: Relevant quantities that are used in the inductive step. $v = v_{l+1}$ and $B = B_l$.

Proof. By construction, N_t is a subset of the set of all columns of $M_{t_1}, M_{t_2}, \ldots, M_{t_{i(t)}}$. Thus, we have that

$$\sum_{i=1}^{i(t)} \left\| M_{t_i}^\top x \right\|^2 \geq x^\top (v_1 v_1^\top + \cdots + v_{c(t)} v_{c(t)}^\top) x \,,$$

which shows that we can obtain (F.3) from (F.2).

The inequality (F.2) is proven by induction. First, we prove the induction base for $j = 1$. Without loss of generality, assume that $x = Cv_1$ for some constant C. From condition $H_1 > 16$, we get that $16^{-1} H_1 (1 \vee S^{-1}) \geq 1$. Thus,

$$\epsilon^2 \geq \frac{\epsilon^2}{16^{-1} H_1 (1 \vee S^{-1})} \,.$$

Thus,

$$C^2 \left\| v_1 \right\|^4 \geq \frac{\epsilon^2 C^2 \left\| v_1 \right\|^2}{16^{-1} H_1 (1 \vee S^{-1})} \,,$$

where we have used the fact that $\|v_1\| \geq \epsilon$ (see definition of v_1 in Figure F.1). Finally, by noting that $C^2 \left\| v_1 \right\|^4 = \langle v_1, x \rangle^2$ and $C^2 \left\| v_1 \right\|^2 = \left\| \pi(x, B_1) \right\|^2$, we get

$$\langle v_1, x \rangle^2 \geq \frac{\epsilon^4}{H_1} \frac{\epsilon^{-2}}{16^{-1}(1 \vee S^{-2})} \left\| \pi(x, B_1) \right\|^2 \,,$$

which establishes the base of induction.

Next, we prove that if the inequality (F.2) holds for $j = l$, then it also holds for $j = l+1$. Figure F.2 contains all relevant quantities that are used in the following argument.

Assume that the inequality (F.2) holds for $j = l$. Without loss of generality, assume that x is in B_{l+1}, and thus $\|\pi(x, B_{l+1})\| = \|x\|$. Let $P \subset B_{l+1}$ be the 2-dimensional subspace that passes through x and v_{l+1}. The 2-dimensional subspace P and the l-dimensional subspace B_l can, respectively, be identified by $l-1$ and one equations in B_{l+1}. Because P is not a subset of B_l, the intersection of P and B_l is a line in B_{l+1}. Let's call this line L. The line L creates two half-planes on P. Without loss of generality, assume that x and v_{l+1} are on the same half-plane. (notice that we can always replace x by $-x$ in (F.2).)

Let $0 \leq \beta \leq \pi/2$ be the angle between v_{l+1} and L. Let $0 < \lambda < \pi/2$ be the orthogonal angle between v_{l+1} and B_l; λ is the angle between v_{l+1} and $\pi(v_{l+1}, B_l)$. We know that $\beta > \lambda$. Recall that u_{l+1} and w_{l+1} are defined such that $v_{l+1} = w_{l+1} + u_{l+1}$, $w_{l+1} \in B_l$, u_{l+1} is orthogonal to B_l, $\|u_{l+1}\| \geq \epsilon$, and $\|v_{l+1}\| \leq 2S$. Thus, $\beta \geq \arcsin(\epsilon / \|v_{l+1}\|)$. Let $0 \leq \alpha \leq \pi$ be the angle between x and L. ($\alpha < \pi$, because x and v_{l+1} are on the same half-plane.) The direction of α is chosen so that it is consistent with the direction of β. Finally, let $0 \leq \eta \leq \pi/2$ be the angle between x and $\pi(x, B_l)$ (see Figure F.2).

By the induction assumption

$$\sum_{k=1}^{l+1}\langle v_k, x\rangle^2 = \langle v_{l+1}, x\rangle^2 + \sum_{k=1}^{l}\langle v_k, x\rangle^2$$
$$\geq \langle v_{l+1}, x\rangle^2 + \frac{\epsilon^4}{H}\frac{\epsilon^{2(l-2)}}{16^{l-2}(1\vee S^{2(l-2)})}\left\|\pi(x, B_l)\right\|^2. \tag{F.4}$$

There are two possibilities:

(i) If $\alpha < \pi/2 + \beta/2$ or $\alpha > \pi/2 + 3\beta/2$, then

$$|\langle v_{l+1}, x\rangle| = |\|v_{l+1}\|\,\|x\|\cos\angle(v_{l+1}, x)| \geq \left|\|v_{l+1}\|\,\|x\|\sin\left(\frac{\beta}{2}\right)\right| \geq \frac{\epsilon\,\|x\|}{4}. \tag{F.5}$$

