23 Stochastic Linear Bandits with Sparsity

In Chapter 19 we showed the linear variant of UCB has regret bounded by

$$R_n = O(d\sqrt{n}\log(n))\,,$$

which for fixed finite action sets can be improved to

$$R_n = \tilde O(\sqrt{dn\log(nK)})\,.$$

For moderately sized action sets these approaches lead to a big improvement over what could be obtained by using the policies that do not make use of the linear structure.

The situation is still not perfect though. In typical applications d is the dimension of the feature space in which the actions are embedded. The features are chosen by the users of the system and one can easily imagine the user has many candidate features and little knowledge about which will be most useful. This presents the user with a challenging tradeoff. If they include many features, then the regret bound will be large. But if a useful feature is omitted, then the linear model will almost certainly be quite wrong. Ideally, one should be able to add features without suffering much additional regret if the feature added does not contribute in a significant way. This can be captured by the notion of sparsity, which is the central theme of this chapter.

23.1 Sparse linear stochastic bandits

The sparse linear stochastic bandit problem is the same as the stochastic linear bandit problem with a small difference. Just like in the standard setting, at the beginning of a round with index t the learner receives a decision set $\mathcal{A}_t \subset \mathbb{R}^d$. They then choose an action $A_t \in \mathcal{A}_t$ and receives the reward

$$X_t = \langle A_t, \theta_* \rangle + \eta_t\,, \tag{23.1}$$

where $(\eta_t)_t$ is zero-mean noise and $\theta_* \in \mathbb{R}^d$ is an unknown vector. The only difference in the sparse setting is that the parameter vector θ is assumed to have many zero entries. Given $\theta \in \mathbb{R}^d$ let

$$\|\theta\|_0 = \sum_{i=1}^d \mathbb{I}\left\{\theta_i \neq 0\right\}\,,$$

which is sometimes call the 0-“norm” (quotations because it is not really a norm, see Exercise 23.1). For the remainder of this chapter we will assume that

1 *(Sparse parameter)* There exist known constants M_0 and M_2 such that $\|\theta_*\|_0 \leq M_0$ and $\|\theta_*\|_2 \leq M_2$.
2 *(Bounded mean rewards):* $\langle a, \theta_* \rangle \leq 1$ for all $a \in \mathcal{A}_t$ and all rounds t.
3 *(Subgaussian noise):* The reward is $X_t = \langle A_t, \theta_* \rangle + \eta_t$ where $\eta_t | \mathcal{F}_{t-1} \sim \text{subG}(1)$ for $\mathcal{F}_t = \sigma(A_1, \eta_1, \dots, A_t, \eta_t)$.

Much ink has been spilled on what can be said about the speed of learning in linear models like (23.1) when $(A_t)_t$ are passively generated and the parameter vector is known to be sparse. Most results are phrased about recovering θ_*, but there also exist a few results that quantify the speed at which good predictions can be made. The ideal outcome would be that the learning speed depends mostly on M_0, while the dependence on d becomes less severe. Almost all the results come under the assumption that the Grammian of the actions $(A_t)_t$ is well-conditioned.

The **condition number** of a positive definite matrix A is the ratio of its largest and smallest eigenvalues. A matrix is **well conditioned** if it has a small condition number.

The details are a bit more complicated than just the conditioning, but the main point is that the usual assumptions imposed on the Grammian for passive learning are never satisfied when the actions are chosen by a good bandit policy. The reason is simple. Bandit algorithms want to choose the optimal action as often as possible, which means the Grammian will have an eigenvector that points (approximately) towards to optimal action with a large corresponding eigenvalue. We need some approach that does not rely on such strong assumptions.

23.2 Elimination on the hypercube

As a warmup problem we consider the special case where the action set is the d-dimensional hypercube: $\mathcal{A} = [-1, 1]^d$. To reduce clutter we will denote the true parameter vector by θ. As usual, in each round t the learner chooses $A_t \in \mathcal{A}$ and receives reward $X_t = \langle A_t, \theta \rangle + \eta_t$. We make the following standard assumptions:

1 *(Bounded mean rewards):* $\|\theta\|_1 \leq 1$, which ensures that $|\langle a, \theta \rangle| \leq 1$ for all $a \in \mathcal{A}$.
2 *(Subgaussian noise):* η_t is conditionally 1-subgaussian given the past:

$$\mathbb{E}\left[\exp(\lambda \eta_t) | \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2}{2}\right) \quad \text{almost surely for all } \lambda \in \mathbb{R}\,,$$

where $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t)$.

