20 Confidence Bounds for Least Squares Estimators

In the last chapter we derived a regret bound for a version of the upper confidence bound algorithm that depended on a particular kind of confidence set. The purpose of this chapter is to justify these choices.

Suppose that at the end of round t a bandit algorithm has chosen actions $A_1, \dots, A_t \in \mathbb{R}^d$ and received the respective payoffs $X_1, \dots, X_t$. Recall from the previous chapter that the **penalized least-squares** (or **ridge regression**) estimate of θ_* is the minimizer of the penalized squared empirical loss,

$$L_t(\theta) = \sum_{s=1}^{t} (X_s - \langle A_s, \theta \rangle)^2 + \lambda \|\theta\|_2^2 ,$$

where $\lambda \geq 0$ is the penalty factor. This is minimized by

$$\hat{\theta}_t = V_t(\lambda)^{-1} \sum_{s=1}^{t} X_s A_s \quad \text{with } V_t(\lambda) = \lambda I + \sum_{s=1}^{t} A_s A_s^\top . \tag{20.1}$$

It is convenient for the remainder to abbreviate $V_t = V_t(0)$.

Designing a confidence set about $\hat{\theta}_t$ when $A_1, \dots, A_t$ have been chosen by a bandit algorithm is a surprisingly delicate matter. The difficulty stems from the fact that the actions $(A_s)_{s<t}$ are neither fixed nor independent, but are intricately correlated via the rewards. We spend the first section of this chapter building intuition by making some simplifying assumptions. Eager readers may skip directly to Section 20.1. For the rest of this section we assume that:

1 *Nonsingular Grammian:* $\lambda = 0$ and V_t is invertible.
2 *Independent subgaussian noise:* $(\eta_s)_s$ are independent and 1-subgaussian.
3 *Fixed design:* $A_1, \dots, A_t$ are deterministically chosen without the knowledge of $X_1, \dots, X_t$.

None of these assumptions is plausible in the bandit setting, but the simplification eases the analysis and provides insight. To emphasize that $A_1, \dots, A_t$ are deterministic we use a_s in place of A_s so that

$$V_t = \sum_{s=1}^{t} a_s a_s^\top \qquad \text{and} \qquad \hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} a_s X_s .$$

Note that for V_t to be non-singular it is neccessary that the actions $(a_s)_{s=1}^t$ span

$\mathbb{R}^d$, which of course implies that $t \geq d$. Comparing θ_* and $\hat{\theta}_t$ in the direction $x \in \mathbb{R}^d$ we have

$$\begin{aligned}\langle x, \hat{\theta}_t - \theta_* \rangle &= \left\langle x, V_t^{-1} \sum_{s=1}^t a_s X_s - \theta_* \right\rangle = \left\langle x, V_t^{-1} \sum_{s=1}^t a_s \left(a_s^\top \theta_* + \eta_s\right) - \theta_* \right\rangle \\ &= \left\langle x, V_t^{-1} \sum_{s=1}^t \eta_s a_s \right\rangle = \sum_{s=1}^t \left\langle x, V_t^{-1} a_s \right\rangle \eta_s \,.\end{aligned}$$

Since $(\eta_s)_s$ are independent and 1-subgaussian, by Lemma 5.2 and Theorem 5.1,

$$\mathbb{P}\left(\left\langle x, \hat{\theta}_t - \theta_* \right\rangle \geq \sqrt{2 \sum_{s=1}^t \left\langle x, V_t^{-1} a_s \right\rangle^2 \log\left(\frac{1}{\delta}\right)}\right) \leq \delta \,.$$

A little linear algebra shows that $\sum_{s=1}^t \left\langle x, V_t^{-1} a_s \right\rangle^2 = \|x\|^2_{V_t^{-1}}$, which means that

$$\mathbb{P}\left(\langle x, \hat{\theta}_t - \theta_* \rangle \geq \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)}\right) \leq \delta \,. \tag{20.2}$$