From $H_1 > 16$, we obtain that for any $l \geq 1$,

$$H_1 > \frac{1}{16^{l-2}(1\vee S^{2(l-1)})},$$

which also implies that

$$H_1 > \frac{\epsilon^{2l}}{16^{l-2}(1\vee S^{2(l-1)})}. \tag{F.6}$$

By (F.5) and (F.6) and noting that $x \in B_{l+1}$, we get

$$\langle v_{l+1}, x\rangle^2 \geq \frac{\epsilon^2\,\|x\|^2}{16} \geq \frac{\epsilon^4}{H_1}\frac{\epsilon^{2(l-1)}}{16^{l-1}(1\vee S^{2(l-1)})}\left\|\pi(x, B_{l+1})\right\|^2. \tag{F.7}$$

(ii) If $\pi/2 + \beta/2 < \alpha < \pi/2 + 3\beta/2$, then $\eta < \pi/2 - \beta/2$. Thus,

$$\|\pi(x, B_l)\| = \|x\|\,|\cos(\eta)| \geq \|x\|\left|\sin\left(\frac{\beta}{2}\right)\right| \geq \frac{\epsilon\,\|x\|}{4S}.$$

Thus,

$$\|\pi(x, B_l)\|^2 \geq \frac{\epsilon^2\,\|x\|^2}{16S^2},$$

and

$$\frac{\epsilon^4}{H_1}\frac{\epsilon^{2(l-2)}}{16^{l-2}(1\vee S^{2(l-2)})}\left\|\pi(x, B_l)\right\|^2 \geq \frac{\epsilon^4}{H}\frac{\epsilon^{2(l-1)}}{16^{l-1}(1\vee S^{2(l-1)})}\|x\|^2,$$

which, together with (F.4) and (F.7), finishes the proof of the induction step.

□

Next we prove a simple bound on $\left\|M_t^\top z_s\right\|$.

Lemma F.2 For all $t \leq T$,

$$\max_{1\leq s\leq t-1}\left\|(\Theta_* - \widetilde{\Theta}_t)^\top z_s\right\| \leq \beta_t(\delta/4)^{1/2}.$$

Proof. On event E, for any $t \leq T$ we have that

$$\operatorname{trace}\left(M_t^\top\left(\lambda I + \sum_{s=1}^{t-1} z_s z_s^\top\right)M_t\right) \leq \beta_t(\delta/4).$$

Because $\lambda > 0$ we get that,

$$\operatorname{trace}\left(\sum_{s=1}^{t-1} M_t^\top z_s z_s^\top M_t\right) \leq \beta_t(\delta/4).$$

Thus,

$$\sum_{s=1}^{t-1} \left\|M_t^\top z_s\right\|^2 = \sum_{s=1}^{t-1} \operatorname{trace}(M_t^\top z_s z_s^\top M_t) \le \beta_t(\delta/4) \,.$$

Thus, for all $t \le T$,

$$\max_{1\le s\le t-1} \left\|M_t^\top z_s\right\| \le \beta_t(\delta/4)^{1/2} \,.$$

□

Now we are ready to show that $\left\|M_t^\top z_t\right\|$ is well-controlled except when $t \in \mathcal{T}_T$.

Lemma F.3 We have that for any $1 \le t \le T$,

$$\max_{s\le t,\, s\notin\mathcal{T}_t} \left\|M_s^\top z_s\right\| \le G Z_t^{m/m+1} \beta_t(\delta/4)^{1/2(m+1)},$$

where

$$G = 2\left(2Sm^m\sqrt{mH_1H_3}\right)^{1/m+1},$$

and

$$Z_t = \max_{s\le t} \|z_s\| \,.$$

Proof. From Lemma F.1 and Definition 5.7 we have that

$$\epsilon^m \|\pi(z_s, \mathcal{B}_s)\| \le \sqrt{i(s)H_1H_3} \max_{1\le i\le i(s)} \left\|M_{t_i}^\top z_s\right\|,$$

which implies that

$$\epsilon^m \|\pi(z_s, \mathcal{B}_s)\| \le \sqrt{mH_1H_3} \max_{1\le i\le i(s)} \left\|M_{t_i}^\top z_s\right\|. \tag{F.8}$$