Since conditional subgaussianity comes up frequently, we introduce a notation for it. When X is σ-subgaussian given some σ-field $\mathcal{F}$ we will write $X|\mathcal{F} \sim \operatorname{subG}(\sigma)$. Our earlier statement that the sum of independent subgaussian random variables is subgaussian with a subgaussianity factor that is the sum of the two factors also holds for conditionally subgaussian random variables.

The hypercube is notable as an action set because it enjoys perfect separability. For each dimension $i \in [d]$ the value of $A_{ti} \in [-1, 1]$ can be chosen without regard to the choice of A_{tj} for other dimensions j. The first consequence of this is that the optimal action is $a^* = \operatorname{sign}(\theta)$ where

$$\operatorname{sign}(\theta)_i = \operatorname{sign}(\theta_i) = \begin{cases} 1 & \text{if } \theta_i > 0 \\ 0 & \text{if } \theta_i = 0 \\ -1 & \text{if } \theta_i < 0\,. \end{cases}$$

So learning the optimal action amounts to learning the sign of θ_i for each dimension. A disadvantage of this structure is that in the worst case the sign of each θ_i must be learned independently, which in Chapter 24 we show leads to a worst case regret of $R_n = \Omega(d\sqrt{n})$. On the positive side, the seperability means that θ_i can be estimated in each dimension independently while paying absolutely no price for this experimentation when $\theta_i = 0$. It turns out that this allows us to design a policy whose regret scales with $O(\|\theta\|_0 \sqrt{n})$ even without knowing the value of $\|\theta\|_0$.

Let $\mathcal{G}_t = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ be the σ-algebra containing information up to time $t-1$ (this differs from $\mathcal{F}_t$, which also includes information about the action chosen). Now suppose that $(A_{ti})_i$ are chosen to be conditionally independent given $\mathcal{G}_t$ and further assume for some specific $i \in [d]$ that A_{ti} is sampled from a Rademacher distribution so that $\mathbb{P}(A_{ti} = 1 \mid \mathcal{G}_t) = \mathbb{P}(A_{ti} = -1 \mid \mathcal{G}_t) = 1/2$. Then

$$\begin{aligned} \mathbb{E}[A_{ti} X_t \mid \mathcal{G}_t] &= \mathbb{E}\left[A_{ti}\left(\sum_{j=1}^d A_{tj}\theta_j + \eta_t\right)\right] \\ &= \theta_i \mathbb{E}[A_{ti}^2 \mid \mathcal{G}_t] + \sum_{j \neq i} \theta_j \mathbb{E}[A_{tj} A_{ti} \mid \mathcal{G}_t] + \mathbb{E}[\eta_t \mid \mathcal{G}_t] = \theta_i\,, \end{aligned}$$

where the first equality is the definition of $X_t = \langle A_t, \theta\rangle + \eta_t$, the second by linearity of expectation and the third by the conditional independence of $(A_{ti})_i$ and the fact that $\mathbb{E}[A_{ti} \mid \mathcal{G}_t] = 0$ and $\mathbb{E}[A_{ti}^2 \mid \mathcal{G}_t] = 1$. This looks quite promising, but we should also check the variance. Using our assumptions we have: $\mathbb{E}[\eta] = 0$ and $\mathbb{E}[\eta^2] \leq 1$ and $\langle a, \theta\rangle \leq 1$ for all actions a we have

$$\mathbb{V}[A_{ti} X_t \mid \mathcal{G}_t] = \mathbb{E}[A_{ti}^2 X_t^2 \mid \mathcal{G}_t] - \theta_i^2 = \mathbb{E}[(\langle A_t, \theta\rangle + \eta)^2 \mid \mathcal{G}_t] - \theta_i^2 \leq 2\,. \tag{23.2}$$

