An introduction to the Concentration of Measure and Empirical Process Theory

Peadar Coyle

January 23, 2014

Contents

List of Figures

1 Acknowledgements

Writing a thesis is a difficult endeavour, and one often wonders if the work is worth it. However seeing the fruits of your labour and remembering that you have learned some wonderful Mathematics - makes it all worth it. I ended up doing this thesis - in Mathematics due to what seems like a stochastic process - ending up in Luxembourg was not in my plan. However I am grateful to the staff at the University, for all their guidance and support over the past few years. I am also grateful to the Machine Learning experts I spoke to during my Internship at Amazon.com and Import.io, I would never have written this thesis if I hadn't been exposed to the reality of *beautiful industrial applications* of statistics and probability theory. I am deeply grateful to Professor Pecatti for his sense of humour (essential when working with me) and inspiring me to learn about Sudakov-Fernique Inequalities[1].

Early drafts of this thesis were read by a Mr Daniel Rowlands and a Dr Robert Horton - I am very grateful that my statistically literate friends could help me solidify my ideas and presentation, writing often feels like a lonely process and community support is invaluable. I also thank my various mentors and friends - Ben, Gregorio, Usman, Matt, Alan, Pierre, Miles, Mike, Alexander, Chris, Danielle, Arthur and all the others who I've forgotten to mention - I am very grateful for your discussions about Mathematics, Physics, Philosophy and life over the past few years. Finally - I thank my girlfriend and family for all their help and support - it is always good to have someone remind you that there are more important things than work.

[1] The acorn of this thesis was in a Mathematics Seminar given by the author last year on a particular Concentration Inequality called the Sudakov-Fernique Inequality[16]

For Audrey
Amare et sapere vix deo conceditur

2 Notation

We introduce some notation that is used throughout the paper. We assume that $X_1, \cdots, X_n$ are independent random variables taking values in a measurable space $\mathcal{X}$. Denote by X_1^n the vector of these n random variables. Let $f : \mathcal{X}^n \to \mathbb{R}$be some measurable function.

- almost surely is denoted by a.s.
- X,Y,Z refer to random variables.
- $\mathbb{P}$ Probability distribution
- $\mathbb{R}, \mathbb{C}$ field of real and complex numbers respectively

3 Introduction

In this thesis we will introduce some aspects from Empirical process theory and Concentration of Measure. We are interested in applications in Learning theory but do not have the time nor the scope to introduce model theory and VC-theory. Our aims are more modest, we want to introduce some concentration inequalities, show their applications in Empirical process theory and mention in passing some Statistical Learning theory. We will not be able to introduce all of these complex notions and we refer to the literature when we feel it is good to do so. Our first subsection will introduce Concentration of Measure and in passing we will see some of the celebrated theorems and inequalities. Elsewhere in thesis we will examine these theorems in more detail readers who are already familiar with the theorems - or perhaps Computer Scientists can just read the introduction and move to the end. The style of this paper will be mostly aimed at pure mathematicians - so the standard of rigor is high. I have endeavoured to add my own opinions about what is difficult and what is easy, and elucidate some of these *unoriginal* and classical proofs. Most of what is reproduced here can be found in other textbooks or seminars, but the aim here is to condense as many of these as possible and rewrite them for an applied Statistician or Machine Learning expert audience.

Let us suppose that we have a large number of scalar random variables $X_1, \cdots, X_n$, which have a bounded size on average (e.g. their mean and variance would be O(1)). An interesting and important question is what can one say about their sum? $S = X_1 + \cdots + X_n$? If each individual sumand X_i varies in an interval of size O(1), then their sum of course varies in an interval of size $0(n)$. However a remarkable phenomenon, known as *concentration of measure*, asserts that assuming a sufficient amount of independendence between the component variables $X_1, \cdots, X_n$, this sum sharply concentrates to a much narrower range, typically an interval of size $O(\sqrt{n})$. This phenomenon is quantified by a variety of *large deviation inequalities*[2] that give upper bounds (often exponential in nature) on the probability that such a combined random variable deviates significantly from its mean. The same phenomenon applies not only to linear expressions

[2] Roughly speaking , large deviations theory concerns itself with the exponential decline of probability measures of certain kinds of extreme or tail events as the number of observations grows large see [53, 59] for further details. Readers interested in a thorough introduction can read from the *"Mozart"* of large deviations theory [57]

such as $S = X_1 + \cdots + X_n$, but more generally to nonlinear combinations $F(X_1, \cdots, X_n)$ of such variables, provided that the nonlinear function F is sufficiently regular - for example Lipschitz.

The basic intuition is that independent random variables find it difficult to "work together" to simultaneously pull a sum $S = X_1 + \cdots + X_n$ or a more general combination $F(X_1, \cdots, X_n)$ too far from its' mean. Independendence is the key here; concentration of measure results typically fail if the X_i are too highly concentrated to each other. Although such results do exist see for example the work by Kontorovich on Mixing Phenomenoa [33] or [39] for work using time series data.

There are many applications of concentration of the concentration of measure phenomenon, such as random matrix theory, but we will mostly focus on applications in Computer Science not Physics[3]. We wholeheartedly recommend the excellent book by Terry Tao on Random Matrix Theory [52]. [4]

Assuming that one has a sufficient amount of independence, the concentration of measure tends to be sub-gaussian in nature, this probability that one is at least λ standard deviations (s.d.) from the mean tends to drop off like $C\exp(-c\lambda^2)$ for some C,c > 0. In particular, one is $O(\log^{\frac{1}{2}} n)$ standard deviations from the mean with overwhelming probability. Indeed, concentration of measure is our primary tool for ensuring that various events hold with overwhelming probability.

3.1 What does the word Application mean?

The word *applications* means different things to different people. If we wear our Mathematican hats we are always looking for a new tool to prove some theorem. Or reprove a celebrated theorem in a new way. Computer Scientists are always concerned with algorithms and performanace analysis and statisticians or *data scientists*[5] are after getting a handle on the

[3] Ironic perhaps, since the author once studied in a Physics department

[4] For an excellent result relating to concentration inequalities in Random Matrix theory I recommend [54], it truly is amazing to see the power of Concentration of Measure

[5] It will be an interesting test case for the durability of job titles in industrial applications of Mathematics to see if the term data scientist is dated in 5 years.

asymptotics of estimators and samplers. In short, each specialist wants to know what a given tool will contribute to his field.

The remarkable thing about concentration of measure is that is uses span a wide range - from something practical as decoding neural signals to esoteric topics such as analyzing convex bodies in Banach spaces. It is too much to go beyond one or two applications. So I'll cite some here

- Milman's proof of Dvoretzky's theorem on sections of convex bodies[41]
- a widely cited lemma of Johnson and Lindenstrauss concerning low-distortion dimensionality reduction in $\mathbb{R}^n$ by random projections[31].
- Applications in Economic Forecasting[39].
- Statistics and empirical processes and machine learning[11]

In the interest of keeping this document short we shall only look at the last example.

3.2 What is in this thesis?

In section 4 we give a rapid overview of concentration of measure including some of the more advanced methods like the Herbst argument which will not be used in the thesis. In section 4.3 we will introduce someof the language from applications in statistical learning theory including in particular empirical process theory. In the following section we include some remarks about Machine Learning and Artificial Intelligence, this is just to motivate the thesis4.4.

In 5 we introduce the major concentration inequalities (we refer the reader for further details to [34]) which we shall use - we start with Hoeffding and then introduce a few variants and finally the celebrated McDiarmid Inequality[37] - which we will use to prove results in the applications section. In 6 we return to our applications - and give a more rigorous and complete introduction to our chosen subfield of Statistical Learning Theory - this will be of particular interest to Mathematicans who may not be familiar with the language. The remainder of the 6 introduces the celebrated Vapnik-Chervonenkis dimnesion. In 7 we introduce the meat of the applications through the technique of Rademacher averages, we end the section by elucidating on the difficulty of calculating such averages - in learning theory through the lens of empirical risk. We follow this up by considering dependent data in 8. After reminding the reader of the language of time series [40, 48] we explicitly calculate some Risk bounds using the language of VC theory and Rademacher 9 and then consider this for a fixed-memory model from Econometrics 9.2. Finally in 10 we end the thesis with some discussion and bibliographic remarks.

4 What are other examples of Concentration of Measure?

We have already met the intuitive idea of *concentration of measure* let us make this more mathematical. Let X be a random variable taking values in some metric space X. Then we say that the distribution of X has the concentration property if, for any set $A \subset \mathsf{X}$ such that $\mathbb{P}(X \in A) \geq 1/2$, we have

$$\mathbb{P}\left(d(X,A) \leq r\right) \xrightarrow{r\to\infty} 1. \tag{1}$$

where r is a constant and $\mathbb{P}$ is some unknown probability distribution. Here, $d(x,A)$ is the distance from the point $x \in \mathsf{X}$ to the set A:

$$d(x,A) := \inf_{y\in A} d(x,y).$$

Another way to express 1 is as follows: for any set $A \subset \mathsf{X}$ and any $r \geq 0$, define the r-blowup of A by

$$A_r := \{y \in \mathsf{X} : d(y,A) \leq r\} \equiv \{y \in \mathsf{X} : \exists x \in A \text{ such that } d(x,y) \leq r\}.$$

Then X has the concentration property if

$$\mathbb{P}(X \in A) \geq 1/2 \quad \Longrightarrow \quad \lim_{r\to\infty} \mathbb{P}(X \in A_r) = 1.$$

In other words, X has the concentration property if any set containing X with not too small a probability can be blown up to contain X with near-certainty.

