4

Detecting stable patterns

As discussed in Chapter 1 perhaps the most important property of a pattern analysis algorithm is that it should identify statistically stable patterns. A stable relation is one that reflects some property of the source generating the data, and is therefore not a chance feature of the particular dataset. Proving that a given pattern is indeed significant is the concern of 'learning theory', a body of principles and methods that estimate the reliability of pattern functions under appropriate assumptions about the way in which the data was generated. The most common assumption is that the individual training examples are generated independently according to a fixed distribution, being the same distribution under which the expected value of the pattern function is small. Statistical analysis of the problem can therefore make use of the law of large numbers through the 'concentration' of certain random variables.

Concentration would be all that we need if we were only to consider one pattern function. Pattern analysis algorithms typically search for pattern functions over whole classes of functions, by choosing the function that best fits the particular training sample. We must therefore be able to prove stability not of a pre-defined pattern, but of one deliberately chosen for its fit to the data.

Clearly the more pattern functions at our disposal, the more likely that this choice could be a spurious pattern. The critical factor that controls how much our choice may have compromised the stability of the resulting pattern is the 'capacity' of the function class. The capacity will be related to tunable parameters of the algorithms for pattern analysis, hence making it possible to directly control the risk of overfitting the data. This will lead to close parallels with regularisation theory, so that we will control the capacity by using different forms of 'regularisation'.

4.1 Concentration inequalities

In Chapter 1 we introduced the idea of a statistically stable pattern function f as a non-negative function whose expected value on an example drawn randomly according to the data distribution $\mathcal{D}$ is small

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} f(\mathbf{x}) \approx 0.$$

Since we only have access to a finite sample of data, we will only be able to make assertions about this expected value subject to certain assumptions. It is in the nature of a theoretical model that it is built on a set of precepts that are assumed to hold for the phenomenon being modelled. Our basic assumptions are summarised in the following definition of our data model.

Definition 4.1 The model we adopt will make the assumption that the distribution $\mathcal{D}$ that provides the quality measure of the pattern is the same distribution that generated the examples in the finite sample used for training purposes. Furthermore, the model assumes that the individual training examples are independently and identically distributed (i.i.d.). We will denote the probability of an event A under distribution $\mathcal{D}$ by $P_{\mathcal{D}}(A)$. The model makes no assumptions about whether the examples include a label, are elements of $\mathbb{R}^n$, though some mild restrictions are placed on the generating distribution, albeit with no practical significance. ■

We gave a definition of what was required of a pattern analysis algorithm in Definition 1.7, but for completeness we repeat it here with some embellishments.

Definition 4.2 A pattern analysis algorithm takes as input a finite set S of ℓ data items generated i.i.d. according to a fixed (but unknown) distribution $\mathcal{D}$ and a confidence parameter $\delta \in (0, 1)$. Its output is either an indication that no patterns were detectable, or a *pattern function* f that with probability $1 - \delta$ satisfies

$$\mathbb{E}_{\mathcal{D}} f(\mathbf{x}) \approx 0.$$

The value of the expectation is known as the *generalisation error* of the pattern function f. ■

In any finite dataset, even if it comprises random numbers, it is always possible to find relations if we are prepared to create sufficiently complicated functions.

Example 4.3 Consider a set of ℓ people each with a credit card and mobile phone; we can find a degree $\ell - 1$ polynomial $g(t)$ that given a person's telephone number t computes that person's credit card number $c = g(t)$, making $|c - g(t)|$ look like a promising pattern function as far as the sample is concerned. This follows from the fact that a degree $\ell - 1$ polynomial can interpolate ℓ points. However, what is important in pattern analysis is to find relations that can be used to make predictions on unseen data, in other words relations, that capture some properties of the source generating the data. It is clear that $g(\cdot)$ will not provide a method of computing credit card numbers for people outside the initial set. ■

The aim of this chapter is to develop tools that enable us to distinguish between relations that are the effect of chance and those that are 'meaningful'. Intuitively, we would expect a statistically stable relation to be present in different randomly generated subsets of the dataset, in this way confirming that the relation is not just the property of the particular dataset.

Example 4.4 The relation found between card and phone numbers in Example 4.3 would almost certainly change if we were to generate a second dataset. If on the other hand we consider the function that returns 0 if the average height of the women in the group is less than the average height of the men and 1 otherwise, we would expect different subsets to usually return the same value of 0. ■

Another way to ensure that we have detected a significant relation is to check whether a similar relation could be learned from scrambled data: if we randomly reassign the height of all individuals in the sets of Example 4.4, will we still find a relation between height and gender? In this case the probability that this relation exists would be a half since there is equal chance of different heights being assigned to women as to men. We will refer to the process of randomly reassigning labels as *randomisation* of a labelled dataset. It is also sometimes referred to as *permutation testing*. We will see that checking for patterns in a randomised set can provide a lodestone for measuring the stability of a pattern function.

Randomisation should not be confused with the concept of a random variable. A *random variable* is any real-valued quantity whose value depends on some random generating process, while a *random vector* is such a vector-valued quantity. The starting point for the analysis presented in this chapter is the assumption that the data have been generated by a random process. Very little is assumed about this generating process, which can be thought of as the distribution governing the natural occurrence of the data. The

only restricting assumption about the data generation is that individual examples are generated independently of one another. It is this property of the randomly-generated dataset that will ensure the stability of a significant pattern function in the original dataset, while the randomisation of the labels has the effect of deliberately removing any stable patterns.

Concentration of one random variable The first question we will consider is that of the stability of a fixed function of a finite dataset. In other words how different will the value of this same function be on another dataset generated by the same source? The key property that we will require of the relevant quantity or random variable is known as *concentration*. A random variable that is concentrated is very likely to assume values close to its expectation since values become exponentially unlikely away from the mean. For a concentrated quantity we will therefore be confident that it will assume very similar values on new datasets generated from the same source. This is the case, for example, for the function 'average height of the female individuals' used above. There are many results that assert the concentration of a random variable provided it exhibits certain properties. These results are often referred to as *concentration inqualities*. Here we present one of the best-known theorems that is usually attributed to McDiarmid.