And now we have cause for celebration. The value of θ_i can be estimated by choosing A_{ti} to be a Rademacher random variable independent of the choices in other dimensions. All the policy does is treat all dimensions independently. For a particular dimension (say i) it explores by choosing $A_{ti} \in \{-1, 1\}$ uniformly at random until its estimate is sufficiently accurate to commit to either $A_{ti} = 1$ or $A_{ti} = -1$ for all future rounds. How long this takes depends on $|\theta_i|$, but note that if $|\theta_i|$ is small, then the price of exploring is also limited. The policy that results from this idea is called Selective Explore-Then-Commit (Algorithm 12, SETC).

```
Input $n$ and $d$
Set $E_{1,i} = 1$ and $\mathcal{C}_{1,i} = \mathbb{R}$ for all $i \in [d]$
for $t = 1, \dots, n$ do
    For each $i \in [d]$ sample $B_{ti} \sim \textsc{Rademacher}$
    Choose action:
        $(\forall i) \qquad A_{ti} = \begin{cases} B_{ti} & \text{if } 0 \in \mathcal{C}_{ti} \\ 1 & \text{if } \mathcal{C}_{ti} \subset (0, \infty] \\ -1 & \text{if } \mathcal{C}_{ti} \subset [-\infty, 0)\,. \end{cases}$
    Play $A_t$ and observe $X_t$
    Construct empirical estimators:
        $(\forall i) \qquad T_i(t) = \sum_{s=1}^t E_{si} \qquad \hat\theta_{ti} = \frac{\sum_{s=1}^t E_{si} A_{si} X_s}{T_i(t)}$
    Construct confidence intervals:
        $(\forall i) \qquad W_{ti} = 2\sqrt{\left(\frac{1}{T_i(t)} + \frac{1}{T_i(t)^2}\right) \log\left(n\sqrt{2T_i(t) + 1}\right)}$
        $(\forall i) \qquad \mathcal{C}_{t+1,i} = \left[\hat\theta_{ti} - W_{ti}, \hat\theta_{ti} + W_{ti}\right]$
    Update exploration parameters:
        $(\forall i) \qquad E_{t+1,i} = \begin{cases} 0 & \text{if } 0 \notin \mathcal{C}_{t+1,i} \text{ or } E_{ti} = 0 \\ 1 & \text{otherwise}\,. \end{cases}$
end for
```

Algorithm 12: Selective Explore-Then-Commit

THEOREM 23.1 *There exists a universal constant $C > 0$ such that the regret of SETC satisfies:*

$$R_n \leq 3\|\theta\|_1 + C \sum_{i:\theta_i \neq 0} \frac{\log(n)}{|\theta_i|}\,.$$

Furthermore $R_n \leq C\|\theta\|_0 \sqrt{n \log(n)}$.

By appealing to the central limit theorem and the variance calculation in Eq. (23.2) we should be hopeful that the confidence intervals used by the algorithm are sufficiently large to contain the true θ_i with high probability, but this still needs to be proven.

LEMMA 23.1 *Define $\tau_i = n \wedge \max\{t : E_{ti} = 1\}$ and $F_i = \mathbb{I}\{\tau_i \leq n \wedge \theta_i \notin \mathcal{C}_{\tau_i+1,i}\}$ be the event that θ_i is not in the confidence interval constructed at time τ_i. Then $\mathbb{P}(F_i) \leq 1/n$.*

Leaving the proof of Lemma 23.1 to the next section, we first use it to prove Theorem 23.1.

Proof of Theorem 23.1 Recalling the definition of the regret and using the fact that the optimal action is $a^* = \operatorname{sign}(\theta)$ we have the following regret decomposition.

$$R_n = \max_{a\in\mathcal{A}}\langle a, \theta\rangle - \mathbb{E}\left[\sum_{t=1}^n \langle A_t, \theta\rangle\right] = \sum_{i=1}^d \underbrace{\left(n|\theta_i| - \mathbb{E}\left[\sum_{t=1}^n A_{ti}\theta_i\right]\right)}_{R_{ni}} . \tag{23.3}$$

Clearly if $\theta_i = 0$, then $R_{ni} = 0$. And so it suffices to bound R_{ni} for each i with $|\theta_i| > 0$. Suppose that $|\theta_i| > 0$ for some i and the failure event F_i given in Lemma 23.1 does not occur. Then $\theta_i \in \mathcal{C}_{\tau_i+1,t}$ and by definition of the algorithm $A_{ti} = \operatorname{sign}(\theta_i)$ for all $t \geq \tau_i$. Therefore

$$\begin{aligned} R_{ni} &= n|\theta_i| - \mathbb{E}\left[\sum_{t=1}^n A_{ti}\theta_i\right] = \mathbb{E}\left[\sum_{t=1}^n |\theta_i|\left(1 - A_{ti}\operatorname{sign}(\theta_i)\right)\right] \\ &\leq 2n|\theta_i|\mathbb{P}(F_i) + 2|\theta_i|\mathbb{E}\left[\mathbb{I}\{F_i^c\}\tau_i\right] \end{aligned} \tag{23.4}$$