Here are two classic examples of concentration:

- **Gaussian distribution in Euclidean space.** Let $\mathsf{X} = \mathbb{R}^n$ and take $d(x,y) = \|x-y\|_2$ the usual Euclidean distance. Let X be a standard n-dimensional Gaussian random variable, i.e., $X \sim N(0, I_n)$, where I_n is the $n \times n$ identity matrix. Then for any $r \geq 0$ we have

  $$\mathbb{P}(X \in A) \geq 1/2 \quad \Longrightarrow \quad \mathbb{P}(X \in A_r) \geq 1 - e^{-r^2/2}.$$

- **Uniform distribution in Hamming space.** Let X be the Hamming cube $\{0,1\}^n$ equipped with the normalized Hamming distance

  $$d(x,y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \neq y_i\}}$$

  that counts the fraction of bits in which $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ disagree. Let X have the uniform distribution on $\{0,1\}^n$, i.e., $\mathbb{P}(X = x) = 2^{-n}$ for all x. Then

  $$\mathbb{P}(X \in A) \geq 1/2 \quad \Longrightarrow \quad \mathbb{P}(X \in A_r) \geq 1 - e^{-2nr^2}.$$

  1

These two examples suggest that we should aim for hard statements in the form of sharp bounds on the concentration function

$$\alpha_X(r) := \sup_{A:\, \mathbb{P}(X \in A) \geq 1/2} \mathbb{P}(X \notin A_r)$$

as opposed to merely soft statements of the form $\alpha_X(r) \to 0$ as $r \to \infty$. The 64,000 pounds question is: how do we get such bounds?
There are two ways to accomplish this goal, and the main idea underlying these two ways is to replace sets with some other objects that are hopefully easier to handle. The first way is to replace sets by probability measures, the second is to replace them by functions.
Fix a set $A \subset \mathsf{X}$ with $\Pr(X \in A) > 0$. Let $\mathbb{P}$ denote the distribution of X, and let P_A denote the conditional distribution of X given $X \in A$. That is, for any (measurable) set $B \subset \mathsf{X}$ we have

$$P_A(B) := \frac{P(A \cap B)}{P(A)}.$$

I am using the subscript notation $\mathbb{P}_A$ instead of the more usual $\mathbb{P}(\cdot|A)$ to indicate the fact that $\mathbb{P}_A$ is a probability measure in its own right. In this way, we can associate to each non-null set A a probability measure $\mathbb{P}_A$. Now, here is a very simple observation that turns out to be very consequential[6]:

$$D(\mathbb{P}_A \| \mathbb{P}) = \log \frac{1}{\mathbb{P}(A)}. \tag{2}$$

This is very easy to prove: for any set B we have

$$\mathbb{P}_A(B) = \frac{1}{\mathbb{P}(A)} \int_B 1_A(x) \mathbb{P}(dx), \tag{3}$$

so $\mathbb{P}_A$ is absolutely continuous with respect to $\mathbb{P}$ with the Radon-Nikodym derivative

$$d\mathbb{P}_A / d\mathbb{P} = 1_A / \mathbb{P}(A)$$

. Therefore, by definition of the divergence,

$$D(\mathbb{P}_A \| \mathbb{P}) = \int d\mathbb{P}_A \log \frac{d\mathbb{P}_A}{d\mathbb{P}} = \frac{1}{\mathbb{P}(A)} \int_A d\mathbb{P} \log \frac{1}{\mathbb{P}(A)} = \log \frac{1}{\mathbb{P}(A)}.$$

So if we are interested in bounding the probabilities of various sets A, we may hope to get some mileage out of the relationship 2.

On the other hand, we may also associate to a set A with $\mathbb{P}(A) > 0$ the function

$$f_A(x) := d(x, A) \equiv \inf_{y \in A} d(x, y).$$

This function is Lipschitz: for any $x, x' \in \mathsf{X}$,

$$f_A(x) - f_A(x') = \inf_{y \in A} d(x, y) - \inf_{y \in A} d(x', y) \leq \sup_{y \in A} [d(x, y) - d(x', y)] \leq d(x, x'),$$

where the last step is by the triangle inequality. Interchanging the roles of x and x', we get the Lipschitz property. Moreover, let us consider the random variable $Z = f_A(X)$, where X is our original X-valued random variable. Then we immediately notice two things:

For any $r \geq 0$, $\mathbb{P}(Z \leq r) = \mathbb{P}(d(X, A) \leq r) = \mathbb{P}(A_r)$. If $P(A) = \mathbb{P}(X \in A) \geq 1/2$, then 0 is a median of Z, in the sense that

[6] In fact we call the following equation *relative entropy* or *Kullback-Leibler* distance - we recommend that readers interested in Information Theory look at [20]. Unfortunately it is beyond the scope of this thesis to introduce all the nomeclature from Information Theory.

$$\mathbb{P}(Z \le 0) = \mathbb{P}(A) \ge 1/2 \qquad \text{and} \qquad \mathbb{P}(Z > 0) \ge 1/2$$

(the second inequality is obviously true since Z is nonnegative with probability 1).
These two observations suggest that we may obtain concentration bounds by bounding the probability that a given Lipschitz function of X deviates from its median by more than r. In fact, it is easy to derive an alternative expression for the concentration function α_X:

$$\alpha_X(r) = \sup_{1-\text{Lipschitz } f} \mathbb{P}\Big(f(X) > m_f + r\Big), \tag{4}$$

where m_f denotes any median of $f(X)$. We already showed, by passing from A to $f_A = d(\cdot, A)$, that α_X is bounded from above by the quantity on the right-hand side of (4):

$$\alpha_X(r) = \sup_{A:\, P(A)\ge 1/2} \mathbb{P}\Big(f_A(X) > \underbrace{m_{f_A}}_{=0} + r\Big) \le \sup_{1-\text{Lipschitz } f} \mathbb{P}\Big(f(X) > m_f + r\Big)$$

To prove the reverse inequality, fix any 1-Lipschitz function f and consider the set $A_f := \{x \in \mathsf{X} : f(x) \le m_f\}$, where m_f is any median of f. Then, by definition,

$$\mathbb{P}(X \in A_f) = \mathbb{P}\Big(f(X) \le m_f\Big) \ge 1/2.$$

Moreover, if we consider the r-blowup

$$[A_f]_r = \Big\{x \in \mathsf{X} : d(x, A_f) \le r\Big\},$$

then for any $x \in \mathsf{X}$ and any $y \in [A_f]_r$ we must have

$$f(x) - m_f \le f(x) - f(y) \le d(x, y),$$

where the last step is by the Lipschitz property of f. Consequently, by definition of the concentration function,

$$\mathbb{P}\Big(f(X) > m_f + r\Big) \le \mathbb{P}\Big(d(X, A_f) > r\Big) = 1 - P\left([A_f]_r\right) \le \alpha_X(r).$$

By passing to the functional viewpoint, we obtain another equivalent characterization of the concentration property: a random variable X taking values in a metric space (X, d) has the concentration property if real-valued Lipschitz functions X are nearly constant.

4.1 Probabilistic viewpoint

Lets look at the first, probabilistic viewpoint, which was born out of a 1996 breakthrough paper by Marton[36]. Given a metric space (X, d), let us define the L_1 Wasserstein distance (or transportation distance) between any two probability measures P and Q on it:

$$W_1(P, Q) := \inf_{X \sim P, Y \sim Q} \mathbb{E}[d(X, Y)],$$

where the infimum is over all jointly distributed random variables $X, Y \in \mathsf{X}$, such that $P_X = P$ and $P_Y = Q$. Now consider a random variable $X \in \mathsf{X}$, for which we wish to establish concentration. What Marton showed is the following: Suppose the distribution P of X satisfies the L_1 transportation inequality

$$W_1(P, Q) \leq \sqrt{2c\, D(Q \| P)} \tag{5}$$

for some constant $c > 0$. Then X has the concentration property, and moreover

$$P(A) \geq 1/2 \quad \Longrightarrow \quad P(A_r) \geq 1 - \exp\left(-\frac{1}{2c}\left(r - \sqrt{2c \log 2}\right)^2\right), \forall r > \sqrt{2c \log 2}.$$

Martons proof is breathtakingly beautiful. Consider any two sets A, B with $P(A), P(B) \neq 0$. Recalling our notation for conditional distributions, we can write

$$\begin{aligned} W_1(P_A, P_B) &\leq W_1(P_A, P) + W_1(P_B, P) \\ &\leq \sqrt{2c\, D(P_A \| P)} + \sqrt{2c\, D(P_B \| P)} \\ &= \sqrt{2c \log \tfrac{1}{P(A)}} + \sqrt{2c \log \tfrac{1}{P(B)}}, \end{aligned}$$

where in the first step we have used the triangle inequality, in the second we have used the fact that P satisfies the transportation inequality (5), and in the last step we have used the formula (2). Now suppose that $P(A) \geq 1/2$ and let $B = A_r^c$ for some r, where c denotes set-theoretic complement. Then we can show that $W_1(P_A, P_B) \geq d(A, B) \geq r$. On the other hand,

$$\log \frac{1}{P(A)} \leq \log 2 \qquad \text{and} \qquad \log \frac{1}{P(B)} = \log \frac{1}{1 - P(A_r)}.$$

