# 19 Stochastic Linear Bandits

Contextual bandits generalize the finite-armed setting by allowing the learner to make use of side information. This chapter focusses on a specific type of contextual bandit problem in the stochastic setup where the reward is assumed to have a linear structure that allows for learning to transfer from one context to another. This leads to a useful and rich model that will be the topic of the next few chapters. To begin we describe the **stochastic linear bandit** problem and start the process of generalizing the upper confidence bound algorithm.

## 19.1 Stochastic contextual bandits

The stochastic contextual bandit problem mirrors the adversarial contextual bandit setup discussed in Chapter 18. At the beginning of round $t$ the learner observes a context $C_t \in \mathcal{C}$, which may be random or not. Having observed the context, the learner chooses their action $A_t \in [K]$ based on the information available. So far everything is the same as the adversarial setting. The difference comes from the assumption that the reward $X_t$ satisfies

$$X_t = r(C_t, A_t) + \eta_t \,,$$

where $r : \mathcal{C} \times [K] \to \mathbb{R}$ is called the **reward function** and $\eta_t$ is the noise, which we will assume is conditionally 1-subgaussian. Precisely, let

$$\mathcal{F}_t = \sigma(C_1, A_1, X_1, \ldots, C_{t-1}, A_{t-1}, X_{t-1}, C_t, A_t)$$

be the $\sigma$-field summarizing the information available just before $X_t$ is observed. Then we assume that

$$\mathbb{E}\left[\exp(\lambda \eta_t) \mid \mathcal{F}_t\right] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{almost surely}\,.$$

The noise could have been chosen to be $\sigma$-subgaussian for any known $\sigma^2$, but like in earlier chapters we save ourselves some ink by fixing its value to $\sigma^2 = 1$. Remember from Chapter 5 that subgaussian random variables have zero mean, so the assumption also implies that $\mathbb{E}\left[\eta_t \mid \mathcal{F}_t\right] = 0$ and $\mathbb{E}\left[X_t \mid \mathcal{F}_t\right] = r(C_t, A_t)$.

If $r$ was given, then the action in round $t$ with the largest expected return is $A_t^* \in \operatorname{argmax}_{a \in [K]} r(C_t, a)$. Notice that this action is now a random variable

because it depends on the context $C_t$. The loss due to the lack of knowledge of $r$ makes the learner incur the (expected) regret

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a \in [K]} r(C_t, a) - \sum_{t=1}^{n} X_t\right].$$

Like in the adversarial setting, there is one big caveat in this definition of the regret. Since we did not make any restrictions on how the contexts are chosen, it could be that choosing a low-rewarding action in the first round might change the contexts observed in subsequent rounds. Then the learner could potentially achieve an even higher cumulative reward by choosing a 'suboptimal' arm initially. As a consequence, this definition of the regret is most meaningful when the actions of the learner do not greatly affect subsequent contexts.

One way to eventually learn an optimal policy is to estimate $r(c, a)$ for each $(c, a) \in \mathcal{C} \times [K]$ pair. As in the adversarial setting, this is ineffective when the number of context-action pairs is large. In particular, the worst-case regret over all possible contextual problems with $M$ contexts and mean reward in $[0, 1]$ is at least $\Omega(\sqrt{nMK})$. While this may not look bad, $M$ is often exponentially large (for example, $2^{100}$). The argument for proving such worst-case lower bounds relies on designing a problem where knowledge of $r(c, \cdot)$ for context $c$ provides no useful information about $r(c', \cdot)$ for some different context $c'$. Fortunately, in most interesting applications the set of contexts is highly structured, which can often be captured by some kind of smoothness of $r(\cdot, \cdot)$.

A very simple idea is to assume the learner has access to a map $\psi : \mathcal{C} \times [K] \to \mathbb{R}^d$ and that there exists an unknown parameter vector $\theta_* \in \mathbb{R}^d$ such that

$$r(c, a) = \langle \psi(c, a), \theta_* \rangle, \qquad \forall (c, a) \in \mathcal{C} \times [K]. \tag{19.1}$$

The map $\psi$ is called a **feature-map**, which is the standard nomenclature in machine learning. The idea of feature maps is best illustrated with an example. Suppose the context denotes the visitor of a website selling books, the actions are books to recommend and the reward is the revenue on a book sold. The features could indicate the interests of the visitors as well as the domain and topic of the book. If the visitors and books are assigned to finitely many categories, indicator variables of all possible combinations of these categories could be used to create the feature map. Of course, many other possibilities exist. For example you can train a neural network (deep or not) on historical data to predict the revenue and use the nonlinear map that we obtained by removing the last layer of the neural network. The subspace $\Psi$ spanned by the **feature vectors** $\{\psi(c, a)\}_{c,a}$ in $\mathbb{R}^d$ is called the **feature-space**.