Theorem 4.5 (McDiarmid) *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\ \hat{x}_i \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

Then for all $\epsilon > 0$

$$P\left\{f\left(X_1, \ldots, X_n\right) - \mathbb{E}f\left(X_1, \ldots, X_n\right) \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

The proof of this theorem is given in Appendix A.1.

Another well-used inequality that bounds the deviation from the mean for the special case of sums of random variables is Hoeffding's inequality. We quote it here as a simple special case of McDiarmid's inequality when

$$f(X_1, \ldots, X_n) = \sum_{i=1}^n X_i.$$

Theorem 4.6 (Hoeffding's inequality) *If $X_1, \ldots, X_n$ are independent random variables satisfying $X_i \in [a_i, b_i]$, and if we define the random variable*

$S_n = \sum_{i=1}^{n} X_i$, *then it follows that*

$$P\{|S_n - \mathbb{E}[S_n]| \geq \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Estimating univariate means As an example consider the average of a set of ℓ independent instances $r_1, r_2, \ldots, r_\ell$ of a random variable R given by a probability distribution P on the interval $[a, b]$. Taking $X_i = r_i/\ell$ it follows, in the notation of Hoeffding's Inequality, that

$$S_\ell = \frac{1}{\ell}\sum_{i=1}^{\ell} r_i = \hat{\mathbb{E}}[R],$$

where $\hat{\mathbb{E}}[R]$ denotes the sample average of the random variable R. Furthermore

$$\mathbb{E}[S_n] = \mathbb{E}\left[\frac{1}{\ell}\sum_{i=1}^{\ell} r_i\right] = \frac{1}{\ell}\sum_{i=1}^{\ell}\mathbb{E}\left[r_i\right] = \mathbb{E}[R],$$

so that an application of Hoeffding's Inequality gives

$$P\{|\hat{\mathbb{E}}[R] - \mathbb{E}[R]| \geq \varepsilon\} \leq 2\exp\left(-\frac{2\ell\varepsilon^2}{(b-a)^2}\right),$$

indicating an exponential decay of probability with the difference between observed sample average and the true average. Notice that the probability also decays exponentially with the size of the sample. If we consider Example 4.4, this bound shows that for moderately sized randomly chosen groups of women and men, the average height of the women will, with high probability, indeed be smaller than the average height of the men, since it is known that the true average heights do indeed differ significantly.

Estimating the centre of mass The example of the average of a random variable raises the question of how reliably we can estimate the average of a random vector $\phi(\mathbf{x})$, where ϕ is a mapping from the input space X into a feature space F corresponding to a kernel $\kappa(\cdot,\cdot)$. This is equivalent to asking how close the centre of mass of the projections of a training sample

$$S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\}$$

will be to the true expectation

$$\mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})] = \int_X \boldsymbol{\phi}(\mathbf{x}) dP(\mathbf{x}).$$

We denote the centre of mass of the training sample by

$$\boldsymbol{\phi}_S = \frac{1}{\ell}\sum_{i=1}^{\ell}\boldsymbol{\phi}(\mathbf{x}_i).$$

We introduce the following real-valued function of the sample S as our measure of the accuracy of the estimate

$$g(S) = \left\|\boldsymbol{\phi}_S - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\right\|.$$

We can apply McDiarmid's theorem to the random variable $g(S)$ by bounding the change in this quantity when $\mathbf{x}_i$ is replaced by $\hat{\mathbf{x}}_i$ to give $\hat{S}$

$$\begin{aligned}|g(S) - g(\hat{S})| &= \left|\left\|\boldsymbol{\phi}_S - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\right\| - \left\|\boldsymbol{\phi}_{S'} - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\right\|\right| \\ &\leq \left\|\boldsymbol{\phi}_S - \boldsymbol{\phi}_{S'}\right\| = \frac{1}{\ell}\left\|\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\mathbf{x}_i')\right\| \leq \frac{2R}{\ell},\end{aligned}$$

where $R = \sup_{\mathbf{x}\in X}\|\phi(\mathbf{x})\|$. Hence, applying McDiarmid's theorem with $c_i = 2R/\ell$, we obtain

$$P\left\{g(S) - \mathbb{E}_S[g(S)] \geq \epsilon\right\} \leq \exp\left(-\frac{2\ell\epsilon^2}{4R^2}\right). \tag{4.1}$$

We are now at the equivalent point after the application of Hoeffding's inequality in the one-dimensional case. But in higher dimensions we no longer have a simple expression for $\mathbb{E}_S[g(S)]$. We need therefore to consider the more involved argument. We present a derivation bounding $\mathbb{E}_S[g(S)]$ that will be useful for the general theory we develop below. The derivation is not intended to be optimal as a bound for $\mathbb{E}_S[g(S)]$. An explanation of the individual steps is given below

$$\begin{aligned}\mathbb{E}_S[g(S)] &= \mathbb{E}_S\left[\left\|\boldsymbol{\phi}_S - \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]\right\|\right] = \mathbb{E}_S\left[\left\|\boldsymbol{\phi}_S - \mathbb{E}_{\tilde{S}}[\boldsymbol{\phi}_{\tilde{S}}]\right\|\right] \\ &= \mathbb{E}_S\left[\left\|\mathbb{E}_{\tilde{S}}[\boldsymbol{\phi}_S - \boldsymbol{\phi}_{\tilde{S}}]\right\|\right] \leq \mathbb{E}_{S\tilde{S}}\left[\left\|\boldsymbol{\phi}_S - \boldsymbol{\phi}_{\tilde{S}}\right\|\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}S\tilde{S}}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}\sigma_i\left(\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\phi}(\tilde{\mathbf{x}}_i)\right)\right\|\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}S\tilde{S}}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i) - \sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\tilde{\mathbf{x}}_i)\right\|\right] \qquad (4.2) \\ &\leq 2\mathbb{E}_{S\boldsymbol{\sigma}}\left[\frac{1}{\ell}\left\|\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i)\right\|\right] \qquad (4.3) \\ &= \frac{2}{\ell}\mathbb{E}_{S\boldsymbol{\sigma}}\left[\left(\left\langle\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{\ell}\sigma_j\boldsymbol{\phi}(\mathbf{x}_j)\right\rangle\right)^{1/2}\right]\end{aligned}$$

$$\leq \frac{2}{\ell}\left(\mathbb{E}_{S\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{\ell}\sigma_i\sigma_j\kappa(\mathbf{x}_i,\mathbf{x}_j)\right]\right)^{1/2}$$

$$= \frac{2}{\ell}\left(\mathbb{E}_{S}\left[\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i,\mathbf{x}_i)\right]\right)^{1/2} \tag{4.4}$$

$$\leq \frac{2R}{\sqrt{\ell}}. \tag{4.5}$$

It is worth examining the stages in this derivation in some detail as they will form the template for the main learning analysis we will give below.