The next step is to convert the above bound on $\langle x, \hat{\theta}_t - \theta^* \rangle$ to a bound on $\|\hat{\theta}_t - \theta^*\|_{V_t}$. To begin this process notice that

$$\|\hat{\theta}_t - \theta^*\|_{V_t} = \langle V_t^{1/2} X, \hat{\theta}_t - \theta_* \rangle \,, \text{ where } X = \frac{V_t^{1/2}(\hat{\theta}_t - \theta_*)}{\|\hat{\theta}_t - \theta_*\|_{V_t}} \,.$$

The problem is that X is random while we have only proven (20.2) for deterministic x. The standard way of addressing problems like this is to use a **covering argument**. First we identify a finite set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ such that whatever value X takes, there exists some $x \in \mathcal{C}_\varepsilon$ that is ε-close to X. Then a union bound and a triangle inequality allows one to finish. By its definition we have $\|X\|_2^2 = X^\top X = 1$, which means that $X \in S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Using that $X \in S^{d-1}$ we see it suffices to 'cover' S^{d-1}. For this we have the following result:

LEMMA 20.1 *There exists a set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ with $|\mathcal{C}_\varepsilon| \leq (3/\varepsilon)^d$ such that for all $x \in S^{d-1}$ there exists a $y \in \mathcal{C}_\varepsilon$ with $\|x - y\|_2 \leq \varepsilon$.*

The proof of this lemma requires a bit work, but nothing really deep is needed. This work is deferred to Exercises 20.1 and 20.2. Letting $\mathcal{C}_\varepsilon$ be the covering set given by the lemma and applying a union bound and Eq. (20.2) shows that

$$\mathbb{P}\left(\text{exists } x \in \mathcal{C}_\varepsilon : \left\langle V_t^{1/2} x, \hat{\theta}_t - \theta_* \right\rangle \geq \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)}\right) \leq \delta \,,$$

where we used the fact that $\|V_t^{1/2} x\|_{V_t^{-1}} = \|x\|_2 = 1$. Then assuming the event

inside the probability does not occur and using Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\hat{\theta}_t - \theta_*\|_{V_t} &= \max_{x \in S^{d-1}} \left\langle V_t^{1/2} x, \hat{\theta}_t - \theta_* \right\rangle \\
&= \max_{x \in S^{d-1}} \min_{y \in \mathcal{C}_\varepsilon} \left[\left\langle V_t^{1/2}(x-y), \hat{\theta}_t - \theta_* \right\rangle + \left\langle V_t^{1/2} y, \hat{\theta}_t - \theta_* \right\rangle \right] \\
&< \max_{x \in S^{d-1}} \min_{y \in \mathcal{C}_\varepsilon} \left[\|\hat{\theta}_t - \theta_*\|_{V_t} \|x - y\|_2 + \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \right] \\
&\leq \varepsilon \|\hat{\theta}_t - \theta_*\|_{V_t} + \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \,.
\end{aligned}$$

Rearranging yields

$$\|\hat{\theta}_t - \theta_*\|_{V_t} \leq \frac{1}{1-\varepsilon} \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \,.$$

And now there is a tension in the choice of $\varepsilon > 0$. The term in the denominator suggests that ε should be small, but by Lemma 20.1 the cardinality of $\mathcal{C}_\varepsilon$ grows rapidly as ε tends to zero. By lazily choosing $\varepsilon = 1/2$,

$$\mathbb{P}\left(\|\hat{\theta}_t - \theta_*\|_{V_t} \geq 2 \sqrt{2 \left(d \log(6) + \log\left(\frac{1}{\delta}\right) \right)} \right) \leq \delta \,. \tag{20.3}$$

Except for constants and other minor differences, this turns out to be about as good as you can get. Unfortunately, however, this analysis only works because V_t was assumed to be deterministic. In the active case, where $A_1, \ldots, A_n$ are chosen by a bandit algorithm, this assumption does not hold and the ideas need to be modified.