Combining these facts gives us the bound

$$r \leq \sqrt{2c \log 2} + \sqrt{2c \log \frac{1}{1 - P(A_r)}}$$

that holds for all r. If $r > \sqrt{2c\log 2}$, then we get

$$P(A_r) \geq 1 - \exp\left(-\frac{1}{2c}\left(r - \sqrt{2c\log 2}\right)^2\right),$$

so we indeed have concentration and a sharp bound on $\alpha_X(r)$, at least for large enough values of r. The main message here is that, in order to study concentration, it suffices to work on the level of probability measures and to focus ones effort on showing that the distribution of X satisfies a suitable transportation inequality. Since Martons original work, there have been many refinements and extensions, which I will not go into here. One such result, due to Sergey Bobkov and Friedrich Götze[9], says that P satisfying a transportation inequality (5) is equivalent to the Gaussian concentration property

$$\alpha_X(r) \leq e^{-r^2/2c}, \qquad \forall r \geq 0.$$

Now lets look at the complementary functional viewpoint. Recall that we seek tight upper bounds on deviation probabilities of the form

$$\mathbb{P}\Big(f(X) \geq m_f + r\Big), \qquad \forall r > 0.$$

It is easier to work with means instead of medians, and indeed it can be shown that concentration around the mean is equivalent to concentration around any median. So lets focus on the mean. Let X, as before, be a random variable over some metric space (X, d), and consider a Lipschitz function $f : \mathsf{X} \to \mathbb{R}$ such that $\mathbb{E}[f(X)] = 0$. We can apply the well-known Chernoff trick: for any $r, \lambda > 0$ we have

$$\mathbb{P}\Big(f(X) \geq r\Big) = \mathbb{P}\Big(e^{\lambda f(X)} \geq e^{\lambda r}\Big) \leq e^{-\lambda r}\mathbb{E}[e^{\lambda f(X)}].$$

Now the whole affair hinges on the availability of tight upper bounds on the logarithmic moment-generating function $\Lambda(\lambda) := \log \mathbb{E}[e^{\lambda f(X)}]$. The entropy method[47, 34], is the name for a set of techniques for bounding $\Lambda(\lambda)$ by means of various inequalities involving the relative entropy between various tilted distributions derived from P and P itself[7].

The entropy method has roots in the work of Michel Ledoux, who in turn distilled it from some very deep results of Michel Talagrand[51, 42, 34].

The simplest version of the entropy method goes something like this. Let us define, for any $t \in \mathbb{R}$, the tilted distribution $P^{(t)}$ via

[7] We shall use P and $\mathbb{P}$ interchangeable for the rest of this section. They both mean distribution

$$\frac{dP^{(t)}}{dP}(x) = \frac{e^{tf(x)}}{e^{\Lambda(t)}}$$

(assuming, of course, that $\Lambda(t)$ exists and is finite). Then we have

$$\begin{array}{rcl} D(P^{(t)}\|P) & = & \int dP^{(t)}\left[tf - \Lambda(t)\right] \\ & = & \frac{1}{e^{\Lambda(t)}} t\,\mathbb{E}\left[f(X)e^{tf(X)}\right] - \Lambda(t) \\ & = & t\Lambda'(t) - \Lambda(t) \\ & = & t^2 \frac{d}{dt}\left(\frac{\Lambda(t)}{t}\right). \end{array}$$

Integrating and using the fact that $\Lambda(0) = 0$, we get

$$\Lambda(\lambda) = \lambda \int_0^\lambda \frac{D(P^{(t)}\|P)}{t^2} dt. \qquad (6)$$

Now suppose that we can bound $D(P^{(t)}\|P) \leq ct^2/2$ for some $c > 0$. Then from (6) we have

$$\Lambda(\lambda) \leq \frac{c\lambda^2}{2}, \qquad \forall t$$

which in turn gives the Gaussian tail bound

$$\mathbb{P}\Big(f(X) \geq r\Big) \leq \inf_{\lambda>0} \exp\left(-\lambda r + \frac{c\lambda^2}{2}\right) = \exp\left(-\frac{r^2}{2c}\right).$$

This is the so-called Herbst argument[34]. Of course, I have glossed over the most nontrivial part namely, showing that we can bound the relative entropy $D(P^{(t)}\|P)$ by a quadratic function of t. I refer to [35, 34, 5] for further details.

4.2 Some further classic examples

Here is one classic example. Suppose that our function f has the bounded difference property, i.e., there exist some constants $c_1, \ldots, c_n \geq 0$, such that changing the i- th argument of f while keeping others constant will change the value of f by at most c_i:

$$\sup_{x_1,\ldots,x_n} \sup_{x'_i} |f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x'_i,\ldots,x_n)| \leq c_i.$$

We can express this more succinctly as a Lipschitz property of f if we define the weighted Hamming metric [8].

$$d(x, y) = \sum_{i=1}^{n} c_i 1_{\{x_i \neq y_i\}} \tag{7}$$

(we can assume without loss of generality that all the c_is are strictly positive, because we can simply ignore those coordinates of x that do not affect the value of f). Then the bounded difference property is equivalent to saying that f is 1-Lipschitz with respect to this weighted Hamming metric. Moreover, it is possible to show that any product probability measure $P = P_1 \otimes \ldots \otimes P_n$ on the product space $\mathsf{X} = \mathcal{X}^n$ satisfies the transportation inequality

$$W_1(Q, P) \leq \sqrt{2c\, D(Q \| P)},$$

where the Wasserstein distance is computed with respect to the weighted Hamming metric (7), and $c = \frac{1}{4}\sum_{i=1}^{n} c_i^2$. By the BobkovGötze result quoted above, this is equivalent to the concentration bound

$$\mathbb{P}\left(f(X) \geq \mathbb{E}[f(X)] + r\right) \leq \exp\left(-\frac{r^2}{2c}\right) = \exp\left(-\frac{2r^2}{\sum_{i=1}^{n} c_i^2}\right).$$

This is the well-known McDiarmids inequality[38] - which we shall use a lot in this thesis. What is important to state is that this inequality was originally derived using martingale techniques, but here we have arrived at it through a completely different route that took us back to where we started: concentration of Lipschitz functions around their means and/or medians, which (as we saw) is the same thing as the original, stochastic-geometric view of the concentration of measure phenomenon. This *equivalence* between *concentration of Lipshitz functions* and *"stochastic-geometric"* views of the concentration of measure phenomenon is a very important general phenomenon, and important for proving new theorems or finding new insights in Probability Theory or information theory[20]. For a thorough introduction to the uses of Concentration of Measure in Information theory we recommend Raginsky [47]. For other applications of concentration of measure we refer to [10, 34, 35, 11]. As mentioned before this work evolved out of some work on the Sudakov-Fernique inequality (in particular a modern proof by Chaterjee[16], readers who are familiar with Stein inequalities and want to

[8] Readers interested in learning about Hamming metrics and Coding theory should refer to [20]

see some applications of Concentration Inequalities to Statistical Mechanics can refer to [17]

4.3 What is Empirical Process theory

The simplest example of an empirical process arises when trying to estimate a probability distribution from sample data. For those interested in applications it is worth highlighting that the *data* could be from Economics - say we wanted to model when the next recession would occur based on historical time-series data [39] or it could be data from a social media website. The difference between the empirical distribution function $F_n(x)$ and the true distribution function $F(x)$ converges to zero everywhere (by the law of large numbers), and this is non- trivial the maximum difference between the empirical and true distribution functions converges to zero, too (by the Glivenko-Cantelli theorem, a uniform law of large numbers). The "empirical process" $E_n(x)$ is the re-scaled difference, $n^{1/2}[F_n(x)F(x)]$, and it converges to a Gaussian stochastic process that only depends on the true distribution (by the functional central limit theorem).

Empirical process theory is concerned with generalizing this sort of material to other stochastic processes determined by random samples, and indexed by infinite classes (like the real line, or the class of all Borel sets on the line, or some space parameterizing a regression model). The typical objects of concern are proving uniform limit theorems, and with establishing distributional limits. (For instance, one might one want to prove that the errors of all possible regression models in some class will come close to their expected errors, so that maximum-likelihood or least-squares estimation is consistent.) This endeavor is closely linked to Vapnik-Chervonenkis-style learning theory, and in fact one can see VC theory as an application of empirical process theory. So it is very difficult to untangle Vapnik-Chervonenkis(VC)-style learning theory from Concentration of Measure and Empirical Process theory. Hence throughout this thesis you will see references to VC theory, and some of the celebrated theorems proven. It is worth highlighting that most of these areas are still active research areas - and you may encounter some 'simple' questions which are not easy to prove, or are in fact open questions. But then we do live in the *Golden age* of statistics and stochastic processes! Readers interested in more about Empirical Process theory should read [55, 26, 45, 10] or the famous monograph [46]. these all introduce the topic in more depth.

4.4 So why should we care about empirical process theory?

Well let's say that you the reader are statistically literate, and you've learned some concentration of measure, and you have an interest in model selection - that is selecting among different mathematical model which all purport to describe the same data set. The basic strategy is to find conditions under which every model in a reasonable class will, with high probability, perform about as well on sample data as they can be expected to do on new data; this involves constraining the richness or flexibility of the model class.
One of the *buzzwords* in contemporary science is Machine Learning. We shall consider statistical learning a subtype of machine learning and leave the reader to form their own impressions of what machine learning is. Let us use the Vapnik notion of statistical learning for pedagogical reasons. We want to estimate some functional which depends on an unknown distribution over a probability space X –it could be a *concept*[9],regression coefficients, moments of the distribution, Shannon entropy, etc., even the distribution itself. We have a class of admissible distributions, called hypotheses, and a "loss functional," an integral over X which tells us, for each hypothesis, how upset we should be when we guess wrong; this implicitly depends on the true distribution. Clearly we want the best hypothesis, the one which minimizes the loss functional — but to explicitly calculate that we'd need to know the true distribution. Vapnik assumes that we have access to a sequence of independent random variables, all drawn from the (stationary) true distribution. What then are we to do?