- The second equality introduces a second random sample $\tilde{S}$ of the same size drawn according to the same distribution. Hence the expectation of its centre of mass is indeed the true expectation of the random vector.
- The expectation over $\tilde{S}$ can now be moved outwards in two stages, the second of which follows from an application of the triangle inequality.
- The next equality makes use of the independence of the generation of the individual examples to introduce random exchanges of the corresponding points in the two samples. The random variables $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_\ell\}$ assume values -1 and $+1$ independently with equal probability 0.5, hence either leave the effect of the examples $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ as it was or effectively interchange them. Since the points are generated independently such a swap gives an equally likely configuration, and averaging over all possible swaps leaves the overall expectation unchanged.
- The next steps split the sum and again make use of the triangle inequality together with the fact that the generation of S and $\tilde{S}$ is identical.
- The movement of the square root function through the expectation follows from Jensen's inquality and the concavity of the square root.
- The disappearance of the mixed terms $\sigma_i\sigma_j\kappa(\mathbf{x}_i,\mathbf{x}_j)$ for $i \neq j$ follows from the fact that the four possible combinations of -1 and $+1$ have equal probability with two of the four having the opposite sign and hence cancelling out.

Hence, setting the right-hand side of inequality (4.1) equal to δ, solving for ϵ, and combining with inequality (4.4) shows that with probability at least $1-\delta$ over the choice of a random sample of ℓ points, we have

$$g(S) \leq \frac{R}{\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right). \tag{4.6}$$

This shows that with high probability our sample does indeed give a good estimate of $\mathbb{E}[\phi(\mathbf{x})]$ in a way that does not depend on the dimension of the feature space. This example shows how concentration inequalities provide mechanisms for bounding the deviation of quantities of interest from their expected value, in the case considered this was the function g that measures the distance between the true mean of the random vector and its sample estimate. Figures 4.1 and 4.2 show two random samples drawn from a 2-dimensional Gaussian distribution centred at the origin. The sample means are shown with diamonds.

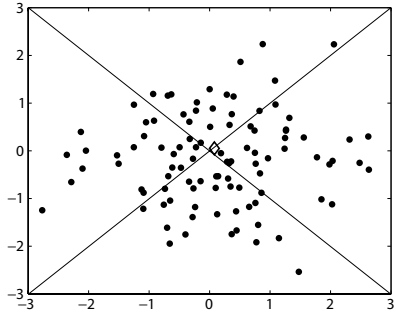

Fig. 4.1. The empirical centre of mass based on a random sample

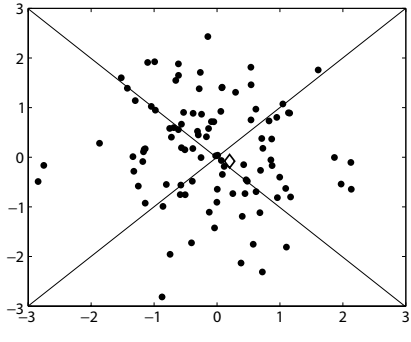

Fig. 4.2. The empirical centre of mass based on a second random sample.

Rademacher variables As mentioned above, the derivation of inequalities (4.2) to (4.4) will form a blueprint for the more general analysis described below. In particular the introduction of the random $\{-1, +1\}$ variables σ_i will play a key role. Such random numbers are known as Rademacher variables. They allow us to move from an expression involving two samples

in equation (4.2) to twice an expression involving one sample modified by the Rademacher variables in formula (4.3).

The result motivates the use of samples as reliable estimators of the true quantities considered. For example, we have shown that the centre of mass of the training sample is indeed a good estimator for the true mean. In the next chapter we will use this result to motivate a simple novelty-detection algorithm that checks if a new datapoint is further from the true mean than the furthest training point. The chances of this happening for data generated from the same distribution can be shown to be small, hence when such points are found there is a high probability that they are outliers.

4.2 Capacity and regularisation: Rademacher theory

In the previous section we considered what were effectively fixed pattern functions, either chosen beforehand or else a fixed function of the data. The more usual pattern analysis scenario is, however, more complex, since the relation is chosen from a set of possible candidates taken from a function class. The dangers inherent in this situation were illustrated in the example involving phone numbers and credit cards. If we allow ourselves to choose from a large set of possibilities, we may find something that 'looks good' on the dataset at hand but does not reflect a property of the underlying process generating the data. The distance between the value of a certain function in two different random subsets does not only depend therefore on its being concentrated, but also on the richness of the class from which it was chosen. We will illustrate this point with another example.