20.1 Martingale noise and Laplace's method

We now remove all of the limiting assumptions in the previous section. Of course we still need some conditions on the noise. In particular, we assume that $\eta_1, \ldots, \eta_t$ are conditionally 1-subgaussian:

$$\mathbb{E}\left[\exp(\alpha \eta_s) \mid \eta_1, \ldots, \eta_{s-1}\right] \leq \exp\left(\frac{\alpha^2}{2}\right) \,, \qquad \text{for all } \alpha \in \mathbb{R} \text{ and } s \in [t] \,. \tag{20.4}$$

We have now dropped the assumption that $A_1, A_2, \ldots$ are fixed in advance and so return to the usual capitalization. We also allow arbitrary penalty factors $\lambda > 0$ and relax the assumption that V_t be invertible (though $V_t(\lambda)$ is now invertible because $\lambda > 0$). Can we still get a confidence set like what appears in (20.3)? Before diving in, we need to introduce another concept from probability theory.

DEFINITION 20.1 (Martingale difference process) Let $\mathbb{F} = (\mathcal{F}_s)_s$ be a filtration over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A sequence of random variables $(U_s)_s$ is an

$\mathbb{F}$-adapted martingale difference process if for all s it holds that $\mathbb{E}[U_s]$ exists and U_s is $\mathcal{F}_s$-measurable and $\mathbb{E}[U_s \mid \mathcal{F}_{s-1}] = 0$.

As usual, the filtration is often not explicitly mentioned if it is obvious from the context. The name is justified by the fact that if $(U_s)_s$ is a martingale difference process, then the partial sums $M_t = \sum_{s=1}^t U_s$ define a martingale. A more descriptive and informal name for a martingale difference process is **martingale noise**.

In the linear bandit model $(\eta_s)_s$ is martingale noise with the filtration given by $\mathcal{F}_s = \{A_1, X_1, \ldots, A_{s-1}, X_{s-1}, A_s\}$. Note the inclusion of A_s in the definition of $\mathcal{F}_s$. The martingale noise assumption allows the noise η_s to depend on past choices, including the most recent action. This is often essential. For example, if the rewards are Bernoulli the distribution of the noise depends on the mean reward of the action. Let us return to the construction of confidence sets. Since we want exponentially decaying tail probabilities one is tempted to try Chernoff's method:

$$\mathbb{P}\left(\|\hat{\theta}_t - \theta_*\|_{V_t} \geq u\right) \leq \inf_{\lambda > 0} \mathbb{E}\left[\exp\left(\lambda \|\hat{\theta}_t - \theta_*\|_{V_t} - \lambda u\right)\right] .$$

Sadly, we do not know how to bound this expectation. Can we still somehow use Chernoff's method? Let $S_t = \sum_{s=1}^t \eta_s A_s$ and apply the 'linearization trick' to show that

$$\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 = \max_{x \in \mathbb{R}^d} \left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{V_t}^2\right) .$$

The exponential of the term inside the maximum is a supermartingale.

LEMMA 20.2 *For all $x \in \mathbb{R}^d$ the process $M_t(x) = \exp(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{V_t}^2)$ is a $\mathbb{F}$-adapted supermartingale.*

Proof of Lemma 20.2 That $M_t(x)$ is $\mathcal{F}_t$-measurable for all t is immediate from the definition. We need to show that $\mathbb{E}[M_t(x) \mid \mathcal{F}_{t-1}] \leq M_{t-1}(x)$ almost surely. The fact that (η_s) is martingale noise with respect to filtration $(\mathcal{F}_s)$ means that

$$\mathbb{E}\left[\exp\left(\eta_s \langle x, A_s\rangle\right) \mid \mathcal{F}_s\right] \leq \exp\left(\frac{\langle x, A_s\rangle^2}{2}\right) = \exp\left(\frac{\|x\|_{A_s A_s^\top}^2}{2}\right) \quad \text{a.s.}$$