If $\|\cdot\|$ is a norm on $\mathbb{R}^d$, then an assumption on $\|\theta_*\|$ encodes **smoothness** of $r$. In particular, from Hölder's inequality,

$$|r(c, a) - r(c', a')| \le \|\theta_*\| \|\psi(c, a) - \psi(c', a')\|_*,$$

where $\|\cdot\|_*$ denotes the dual of $\|\cdot\|$. Restrictions on $\|\theta_*\|$ have a similar effect to assuming that the dimensionality $d$. In fact, one may push this to the extreme

and allow $d$ to be infinite, an approach which can buy tremendous flexibility and makes the linearity assumption less limiting.

## 19.2 Stochastic linear bandits

Stochastic linear bandits arise from realizing that when the reward is given by Eq. (19.1), then the identity of the actions becomes secondary. All that matters is the feature vector that results from choosing a given action. This justifies studying the following simplified model: In round $t$, the learner is given the decision set $\mathcal{A}_t \subset \mathbb{R}^d$ from which it chooses its action $A_t \in \mathcal{A}_t$ and receives reward

$$X_t = \langle A_t, \theta_* \rangle + \eta_t \,,$$

where $\eta_t$ is 1-subgaussian given $\mathcal{A}_1, A_1, X_1, \ldots, \mathcal{A}_{t-1}, A_{t-1}, X_{t-1}, \mathcal{A}_t$ and $A_t$. The random regret and regret are defined by

$$\hat{R}_n = \sum_{t=1}^{n} \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - \sum_{t=1}^{n} X_t \,.$$

$$R_n = \mathbb{E}\left[\hat{R}_n\right] = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - \sum_{t=1}^{n} X_t\right] \,.$$

Different choices of $\mathcal{A}_t$ lead to different settings, some of which we have seen before. For example, if $(e_i)_i$ are the unit vectors and $\mathcal{A}_t = \{e_1, \ldots, e_d\}$, then the resulting stochastic linear bandit problem reduces to the finite-armed setting. On the other hand, if $\mathcal{A}_t = \{\psi(C_t, k) : k \in [K]\}$, then we have a contextual linear bandit. Yet another possibility is a **combinatorial action set** like $\mathcal{A}_t \subseteq \{0,1\}^d$. Many combinatorial problems (such as matching, least-cost problems in directed graphs and choosing spanning trees) can be written as linear optimization problems over some combinatorial set $\mathcal{A}$ obtained from considering incidence vectors often associated with some graph. Some of these topics will be covered later in Chapter 30.

As we have seen in earlier chapters, the UCB algorithm is an attractive approach for finite-action stochastic bandits. Its best variants are nearly minimax optimal, instance optimal and exactly optimal asymptotically. With these merits in mind, it seems quite natural to try and generalize the idea to the linear setting.

The generalization is based on the view that UCB implements the 'optimism in the face of uncertainty' principle, which is to act in each round as if the environment is as nice as plausibly possible. In finite-action stochastic bandits this means choosing the action with the largest upper confidence bound. In the case of linear bandits the idea remains the same, but the form of the confidence bound is more complicated because rewards received yield information about more than just the arm played.

The first step is to construct a confidence set $\mathcal{C}_t \subset \mathbb{R}^d$ based on $(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ that contains the unknown parameter vector $\theta_*$ with

high probability. Leaving the details of how the confidence set is constructed aside for a moment and assuming that the confidence set indeed contains $\theta_*$, then for any given action $a \in \mathbb{R}^d$,

$$\mathrm{UCB}_t(a) = \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle \tag{19.2}$$

will be an upper bound on the mean payoff of $a$, which is $\langle a, \theta_* \rangle$. The UCB algorithm that uses the confidence set $\mathcal{C}_t$ at time $t$ then selects

$$A_t = \mathrm{argmax}_{a \in \mathcal{A}_t} \mathrm{UCB}_t(a)\,. \tag{19.3}$$

Depending on the authors, UCB applied to linear bandits is known by many names, including LinRel (**Lin**ear **R**einforcement **L**earning), LinUCB and OFUL (**O**ptimism in the **F**ace of **U**ncertainty for **L**inear bandits).