Example 4.7 [Birthday paradox] Given a random set of N people, what is the probability that two of them have the same birthday? This probability depends of course on N and is surprisingly high even for small values of N. Assuming that the people have equal chance of being born on all days, the probability that a pair have the same birthday is 1 minus the probability that all N have different birthdays

$$\begin{aligned} P(\text{same birthday}) &= 1 - \prod_{i=1}^{N} \frac{365 - i + 1}{365} = 1 - \prod_{i=1}^{N} \left(1 - \frac{i-1}{365}\right) \\ &\geq 1 - \prod_{i=1}^{N} \exp\left(-\frac{i-1}{365}\right) = 1 - \exp\left(-\sum_{i=1}^{N} \frac{(i-1)}{365}\right) \\ &= 1 - \exp\left(-\frac{N(N-1)}{730}\right). \end{aligned}$$

It is well-known that this increases surprisingly quickly. For example taking $N = 28$ gives a probability greater than 0.645 that there are two people in the group that share a birthday. If on the other hand we consider a pre-fixed day, the probability that two people in the group have their birthday on that day is

$$P(\text{same birthday on a fixed day}) = \sum_{i=2}^{N} \binom{N}{i} \left(\frac{1}{365}\right)^i \left(\frac{364}{365}\right)^{N-i}.$$

If we evaluate this expression for $N = 28$ we obtain 0.002 7. The difference between the two probabilities follows from the fact that in the one case we fix the day after choosing the set of people, while in the second case it is chosen beforehand. In the first case we have much more freedom, and hence it is more likely that we will find a pair of people fitting our hypothesis. We will expect to find a pair of people with the same birthday in a set of 28 people with more than even chance, so that no conclusions could be drawn from this observation about a relation between the group and that day. For a pre-fixed day the probability of two or more having a birthday on the same day would be less than 0.3%, a very unusual event. As a consequence, in the second case we would be justified in concluding that there is some connection between the chosen date and the way the group was selected, or in other words that we have detected a significant pattern.

Our observation shows that if we check for one property there is unlikely to be a spurious match, but if we allow a large number of properties such as the 365 different days there is a far higher chance of observing a match. In such cases we must be careful before drawing any conclusions. ■

Uniform convergence and capacity What we require if we are to use a finite sample to make inferences involving a whole class of functions is that the difference between the sample and true performance should be small for every function in the class. This property will be referred to as *uniform convergence* over a class of functions. It implies that the concentration holds not just for one function but for all of the functions at the same time.

If a set is so rich that it always contains an element that fits any given random dataset, then the patterns found may not be significant and it is unlikely that the chosen function will fit a new dataset even if drawn from the same distribution. The example given in the previous section of finding a polynomial that maps phone numbers to credit card numbers is a case in point. The capability of a function class to fit different data is known as its *capacity*. Clearly the higher the capacity of the class the greater the risk of

overfitting the particular training data and identifying a spurious pattern. The critical question is how one should measure the capacity of a function class. For the polynomial example the obvious choice is the degree of the polynomial, and keeping the degree smaller than the number of training examples would lessen the risk described above of finding a spurious relation between phone and credit card numbers. Learning theory has developed a number of more general measures that can be used for classes other than polynomials, one of the best known being the Vapnik–Chervonenkis dimension.

The approach we adopt here has already been hinted at in the previous section and rests on the intuition that we can measure the capacity of a class by its ability to fit random data. The definition makes use of the Rademacher variables introduced in the previous section and the measure is therefore known as the Rademacher complexity.

Definition 4.8 [Rademacher complexity] For a sample $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ generated by a distribution $\mathcal{D}$ on a set X and a real-valued function class $\mathcal{F}$ with domain X, the *empirical Rademacher complexity* of $\mathcal{F}$ is the random variable

$$\hat{R}_\ell(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(\mathbf{x}_i)\right| \middle| \mathbf{x}_1, \ldots, \mathbf{x}_\ell\right],$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_\ell\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The *Rademacher complexity* of $\mathcal{F}$ is

$$R_\ell(\mathcal{F}) = \mathbb{E}_S\left[\hat{R}_\ell(\mathcal{F})\right] = \mathbb{E}_{S\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(\mathbf{x}_i)\right|\right].$$

■

The sup inside the expectation measures the best correlation that can be found between a function of the class and the random labels. It is important to stress that pattern detection is a probabilistic process, and there is therefore always the possibility of detecting a pattern in noise. The Rademacher complexity uses precisely the ability of the class to fit noise as its measure of capacity. Hence controlling this measure of capacity will intuitively guard against the identification of spurious patterns. We now give a result that formulates this insight as a precise bound on the error of pattern functions in terms of their empirical fit and the Rademacher complexity of the class.

Note that we denote the input space with Z in the theorem, so that in the case of supervised learning we would have $Z = X \times Y$. We use $\mathbb{E}_{\mathcal{D}}$ for

the expectation with respect to the underlying distribution, while $\hat{\mathbb{E}}$ denotes the empirical expectation measured on a particular sample.

Theorem 4.9 *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from Z to $[0,1]$. Let $(\mathbf{z}_i)_{i=1}^{\ell}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies*

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[f(\mathbf{z})\right] &\leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + R_\ell(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2\ell}} \\ &\leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}.\end{aligned}$$

Proof For a fixed $f \in \mathcal{F}$ we have

$$\mathbb{E}_{\mathcal{D}}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}h - \hat{\mathbb{E}}h\right).$$

We now apply McDiarmid's inequality bound to the second term on the right-hand side in terms of its expected value. Since the function takes values in the range $[0,1]$, replacing one example can change the value of the expression by at most $1/\ell$. Subsituting this value of c_i into McDiarmid's inequality, setting the right-hand side to be $\delta/2$, and solving for ϵ, we obtain that with probability greater than $1-\delta/2$

$$\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}h - \hat{\mathbb{E}}h\right) \leq \mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}h - \hat{\mathbb{E}}h\right)\right] + \sqrt{\frac{\ln(2/\delta)}{2\ell}}.$$

giving

$$\mathbb{E}_{\mathcal{D}}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}h - \hat{\mathbb{E}}h\right)\right] + \sqrt{\frac{\ln(2/\delta)}{2\ell}}.$$