We can write

$$\begin{aligned}
\left\|M_s^\top z_s\right\| &= \left\|(\pi(M_s, \mathcal{B}_s^\perp) + \pi(M_s, \mathcal{B}_s))^\top (\pi(z_s, \mathcal{B}_s^\perp) + \pi(z_s, \mathcal{B}_s))\right\| \\
&= \left\|\pi(M_s, \mathcal{B}_s^\perp)^\top \pi(z_s, \mathcal{B}_s^\perp) + \pi(M_s, \mathcal{B}_s)^\top \pi(z_s, \mathcal{B}_s)\right\| \\
&\le \left\|\pi(M_s, \mathcal{B}_s^\perp)^\top \pi(z_s, \mathcal{B}_s^\perp)\right\| + \left\|\pi(M_s, \mathcal{B}_s)^\top \pi(z_s, \mathcal{B}_s)\right\| \\
&\le \epsilon m \|z_s\| + \frac{2S}{\epsilon^m}\sqrt{mH_1H_3} \max_{1\le i\le i(s)} \left\|M_{t_i}^\top z_s\right\|. \qquad \text{by (F.8) and (F.1)}
\end{aligned}$$

Thus,

$$\max_{s\le t, s\notin\mathcal{T}_t} \left\|M_s^\top z_s\right\| \le \epsilon m Z_t + \frac{2S}{\epsilon^m}\sqrt{mH_1H_3} \max_{s\notin\mathcal{T}_t,\, s\le t} \max_{1\le i\le i(s)} \left\|M_{t_i}^\top z_s\right\|.$$

From $1 \le i \le i(s), s \notin \mathcal{T}_t$, we conclude that $s < t_i$. Thus,

$$\max_{s\le t, s\notin\mathcal{T}_t} \left\|M_s^\top z_s\right\| \le \epsilon m Z_t + \frac{2S}{\epsilon^m}\sqrt{mH_1H_3} \max_{1\le s<t} \left\|M_t^\top z_s\right\|.$$

By Lemma F.2, we upper bound the second term on the RHS and get that

$$\max_{s\le t, s\notin\mathcal{T}_t} \left\|M_s^\top z_s\right\| \le \epsilon m Z_t + \frac{2S}{\epsilon^m}\sqrt{mH_1H_3\beta_t(\delta/4)} \,.$$

Finally, if we choose

$$\epsilon = \left(\frac{2S}{Z_t}\sqrt{\frac{\beta_t(\delta/4)H_3}{mH_1}}\right)^{1/m+1}$$

we get that

$$\max_{s\le t, s\notin\mathcal{T}_t} \left\|M_s^\top z_s\right\| \le 2\left(2S(Z_t m)^m \sqrt{m\beta_t(\delta/4)H_1H_3}\right)^{1/m+1}$$
$$= GZ_t^{m/m+1}\beta_t(\delta/4)^{1/2(m+1)},$$

which is the statement of the lemma. It remains to show that the choice of ϵ satisfies $\epsilon < 1$. From the definition of H_1, we have that

$$H_1 > \frac{4S^2H_2^2H_3}{m}\,.$$

Thus,

$$\left(\frac{4S^2H_2^2H_3}{mH_1}\right)^{1/2(m+1)} < 1$$

and so we also have that

$$\epsilon = \left(\frac{2S}{Z_t}\sqrt{\frac{\beta_t(\delta/4)H_3}{mH_1}}\right)^{1/m+1} < 1\,,$$

finishing the proof. □

Now we are ready to prove Lemma 5.8. We show that the event $E \cap F$ holds with high probability.

Proof of Lemma 5.8. We can write the state update as

$$x_{t+1} = \Gamma_t x_t + r_{t+1},$$

where

$$\Gamma_{t+1} = \begin{cases} \widetilde{A}_t + \widetilde{B}_t K(\widetilde{\Theta}_t) & t \notin \mathcal{T}_T \\ A_* + B_* K(\widetilde{\Theta}_t) & t \in \mathcal{T}_T \end{cases}$$

and

$$r_{t+1} = \begin{cases} M_t^\top z_t + w_{t+1} & t \notin \mathcal{T}_T \\ w_{t+1} & t \in \mathcal{T}_T \end{cases}$$

Thus, we obtain

$$\begin{aligned} x_t &= \Gamma_{t-1}x_{t-1} + r_t = \Gamma_{t-1}(\Gamma_{t-2}x_{t-2} + r_{t-1}) + r_t = \Gamma_{t-1}\Gamma_{t-2}x_{t-2} + r_t + \Gamma_{t-1}r_{t-1} \\ &= \Gamma_{t-1}\Gamma_{t-2}\Gamma_{t-3}x_{t-3} + r_t + \Gamma_{t-1}r_{t-1} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} = \cdots = \Gamma_{t-1}\ldots\Gamma_{t-t}x_{t-t} \\ &+ r_t + \Gamma_{t-1}r_{t-1} + \Gamma_{t-1}\Gamma_{t-2}r_{t-2} + \cdots + \Gamma_{t-1}\Gamma_{t-2}\ldots\Gamma_{t-(t-1)}r_{t-(t-1)} \\ &= \sum_{k=1}^{t}\left(\prod_{s=k}^{t-1}\Gamma_s\right) r_k\,. \end{aligned}$$