Hence

$$\begin{aligned}
\mathbb{E}[M_t(x) \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{V_t}^2\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= M_{t-1}(x)\mathbb{E}\left[\exp\left(\eta_t \langle x, A_t\rangle - \frac{1}{2}\|x\|_{A_t A_t^\top}^2\right) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leq M_{t-1}(x) \quad \text{a.s.}
\end{aligned}$$

□

Combining the lemma and the linearization idea almost works. Chernoff's

method leads to

$$\begin{aligned}\mathbb{P}\left(\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 \geq \log(1/\delta)\right) &= \mathbb{P}\left(\exp\left(\max_{x\in\mathbb{R}^d}\langle x, S_t\rangle - \frac{1}{2}\|x\|_{V_t}^2\right) \geq 1/\delta\right) \\ &\leq \delta\mathbb{E}\left[\exp\left(\max_{x\in\mathbb{R}^d}\langle x, S_t\rangle - \frac{1}{2}\|x\|_{V_t}^2\right)\right] \\ &= \delta\mathbb{E}\left[\max_{x\in\mathbb{R}^d} M_t(x)\right]. \end{aligned} \tag{20.5}$$

Now Lemma 20.2 shows that $\mathbb{E}[M_t(x)] \leq 1$. This seems quite promising, but the presence of the maximum is a setback because Jensen's inequality implies that $\mathbb{E}[\max_{x\in\mathbb{R}^d} M_t(x)] \geq \max_{x\in\mathbb{R}^d} \mathbb{E}[M_t(x)]$, which is the wrong direction to be used above. This means we cannot directly use the lemma to bound Eq. (20.5). There are two natural ways to attack this problem. The first idea is to define a finite covering set $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ so that

$$\begin{aligned}\mathbb{E}\left[\max_{x\in\mathbb{R}^d} M_t(x)\right] &= \mathbb{E}\left[\max_{x\in\mathbb{R}^d}\min_{y\in\mathcal{C}_\varepsilon} M_t(x) - M_t(y) + M_t(y)\right] \\ &\leq \mathbb{E}\left[\max_{x\in\mathbb{R}^d}\min_{y\in\mathcal{C}_\varepsilon} |M_t(x) - M_t(y)|\right] + \mathbb{E}\left[\max_{y\in\mathcal{C}_\varepsilon} M_t(y)\right] \\ &\leq \mathbb{E}\left[\max_{x\in\mathbb{R}^d}\min_{y\in\mathcal{C}_\varepsilon} |M_t(x) - M_t(y)|\right] + \sum_{y\in\mathcal{C}_\varepsilon}\mathbb{E}\left[M_t(y)\right] \\ &\leq \varepsilon + |\mathcal{C}_\varepsilon|. \end{aligned} \tag{20.6}$$

The last inequality follows from a careful choice of $\mathcal{C}_\varepsilon$, and as usual the size of the covering set must be balanced against the required accuracy. Choosing $\mathcal{C}_\varepsilon$ is quite non-trivial because $M_t(x) - M_t(y)$ is random, even for fixed x and y. We leave the 'last few steps' as an exercise (see Exercise 20.3). The second approach actually does not require us to bound Eq. (20.5), but uses it for inspiration when combined with Laplace's method for approximating integrals of well-behaved exponentials.