We must now bound the middle term of the right-hand side. This is where we follow the technique applied in the previous section to bound the deviation of the mean of a random vector

$$\begin{aligned}\mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}h - \hat{\mathbb{E}}h\right)\right] &= \mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{\ell}\sum_{i=1}^{\ell}h(\tilde{\mathbf{z}}_i) - \frac{1}{\ell}\sum_{i=1}^{\ell}h(\mathbf{z}_i)\,\middle|\, S\right]\right] \\ &\leq \mathbb{E}_S\mathbb{E}_{\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{\ell}\sum_{i=1}^{\ell}\left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i)\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}S\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{\ell}\sum_{i=1}^{\ell}\sigma_i\left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i)\right)\right]\end{aligned}$$

$$\begin{aligned} &\leq 2\mathbb{E}_{S\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{F}}\left|\frac{1}{\ell}\sum_{i=1}^{\ell}\sigma_i h(\mathbf{z}_i)\right|\right] \\ &= R_\ell(\mathcal{F}). \end{aligned}$$

Finally, with probability greater than $1-\delta/2$, we can bound the Rademacher complexity in terms of its empirical value by a further application of McDiarmid's theorem for which $c_i = 2/\ell$. The complete results follows. □

The only additional point to note about the proof is its use of the fact that the sup of an expectation is less than or equal to the expectation of the sup in order to obtain the second line from the first. This follows from the triangle inequality for the ℓ_∞ norm.

The theorem shows that modulo the small additional square root factor the difference between the empirical and true value of the functions or in our case with high probability the difference between the true and empirical error of the pattern function is bounded by the Rademacher complexity of the pattern function class. Indeed we do not even need to consider the full Rademacher complexity, but can instead use its empirical value on the given training set. In our applications of the theorem we will invariably make use of this empirical version of the bound.

In the next section we will complete our analysis of stability by computing the (empirical) Rademacher complexities of the kernel-based linear classes that are the chosen function classes for the majority of the methods presented in this book. We will also give an example of applying the theorem for a particular pattern analysis task.

4.3 Pattern stability for kernel-based classes

Clearly the results of the previous section can only be applied if we are able to bound the Rademacher complexities of the corresponding classes of pattern functions. As described in Chapter 1, it is frequently useful to decompose the pattern functions into an underlying class of functions whose outputs are fed into a so-called loss function. For example, for binary classification the function class $\mathcal{F}$ may be a set of real-valued functions that we convert to a binary value by thresholding at 0. Hence a function $g \in \mathcal{F}$ is converted to a binary output by applying the sign function to obtain a classification function h

$$h(\mathbf{x}) = \operatorname{sgn}(g(\mathbf{x})) \in \{\pm 1\}.$$

We can therefore express the pattern function using the discrete loss function $\mathcal{L}$ given by

$$\mathcal{L}(\mathbf{x}, y) = \frac{1}{2}|h(\mathbf{x}) - y| = \begin{cases} 0, & \text{if } h(\mathbf{x}) = y; \\ 1, & \text{otherwise.} \end{cases}$$

Equivalently we can apply the *Heaviside function*, $\mathcal{H}(\cdot)$ that returns 1 if its argument is greater than 0 and zero otherwise as follows

$$\mathcal{L}(\mathbf{x}, y) = \mathcal{H}(-yg(\mathbf{x})).$$

Hence, the pattern function is $\mathcal{H} \circ f$, where $f(\mathbf{x}, y) = -yg(\mathbf{x})$. We use the notation $\hat{\mathcal{F}}$ to also denote the class

$$\hat{\mathcal{F}} = \{(\mathbf{x}, y) \mapsto -yg(\mathbf{x}) : g \in \mathcal{F}\}.$$

Using this loss implies that

$$\mathbb{E}_{\mathcal{D}}[\mathcal{H}(-yg(\mathbf{x}))] = \mathbb{E}_{\mathcal{D}}[\mathcal{H}(f(\mathbf{x}, y))] = P_{\mathcal{D}}(y \neq h(\mathbf{x})).$$

This means we should consider the Rademacher complexity of the class

$$\mathcal{H} \circ \hat{\mathcal{F}} = \left\{\mathcal{H} \circ f : f \in \hat{\mathcal{F}}\right\}.$$

Since we will bound the complexity of such classes by assuming the loss function satisfies a Lipschitz condition, it is useful to introduce an auxiliary loss function $\mathcal{A}$ that has a better Lipschitz constant and satisfies

$$\mathcal{H}(f(\mathbf{x}, y)) \leq \mathcal{A}(f(\mathbf{x}, y)), \tag{4.7}$$

where the meaning of the Lipschitz condition is given in the following definition. A function $\mathcal{A}$ satisfying equation (4.7) will be known as a *dominating* cost function.

Definition 4.10 A loss function $\mathcal{A} : \mathbb{R} \to [0, 1]$ is Lipschitz with constant L if it satisfies

$$\left|\mathcal{A}(a) - \mathcal{A}(a')\right| \leq L\left|a - a'\right| \text{ for all } a, a' \in \mathbb{R}.$$

■

We use the notation $(\cdot)_+$ for the function

$$(x)_+ = \begin{cases} x, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

The binary classification case described above is an example where such a function is needed, since the true loss is not a Lipschitz function at all. By taking $\mathcal{A}$ to be the *hinge loss* given by

$$\mathcal{A}(f(\mathbf{x}, y)) = (1 + f(\mathbf{x}, y))_+ = (1 - yg(\mathbf{x}))_+ ,$$

we get a Lipschitz constant of 1 with $\mathcal{A}$ dominating $\mathcal{H}$.

Since our underlying class will usually be linear functions in a kernel-defined feature space, we first turn our attention to bounding the Rademacher complexity of these functions. Given a training set S the class of functions that we will primarily be considering are linear functions with bounded norm

$$\left\{ \mathbf{x} \to \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) : \boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha} \leq B^2 \right\} \subseteq \{ \mathbf{x} \to \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq B \} = \mathcal{F}_B,$$

where $\boldsymbol{\phi}$ is the feature mapping corresponding to the kernel κ and $\mathbf{K}$ is the kernel matrix on the sample S. Note that although the choice of functions appears to depend on S, the definition of $\mathcal{F}_B$ does not depend on the particular training set.