One answer which I term the *Vapnik answer* takes two parts. The first has to do with "empirical risk minimization": approximate the true, but unknown, loss functional, which is an integral over the whole space X, with a sum over the observed data-points, and go with the hypothesis that minimizes this "empirical risk"; call this, though Vapnik doesn't, the ERM hypothesis. It's possible that the ERM hypothesis will do badly in the future, because we blundered into unrepresentative data, but we can show necessary and sufficient conditions for the loss of the ERM hypothesis to converge in probability to the loss of the best hypothesis. Moreover, we can prove that under certain very broad conditions, that if we just collect enough data-points, then the loss of the ERM hypothesis is, with high

[9] In the machine learning sense

probability, within a certain additive distance (“confidence interval” — Vapnik’s scare-quotes) of the loss of the best hypothesis. These conditions involve the Vapnik-Chervonenkis dimension, and a related quantity called the Vapnik-Chervonenkis entropy. Very remarkably, we can even calculate how much data we need to get a given approximation, at a given level of confidence, regardless of what the true distribution is, i.e. we can calculate distribution-independent bounds. (They do, however, depend on the nature of the integrands in the loss functional.)

These results about convergence, approximation, etc. are in essence extensions of the Law of Large Numbers to spaces of functions. As such the assumption that successive data-points are independent and identically distributed is key to the whole exercise. While it is possible to consider dependant data in Statistical learning we shall not consider it in this thesis. The second part of Vapnik’s procedure is an elaboration of the first: For a given amount of data, we pick the hypothesis which minimizes the sum of the empirical risk and the “confidence interval” about it. This is termed by Vapnik - *structural risk minimization* and shall not be considered in this thesis. I recommend [55] for further details.

5 Basic concentration inequalities via the martingale approach

In the following section, some basic inequalities that are widely used for proving concentration inequalities are presented, whose derivation relies on the martingale approach. Their proofs convey the main concepts of the martingale approach for proving concentration. Their presentation also motivates some further refinements that are considered in the continuation of this chapter.

5.1 The Azuma-Hoeffding inequality

The Azuma-Hoeffding inequality[10] is a useful concentration inequality for bounded-difference martingales. It was proved in [30] for independent bounded random variables, followed by a discussion on sums of dependent random variables; this inequality was later derived in [3] for the more general setting of bounded-difference martingales. In the following, this inequality is introduced.

Theorem 5.1. [Azuma-Hoeffding inequality] *Let $\{X_k, \mathcal{F}_k\}_{k=0}^n$ be a discrete-parameter real-valued martingale sequence. Suppose that, for every $k \in \{1, \ldots, n\}$, the condition $|X_k - X_{k-1}| \leq d_k$ holds almost surely (a.s.) for a real-valued sequence $\{d_k\}_{k=1}^n$ of non-negative numbers. Then, for every $\alpha > 0$,*

$$\mathbb{P}(|X_n - X_0| \geq \alpha) \leq 2 \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^n d_k^2}\right). \tag{8}$$

It is noted that Azuma's concentration inequality is typically interpreted as $\frac{X_n - X_0}{\sqrt{n}}$ being sub-Gaussian. The proof of the Azuma-Hoeffding inequality serves also to present the basic principles on which the martingale approach for proving concentration results is based. Therefore, we present in the following the proof of this inequality.

Proof. For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha) = \mathbb{P}(X_n - X_0 \geq \alpha) + \mathbb{P}(X_n - X_0 \leq -\alpha). \tag{9}$$

Let $\xi_i \triangleq X_i - X_{i-1}$ for $i = 1, \ldots, n$ designate the jumps of the martingale sequence. Then, it follows by assumption that $|\xi_k| \leq d_k$ and $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \{1, \ldots, n\}$.

[10] The Azuma-Hoeffding inequality is also known as Azuma's inequality. Since it is referred numerous times in this chapter, it will be named Azuma's inequality for the sake of brevity.

From Chernoff's inequality,

$$\begin{aligned}
&\mathbb{P}(X_n - X_0 \geq \alpha) \\
&= \mathbb{P}\left(\sum_{i=1}^{n} \xi_i \geq \alpha\right) \\
&\leq e^{-\alpha t}\, \mathbb{E}\left[\exp\left(t \sum_{i=1}^{n} \xi_i\right)\right], \quad \forall\, t \geq 0. \qquad (10)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right) \mid \mathcal{F}_{n-1}\right]\right] \\
&= \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \mathbb{E}\big[\exp(t\xi_n) \mid \mathcal{F}_{n-1}\big]\right] \qquad (11)
\end{aligned}$$

where the last equality holds since $Y \triangleq \exp\big(t \sum_{k=1}^{n-1} \xi_k\big)$ is $\mathcal{F}_{n-1}$-measurable; this holds due to fact that $\xi_k \triangleq X_k - X_{k-1}$ is $\mathcal{F}_k$-measurable for every $k \in \mathbb{N}$, and $\mathcal{F}_k \subseteq \mathcal{F}_{n-1}$ for $0 \leq k \leq n-1$ since $\{\mathcal{F}_k\}_{k=0}^{n}$ is a filtration. Hence, the RV $\sum_{k=1}^{n-1} \xi_k$ and Y are both $\mathcal{F}_{n-1}$-measurable, and $\mathbb{E}[XY|\mathcal{F}_{n-1}] = Y\, \mathbb{E}[X|\mathcal{F}_{n-1}]$.

Due to the convexity of the exponential function, the straight line connecting the end points of the function over the interval $[-d_k, d_k]$ lies above this function. Since $|\xi_k| \leq d_k$ for every k (note that $\mathbb{E}[\xi_k \mid \mathcal{F}_{k-1}] = 0$), it follows that

$$\begin{aligned}
&\mathbb{E}\big[e^{t\xi_k} \mid \mathcal{F}_{k-1}\big] \\
&\leq \mathbb{E}\Big[\frac{(d_k + \xi_k)e^{td_k} + (d_k - \xi_k)e^{-td_k}}{2d_k} \mid \mathcal{F}_{k-1}\Big] \\
&= \frac{1}{2}\left(e^{td_k} + e^{-td_k}\right) \\
&= \cosh(td_k). \qquad (12)
\end{aligned}$$

Since, for every integer $m \geq 0$,

$$(2m)! \geq (2m)(2m-2)\dots 2 = 2^m \, m!$$

then, due to the power series expansions of the hyperbolic cosine and exponential functions,

$$\cosh(td_k) = \sum_{m=0}^{\infty} \frac{(td_k)^{2m}}{(2m)!} \leq \sum_{m=0}^{\infty} \frac{(td_k)^{2m}}{2^m \, m!} = e^{\frac{t^2 \, d_k^2}{2}}$$

which therefore implies that

$$\mathbb{E}\big[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}\big] \leq e^{\frac{t^2 \, d_k^2}{2}}.$$

Consequently, by repeatedly using the recursion in 11, it follows that

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \leq \prod_{k=1}^{n} \exp\left(\frac{t^2 \, d_k^2}{2}\right) = \exp\left(\frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right)$$

which then gives (see 10) that

$$\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp\left(-\alpha t + \frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right), \quad \forall \, t \geq 0.$$

An optimization over the free parameter $t \geq 0$ gives that $t = \alpha \left(\sum_{k=1}^{n} d_k^2\right)^{-1}$, and

$$\mathbb{P}(X_n - X_0 \geq \alpha) \leq \exp\left(-\frac{\alpha^2}{2\sum_{k=1}^{n} d_k^2}\right). \tag{13}$$

Since, by assumption, $\{X_k, \mathcal{F}_k\}$ is a martingale with bounded jumps, so is $\{-X_k, \mathcal{F}_k\}$ (with the same bounds on its jumps). This implies that the same bound is also valid for the probability $\mathbb{P}(X_n - X_0 \leq -\alpha)$ and together with 9 it completes the proof of Theorem 5.1. ■

The proof of this inequality will be revisited later in this chapter for the derivation of some refined versions, whose use and advantage will be also exemplified.

Remark 5.2. In [38, Theorem 3.13], Azuma's inequality is stated as follows: Let $\{Y_k, \mathcal{F}_k\}_{k=0}^n$ be a martingale-difference sequence with $Y_0 = 0$ (i.e., Y_k is $\mathcal{F}_k$-measurable, $\mathbb{E}[|Y_k|] < \infty$ and $\mathbb{E}[Y_k|\mathcal{F}_{k-1}] = 0$ a.s. for every $k \in \{1, \ldots, n\}$). Assume that, for every k, there exist some numbers $a_k, b_k \in \mathbb{R}$ such that a.s. $a_k \leq Y_k \leq b_k$. Then, for every $r \geq 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^n Y_k\right| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right). \tag{14}$$

As a consequence of this inequality, consider a discrete-parameter real-valued martingale sequence $\{X_k, \mathcal{F}_k\}_{k=0}^n$ where $a_k \leq X_k - X_{k-1} \leq b_k$ a.s. for every k. Let $Y_k \triangleq X_k - X_{k-1}$ for every $k \in \{1, \ldots, n\}$, so since $\{Y_k, \mathcal{F}_k\}_{k=0}^n$ is a martingale-difference sequence and $\sum_{k=1}^n Y_k = X_n - X_0$, then

$$\mathbb{P}\left(|X_n - X_0| \geq r\right) \leq 2\exp\left(-\frac{2r^2}{\sum_{k=1}^n (b_k - a_k)^2}\right), \quad \forall\, r > 0. \tag{15}$$

Example 5.3. Let $\{Y_i\}_{i=0}^\infty$ be i.i.d. binary random variables which get the values $\pm d$, for some constant $d > 0$, with equal probability. Let $X_k = \sum_{i=0}^k Y_i$ for $k \in \{0, 1, \ldots, \}$, and define the natural filtration $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots$ where

$$\mathcal{F}_k = \sigma(Y_0, \ldots, Y_k), \quad \forall\, k \in \{0, 1, \ldots, \}$$

is the σ-algebra that is generated by the random variables $Y_0, \ldots, Y_k$. Note that $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ is a martingale sequence, and (a.s.) $|X_k - X_{k-1}| = |Y_k| = d$, $\forall\, k \in \mathbb{N}$. It therefore follows from Azuma's inequality that

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right). \tag{16}$$

for every $\alpha \geq 0$ and $n \in \mathbb{N}$. From the central limit theorem (CLT), since the RVs $\{Y_i\}_{i=0}^\infty$ are i.i.d. with zero mean and variance d^2, then

$\frac{1}{\sqrt{n}}(X_n - X_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^n Y_k$ converges in distribution to $\mathcal{N}(0, d^2)$. Therefore, for every $\alpha \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\, Q\Big(\frac{\alpha}{d}\Big) \tag{17}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\Big(-\frac{t^2}{2}\Big) \mathrm{d}t, \quad \forall\, x \in \mathbb{R} \tag{18}$$

is the probability that a zero-mean and unit-variance Gaussian Random Variable(RV) is larger than x. Since the following exponential upper and lower bounds on the Q-function hold

$$\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot e^{-\frac{x^2}{2}} < Q(x) < \frac{1}{\sqrt{2\pi}\, x} \cdot e^{-\frac{x^2}{2}}, \;\; \forall\, x > 0 \tag{19}$$

then it follows from 17 that the exponent on the right-hand side of 16 is the exact exponent in this example.