Laplace's method (†)

We briefly review Laplace's method for one-dimension functions. Assume that $f : [a, b] \to \mathbb{R}$ is twice differentiable and has a unique maximum at $x_0 \in (a, b)$ with $-q = f''(x_0) < 0$. Laplace's method for approximating $f(x_0)$ is to compute the integral

$$I_s = \int_a^b \exp(sf(x))dx$$

for some large value of $s > 0$. From a Taylor expansion we may write

$$f(x) = f(x_0) - \frac{q}{2}(x - x_0)^2 + R(x),$$

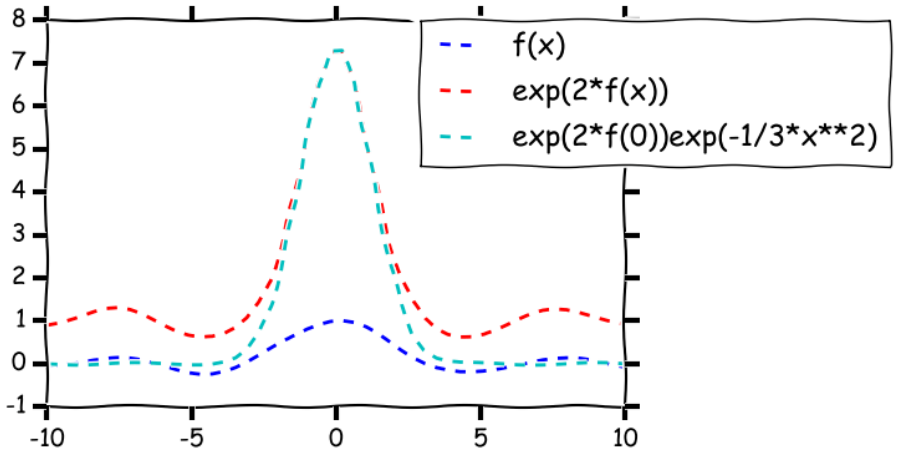

Figure 20.1 Laplace's method

where $R(x) = o((x - x_0)^2)$. Under appropriate technical assumptions,

$$I_s \sim \exp(sf(x_0)) \int_a^b \exp\left(-\frac{sq(x - x_0)^2}{2}\right) dx \qquad \text{as } s \to \infty\,.$$

Furthermore, as s gets large

$$\int_a^b \exp\left(-\frac{sq(x - x_0)^2}{2}\right) dx \sim \int_{-\infty}^{\infty} \exp\left(-\frac{sq(x - x_0)^2}{2}\right) dx = \sqrt{\frac{2\pi}{sq}}$$

and hence

$$I_s \sim \exp(sf(x_0))\sqrt{\frac{2\pi}{sq}}\,.$$

Intuitively, the dominant term in the integral I_s is $\exp(sf(x_0))$. It should also be clear that the fact that we integrate with respect the Lebesgue measure does not matter much. We could have integrated with respect to any other measure as long as that measure puts a positive mass on the neighborhood of the maximizer. The method is illustrated on the figure shown below. The take home message of this is that if we integrate the exponential of a function that has a pronounced maximum then we can expect that the integral will be close to the exponential function of the maximum.

Method of mixtures

Laplace's approximation suggests that

$$\max_x M_t(x) \approx \int_{\mathbb{R}^d} M_t(x) dh(x)\,, \tag{20.7}$$

where h is some measure on $\mathbb{R}^d$ chosen so that the integral can be calculated in closed form. This is not a requirement of the method, but it does make the

argument shorter. The main benefit of replacing the maximum with an integral is that we obtain the following lemma, which you will prove in Exercise 20.4.

LEMMA 20.3 *Let h be a probability measure on $\mathbb{R}^d$, then $\bar{M}_t = \int_{\mathbb{R}^d} M_t(x) dh(x)$ is a $\mathbb{F}$-adapted supermartingale.*

THEOREM 20.1 *For all $\lambda > 0$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left(\textit{exists } t \le n : \|S_t\|^2_{V_t(\lambda)^{-1}} \ge 2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det(V_t(\lambda))}{\lambda^d}\right)\right) \le \delta\,.$$