Since τ_i is the last round t when $0 \notin \mathcal{C}_{t+1,i}$ it follows that if F_i does not occur, then $\theta \in \mathcal{C}_{\tau_i,i}$ and $0 \in \mathcal{C}_{\tau_i,i}$. Thus the width of the confidence interval $\mathcal{C}_{\tau_i,i}$ must be at least $|\theta_i|$ and so

$$2W_{\tau_i-1} = 4\sqrt{\left(\frac{1}{\tau_i-1} + \frac{1}{(\tau_i-1)^2}\right)\log\left(n\sqrt{2\tau_i-1}\right)} \geq |\theta_i| ,$$

which after rearranging shows for some universal constant $C > 0$ that

$$\mathbb{I}\{F_i^c\}(\tau_i - 1) \leq 1 + \frac{C\log(n)}{\theta_i^2} .$$

Combining this result with Eq. (23.4) leads to

$$R_{ni} \leq 2n|\theta_i|\mathbb{P}(F_i) + 2|\theta_i| + \frac{C\log(n)}{|\theta_i|} .$$

Using Lemma 23.1 to bound $\mathbb{P}(F_i)$ and substituting into the decomposition Eq. (23.3) completes the proof of the first part. The second part is left as an exercise to the reader. □

23.3 Proof of technical lemma

We start with a simple variation on the self-normalized concentration inequality of Theorem 20.1.

LEMMA 23.2 *Let $\delta \in (0,1)$ and $(\mathcal{F}_t)_{t\in[n]}$ be a filtration and $(Z_t)_{t\in[n]}$ be $\mathcal{F}_t$-adapted such that $Z_t|\mathcal{F}_{t-1} \sim \operatorname{subG}(\sigma)$. Then for any stopping time $\tau \in [n]$ it holds that*

$$\mathbb{P}\left(\text{exists } t \leq \tau : |S_t| \geq \sqrt{2\sigma^2(t+1)\log\left(\frac{\sqrt{t\sigma^2+1}}{\delta}\right)}\right) \leq \delta\,.$$

Proof Let $f(\lambda) = \frac{1}{\sqrt{2\pi}}\exp(-\lambda^2/2)$ be the density of the standard Gaussian and define supermartingale M_t by

$$M_t = \int_{\mathbb{R}} f(\lambda)\exp\left(\lambda S_t - \frac{t\sigma^2\lambda^2}{2}\right)d\lambda = \frac{1}{\sqrt{t\sigma^2+1}}\exp\left(\frac{S_t^2}{2\sigma^2(t+1)}\right)\,.$$

Since $\mathbb{E}[M_\tau] = M_0 = 1$, the maximal inequality shows that $\mathbb{P}\left(\sup_{t\leq\tau} M_t \geq 1/\delta\right) \leq \delta$, which after rearranging the previous display completes the result. □

One might question whether or not the choice of Gaussian for mixing distribution f is optimal. In fact it is nothing more than a convenient choice that allows for an easy evaluation of the integral. By selecting a more appropriate mixing distribution one can show a result that is reminiscent of the law of the iterated logarithm. For details see Exercise 23.6.

Proof of Lemma 23.1 Let $Z_{ti} = A_{ti}\eta_t + A_{ti}\sum_{j\neq i} A_{tj}\theta_j$. Setting $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$, we see that Z_{ti} is $\mathcal{F}_t$-adapted. Letting $S_{ti} = \sum_{j\neq i} A_{tj}\theta_j$ and expanding $Z_{ti} = A_{ti}S_{ti} + A_{ti}\eta_t$. The first step is to check that $Z_{ti}|\mathcal{F}_{t-1} \sim \operatorname{subG}(\sqrt{2})$.

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda Z_{ti}) \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda Z_{ti}) \mid \mathcal{F}_{t-1}, A_t\right] \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\exp(\lambda A_{ti}S_{ti})\mathbb{E}\left[\exp(\lambda A_{ti}\eta_t) \mid \mathcal{F}_{t-1}, A_t\right] \mid \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[\exp(\lambda A_{ti}S_{ti})\exp\left(\frac{\lambda^2}{2}\right) \middle| \mathcal{F}_{t-1}\right] \\
&= \exp\left(\frac{\lambda^2}{2}\right)\mathbb{E}\left[\mathbb{E}\left[\exp(\lambda A_{ti}S_{ti}) \mid \mathcal{F}_{t-1}, S_{ti}\right] \mid \mathcal{F}_{t-1}\right] \\
&\leq \exp\left(\frac{\lambda^2}{2}\right)\mathbb{E}\left[\exp\left(\frac{\lambda^2 S_{ti}^2}{2}\right) \middle| \mathcal{F}_{t-1}\right] \\
&\leq \exp(\lambda^2)\,,
\end{aligned}$$

where the first inequality used that η_t and A_t are conditionally independent given