Example 5.4. In continuation to Example 5.3, let $\gamma \in (0, 1]$, and let us generalize this example by considering the case where the i.i.d. binary RVs $\{Y_i\}_{i=0}^\infty$ have the probability law

$$\mathbb{P}(Y_i = +d) = \frac{\gamma}{1+\gamma}, \quad \mathbb{P}(Y_i = -\gamma d) = \frac{1}{1+\gamma}\,.$$

Hence, it follows that the i.i.d. RVs $\{Y_i\}$ have zero mean and variance $\sigma^2 = \gamma d^2$. Let $\{X_k, \mathcal{F}_k\}_{k=0}^\infty$ be defined similarly to Example 5.3, so that it forms a martingale sequence. Based on the CLT, $\frac{1}{\sqrt{n}}\,(X_n - X_0) = \frac{1}{\sqrt{n}}\, \sum_{k=1}^n Y_k$ converges weakly to $\mathcal{N}(0, \gamma d^2)$, so for every $\alpha \geq 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\, Q\Big(\frac{\alpha}{\sqrt{\gamma}\, d}\Big). \tag{20}$$

From the exponential upper and lower bounds of the Q-function in 19, the right-hand side of 20 scales exponentially like $e^{-\frac{\alpha^2}{2\gamma d^2}}$. Hence, the exponent in this example is improved by a factor $\frac{1}{\gamma}$ as compared Azuma's inequality (that is the same as in Example 5.3 since $|X_k - X_{k-1}| \leq d$ for every $k \in \mathbb{N}$). This indicates on the possible refinement of Azuma's inequality by introducing an additional constraint on the second moment. This route was studied extensively in the probability literature, and it is studied along with Bernstein inequalities[8] (that is inequalities which focus on the second moment) in [42, 10]. We shall not consider such examples in this thesis.

5.2 McDiarmid's inequality

The following useful inequality is due to McDiarmid (see Theorem 3.1 in([37] or [38]), and its original derivation uses the martingale approach for its derivation. We will relate, in the following, the derivation of this inequality to the derivation of the Azuma-Hoeffding inequality (see the preceding subsection). In some sense we can say that McDiarmid's inequality generalizes the simple concentration inequality which we've seen before like [3] to functions that depend on independently identically distributed (i.i.d.) random variables.

Theorem 5.5. [McDiarmid's inequality] *Let $\{X_k\}_{k=1}^n$ be independent real-valued random variables, taking values in the set $\mathcal{X} := \prod_{k=1}^n \mathcal{X}_k \subseteq \mathbb{R}^n$. Let $g : \mathcal{X} \to \mathbb{R}$ be a measurable function such that, for some constants $\{d_k\}_{k=1}^n$,*

$$\big|g(\underline{x}) - g(\underline{x}')\big| \leq d_k, \quad \forall\, k \in \{1, \ldots, n\} \tag{21}$$

where $\underline{x} = (x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n)$ and $\underline{x}' = (x_1, \ldots, x_{k-1}, x_k', x_{k+1}, \ldots, x_n)$ are two arbitrary points in the set $\mathcal{X}$ that may only differ in their k-th coordinate (this is equivalent to saying that the variation of the function g w.r.t. its k-th coordinate is upper bounded by d_k). Then, for every $\alpha \geq 0$,

$$\mathbb{P}(\big|g(X_1, \ldots, X_n) - \mathbb{E}\big[g(X_1, \ldots, X_n)\big]\big| \geq \alpha) \leq 2\exp\left(-\frac{2\alpha^2}{\sum_{k=1}^n d_k^2}\right). \tag{22}$$

Remark 5.6. The following proof provides in this setting an improvement by a factor of 4 in the exponent of the bound, over using the Azuma-Hoeffding inequality.

Proof. For $k \in \{1, \dots, n\}$, let $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ be the σ-algebra that is generated by $X_1, \dots, X_k$ with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Define

$$\xi_k \triangleq \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_k\big] - \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_{k-1}\big], \quad \forall\, k \in \{1, \dots, n\}. \tag{23}$$

Note that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \ldots \subseteq \mathcal{F}_n$ is a filtration, and

$$\begin{aligned} \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_0\big] &= \mathbb{E}\big[g(X_1, \dots, X_n)\big] \\ \mathbb{E}\big[g(X_1, \dots, X_n) \,|\, \mathcal{F}_n\big] &= g(X_1, \dots, X_n). \end{aligned} \tag{24}$$

Hence, it follows from the last three equalities that

$$g(X_1, \dots, X_n) - \mathbb{E}\big[g(X_1, \dots, X_n)\big] = \sum_{k=1}^{n} \xi_k.$$

In the following, we need a lemma:

Lemma 5.7. *For every $k \in \{1, \dots, n\}$, the following properties hold a.s.:*

1. $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$*, so* $\{\xi_k, \mathcal{F}_k\}$ *is a martingale-difference and* ξ_k *is* $\mathcal{F}_k$*-measurable.*
2. $|\xi_k| \leq d_k$
3. $\xi_k \in [A_k, A_k + d_k]$ *where* A_k *is some non-positive and* $\mathcal{F}_{k-1}$*-measurable random variable.*

Proof. The random variable ξ_k is $\mathcal{F}_k$-measurable since $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, and ξ_k is a difference of two functions where one is $\mathcal{F}_k$-measurable and the other is $\mathcal{F}_{k-1}$-measurable. Furthermore, it is easy to verify that $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$.

This proves the first item. The second item follows from the first and third items. To prove the third item, note that $\xi_k = f_k(X_1, \ldots, X_k)$ holds a.s. for some function $f_k : \mathcal{X}_1 \times \ldots \times \mathcal{X}_k \to \mathbb{R}$ that is $\mathcal{F}_k$-measurable. Let us define, for every $k \in \{1, \ldots, n\}$,

$$\begin{aligned} A_k &\triangleq \inf_{x \in \mathcal{X}_k} f_k(X_1, \ldots, X_{k-1}, x), \\ B_k &\triangleq \sup_{x \in \mathcal{X}_k} f_k(X_1, \ldots, X_{k-1}, x) \end{aligned}$$

which are $\mathcal{F}_{k-1}$-measurable, and by definition $\xi_k \in [A_k, B_k]$ holds almost surely. Furthermore, for every point $(x_1, \ldots, x_{k-1}) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_{k-1}$, we have

$$\begin{aligned} &\sup_{x \in \mathcal{X}_k} f_k(x_1, \ldots, x_{k-1}, x) - \inf_{x' \in \mathcal{X}_k} f_k(x_1, \ldots, x_{k-1}, x') \\ &= \sup_{x, x' \in \mathcal{X}_k} \big\{ f_k(x_1, \ldots, x_{k-1}, x) - f_k(x_1, \ldots, x_{k-1}, x') \big\} \\ &= \sup_{x, x' \in \mathcal{X}_k} \Big\{ \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_k = x\big] \\ &\quad - \mathbb{E}\big[g(X_1, \ldots, X_n) \,|\, X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_k = x'\big] \Big\} \\ &= \sup_{x, x' \in \mathcal{X}_k} \Big\{ \mathbb{E}\big[g(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n)\big] - \mathbb{E}\big[g(x_1, \ldots, x_{k-1}, x', X_{k+1}, \ldots, X_n)\big] \Big\} \qquad (25) \\ &= \sup_{x, x' \in \mathcal{X}_k} \Big\{ \mathbb{E}\big[g(x_1, \ldots, x_{k-1}, x, X_{k+1}, \ldots, X_n) - g(x_1, \ldots, x_{k-1}, x', X_{k+1}, \ldots, X_n)\big] \Big\} \\ &\leq d_k \qquad (26) \end{aligned}$$

where 25 follows from the independence of the random variables $\{X_k\}_{k=1}^n$, and 26 follows from the condition in 21. Hence, it follows that $B_k - A_k \leq d_k$ a.s., which then implies that $\xi_k \in [A_k, A_k + d_k]$. Since $\mathbb{E}[\xi_k \,|\, \mathcal{F}_{k-1}] = 0$ then a.s. the $\mathcal{F}_{k-1}$-measurable function A_k is non-positive. It is noted that the third item of the lemma makes it different from the proof of the Azuma-Hoeffding inequality (in that case, it implies that $\xi_k \in [-d_k, d_k]$ where the length of the interval is twice larger.) ■