The main question is how to choose the confidence set $\mathcal{C}_t \subset \mathbb{R}^d$. As usual, there are conflicting desirable properties:

(a) $\mathcal{C}_t$ should contain $\theta_*$ with high probability.
(b) $\mathcal{C}_t$ should be as small as possible.

At first sight it is not at all obvious what $\mathcal{C}_t$ should look like. After all, it is a subset of $\mathbb{R}^d$, not just an interval like the confidence intervals about the empirical estimate of the mean reward for a single action that we saw in the previous chapters. We will leave the details to the next chapter, but sketch the basic approach here. Following the idea for UCB, we need an analogue for the empirical estimate of the unknown quantity, which in this case is $\theta^*$. There are several principles one might use for deriving such an estimate. For now we use the **regularized least-squares estimator**, which is

$$\hat{\theta}_t = \mathrm{argmin}_{\theta \in \mathbb{R}^d} \left( \sum_{s=1}^{t} (X_s - \langle A_s, \theta \rangle)^2 + \lambda \|\theta\|_2^2 \right), \tag{19.4}$$

where $\lambda \geq 0$ is called the **penalty factor**. Choosing $\lambda > 0$ helps because it ensures that the loss function has a unique minimizer even when $A_1, \ldots, A_t$ do not span $\mathbb{R}^d$, which simplifies the math. The solution to Eq. (19.4) is obtained easily by differentiation and is

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s\,, \tag{19.5}$$

where $V_t$ is a $d \times d$ matrices given by

$$V_0 = \lambda I \qquad \text{and} \qquad V_t = V_0 + \sum_{s=1}^{t} A_s A_s^\top\,.$$

The matrix $V_t - V_0$ is called the **Grammian** while $V_t$ is sometimes called the **regularized Grammian**. So $\hat{\theta}_t$ is an estimate of $\theta_*$, which makes it natural to

choose $\mathcal{C}_t$ to be centered at $\hat{\theta}_{t-1}$. For what follows we will simply assume that the confidence set $\mathcal{C}_t$ is closed and satisfies

$$\mathcal{C}_t \subseteq \mathcal{E}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|^2_{V_{t-1}} \leq \beta_t \right\}, \tag{19.6}$$

where $(\beta_t)_t$ is a sequence of monotone nondecreasing constants with $\beta_1 \geq 1$. The set $\mathcal{E}_t$ is an ellipsoid centered at $\hat{\theta}_{t-1}$ and with principle axis being the eigenvectors of $V_t$ with corresponding lengths being the reciprocal of the eigenvalues. Notice that as $t$ grows the matrix $V_t$ has increasing eigenvalues, which means the volume of the ellipse is also shrinking (at least, provided $\beta_t$ does not grow too fast). In the next chapter we will see that $\mathcal{C}_t = \mathcal{E}_t$ is a natural choice for carefully chosen $\beta_t$, but for the rest of this chapter we will simply examine the consequence of using a confidence set satisfying Eq. (19.6) and assume all the desirable properties.

## 19.3 Regret analysis

We prove a regret bound for LinUCB under the assumption that the confidence intervals indeed contain the true parameter with high probability and boundedness conditions on the action set and rewards.

ASSUMPTION 19.1 The following hold:

(a) $|\langle a, \theta_* \rangle| \leq 1$ for any $a \in \cup_t \mathcal{A}_t$.
(b) For any $a \in \cup_t \mathcal{A}_t$, $\|a\|_2 \leq L$.
(c) There exists a $\delta \in (0,1)$ such that with probability $1-\delta$, for all $t \in [n]$, $\theta_* \in \mathcal{C}_t$ where $\mathcal{C}_t$ satisfies Eq. (19.6).

THEOREM 19.1 *Under the conditions of Assumption 19.1 with probability* $1-\delta$ *the regret of LinUCB satisfies*

$$\hat{R}_n \leq \sqrt{8 n \beta_n \log\left(\frac{\det V_n}{\det V_0}\right)} \leq \sqrt{8 d n \beta_n \log\left(\frac{\mathrm{trace}(V_0) + nL^2}{d \det^{\frac{1}{d}}(V_0)}\right)}.$$

Provided that $\beta_n$ has polylogarithmic growth, then $\hat{R}_n = \tilde{O}(\sqrt{n})$, which matches the worst-case rate for finite-armed bandits except for logarithmic factors. We can also get a bound on the (expected) regret $R_n$ if $\delta \leq c/\sqrt{n}$ and by combining the theorem with trivial fact that $\hat{R}_n \leq 2n$, which follows from our assumption that the magnitude of the immediate reward is bounded by one. The proof of Theorem 19.1 depends on the following lemma.