$\mathcal{F}_{t-1}$ and that $\eta_t | \mathcal{F}_{t-1} \sim \mathrm{subG}(1)$, the second to last inequality used that S_{ti} and A_{ti} are conditionally independent given $\mathcal{F}_{t-1}$ and that $A_{ti}|\mathcal{F}_{t-1} \sim \mathrm{subG}(1)$, the last step used that $|S_{ti}| \le 1$. From this we conclude that $Z_{ti}|\mathcal{F}_{t-1} \sim \mathrm{subG}(\sqrt{2})$. The result follows by applying Lemma 23.2 with stopping time $\tau_i = n \wedge \max\{t : E_{ti} = 1\}$. Then $\mathbb{P}(F_i) = \mathbb{P}(\theta_i \notin \mathcal{C}_{\tau_i,i}) \le 1/n$. □

23.4 UCB with sparsity

A new plan is needed to relax the assumption that the action set is a hypercube. The idea is to modify the ellipsoidal confidence set used in Chapter 19 to have a smaller radius, which is made possible by exploiting the lower variance of the least-squares estimator when the unknown parameter is sparse. We will see that modifying the algorithm in Chapter 19 to use the smaller confidence intervals improves the regret to $R_n = O(\sqrt{dpn}\log(n))$.

⚠ Without assumptions on the action-set one cannot hope to have a regret smaller than $O(\sqrt{dn})$. To see this, recall that d-armed bandits can be represented as linear bandits with $\mathcal{A}_t = \{e_1, \dots, e_d\}$. For these problems Theorem 15.1 shows that for any policy there exists a d-armed bandit for which $R_n = \Omega(\sqrt{dn})$. Checking the proof reveals that when adapted to the linear setting the parameter vector is 2-sparse.

23.5 Online to confidence set conversion

The construction that follows makes use of a kind of duality between online prediction and confidence sets. While we will only apply the idea to the sparse linear case, the approach is generic. Unless otherwise mentioned, for the remainder of the chapter we make the following assumptions:

The prediction problem considered is **online linear prediction** where the prediction error is measured by the squared loss. This is also known as **online linear regression**. The learner interacts with an environment in a sequential manner where in each round $t \in \mathbb{N}^+$:

1 The environment chooses $X_t \in \mathbb{R}$ and $A_t \in \mathbb{R}^d$ in an arbitrary fashion.
2 The value of A_t is revealed to the learner (but not X_t).
3 The learner produces a real-valued prediction $\hat{X}_t$ in some way.
4 The environment reveals X_t to the learner and the loss is $(X_t - \hat{X}_t)^2$.

The learner's goal is to produce predictions whose total loss is not much worse than the loss suffered by any of the linear predictors in some set $\Theta \subset \mathbb{R}^d$. The regret of the learner relative to a linear predictor that uses the weights $\theta \in \mathbb{R}^d$ is

$$\rho_n(\theta) = \sum_{t=1}^n (X_t - \hat{X}_t)^2 - \sum_{t=1}^n (X_t - \langle A_t, \theta \rangle)^2 \,. \tag{23.5}$$

We say that the learner enjoys a regret guarantee B_n relative to Θ if for any strategy of the environment,

$$\sup_{\theta \in \Theta} \rho_n(\theta) \leq B_n \,. \tag{23.6}$$

The online learning literature in machine learning has a number of powerful algorithms for this learning problem with equally powerful regret guarantees. Later we will give a specific result for the sparse case when $\Theta = \{x : \|x\|_0 \leq M_0\}$, but first we show how to use such a learning algorithm to construct a confidence set. Take any learner for online linear regression and assume the environment generates X_t in a stochastic manner like in linear bandits:

$$X_t = \langle A_t, \theta_* \rangle + \eta_t \,, \tag{23.7}$$

Combining Eqs. (23.5) to (23.7) with elementary algebra,

$$\begin{aligned} Q_t = \sum_{t=1}^n (\hat{X}_t - \langle A_t, \theta_* \rangle)^2 &= \rho_n(\theta_*) + 2 \sum_{t=1}^n \eta_t (\hat{X}_t - \langle A_t, \theta_* \rangle) \\ &\leq B_n + 2 \sum_{t=1}^n \eta_t (\hat{X}_t - \langle A_t, \theta_* \rangle) \,, \end{aligned} \tag{23.8}$$