Applying the convexity of the exponential function (similarly to the derivation of the Azuma-Hoeffding inequality, but this time w.r.t. the interval $[A_k, A_k + d_k]$) implies that for every $k \in \{1, \dots, n\}$

$$\begin{aligned}
&\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \\
&\leq \mathbb{E}\left[\frac{(\xi_k - A_k)e^{t(A_k+d_k)} + (A_k + d_k - \xi_k)e^{tA_k}}{d_k} \,\Big|\, \mathcal{F}_{k-1}\right] \\
&= \frac{(A_k + d_k)e^{tA_k} - A_k e^{t(A_k+d_k)}}{d_k}.
\end{aligned}$$

Let $P_k \triangleq -\frac{A_k}{d_k} \in [0, 1]$, then

$$\begin{aligned}
&\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \\
&\leq P_k e^{t(A_k+d_k)} + (1 - P_k)e^{tA_k} \\
&= e^{tA_k}\big(1 - P_k + P_k e^{td_k}\big) \\
&= e^{H_k(t)} \qquad (27)
\end{aligned}$$

where

$$H_k(t) \triangleq tA_k + \ln\big(1 - P_k + P_k e^{td_k}\big), \quad \forall\, t \in \mathbb{R}. \qquad (28)$$

Since $H_k(0) = H_k'(0) = 0$ and the geometric mean is less than or equal to the arithmetic mean then, for every t,

$$H_k''(t) = \frac{d_k^2 P_k(1 - P_k)e^{td_k}}{(1 - P_k + P_k e^{td_k})^2} \leq \frac{d_k^2}{4}$$

which implies by Taylor's theorem that

$$H_k(t) \leq \frac{t^2 d_k^2}{8} \qquad (29)$$

so, from 27,

$$\mathbb{E}[e^{t\xi_k} \,|\, \mathcal{F}_{k-1}] \leq e^{\frac{t^2 d_k^2}{8}}\ .$$

Similarly to the proof of the Azuma-Hoeffding inequality, by repeatedly using the recursion in 11, the last inequality implies that

$$\mathbb{E}\left[\exp\left(t\sum_{k=1}^{n}\xi_k\right)\right] \leq \exp\left(\frac{t^2}{8}\sum_{k=1}^{n}d_k^2\right) \tag{30}$$

which then gives from 10 that, for every $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}(g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)] \geq \alpha) \\
&= \mathbb{P}\left(\sum_{k=1}^{n}\xi_k \geq \alpha\right) \\
&\leq \exp\left(-\alpha t + \frac{t^2}{8}\sum_{k=1}^{n}d_k^2\right).
\end{aligned} \tag{31}$$

An optimization over the free parameter $t \geq 0$ gives that $t = 4\alpha\left(\sum_{k=1}^{n}d_k^2\right)^{-1}$, so

$$\mathbb{P}(g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)] \geq \alpha) \leq \exp\left(-\frac{2\alpha^2}{\sum_{k=1}^{n}d_k^2}\right). \tag{32}$$

By replacing g with $-g$, it follows that this bound is also valid for the probability

$$\mathbb{P}\big(g(X_1,\ldots,X_n) - \mathbb{E}[g(X_1,\ldots,X_n)] \leq -\alpha\big)$$

which therefore gives the bound in 22. This completes the proof of Theorem 5.5. ■

5.3 Hoeffding's inequality, and its improved version (the Kearns-Saul inequality)

In the following, we derive a concentration inequality for sums of independent and bounded random variables as a consequence of McDiarmid's inequality. This inequality is due to Hoeffding (see [30, Theorem 2]). An improved version of Hoeffding's inequality, due to Kearns and Saul [32], is also introduced in the following.

Theorem 5.8 (Hoeffding)**.** *Let $\{U_k\}_{k=1}^n$ be a sequence of independent and bounded random variables such that, for every $k \in \{1, \ldots, n\}$, $U_k \in [a_k, b_k]$ holds a.s. for some constants $a_k, b_k \in \mathbb{R}$. Let $\mu_n \triangleq \sum_{k=1}^n \mathbb{E}[U_k]$. Then,*

$$\mathbb{P}\left(\left|\sum_{k=1}^n U_k - \mu_n\right| \geq \alpha\sqrt{n}\right) \leq 2\exp\left(-\frac{2\alpha^2\, n}{\sum_{k=1}^n (b_k - a_k)^2}\right), \quad \forall\, \alpha \geq 0. \quad (33)$$

Proof. Apply Theorem 5.5 to $g(\underline{u}) \triangleq \sum_{k=1}^n u_k$ for every
$\underline{u} \in \prod_{k=1}^n [a_k, b_k]$. ■

6 Statistical Learning Theory

We want to introduce some notions from Statistical Learning Theory (an application in Computer Science/ Statistics) to descripe how risk is controlled in predictive models, i.e., their expected inaccuracy on new data from the same source as that used to fit the model. Or to say this in another way, the goal of statistical learning theory is to study, in a statistical framework, the properties of learning algorithms. In this chapter, I summarize the basic forms of these results in the literature, sacrificing some rigour for brevity.

6.1 Notations from learning theory

Before we explore the traditional setup. It is useful at this point to introduce the formal language of *learning* while we are only studying Statistical Learning theory - there are other kinds of learning theory. As ever we are strongly influenced in this presentation by [11, 55, 4] and we refer interested readers to those sources for more information or perhaps to use your favourite Machine Learning textbook such as [29].

What is the formal defintion of learning? Learning is formally define as finding a hypothesis h based on observed samples $Z_1, \cdots, Z_n$. To evaluate the quality of h, a bounded real-valued loss (cost) function l is introduced such that $l(h; z)$ indicates how well h explains (or fits) Z. We assume that $-M \leq l \leq M$ for some $M > 0$. We introduce some of the nomeclature from Machine Learning and describe some classic problems. These are covered in more depth in [29].

6.1.1 Classification

A classification problem is loosely when you try to *classify* things. Z is defined as the product $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is a discrete output space denoting the labels of inputs. In the case of binary classification (an example used in this thesis) $\mathcal{Y} = \{-1, 1\}$, corresponding to the labels of the two classes. The loss function l takes the form $l(h; z) = l(gh(x))$ and h is called a **binary classifier**.The basic example is:

$$l(yh^{'}(x)) = I(yh^{'}(x) < 0) = I(u \neq \text{sign}(h^{'}(x))) \tag{34}$$

6.1.2 Regression

Regression is a powerful method in Statistics and one can write books about types of regression theories. We shall loosely cover an introduction to regression here - and refer our interested readers to [40]. One of the more remarkable applications of statistical learning theory in recent years is by Shalizi and McDonald and is mentioned in [39] - this has been to apply a theory of SLT to *dependent* data, and use this to point out the inadequacy of some classical Economic models such as Dynamic stochastic general equilibrium modeling (DSGE) [48]. We shall not unfortunately have the time or space to go into this, but we do mention it *en passant* as it shows that SLT is still a very active research area.

Definition 6.1 (Regression). *Z is defined as $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is an input space and $\mathcal{Y}$ is a real output space denoting the real-valued lables of inputs. The loss function l often takes the form $l(h; z) = l(y - h(x))$, and the basic example is the squared loss:*

$$l(y - h(x)) = (y - h(x))^2 \tag{35}$$

We include a real example using the [44] library in Python.

Example 6.1. We consider a simple linear model using diabetes data.

This example uses the only the first feature of the diabetes dataset, in order to illustrate a two-dimensional plot of this regression technique. The straight line can be seen in the plot**??**, showing how linear regression attempts to draw a straight line that will best minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. As you can see there is high variance in this data set, so one of the processes we can do with our loss functions is to use *regularization* techniques. These are covered in the canonical textbook[29] or alternatively a short tutorial is in [11].

$$y = X\beta + \epsilon \tag{36}$$

- X: data
- y: target vairable
- β: Coefficients
- ϵ: Observation noise

We include the results from the calcuation. And a figure showing our ordinary least squares example, this is randomly generated data in a Python library as you can see the Pearson coefficient is included here. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. These can be calculated easily in most mathematical packages, and we refer to any book on regression for more on Pearson's correlation coefficient [40].

6.1.3 Density Estimation

The functions h are probability densities over Z, and the loss function takes the form l(h;z) = l(h(z)) For instance,

$$l(h(z)) = -\log h(z) \tag{37}$$

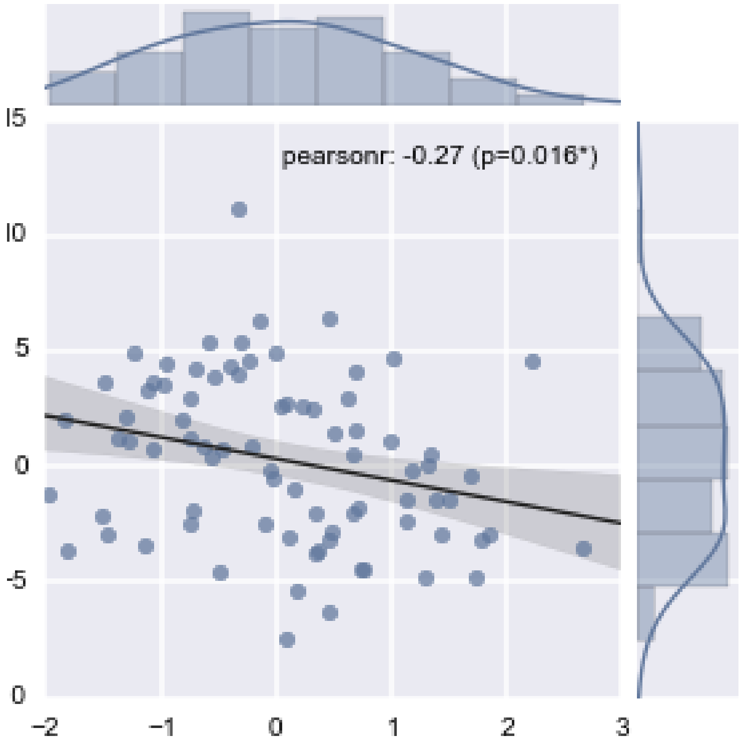

Figure 1: Linear regression example with Pseudo-Randomly generated data

is the likelihood of a point z being generated by h. We introduce some notation -

Definition 6.2. *We call the following the space of loss functions*

$$\mathcal{L}(\mathcal{H}) = \{l(h, \cdot) : h \in \mathcal{H}\}$$

where $\mathcal{H}$is the hypothesis.