Proof Let $H = \lambda I$ and $h = \mathcal{N}(0, H)$ and

$$\begin{aligned}\bar{M}_t &= \int_{\mathbb{R}^d} M_t(x) dh(x) \\ &= \frac{1}{\sqrt{(2\pi)^d \det(H^{-1})}} \int_{\mathbb{R}^d} \exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{V_t} - \frac{1}{2}\|x\|^2_H\right) dx\,.\end{aligned}$$

Completing the square,

$$\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{V_t} - \frac{1}{2}\|x\|^2_H = \frac{1}{2}\|S_t\|^2_{(H+V_t)^{-1}} - \frac{1}{2}\|x - (H+V_t)^{-1}S_t\|^2_{H+V_t}\,.$$

The first term $\|S_t\|^2_{(H+V_t)^{-1}}$ does not depend on x and can be moved outside the integral, which leaves a quadratic 'Gaussian' term that may be integrated exactly and results in

$$\bar{M}_t = \left(\frac{\det(H)}{\det(H+V)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(H+V_t)^{-1}}\right)\,. \tag{20.8}$$

And things have worked out beautifully. Since $\bar{M}_t$ is a nonnegative supermartingale, the maximal inequality (Theorem 3.5) shows that

$$\mathbb{P}\left(\sup_{t\in\{0,\dots,n\}} \log(\bar{M}_t) \ge \log\left(\frac{1}{\delta}\right)\right) = \mathbb{P}\left(\sup_{t\in\{0,\dots,n\}} \bar{M}_t \ge \frac{1}{\delta}\right) \le \delta\,.$$

The result follows by substituting Eq. (20.8) into the above display and rearranging. □

THEOREM 20.2 *Assuming $\delta \in (0,1)$, then with probability at least $1-\delta$ it holds that for all $t \in \{0, 1, \dots, n\}$,*

$$\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} < \sqrt{\lambda}\|\theta_*\| + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_t(\lambda)}{\lambda^d}\right)}\,.$$

Furthermore, if $\|\theta_*\| \le S$, then $\mathbb{P}\left(\textit{exists } t \in [n] : \theta_* \notin \mathcal{C}_t\right) \le \delta$ with

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d \,:\, \|\hat{\theta}_{t-1} - \theta\|_{V_{t-1}(\lambda)} \le \sqrt{\lambda}S + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_{t-1}(\lambda)}{\lambda^d}\right)}\right\}\,.$$

Proof We only have to compare $\|S_t\|_{V_t(\lambda)^{-1}}$ and $\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)}$.

$$\begin{aligned}
\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} &= \|V_t(\lambda)^{-1} S_t + (V_t(\lambda)^{-1} V_t - I)\theta_*\|_{V_t(\lambda)} \\
&\le \|S_t\|_{V_t(\lambda)^{-1}} + (\theta_*^\top (V_t(\lambda)^{-1} V_t - I) V_t(\lambda) (V_t(\lambda)^{-1} V_t - I)\theta_*)^{1/2} \\
&= \|S_t\|_{V_t(\lambda)^{-1}} + \lambda^{1/2} (\theta_*^\top (I - V_t(\lambda)^{-1} V_t)\theta_*)^{1/2} \\
&\le \|S_t\|_{V_t(\lambda)^{-1}} + \lambda^{1/2} \|\theta_*\| \,.
\end{aligned}$$

And the result follows from Theorem 20.1. □

20.2 Notes

1 An alternative to the 2-norm based construction is to use 1-norms. In the fixed design setting, under the independent Gaussian noise assumption, using Chernoff's method this leads to

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d \,:\, \|V^{1/2}(\hat{\theta}_t - \theta)\|_1 \le \sqrt{2\log(2)d^2 + 2d\log(1/\delta)} \right\} . \quad (20.9)$$

2 Supermartingales come up all the time in proofs relying on Chernoff's method. Just one example is the proof of Lemma 12.1. One could rewrite most the proofs involving sums of random variables relying on Chernoff's method in a way that it would become clear that proof hinges on the supermartingale property of an appropriate sequence.