Thus,

$$\begin{aligned} \|x_t\| &\le \sum_{k=1}^{t}\left(\prod_{s=k}^{t-1}\|\Gamma_s\|\right)\|r_k\| \\ &\le \left(\frac{\Psi}{\Lambda}\right)^m \sum_{k=1}^{t}\Lambda^{t-k+1}\|r_k\| \\ &\le \frac{1}{1-\Lambda}\left(\frac{\Psi}{\Lambda}\right)^m \max_{1\le k\le t}\|r_k\|\,. \end{aligned}$$

where Λ and Ψ are defined in Assumption A3 and Definition 5.7. We have that $\|r_{k+1}\| \leq \left\|M_k^\top z_k\right\| + \|w_{k+1}\|$ when $k \notin \mathcal{T}_T$, and $\|r_{k+1}\| = \|w_{k+1}\|$, otherwise. Thus,

$$\max_{k<t} \|r_{k+1}\| \leq \max_{k<t,\, k\notin\mathcal{T}_t} \left\|M_k^\top z_k\right\| + \max_{k<t} \|w_{k+1}\| .$$

The first term on the RHS can be bounded by Lemma F.3. The second term can be bounded as follows: from the sub-Gaussianity Assumption A2, we have that for any index $1 \leq i \leq n$ and any time $k \leq t$, with probability $1 - \delta/(t(t+1))$

$$|w_{k,i}| \leq L\sqrt{2\log\frac{t(t+1)}{\delta}} .$$

As a result, with a union bound argument, on some event H with $\mathbb{P}(H) \geq 1-\delta/4$, $\|w_t\| \leq 2L\sqrt{n\log\frac{4nt(t+1)}{\delta}}$. Thus, on $H \cap E$,

$$\|x_t\| \leq \frac{1}{1-\Lambda}\left(\frac{\Psi}{\Lambda}\right)^m \left(GZ_t^{m/m+1}\beta_t(\delta/4)^{1/2(m+1)} + 2L\sqrt{n\log(4nt(t+1)/\delta)}\right) = \Upsilon_t .$$

By the definition of F, $H \cap E \subset F \cap E$. Because, by the union bound, $\mathbb{P}(H \cap E) \geq 1-\delta/2$, $\mathbb{P}(E \cap F) \geq 1-\delta/2$ also holds, finishing the proof. □

Proof of Lemma 5.9. Fix t. On F_t, $\widehat{x}_t \doteq \max_{1\leq s\leq t}\|x_s\| \leq \Upsilon_t$. With appropriate constants, this implies that

$$x \leq D_1\sqrt{\beta_t(\delta)}\log(t)x^{m/m+1} + D_2\sqrt{\log(t/\delta)},$$

or

$$x \leq \left(D_1\sqrt{\beta_t(\delta)}\log(t) + D_2\sqrt{\log(t/\delta)}\right)^{m+1}, \tag{F.9}$$

holds for $x = \widehat{x}_t$. Let X_t be the largest value of $x \geq 0$ that satisfies (F.9). Thus,

$$X_t \leq \left(D_1\sqrt{\beta_t(\delta)}\log(t) + D_2\sqrt{\log(t/\delta)}\right)^{m+1}. \tag{F.10}$$

Clearly, $\widehat{x}_t \leq X_t$. Because $\beta_t(\delta)$ is a function of $\log\det(\overline{V}_t)$, (F.10) has the form of

$$X_t \leq f(\log(X_t))^{m+1}. \tag{F.11}$$

Let $a_t = X_t^{1/(m+1)}$. Then, (F.11) is equivalent to

$$a_t \leq f(\log a_t^{m+1}) = f((m+1)\log a_t).$$

Let $c = \max(1, \max_{1\leq s\leq t}\|a_s\|)$. Assume that $t \geq \lambda m$. By the construction of F_t, Lemma E.2, tedious, but elementary calculations, it can then be shown that

$$c \leq A\log^2(c) + B_t, \tag{F.12}$$

where $A = G_1\log(1/\delta)$ and $B_t = G_2\log(t/\delta)$. From this, further elementary calculations show that the maximum value that c can take on subject to the constraint (F.12) is bounded from above by Y_t. □