After introducing the traditional setup - we will discuss a particular kind of loss function - in a certain abstract framework we introduce the notion of *risk or generalization error* - an estimate of the expected error which can be computed from the sample.

6.2 The Traditional Setup

Consider predictors $X \in \mathcal{X}$ and responses $Y \in \mathcal{Y}$. Let $\mathcal{F}$be a calss of functions $f : \mathcal{X} \to \mathcal{Y}$ which take predictors as inputs. Define a loss function

$l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ which measures the cost of making poor predictions. Throughout this chapter I make the following assumption on the loss function.

Assumption 1. $\forall f \in \mathcal{F}$

$$0 \leq l(y, y^{'}) \leq M < \infty$$

Then, I can define the risk of any predictor $f \in \mathcal{F}$.

Definition 6.3 (Risk or generalization error)**.**

$$R(f) := \int l(f(X), Y) d\mathbb{P} = \mathbb{E}_{\mathbb{P}}[l(f(X), Y)], \tag{38}$$

where (X,Y) $\sim \mathbb{P}$

The risk or generalization error measures the expected cost of using f to predict Y from X given a new observation. Just to emphasize, the expectation is taken with respect to the distribution $\mathbb{P}$ of the test point (X,Y) which is independent of f; the risk is a deterministic function of f with all the randomness in the data averaged away.

Since the true distribution $\mathbb{P}$is unknown, so is R(f), but one can attempt to estimate it based on only the observed data. Suppose that I observe a random sample $D_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ so that $(X_i, Y_i) \overset{i.i.d}{\sim} \mathbb{P}$, i.e. $D_n \sim \mathbb{P}$. Define the *training error or empirical risk* of f as follows.

Definition 6.2 (Training error or empirical risk)**.**

$$\widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i). \tag{39}$$

In other words, the in-sample training error, $\widehat{R}_n(f)$, is the average loss over the actual training points. It is easy to see that, because the training data D_n and the test point (X,Y) are IID, then given some fixed function f (chosen independently of the sample D_n),

$$\widehat{R}_n(f) = R(f) + \gamma_n(f) \tag{40}$$

where $\gamma_n(f)$ is a mean-zero noise variable that reflects how far the training sample departs from being perfectly representative of the data-generating distribution. Here I should emphasize that $\widehat{R}_n(f)$ is random enough through the training sample D_n. By the law of large numbers, for such fixed f, $\gamma_n(f) \to 0$ as $n \to \infty$, so, with enough data, one has a good idea of how well any given function will generalize to new data.

However, one is rarely interested in the performance of a single function f without adjustable parameters fixed for them in advance by theory. Rather, researchers are interested in a class of plausible functions $\mathcal{F}$, possibly indexed by some possibly infinite parameter $\theta \in \Theta$, which I refer to as a model. One fucntion (one particular parameter point) is chosen from the model class minimizing some criterion funciton. Maximum likelihood, Bayesian maximization *a posteriori*, least squares, regularized methods, and empirical risk minimization (ERM) all have this flavor as do many other estimation methods. In these cases, one can define the empirical risk minimizer for an appropriate loss function l.

Definition 6.4. *[Empirical Risk Minimizer]*

$$\widehat{f} := \operatorname{argmin}_{f \in \mathcal{F}} \widehat{R}_n(f) = \operatorname{argmin}_{f \in \mathcal{F}} (R(f) + \gamma_n(f)). \tag{41}$$

It is important to note that $\widehat{f}$ is random and measurable with respect to the empirical risk process $\widehat{R}_n(f)$ for $f \in \mathcal{F}$. Choosing a predictor $\widehat{f}$ by empirical risk minimization (tuning the adjustable parameters so that $\widehat{f}$ fits the training data well) conflates predicting future data well (low $R(\widehat{f})$, the true risk) with exploiting the accidents and noise of the training data (large negative $\gamma_n(\widehat{f})$, finite-sample noise). The true risk of $\widehat{f}$ will generally be bigger than its in-sample risk precisely because I picked it to match the data well. In doing so, $\widehat{f}$ ends up reproducing some of the noise in the data and therefore will not generalize well. The difference between the true and apparent risk depends on the magnitude of the sampling fluctuations:

$$R(\widehat{f}) - \widehat{R}_n(\widehat{f}) \leq \sup_{f \in \mathcal{F}} \|\gamma_n(f)\| = \Gamma_n(\mathcal{F}) \tag{42}$$

In (42), $R(\widehat{f})$ is random and measurable with respect to $\widehat{f}$.

The main goal of statistical learning theory is to control $\Gamma_n(\mathcal{F})$ while making minimal assumptions about the the data generating process - i.e. to provide bounds on over-fitting. Using more flexible models (allowing more general distributions, adding parameters, etc.) has two contrasting effects. On the one hand, it improves the best possible accuracy, lowering the minimum of the true risk. On the other hand, it increases the ability to, as it were, memorize noise for any fixed sample n. This qualitative observation - a generalization of the bias-variance trade-off from estimation theory - can be made use-fully precise by quantifying the complexity of model classes. A typical result is a confidence bound on Γ_n (and hence on over-fitting), which says that with probability at least $1 - \nu$,

$$\Gamma_n(\mathcal{F}) \leq \Theta(\Delta(\mathcal{F}), n, \nu), \tag{43}$$

where $\Delta(\cdot)$ is some suitable measure of the complexity of the model $\mathcal{F}$. To give specific forms of $\Theta(\cdot)$, I need to show that for a a particular f, $R(f)$ and $\widehat{R}_n(f)$ will be close to each other for any fixed n without knowledge of the distribution of the data. Furthermore, I need the complexity $\Delta(\mathcal{F})$, to claim that $R(f)$ and $\widehat{R}_n(f)$ will be close, not only for a particular f, but uniformly over all $f \in \mathcal{F}$. Together these two results will allow me to show, despite little knowledge fo the data generating process, how bad the $\widehat{f}$ which I choose will be at forecasting future observations.

6.3 Concentration

The first step to controlling the difference between the empirical and expected risk is to develop concentration results for fixed functions. We have already introduced various inequalities. McDiarmid Inequality and Hoefferding's Inequality. These results are extremely important in Statistical learning theory, but it is beyond the scope of this thesis to go into much more detail. In the remainder of this section, I will show how to obtain concentration for the training error around the risk for two different choices of the random variables Z_i. This will lead to two different ways of controlling Γ_n and hence the generalization error of prediction functions.

6.4 Contol by Counting

Let us assume that we let Z_i be the loss of the i^{th} training point for some fixed function f. Then by Hoeffding's inequality we get the following remarkable result

$$\mathbb{P}^n \left(\|R(f) - \widehat{R}_n(f)\| \geq \epsilon \right) \leq 2 \exp\{-\frac{2n\epsilon^2}{M^2}\} \tag{44}$$

This result is quite powerful, it says that the probability of observing data which will result in a training error much different from the expected risk goes to zero exponentially with the size of the training set. The only assumption necessary was $0 \leq l(y, y^{'}) \leq M < \infty$.

6.5 Capacity

For “small” models, we can just count the number of functions in the class and take the union bound. Suppose that $\mathcal{F} = \{f_1, \ldots, f_N\}$. Then we have

$$\mathbb{P}\left(\sup_{1\leq i\leq N} |R(f_i) - \widehat{R}_n(f_i)| > \epsilon\right) \leq \sum_{i=1}^{N} \mathbb{P}\left(|R(f_i) - \widehat{R}_n(f_i)| > \epsilon\right) \tag{45}$$

$$\leq N \exp\left\{-\frac{2n\epsilon^2}{K}\right\}, \tag{46}$$

by Theorem 5.8. Most interesting models are not small in this sense, but similar results hold when model size is measured appropriately.
There are a number of measures for the size or capacity of a model. Algorithmic stability [32, 12] quantifies the sensitivity of the chosen function to small perturbations to the data. Similarly, maximal discrepancy [55] asks how different the predictions could be if two functions are chosen using two separate data sets. A more direct, functional-analytic approach partitions $\mathcal{F}$ into equivalence classes under some metric, leading to covering numbers [46]. Rademacher complexity [4] directly describes a model’s ability to fit random noise. We focus on a measure which is both intuitive and powerful: Vapnik-Chervonenkis (VC) dimension [55].
VC dimension starts as an idea about collections of sets.

Definition 6.5. *Let $\mathbb{U}$ be some (infinite) set and S a finite subset of $\mathbb{U}$. Let $\mathcal{C}$ be a family of subsets of $\mathbb{U}$. We say that $\mathcal{C}$* shatters *S if for every $S' \subseteq S$, $\exists C \in \mathcal{C}$ such that $S' = S \cap C$.*

Essentially, $\mathcal{C}$ can shatter a set S if it can pick out every subset of points in S. This says that the collection $\mathcal{C}$ is very complicated or flexible. The cardinality of the largest set S that can be shattered by $\mathcal{C}$ is the latter’s VC dimension.

Definition 6.6 (VC dimension)**.** *The* Vapnik-Chervonenkis (VC) dimension *of a collection $\mathcal{C}$ of subsets of $\mathbb{U}$ is*

$$\textsc{vcd}(\mathcal{C}) := \sup\{|S| : S \subseteq \mathbb{U} \text{ and } S \text{ is shattered by } \mathcal{C}\}. \tag{47}$$

To see why this is a “dimension”, we need one more notion.