Remark 4.11 [The weight vector norm] Notice that for this class of functions, $f(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle = \left\langle \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}) \right\rangle = \sum_{i=1}^{\ell} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})$, we have made use of the derivation

$$\begin{aligned} \|\mathbf{w}\|^2 &= \langle \mathbf{w}, \mathbf{w} \rangle = \left\langle \sum_{i=1}^{\ell} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{\ell} \alpha_j \boldsymbol{\phi}(\mathbf{x}_j) \right\rangle \\ &= \sum_{i,j=1}^{\ell} \alpha_i \alpha_j \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle = \sum_{i,j=1}^{\ell} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \\ &= \boldsymbol{\alpha}' \mathbf{K} \boldsymbol{\alpha}, \end{aligned}$$

in order to show that $\mathcal{F}_B$ is a superset of our class. We will further investigate the insights that can be made into the structure of the feature space using only information gleaned from the kernel matrix in the next chapter. ■

The proof of the following theorem again uses part of the proof given in the first section showing the concentration of the mean of a random vector. Here we use the techniques of the last few lines of that proof.

Theorem 4.12 *If $\kappa : X \times X \to \mathbb{R}$ is a kernel, and $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ is a sample of points from X, then the empirical Rademacher complexity of the*

class $\mathcal{F}_B$ satisfies

$$\hat{R}_\ell(\mathcal{F}_B) \leq \frac{2B}{\ell}\sqrt{\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i,\mathbf{x}_i)} = \frac{2B}{\ell}\sqrt{\mathrm{tr}\,(\mathbf{K})}$$

Proof The result follows from the following derivation

$$\begin{aligned}
\hat{R}_\ell(\mathcal{F}_B) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_B}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f\left(\mathbf{x}_i\right)\right|\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|\leq B}\left|\left\langle \mathbf{w}, \frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}\left(\mathbf{x}_i\right)\right\rangle\right|\right] \\
&\leq \frac{2B}{\ell}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i)\right\|\right] \\
&= \frac{2B}{\ell}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\left\langle\sum_{i=1}^{\ell}\sigma_i\boldsymbol{\phi}(\mathbf{x}_i), \sum_{j=1}^{\ell}\sigma_j\boldsymbol{\phi}(\mathbf{x}_j)\right\rangle\right)^{1/2}\right] \\
&\leq \frac{2B}{\ell}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{\ell}\sigma_i\sigma_j\kappa(\mathbf{x}_i,\mathbf{x}_j)\right]\right)^{1/2} \\
&= \frac{2B}{\ell}\left(\sum_{i=1}^{\ell}\kappa(\mathbf{x}_i,\mathbf{x}_i)\right)^{1/2}.
\end{aligned}$$

□

Note that in the proof the second line follows from the first by the linearity of the inner product, while to get the third we use the Cauchy–Schwarz inequality. The last three lines mimic the proof of the first section except that the sample is in this case fixed.

Remark 4.13 [Regularisation strategy] When we perform some kernel-based pattern analysis we typically compute a dual representation $\boldsymbol{\alpha}$ of the weight vector. We can compute the corresponding norm B as $\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$ where $\mathbf{K}$ is the kernel matrix, and hence estimate the complexity of the corresponding function class. By controlling the size of $\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$, we therefore control the capacity of the function class and hence improve the statistical stability of the pattern, a method known as *regularisation*. ■

Properties of Rademacher complexity The final ingredient that will be required to apply the technique are the properties of the Rademacher complexity that allow it to be bounded in terms of properties of the loss function. The following theorem summarises some of the useful properties of the empirical Rademacher complexity, though the bounds also hold for the full complexity as well. We need one further definition.

Definition 4.14 Let F be a subset of a vector space. By $\operatorname{conv}(F)$ we denote the set of convex combinations of elements of F. ■

Theorem 4.15 *Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ and $\mathcal{G}$ be classes of real functions. Then:*

(i) *If $\mathcal{F} \subseteq \mathcal{G}$, then $\hat{R}_\ell(\mathcal{F}) \leq \hat{R}_\ell(\mathcal{G})$;*
(ii) $\hat{R}_\ell(\mathcal{F}) = \hat{R}_\ell(\operatorname{conv}\mathcal{F})$;
(iii) *For every $c \in \mathbb{R}$, $\hat{R}_\ell(c\mathcal{F}) = |c|\hat{R}_\ell(\mathcal{F})$;*
(iv) *If $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and satisfies $\mathcal{A}(0) = 0$, then $\hat{R}_\ell(\mathcal{A} \circ \mathcal{F}) \leq 2L\hat{R}_\ell(\mathcal{F})$;*
(v) *For any function h, $\hat{R}_\ell(\mathcal{F} + h) \leq \hat{R}_\ell(\mathcal{F}) + 2\sqrt{\hat{\mathbb{E}}\left[h^2\right]/\ell}$;*
(vi) *For any $1 \leq q < \infty$, let $\mathcal{L}_{\mathcal{F},h,q} = \{|f - h|^q \mid f \in \mathcal{F}\}$. If $\|f - h\|_\infty \leq 1$ for every $f \in \mathcal{F}$, then $\hat{R}_\ell(\mathcal{L}_{\mathcal{F},h,q}) \leq 2q\left(\hat{R}_\ell(\mathcal{F}) + 2\sqrt{\hat{\mathbb{E}}\left[h^2\right]/\ell}\right)$;*
(vii) $\hat{R}_\ell(\sum_{i=1}^n \mathcal{F}_i) \leq \sum_{i=1}^n \hat{R}_\ell(\mathcal{F}_i)$.

Though in many cases the results are surprising, with the exception of (iv) their proofs are all relatively straightforward applications of the definition of empirical Rademacher complexity. For example, the derivation of part (v) is as follows

$$\begin{aligned}
\hat{R}_\ell(\mathcal{F}+h) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i\left(f\left(\mathbf{x}_i\right)+h(\mathbf{x}_i)\right)\right|\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{2}{\ell}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{\ell}\sigma_i f\left(\mathbf{x}_i\right)\right|\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{2}{\ell}\left|\sum_{i=1}^{\ell}\sigma_i h(\mathbf{x}_i)\right|\right] \\
&\leq \hat{R}_\ell(\mathcal{F})+\frac{2}{\ell}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{\ell}\sigma_i h(\mathbf{x}_i)\sigma_j h(\mathbf{x}_j)\right]\right)^{1/2} \\
&= \hat{R}_\ell(\mathcal{F})+\frac{2}{\ell}\left(\sum_{i=1}^{\ell}h(\mathbf{x}_i)^2\right)^{1/2} \\
&= \hat{R}_\ell(\mathcal{F})+\frac{2}{\ell}\left(\ell\hat{\mathbb{E}}\left[h^2\right]\right)^{1/2}.
\end{aligned}$$

The proof of (iv) is discussed in Section 4.6.