LEMMA 19.1 *Let* $V_0$ *be positive definite and* $v_0 = \mathrm{trace}(V_0)$ *and* $x_1, \ldots, x_n \in \mathbb{R}^d$ *be a sequence of vectors with* $\|x_t\|_2 \leq L < \infty$ *for all* $t \in [n]$. *Then*

$$\sum_{t=1}^{n} \left(1 \wedge \|x_t\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(\frac{\det V_n}{\det V_0}\right) \leq 2d \log\left(\frac{v_0 + nL^2}{d \det^{1/d}(V_0)}\right).$$

*Proof* Using that for any $u \in [0,1]$, $u \wedge 1 \le 2\ln(1+u)$, we get

$$\sum_{t=1}^{n} \left(1 \wedge \|x_t\|^2_{V_{t-1}^{-1}}\right) \le 2 \sum_{t} \log\left(1 + \|x_t\|^2_{V_{t-1}^{-1}}\right) .$$

We now argue that this last expression is $\log\left(\frac{\det V_n}{\det V_0}\right)$. For $t \ge 1$ we have

$$V_t = V_{t-1} + x_t x_t^\top = V_{t-1}^{1/2}(I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2}) V_{t-1}^{1/2} .$$

Hence

$$\det(V_t) = \det(V_{t-1}) \det\left(I + V_{t-1}^{-1/2} x_t x_t^\top V_{t-1}^{-1/2}\right) = \det(V_{t-1})\left(1 + \|x_t\|^2_{V_{t-1}^{-1}}\right) ,$$

where the second equality follows because the matrix $I + yy^\top$ has eigenvalues $1 + \|y\|_2^2$ and 1 as well as the fact that the determinant of a matrix is the product of its eigenvalues. Putting things together we see that

$$\det(V_n) = \det(V_0) \prod_{t=1}^{n} \left(1 + \|x_t\|^2_{V_{t-1}^{-1}}\right) ,$$

which is equivalent to the first inequality that we wanted to prove. To get the second inequality note that by the inequality of arithmetic and geometric means,

$$\det(V_n) = \prod_{i=1}^{d} \lambda_i \le \left(\frac{1}{d} \operatorname{trace} V_n\right)^d \le \left(\frac{v_0 + nL^2}{d}\right)^d ,$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $V_n$. □

*Proof of Theorem 19.1* By part (c) of Assumption 19.1 it suffices to prove the bound on the event that $\theta_* \in \mathcal{C}_t$ for all rounds $t \in [n]$. Let $A_t^* = \operatorname{argmax}_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle$ be an optimal action for round $t$ and $r_t$ be the instantaneous regret in round $t$ defined by

$$r_t = \langle A_t^* - A_t, \theta_* \rangle .$$

Let $\tilde{\theta}_t \in \mathcal{C}_t$ be the parameter in the confidence set for which $\langle A_t, \tilde{\theta}_t \rangle = \mathrm{UCB}_t(A_t)$. Then using the fact that $\theta_* \in \mathcal{C}_t$ and the definition of the algorithm leads to

$$\langle A_t^*, \theta_* \rangle \le \mathrm{UCB}_t(A_t^*) \le \mathrm{UCB}_t(A_t) = \langle A_t, \tilde{\theta}_t \rangle .$$

Using Cauchy-Schwartz inequality and the assumption that $\theta_* \in \mathcal{C}_t$ and facts that $\tilde{\theta}_t \in \mathcal{C}_t$ and $\mathcal{C}_t \subseteq \mathcal{E}_t$ leads to

$$r_t = \langle A_t^* - A_t, \theta_* \rangle \le \langle A_t, \tilde{\theta}_t - \theta_* \rangle \le \|A_t\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \theta_*\|_{V_{t-1}} \le 2\|A_t\|_{V_{t-1}^{-1}} \sqrt{\beta_t} .$$

By part (a) we also have $r_t \le 2$, which combined with $\beta_n \ge \max\{1, \beta_t\}$ yields

$$r_t \le 2 \wedge 2\sqrt{\beta_t} \|A_t\|_{V_{t-1}^{-1}} \le 2\sqrt{\beta_n}(1 \wedge \|A_t\|_{V_{t-1}^{-1}}) .$$