where the first equality serves as the definition of Q_t. Let us now take stock for a moment. If we could somehow remove the dependence on the noise η_t in the right-hand side, then we could define a confidence set consisting of all θ that satisfy the equation. Of course the noise has zero mean and is conditionally independent of its multiplier, so the expectation of this term is zero. If we can control the fluctuations with high probability, then we will have made some progress. Let

$$Z_t = \sum_{s=1}^t \eta_t (\hat{X}_t - \langle A_t, \theta_* \rangle)$$

Since $\hat{X}_t$ is chosen based on information available at the beginning of the round, $\hat{X}_t$ is $\mathcal{F}_{t-1}$-measurable and so

$$(Z_t - Z_{t-1}) | \mathcal{F}_{t-1} \sim \text{subG}(\sigma_t) \,, \qquad \text{where } \sigma_t^2 = (\hat{X}_t - \langle A_t, \theta_* \rangle)^2 \,.$$

The uniform self-normalized tail bound (Theorem 20.1) with $\lambda = 1$ implies that,

$$\mathbb{P}\left(\text{exists } t \geq 0 \text{ such that } |Z_t| \geq \sqrt{(1 + Q_t) \log\left(\frac{1 + Q_t}{\delta^2} \right)} \right) \leq \delta \,.$$

Provided this low probability event does not occur, then from Eq. (23.8) we have

$$Q_t \leq B_t + 2\sqrt{(1+Q_t)\log\left(\frac{1+Q_t}{\delta^2}\right)}\,. \tag{23.9}$$

While both sides depend on Q_t, the left-hand side grows linearly, while the right-hand side grows sublinearly in Q_t. This means that the largest value of Q_t that satisfies the above inequality is finite. A tedious calculation then shows this value must be less than

$$\beta_t(\delta) = 1 + 2B_t + 32\log\left(\frac{\sqrt{8}+\sqrt{1+B_t}}{\delta}\right)\,. \tag{23.10}$$

By piecing together the parts we conclude that with probability at least $1-\delta$ the following holds for all t:

$$Q_t = \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta_* \rangle)^2 \leq \beta_t(\delta)\,.$$

We could define $\mathcal{C}_{t+1}$ to be the set of all θ such that the above holds with θ_* replaced by θ, but there is one additionally subtlety, which is that the resulting confidence interval may be unbounded (think about the case that $\sum_{s=1}^t A_s A_s^\top$ is not invertible). In Chapter 19 we overcame this problem by regularizing the least squares estimator. Since we have assumed that $\|\theta_*\|_2 \leq M_2$ the previous display implies that

$$\|\theta_*\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta_* \rangle)^2 \leq M_2^2 + \beta_t(\delta)\,.$$

All together we have the following theorem.

THEOREM 23.2 *Let $\delta \in (0,1)$ and assume that $\theta_* \in \Theta$ and $\sup_{\theta\in\Theta} \rho_t(\theta) \leq B_t$. If*

$$\mathcal{C}_{t+1} = \left\{\theta \in \mathbb{R}^d \,:\, \|\theta\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta \rangle)^2 \leq M_2^2 + \beta_t(\delta)\right\}\,,$$

then $\mathbb{P}\left(\text{exists } t \in \mathbb{N} \text{ such that } \theta_* \notin \mathcal{C}_{t+1}\right) \leq \delta$.

The confidence set in Theorem 23.2 is not in the most convenient form. By defining $V_t = I + \sum_{s=1}^t A_s A_s^\top$ and $S_t = \sum_{s=1}^t A_s \hat{X}_s$ and $\hat{\theta}_t = V_t^{-1} S_t$ and performing an algebraic calculation that we leave to the reader (see Exercise 23.5) one can see that

$$\|\theta\|_2^2 + \sum_{s=1}^t (\hat{X}_s - \langle A_s, \theta \rangle)^2 = \|\theta - \hat{\theta}_t\|_{V_t}^2 + \sum_{s=1}^t (\hat{X}_s - \langle \hat{\theta}_t, A_s \rangle)^2 + \|\hat{\theta}_t\|_2^2\,. \tag{23.11}$$