Margin bound We are now in a position to give an example of an application of the bound. We will take the case of pattern analysis of a classification function. The results obtained here will be used in Chapter 7 where we describe algorithms that optimise the bounds we derive here based involving either the margin or the slack variables.

We need one definition before we can state the theorem. When using the Heaviside function to convert a real-valued function to a binary classification, the margin is the amount by which the real value is on the correct side of the threshold as formalised in the next definition.

Definition 4.16 For a function $g : X \to \mathbb{R}$, we define its *margin* on an example $(\mathbf{x}, y)$ to be $yg(\mathbf{x})$. The *functional margin* of a training set $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$, is defined to be

$$m(S, g) = \min_{1 \le i \le \ell} y_i g(\mathbf{x}_i).$$

Given a function g and a desired margin γ we denote by $\xi_i = \xi((\mathbf{x}_i, y_i), \gamma, g)$ the amount by which the function g fails to achieve margin γ for the example $(\mathbf{x}_i, y_i)$. This is also known as the example's *slack variable*

$$\xi_i = (\gamma - y_i g(\mathbf{x}_i))_+,$$

where $(x)_+ = x$ if $x \ge 0$ and 0 otherwise. ■

Theorem 4.17 *Fix $\gamma > 0$ and let $\mathcal{F}$ be the class of functions mapping from $Z = X \times Y$ to $\mathbb{R}$ given by $f(\mathbf{x}, y) = -yg(\mathbf{x})$, where g is a linear function in a kernel-defined feature space with norm at most 1. Let*

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$$

be drawn independently according to a probability distribution $\mathcal{D}$ and fix $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over samples of size ℓ we have

$$\begin{aligned} P_{\mathcal{D}}(y \neq \operatorname{sgn}(g(\mathbf{x}))) &= \mathbb{E}_{\mathcal{D}}[\mathcal{H}(-yg(\mathbf{x}))] \\ &\le \frac{1}{\ell\gamma}\sum_{i=1}^{\ell}\xi_i + \frac{4}{\ell\gamma}\sqrt{\operatorname{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}, \end{aligned}$$

where $\mathbf{K}$ is the kernel matrix for the training set and $\xi_i = \xi((\mathbf{x}_i, y_i), \gamma, g)$.

Proof Consider the loss function $\mathcal{A} : \mathbb{R} \to [0, 1]$, given by

$$\mathcal{A}(a) = \begin{cases} 1, & \text{if } a > 0; \\ 1 + a/\gamma, & \text{if } -\gamma \leq a \leq 0; \\ 0, & \text{otherwise.} \end{cases}$$

By Theorem 4.9 and since the loss function $\mathcal{A} - 1$ dominates $\mathcal{H} - 1$, we have that

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\left[\mathcal{H}(f(\mathbf{x}, y)) - 1\right] &\leq \mathbb{E}_{\mathcal{D}}\left[\mathcal{A}(f(\mathbf{x}, y)) - 1\right] \\ &\leq \hat{\mathbb{E}}\left[\mathcal{A}(f(\mathbf{x}, y)) - 1\right] + \hat{R}_\ell((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}. \end{aligned}$$

But the function $\mathcal{A}(-y_i g(\mathbf{x}_i)) \leq \xi_i/\gamma$, for $i = 1, \ldots, \ell$, and so

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{H}(f(\mathbf{x}, y))\right] \leq \frac{1}{\ell\gamma}\sum_{i=1}^{\ell}\xi_i + \hat{R}_\ell((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}.$$

Since $(\mathcal{A} - 1)(0) = 0$, we can apply part (iv) of Theorem 4.15 with $L = 1/\gamma$ to give $\hat{R}_\ell((\mathcal{A} - 1) \circ \mathcal{F}) \leq 2\hat{R}_\ell(\mathcal{F})/\gamma$. It remains to bound the empirical Rademacher complexity of the class $\mathcal{F}$

$$\begin{aligned} \hat{R}_\ell(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i f(\mathbf{x}_i, y_i)\right|\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i y_i g(\mathbf{x}_i)\right|\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_1}\left|\frac{2}{\ell}\sum_{i=1}^{\ell}\sigma_i g(\mathbf{x}_i)\right|\right] = \hat{R}_\ell(\mathcal{F}_1) \\ &= \frac{2}{\ell}\sqrt{\mathrm{tr}(\mathbf{K})}, \end{aligned}$$

where we have used the fact that $g \in \mathcal{F}_1$ that is that the norm of the weight vector is bounded by 1, and that multiplying σ_i by a fixed y_i does not alter the expectation. This together with Theorem 4.12 gives the result. □

If the function g has margin γ, or in other words if it satisfies $m(S, g) \geq \gamma$, then the first term in the bound is zero since all the slack variables are zero in this case.

Remark 4.18 [Comparison with other bounds] This theorem mimics the well-known margin based bound on generalisation (see Section 4.6 for details), but has several advantages. Firstly, it does not involve additional $\log(\ell)$ factors in the second term and the constants are very tight. Furthermore it handles the case of slack variables without recourse to additional constructions. It also does not restrict the data to lie in a ball of some

predefined radius, but rather uses the trace of the matrix in its place as an empirical estimate or effective radius. Of course if it is known that the support of the distribution is in a ball of radius R about the origin, then we have

$$\frac{4}{\ell\gamma}\sqrt{\mathrm{tr}(\mathbf{K})} \leq \frac{4}{\ell\gamma}\sqrt{\ell R^2} = 4\sqrt{\frac{R^2}{\ell\gamma^2}}.$$

Despite these advantages it suffers from requiring a square root factor of the ratio of the effective dimension and the training set size. For the classification case this can be avoided, but for more general pattern analysis tasks it is not clear that this can always be achieved. We do, however, feel that the approach succeeds in our aim of providing a unified and transparent framework for assessing stability across a wide range of different pattern analysis tasks. ■

As we consider different algorithms in later chapters we will indicate the factors that will affect the corresponding bound that guarantees their stability. Essentially this will involve specifying the relevant loss functions and estimating the corresponding Rademacher complexities.