Jensen's inequality shows that

$$\hat{R}_n = \sum_{t=1}^{n} r_t \leq \sqrt{n \sum_{t=1}^{n} r_t^2} \leq 2\sqrt{n\beta_n \sum_{t=1}^{n} (1 \wedge \|A_t\|^2_{V_{t-1}^{-1}})} \,.$$

The result is completed using Lemma 19.1, which depends on part (b) of Assumption 19.1. □

## 19.4 Notes

1 It was mentioned that $\psi$ may map its arguments to an infinite dimensional space. There are several issues that arise in this setting. The first is whether or not the algorithm can be computed efficiently, which is usually tackled via the **kernel trick**, which assumes the existence of an efficiently computable **kernel function** $\kappa : (\mathcal{C} \times [K]) \times (\mathcal{C} \times [K]) \to \mathbb{R}$ such that

$$\langle \psi(c, a), \psi(c', a') \rangle = \kappa((c, a), (c', a')) \,.$$

Then all operations are written in terms of the kernel function so that $\psi(c, a)$ never needs to be computed or stored. The second issue is that the statement of Theorem 19.1 depends on the dimension $d$ and becomes vacuous when $d$ is large or infinite. This dependence arises from Lemma 19.1, which must be replaced with a data-dependent quantity that measures the 'effective dimension' of the image of the data under $\phi$. The final challenge is to define an appropriate confidence set. These issues have not yet been resolved in a complete way. See the bibliographic remarks for further references.

2 The bound given in Theorem 19.1 is essentially a worst-case style of bound, with little dependence on the parameter $\theta_*$ or the geometry of the action-set. Instance-dependent bounds for linear bandits are still an open topic of research, and the asymptotics are only understood in the special case where the action set is finite and unchanging (Chapter 25).

3 An obvious question is whether or not the optimization problem in Eq. (19.3) can be solved efficiently. First note that the computation of $A_t$ can also be written as

$$(A_t, \tilde{\theta}_t) = \operatorname{argmax}_{(a,\theta) \in \mathcal{A}_t \times \mathcal{C}_t} \langle a, \theta \rangle \,. \tag{19.7}$$

This is a bilinear optimization problem over the set $\mathcal{A}_t \times \mathcal{C}_t$. In general, nothing much can be said about the computational efficiency of solving this problem. There are two notable special cases:

(c) If the linear optimization problem $\max_{a \in \mathcal{A}_t} \langle a, \theta \rangle$ can be efficiently solved for any $\theta$ and $\mathcal{C}_t$ is the convex hull of a small number of vertices: $\mathcal{C}_t = \operatorname{co}(c_{t1}, \ldots, c_{tp})$. Then it is easy to verify that the solution to Eq. (19.7) has the form $(a, c_{ti})$ for some $i \in [p]$. Hence the solution may be found by solving $\max_{a \in \mathcal{A}_t} \langle a, c_{t1} \rangle, \ldots, \max_{a \in \mathcal{A}_t} \langle a, c_{tp} \rangle$.

(c) If $\mathcal{C}_t = \mathcal{E}_t$ is the ellipsoid given in Eq. (19.6) and $\mathcal{A}_t$ is a small finite set. Then the action $A_t$ from Eq. (19.7) can be found using

$$A_t = \operatorname{argmax}_a \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \|a\|_{V_{t-1}^{-1}} , \tag{19.8}$$

which may be solved by simply iterating over the arms and calculating the term inside the argmax.

4 The previous note highlights the fact that the algorithm presented in this section has more than just a passing resemblance to the UCB algorithm introduced in earlier chapters on finite-armed bandits. The term $\langle a, \hat{\theta}_t \rangle$ may be interpreted as an empirical estimate of the reward from choosing action $a$ and $\sqrt{\beta_t} \|a\|_{V_{t-1}^{-1}}$ is a bonus term that ensures sufficient exploration. If the penalty term vanishes ($\lambda = 0$) and $\mathcal{A}_t = \{e_1, \ldots, e_d\}$ for all $t \in [n]$, then $\hat{\theta}_i$ becomes the empirical mean of action $e_i$ and the matrix $V_t$ is diagonal with its $i$ diagonal entry being the number of times action $e_i$ is used up to and including round $t$. Then the bonus term has order

$$\sqrt{\beta_t} \|e_i\|_{V_{t-1}^{-1}} = \sqrt{\frac{\beta_t}{T_i(t-1)}} ,$$

where $T_i(t-1)$ is the number of times action $e_i$ has been chosen before the $t$th round. So UCB for finite-armed bandits is recovered by choosing $\beta_t = 2 \log(\cdot)$, where the term inside the logarithm can be chosen in a variety of ways as discussed in earlier chapters. Notice now that the simple analysis given in this chapter leads to a regret bound of $O(\sqrt{dn \log(\cdot)})$, which is quite close to the highly specialized analysis given in Chapters 7 to 9.