20.3 Bibliographic remarks

Laplace's method is also called the 'Method of Mixtures' [Peña et al., 2008] and its use goes back to the work of Robbins and Siegmund [1970]. In practice, the improvement that results from using Laplace's method as compared to the previous ellipsoidal constructions that are based on covering arguments is quite large. A historical account of martingale methods in sequential analysis is by Lai [2009]. A simple proof of Lemma 20.1 appears as Lemma 2.5 in the book by van de Geer [2000]. Calculating covering numbers (or related packing numbers) is a whole field by itself, with open questions even in the most obvious examples. The main reference is by Rogers [1964], which by now is a little old, but still interesting.

20.4 Exercises

For Exercise 20.2 where we ask you to prove Lemma 20.1 a few standard definitions will be useful.

DEFINITION 20.2 (Covering and Packing) Let $\mathcal{A} \subset \mathbb{R}^d$. A subset $\mathcal{C} \subset \mathcal{A}$ is said to be an **ε-cover** of $\mathcal{A}$ if $\mathcal{A} \subset \cup_{x \in \mathcal{C}} B(x, \varepsilon)$, where $B(x, \varepsilon) = \{y \in \mathbb{R}^d : \|x - y\| \leq \varepsilon\}$ is the ε ball centered at x. An **ε-packing** of $\mathcal{A}$ is a subset $\mathcal{P} \subset \mathcal{A}$ such that for any $x, y \in \mathcal{P}$, $\|x - y\| > \varepsilon$ (note the strict inequality). The **ε-covering number** of $\mathcal{A}$ is $N(\mathcal{A}, \varepsilon) = \min\{|\mathcal{C}| : \mathcal{C} \text{ is an } \varepsilon\text{-covering of } \mathcal{A}\}$, while the **$\varepsilon$-packing number** of $\mathcal{A}$ is $M(\mathcal{A}, \varepsilon) = \max\{|\mathcal{P}| : \mathcal{P} \text{ is an } \varepsilon\text{-packing of } \mathcal{A}\}$, where we allow for both the covering and packing numbers to take on the value of $+\infty$.

There are various generalizations of these definitions, which do not change their essence. For one, the definitions can be repeated for arbitrarily pseudo-metric spaces (instead of $\mathbb{R}^d$ with the Euclidean distance, we can consider a set X with a $d : X \times X \to [0, \infty)$ function on it which is symmetric and satisfies the triangle inequality and that $d(x, x) = 0$ for any $x \in X$). The basic results concerning covering and packing stated in the next exercise remain valid with this more general definition. In applications we often need the logarithm of the covering and packing numbers, which are then given the new name the set's **metric entropy** (at a scale ε). As we shall see these are often close no matter whether we consider packing or covering.

We separate a useful set of a results concerning packing and covering:

20.1 [Coverings and Packings] Let $\mathcal{A} \subset \mathbb{R}^d$, B the unit ball of $\mathbb{R}^d$, $\mathrm{vol}(\cdot)$ the usual volume (measure under the Lebesgue measure). For brevity let $N(\varepsilon) = N(\mathcal{A}, \varepsilon)$ and $M(\varepsilon) = M(\mathcal{A}, \varepsilon)$. Show that the following hold:

(a) $\varepsilon \to N(\varepsilon)$ is increasing as $\varepsilon \geq 0$ is decreasing.
(b) $M(2\varepsilon) \leq N(\varepsilon) \leq M(\varepsilon)$.
(c) We have