4.4 A pragmatic approach

There exist many different methods for modelling learning algorithms and quantifying the reliability of their results. All involve some form of capacity control, in order to prevent the algorithm from fitting 'irrelevant' aspects of the data. The concepts outlined in this chapter have been chosen for their intuitive interpretability that can motivate the spirit of all the algorithms discussed in this book. However we will not seek to derive statistical bounds on the generalization of every algorithm, preferring the pragmatic strategy of using the theory to identify which parameters should be kept under control in order to control the algorithm's capacity. For detailed discussions of statistical bounds covering many of the algorithms, we refer the reader to the last section of this and the following chapters, which contain pointers to the relevant literature.

The relations we will deal with will be quite diverse ranging from correlations to classifications, from clusterings to rankings. For each of them, different performance measures can be appropriate, and different cost functions should be optimised in order to achieve best performance. In some cases we will see that we can estimate capacity by actually doing the randomisation ourselves, rather than relying on a priori bounds such as those

given above. Such attempts to directly estimate the empirical Rademacher complexity are likely to lead to much better indications of the generalisation as they can take into account the structure of the data, rather than slightly uninformative measures such as the trace of the kernel matrix.

Our strategy will be to use cost functions that are 'concentrated', so that any individual pattern that has a good performance on the training sample will with high probability achieve a good performance on new data from the same distribution. For this same stability to apply across a class of pattern functions will depend on the size of the training set and the degree of control that is applied to the capacity of the class from which the pattern is chosen. In practice this trade-off between flexibility and generalisation will be achieved by controlling the parameters indicated by the theory. This will often lead to regularization techniques that penalise complex relations by controlling the norm of the linear functions that define them.

We will make no effort to eliminate every tunable component from our algorithms, as the current state-of-the-art in learning theory often does not give accurate enough estimates for this to be a reliable approach. We will rather emphasise the role of any parameters that can be tuned in the algorithms, leaving it for the practitioner to decide how best to set these parameters with the data at his or her disposal.

4.5 Summary

- The problem of determining the stability of patterns can be cast in a statistical framework.
- The stability of a fixed pattern in a finite sample can be reliably verified if it is statistically concentrated, something detectable using McDiarmid's inequality.
- When considering classes of pattern functions, the issue of the capacity of the class becomes crucial in ensuring that concentration applies simultaneously for all functions.
- The Rademacher complexity measures the capacity of a class. It assesses the 'richness' of the class by its ability to fit random noise. The difference between empirical and true estimation over the pattern class can be bounded in terms of its Rademacher complexity.
- Regularisation is a method of controlling capacity and hence ensuring that detected patterns are stable.
- There are natural methods for measuring and controlling the capacity of linear function classes in kernel-defined feature spaces.

4.6 Further reading and advanced topics

The modelling of learning algorithms with methods of empirical processes was pioneered by Vladimir Vapnik and Alexei Chervonenkis (VC) [144], [145] in the 1970s, and greatly extended in more recent years by a large number of other researchers. Their work emphasised the necessity to control the capacity of *a class* of functions, in order to avoid overfitting, and devised a measure of capacity known as VC dimension [142].

Their analysis does not, however, extend to generalisation bounds involving the margin or slack variables. The first papers to develop these bounds were [124] and [8]. The paper [124] developed the so-called luckiness framework for analysing generalisation based on fortuitous observations during training such as the size of the margin. The analysis of generalisation in terms of the slack variables in the soft margin support vector machine is given in [125]. A description of generalisation analysis for support vector machines based on these ideas is also contained in Chapter 4 of the book [32]. In this chapter we have, however, followed a somewhat different approach, still within a related general framework.

The original VC framework was specialised for the problem of classification, and later extended to cover regression problems and novelty-detection. Its extension to general classes of patterns in data is difficult. It is also well-known that traditional VC arguments provide rather loose bounds on the risk of overfitting. A number of new methodologies have been proposed in recent years to overcome some of these problems, mostly based on the notion of concentration inequalities [18], [17], and the use of Rademacher complexity: [80], [9], [82], [10], [80]. At an intuitive level we can think of Rademacher complexity as being an empirical estimate of the VC dimension. Despite the transparency of the results we have described, we have omitted a proof of part (iv) of Theorem 4.15. This is somewhat non-trivial and we refer the interested reader to [85] who in turn refer to [85]. The full proof of the result requires a further theorem proved by X. Fernique.

The analysis we presented in this chapter aims at covering all the types of patterns we are interested in, and therefore needs to be very general. What has remained unchanged during this evolution from VC to Rademacher-type of arguments, is the use of the notion of uniform convergence of the empirical means of a set of random variables to their expectations, although the methods for proving uniform convergence have become simpler and more refined. The rate of such uniform convergence is however still dictated by some measure of richness of such set.

The use of Rademacher Complexity for this purpose is due to [80]. Our

discussion of Rademacher complexity for kernel function classes is based on the paper by Bartlett and Mendelson [10] and on the lectures given by Peter Bartlett at UC Berkeley in 2001. The discussion of concentration inequalities is based on Boucheron, Lugosi and Massart [17] and on the seminar notes of Gabor Lugosi.

More recently tighter bounds on generalisation of SVMs has been obtained using a theoretical linking of Bayesian and statistical learning [84]. Finally, notions of regularizations date back to [138], and certainly have been fully exploited by Wahba in similar contexts [155].

The books [38] and [4] also provide excellent coverage of theoretical foundations of inference and learning.

For constantly updated pointers to online literature and free software see the book's companion website: www.kernel-methods.net