5 A practical extension of the linear model is the **generalized linear model** where the reward is

$$X_t = g^{-1}(\langle A_t, \theta_* \rangle + \eta_t) , \tag{19.9}$$

where $g : \mathbb{R} \to \mathbb{R}$ is called the **link function**. A common choice is $g(p) = \log(p/(1-p))$, which yields the sigmoid function as the inverse: $g^{-1}(x) = 1/(1 + \exp(-x))$. Bandits with rewards from a generalized linear model have been studied by Filippi et al. [2010], who prove a bound with a similar form as Theorem 19.1. Unfortunately, however, the bound depends in a slightly unpleasant manner on the form of the link function and it seems there may be significant room for improvement.

## 19.5 Bibliographic remarks

Stochastic linear bandits were introduced by Abe and Long [1999]. The first paper to consider algorithms based on the optimism principle for linear bandits is by Auer [2002], who considered the case when the number of actions is finite. The core ideas of the analysis of optimistic algorithms (and more) is already present in

this paper. An algorithm based on confidence ellipsoids is described in the papers by Dani et al. [2008], Rusmevichientong and Tsitsiklis [2010], Abbasi-yadkori et al. [2011]. The regret analysis presented here and the discussion of the computational questions is largely based on the former of these works, which also stresses that an expected regret of $\tilde{O}(d\sqrt{n})$ can be achieved regardless of the shape of the decision sets $\mathcal{A}_t$ as long as the means are guaranteed to lie in a bounded interval. Rusmevichientong and Tsitsiklis [2010] consider both optimistic and explore-then-commit strategies which they call "phased exploration and greedy exploitation" (PEGE). They focus on the case where $\mathcal{A}_t$ is the unit ball and show that PEGE is optimal up to logarithmic factors. The observation that explore-then-commit works for the unit ball (and other action sets with a smooth boundary) was independently made by Abbasi-Yadkori et al. [2009], Abbasi-Yadkori [2009a]. Generalized linear models are credited to Nelder and Wedderburn [1972]. We mentioned already that LinUCB was generalized to this model by Filippi et al. [2010]. A more computationally efficient algorithm has recently been proposed by Jun et al. [2017]. Nonlinear structured bandits where the payoff function belongs to a known set has also been studied [Anantharam et al., 1987, Russo and Van Roy, 2013, Lattimore and Munos, 2014]. The kernelized version of UCB is by Valko et al. [2013]. We mentioned early in the chapter that making assumptions on the norm $\theta_*$ is related to smoothness of the reward function with smoother functions leading to stronger guarantees. For an example of where this is done see the paper on 'spectral bandits' by Valko et al. [2014].

## 19.6 Exercises

**19.1** Prove that the solution given in Eq. (19.5) is indeed the minimizer of Eq. (19.4).

**19.2** Let $V_0 = \lambda I$ and $x_1, \dots, x_n \in \mathbb{R}^d$ be a sequence of vectors with $\|x_t\|_2 \leq L$ for all $t \in [n]$. Then let $V_t = V_0 + \sum_{s=1}^{t} x_s x_s^\top$ and show that the number of times $\|x_t\|_{V_{t-1}^{-1}} \geq 1$ is at most

$$\frac{3d}{\log(2)} \log\left(1 + \frac{L^2}{\lambda \log(2)}\right).$$

The proof of Theorem 19.1 depended on part (a) of Assumption 19.1, which asserts that the mean rewards are bounded by 1. Suppose we replace this assumption with the relaxation that there exists a $B > 0$ such that

$$\max_{t \in [n]} \sup_{a,b \in \mathcal{A}_t} \langle a - b, \theta_* \rangle \leq B.$$

Then the previous exercise allows you to bound the number of rounds when $\|x_t\|_{V_{t-1}^{-1}} \geq 1$ and in these rounds the naive bound of $r_t \leq B$ is used. For the

remaining rounds the analysis of Theorem 19.1 goes through unaltered. As a consequence we see that the dependence on $B$ is an additive constant term that does not grow with the horizon.