Using this, the confidence set can be rewritten in the familiar form of an ellipsoid:

$$\mathcal{C}_{t+1} = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t}^2 \leq M_2^2 + \beta_t(\delta) - \|\hat{\theta}_t\|_2^2 - \sum_{s=1}^t (\hat{X}_s^2 - \langle \hat{\theta}_t, A_s \rangle)^2\right\}\,.$$

```
Input  Online linear predictor and regret bound $B_t$, confidence parameter $\delta \in (0, 1)$
for $t = 1, \dots, n$ do
    Receive action set $\mathcal{A}_t$
    Computer confidence set:
        $$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d \,:\, \|\theta\|_2^2 + \sum_{s=1}^{t-1} (\hat X_s - \langle A_s, \theta\rangle)^2 \le M_2^2 + \beta_t(\delta)\right\}$$
    Calculate optimistic action
        $$A_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle a, \theta\rangle$$
    Feed $A_t$ to the online linear predictor and obtain prediction $\hat X_t$
    Play $A_t$ and receive reward $X_t$
    Feed $X_t$ to online linear predictor as feedback
end for
```

Algorithm 13: Online Linear Predictor UCB

It is not obvious that $\mathcal{C}_{t+1}$ is not empty because the radius could be negative. Theorem 23.2 shows, however, that with high probability $\theta_* \in \mathcal{C}_{t+1}$. At last we have established all the conditions required for Theorem 19.1, which implies the following theorem bounding the regret of Algorithm 13.

THEOREM 23.3 *With probability at least $1 - \delta$ the pseudo-regret of OLR-UCB satisfies*

$$\hat R_n \le \sqrt{8dn\,(M_2^2 + \beta_{n-1}(\delta)) \log\left(1 + \frac{n}{d}\right)}\,.$$

23.6 Sparse online linear prediction

THEOREM 23.4 *There exists a strategy π for the learner such that for any $\theta \in \mathbb{R}^d$, the regret $\rho_n(\theta)$ of π against any strategic environment such that $\max_{t \in [n]} \|A_t\|_2 \le L$ and $\max_{t \in [n]} |X_t| \le X$ satisfies*

$$\rho_n(\theta) \le cX^2 \|\theta\|_0 \left\{\log(e + n^{1/2}L) + C_n \log(1 + \tfrac{\|\theta\|_1}{\|\theta\|_0})\right\} + (1 + X^2) C_n\,,$$

where $c > 0$ is some universal constant and $C_n = 2 + \log_2 \log(e + n^{1/2}L)$.

The strategy is a variant of the exponential weights method except that the method is now adjusted so that the set of experts is now $\mathbb{R}^d$. An appropriate sparsity prior is used and when predicting an appropriate truncation strategy is used. The details of the procedure are less important at this stage for our purposes and are thus left out. A reference to the work containing the missing details will be given at the end of the chapter.

Note that $A_n = O(\log\log(n))$. Hence, dropping the dependence on X and L, for $p > 0$, $\sup_{\theta:\|\theta\|_0 \le p, \|\theta\|_2 \le L} \rho_n(\theta) = O(p\log(n))$. Note how strong this is: The guarantee hold no matter what strategy the environment uses!

Now, in sparse linear bandits with subgaussian noise, the noise $(\eta_t)_t$ is not necessarily bounded, and as a consequence the rewards $(X_t)_t$ are also not necessarily bounded. However, the subgaussian property implies with probability $1-\delta$, $|\eta_t| \le \log(2/\delta)$. Now, choosing $\delta = 1/n^2$, we thus see that for problems with bounded mean reward, $\max_{t\in[n]} |X_t| \le X \doteq 1 + \log(2n^2)$ with probability at least $1 - 1/n$. Putting things together then yields the announced result. The expected regret of OLR-UCB when using the strategy π from above satisfies

$$R_n = \tilde{O}(\sqrt{dpn})\,.$$

23.7 Notes

1 The strategy achieving the bound in Theorem 23.4 is not computationally efficient. In fact we do not know of any polynomial time algorithm with logarithmic regret for this problem. The consequence is that Algorithm 13 does not yet have an efficient implementation.

2 While we focused on the sparse case, the results and techniques apply to other settings. For example, we can also get alternative confidence sets from results in online learning even for the standard non-sparse case. Or one may consider additional or different structural assumptions on θ (for example, θ that when reshaped into a matrix, could have a low spectral norm).

3 When the online linear regression results are applied it is important to use the tightest possible, data-dependent regret bounds B_n. In online learning most regret bounds start as tight, data-dependent bounds, which are then loosened to get further insight into the structure of problems. For our application, naturally one should use the tightest available regret bounds (or one should attempt to modify the existing proofs to get tighter data-dependent bounds). The gains from using data-dependent bounds can be significant.