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(B)} \leq N(\varepsilon) \leq M(\varepsilon) \leq \frac{\mathrm{vol}(\mathcal{A} + \frac{\varepsilon}{2}B)}{\mathrm{vol}(\frac{\varepsilon}{2}B)} \overset{(*)}{\leq} \frac{\mathrm{vol}(\frac{3}{2}\mathcal{A})}{\mathrm{vol}(\frac{\varepsilon}{2}B)} \leq \left(\frac{3}{\varepsilon}\right)^d \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(B)},$$

where $(*)$ holds under the assumption that $\varepsilon B \subset \mathcal{A}$ and that $\mathcal{A}$ is convex and for $U, V \subset \mathbb{R}^d$, $c \in \mathbb{R}$, $U + V = \{u + v : u \in U, v \in V\}$ and $cU = \{cu : u \in U\}$;
(d) Fix $\varepsilon > 0$. Then $N(\varepsilon) < +\infty$ if and only if $\mathcal{A}$ is bounded. The same holds for $M(\varepsilon)$.

20.2 Use the results of the previous exercise to prove Lemma 20.1.

20.3 Complete the steps to show Eq. (20.6).

20.4 Prove Lemma 20.3.

20.5 [Hoeffding–Azuma] Let $X_1, \ldots, X_n$ be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$. Suppose that $|X_t| \in [a_t, b_t]$ almost surely for

arbitrary fixed sequences (a_t) and (b_t) with $a_t \le b_t$ for all $t \in [n]$. Show that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{n}(X_t - \mathbb{E}[X_t \mid \mathcal{F}_t]) \ge \varepsilon\right) \le \exp\left(-\frac{2n^2\varepsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right).$$

It may help to recall Hoeffding's lemma from Note 7 in Chapter 5, which states that for random variable $X \in [a, b]$ the moment generating function satisfies

$$M_X(\lambda) \le \exp(\lambda^2(b-a)^2/8).$$

20.6 The following simple extension of Hoeffding–Azuma is often useful. Let $n \in \mathbb{N}^+$ and (a_t) and (b_t) be fixed sequences with $a_t \le b_t$ for all $t \in [n]$. Let $X_1, \ldots, X_n$ be a sequence of random variables adapted to filtration $\mathbb{F} = (\mathcal{F}_t)_t$ and A be an event. Assume that $\mathbb{P}(\text{exists } t \in [n] : A \text{ and } X_t \notin [a_t, b_t]) = 0$ and $\varepsilon > 0$ and show that

(a) $$\mathbb{P}\left(A \cap \sum_{t=1}^{n}(X_t - \mathbb{E}[X_t \mid \mathcal{F}_t]) \ge \varepsilon\right) \le \exp\left(-\frac{2n^2\varepsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right).$$

(b) $$\mathbb{P}\left(\sum_{t=1}^{n}(X_t - \mathbb{E}[X_t \mid \mathcal{F}_t]) \ge \varepsilon\right) \le \mathbb{P}(A^c) + \exp\left(-\frac{2n^2\varepsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right).$$

The utility of this result comes from the fact that very often the range of some adapted sequence is itself random and could be arbitrarily large with low probability (when A does not hold). A reference for the above result is the survey by McDiarmid [1998].

20.7 Let $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$ be a filtration and $X_1, X_2, \ldots, X_n$ be a sequence of $\mathbb{F}$-adapted random variables with $X_t \in \{-1, 0, 1\}$ and $\mu_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}, X_t \neq 0]$, which we define to be zero whenever $\mathbb{P}(X_t \neq 0 \mid \mathcal{F}_{t-1}) = 0$. Then with $S_t = \sum_{s=1}^{t}(X_s - \mu_s|X_s|)$ and $N_t = \sum_{s=1}^{t}|X_s|$,

$$\mathbb{P}\left(\text{exists } t \le n : |S_t| \ge \sqrt{2N_t \log\left(\frac{c\sqrt{N_t}}{\delta}\right)} \text{ and } N_t > 0\right) \le \delta,$$

where $c > 0$ is a universal constant.

This result appeared in a paper by the authors and others with the constant $c = 4\sqrt{2/\pi}/\operatorname{erf}(\sqrt{2}) \approx 3.43$ [Lattimore et al., 2018].