4 We need to emphasize that the sparsity parameter p must be known in advance and that no algorithm can enjoy a regret of $\Omega(\sqrt{dpn})$ for all p simultaneously. This will be seen shortly in Chapter 24 that focuses exclusively on lower bounds for stochastic linear bandits.

23.8 Bibliographical Remarks

The Selective Explore-Then-Commit algorithm is due to the authors [Lattimore et al., 2015]. The construction for the sparse case is from another paper co-authored by one of the authors [Abbasi-Yadkori et al., 2012]. The online linear predictor that competes with sparse parameter vectors and its analysis summarized in Theorem 23.4 is due to [Gerchinovitz, 2013, Thm. 10]. A recent paper by Rakhlin and Sridharan [2017] also discusses relationship between online learning regret bounds and self-normalized tail bounds of the type given here. Interestingly, what they show is that the relationship goes in both directions: Tail inequalities imply regret bounds and regret bounds imply tail inequalities. We are told by Francesco Orabona that techniques similar to used here for constructing confidence bounds have been used earlier in a series of papers by Claudio Gentile and friends. For completeness, here is the list for further exploration: Dekel et al. [2010, 2012], Crammer and Gentile [2013], Gentile and Orabona [2012, 2014]. Carpentier and Munos [2012] have also published a paper on sparse linear stochastic bandits, but with the action-set restricted to the $(d-1)$-sphere. Like the hypercube, it turns out that this makes it possible to avoid the poor dependence on the dimension and their regret bound is $R_n = O(p\sqrt{n}\log(d))$. The online-to-confidence set construction idea has recently been used for designing more efficient algorithms for generalized linear bandits [Jun et al., 2017].

23.9 Exercises

23.1 A norm on $\mathbb{R}^d$ is a function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}$ such that for all $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$ it holds that: (a) $\|x\| = 0$ if and only if $x = 0$ and (b) $\|ax\| = |a|\|x\|$ and (c) $\|x + y\| \leq \|x\| + \|y\|$ and (d) $\|x\| \geq 0$. Show that $\|\cdot\|_0$ given by $\|x\|_0 = \sum_{i=1}^d \mathbb{I}\{x_i \neq 0\}$ is not a norm.

23.2 Prove the second part of Theorem 23.1.

Think about what happens to R_{ni} if $|\theta_i|$ is small.

23.3 Algorithm 12 is not anytime (it requires advance knowledge of the horizon). Design a modified version that does not require this knowledge and prove a comparable regret bound to what was given in Theorem 23.1.

One way is to use the doubling trick, but a more careful approach will lead to a more practical algorithm.

23.4 Complete the calculation to derive Eq. (23.10) from Eq. (23.9).

23.5 Prove the equality in Eq. (23.11).

23.6 Let f be a density function on $[0, \infty)$ so that $\int_0^\infty f(\lambda) d\lambda = 1$ and $f(\lambda) \geq 0$ for all $\lambda \geq 0$. Then define

$$M_n = \int_{\mathbb{R}} f(\lambda) \exp\left(\lambda S_n - \frac{\lambda^2 n}{2}\right) d\lambda\,.$$

(a) Show that $\operatorname{argmax}_{\lambda \in \mathbb{R}} \lambda S_n - \lambda^2 n/2 = S_n/n$.

(b) Suppose that $f(\lambda)$ is monotone decreasing for $\lambda > 0$. Show that for any $\varepsilon > 0$ and $\Lambda_n = S_n/n$ that,

$$M_n \geq \varepsilon \Lambda_n f(\Lambda_n(1+\varepsilon)) \exp\left(\frac{(1-\varepsilon^2)S^2}{2n}\right)$$

(c) Use the previous result to show for any $\delta \in (0,1)$ that

$$\mathbb{P}\left(\text{exists } n : S_n \geq \inf_{\varepsilon > 0} \sqrt{\frac{2n}{(1-\varepsilon^2)}\left(\log\left(\frac{1}{\delta}\right) + \log\left(\frac{1}{\varepsilon \Lambda_n f(\Lambda_n(1+\varepsilon))}\right)\right)}\right) \leq \delta\,.$$

(d) Find an f such that $\int_0^\infty f(\lambda) d\lambda = 1$ and $f(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$ and

$$\log\left(\frac{1}{\lambda f(\lambda)}\right) = (1 + o(1)) \log\log\left(\frac{1}{\lambda}\right)$$

as $\lambda \to 0$.

(e) Use the previous results to show that

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \log\log(n)}} \leq 1\right) = 1\,.$$