Definition 6.7 (Growth function)**.** *The* growth function $G(\mathcal{C}, n)$ *of a collection* $\mathcal{C}$ *of subsets of* $\mathbb{U}$ *is the maximum number of subsets which can be formed by intersecting a set* $S \subset \mathbb{U}$ *of cardinality* n *with* $\mathcal{C}$*,*

$$G(n, \mathcal{C}) := \sup_{S \subset U \ : \ |S| = n} |S \wedge \mathcal{C}| \tag{48}$$

The growth function counts how many *effectively* distinct sets the collection contains, when we can only observe what is going on at n points, not all of $\mathbb{U}$. If $n \leq \text{VCD}(\mathcal{C})$, then from the definitions $G(n, \mathcal{C}) = 2^n$, If the VC dimension is finite, however, and $n > \text{VCD}(\mathcal{C})$, then $G(n, \mathcal{C}) < 2^n$, and in fact it can be shown [56] that

$$G(n, \mathcal{C}) \leq (n+1)^{\text{VCD}(\mathcal{C})}. \tag{49}$$

This polynomial growth of capacity with n is why VCD is a "dimension". Using VC dimension to measure the capacity of function classes is straightforward. Define the indicator function $\mathbf{1}_A(x)$ to take the value 1 if $x \in A$ and 0 otherwise. Suppose that $f \in \mathcal{F}$, $f : \mathbb{U} \to \mathbb{R}$. Each f corresponds to the set

$$C_f = \{(u, a) : \mathbf{1}_{(0,\infty)}(f(u) - b) = 1, \quad u \in \mathbb{U}, \quad b \in \mathbb{R}\}, \tag{50}$$

so $\mathcal{F}$ corresponds to the class $\mathcal{C}_{\mathcal{F}} := \{C_f : f \in \mathcal{F}\}$. Essentially, the growth function $G(n, \text{VCD}(\mathcal{F}))$ counts the effective number of functions in $\mathcal{F}$, i.e., how many can be told apart using only n observations. When $\text{VCD}(\mathcal{F}) < \infty$, this number grows only polynomially with n. This observation lets us control the risk over the entire model, providing one of the pillars of statistical learning theory.

Theorem 6.8 ([56])**.** *Suppose that* $\text{VCD}(\mathcal{F}) < \infty$ *and* $0 \leq \ell(y, y') \leq K < \infty$*. Then,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| > \epsilon\right) \leq 4(2n+1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{n\epsilon^2}{K_1^2}\right\} , \tag{51}$$

where K_1 *depends only on* K *and not* n *or* $\mathcal{F}$*.*

The proof of this theorem has a similar flavor to the union bound argument given in 46.

This theorem has as an immediate corollary a bound for the out-of-sample risk. Since $\sup_{f \in \mathcal{F}}$ is inside the probability statement in 51, it applies to both pre-specified and to data-dependent functions, including any $\widehat{f}$ chosen by fitting a model or minimizing empirical risk.

Corollary 6.9. *When 6.8 applies, for any $\eta > 0$ and any $f \in \mathcal{F}$, with probability at least $1 - \eta$,*

$$R(f) \leq \widehat{R}_n(f) + K_1 \sqrt{\frac{\mathrm{VCD}(\mathcal{F}) \log(2n+1) + \log 4/\eta}{n}}. \tag{52}$$

The factor K_1 can be calculated explicitly but is unilluminating and we will not need it. Conceptually, the right-hand side of this inequality resembles standard model selection criteria, like AIC or BIC, with in-sample fit plus a penalty term which goes to zero as $n \to \infty$. Here however, the bound holds with high probability despite lack of knowledge of $\mathbb{P}$ and it has nothing to do with asymptotic convergence: it holds for each n. It does however hold *only* with high $\mathbb{P}$ probability, not always.

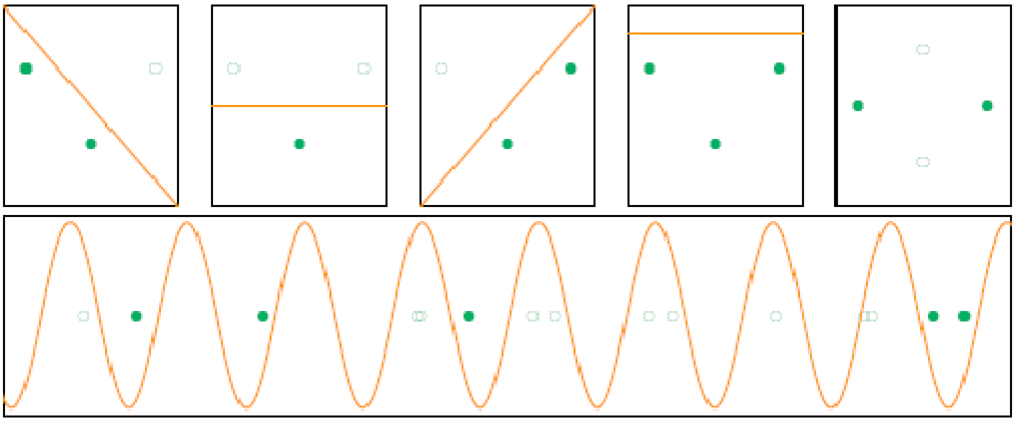

Figure 2: Illustration of VC dimension for some function classes.

VC dimension is well understood for some function classes. For instance, if $\mathcal{F} = \{\mathbf{x} \mapsto \boldsymbol{\gamma} \cdot \mathbf{x} : \boldsymbol{\gamma} \in \mathbb{R}^p\}$ then $\mathrm{VCD}(\mathcal{F}) = p + 1$, i.e. it is the number of free parameters in a linear regression plus 1. VC dimension does not always have such a nice relation to the number of free parameters however; the classic example is the model $\mathcal{F} = \{x \mapsto \sin(\omega x) : \omega \in \mathbb{R}\}$, which has only one free parameter, but $\mathrm{VCD}(\mathcal{F}) = \infty$.[11] These examples are illustrated in 2. At the same time, there are model classes (support vector machines)

[11] This result follows if we can show that for any positive integer J and any binary sequence $(r_1, \ldots, r_J)$, there exists a vector $(x_1, \ldots, x_J)$ such that $\mathbf{1}_{[0,1]}(\sin(\omega x_i)) = r_i$. If we choose $x_i = 2\pi 10^{-i}$, then one can show that taking $\omega = \frac{1}{2}\left(\sum_{i=1}^{J}(1 - r_i)10^i + 1\right)$ solves the system of equations.

which may have infinitely many parameters but finite VC dimension. [21]. This illustrates a further difference between the statistical learning approach and the usual information criteria, which are based on parameter-counting. The concentration results in 6.8 and 6.9 work well for independent data. The first shows how quickly averages concentrate around their expectations: exponentially fast in the size of the data. The second result generalizes the first from a single function to entire function classes. Both results, as stated, depend critically on the independence of the random variables. For time series, we must be able to handle dependent data. In particular, because time-series data are dependent, the length n of a sample path $Y_1, \ldots, Y_n$ exaggerates how much information it contains. Knowing the past allows forecasters to predict future data (at least to some degree), so actually observing those future data points gives less information about the underlying process than in the IID case. Thus, while in 5.8 the probability of large discrepancies between empirical means and their expectations decreases exponentially in n, in the dependent case, the effective sample size may be much less than n resulting in looser bounds.

7 Rademacher Processes

We introduce very quickly some remarks from Rademacher Processes, and some key techniques. This is merely to show the
A key technique in the theory of empirical processes is *Rademacher symmetrization*. This was first introduced into empirical processes in a classical paper by Gine [28] so we'll show how this applies in the context of Talagrand's inequality[50, 51].

Let $\epsilon_i, i = 1, \cdots, n$, be i.i.d Rademacher random signs (taking values -1,1 with probability 1/2), independent of the $X_i' s$, defined on a large product probability space with product probability Pr, denote the joint expectation by E, and the E_ϵ and E_X the corresponding expectations w.r.t the $\epsilon_i' s$ $X_i' s$, respectively. The following symmetrization inequality holds for random variables in arbitrary normed spaces, but we state it for the suprenum norm relevant in empirical process theory: For $\mathcal{F}$ a class of functions on $(S, \mathcal{A})$, define $\|H\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |H(f)|$.

Lemma 7.1. *Let $\mathcal{F}$ be a uniformly bounded P-centered class of functions defined on a measurable space $(S, \mathcal{A})$. Let ϵ_i be i.i.d. Rademachers as above,* *and let $a_i, i = 1, \cdots, n$ be any sequence of real numbers. Then*

$$\frac{1}{2}E\|\sum_{i=1}^{n}\epsilon_i f(X_i)\|_{\mathcal{F}} \le E\|\sum_{i=1}^{n} f(X_i)\|_{\mathcal{F}} \le \|\sum_{i=1}^{n}\epsilon_i(f(X_i)+a_i)\|_{\mathcal{F}} \tag{53}$$

Proof. Let us assume for simplictly that $\mathcal{F}$ is countable (so that we can neglect measurability problems). Since $E_X f(X_i) = 0$ for every f, i, the first inequality follows from

$$E\|\sum_{i=1}^{n}\epsilon_i(f(X_i)\|_{\mathcal{F}} = E_\epsilon E_X \le E_\epsilon E_X\|\sum_{i:\epsilon_i=-1} f(X_i) + E_X\sum_{i:\epsilon_i=1} f(X_i)\|_{\mathcal{F}}$$
$$+E_\epsilon E_X\|\sum_{i:\epsilon_i=1} f(X_i) + E_X\sum_{i:\epsilon_i=-1} f(X_i)\|_{\mathcal{F}} \le 2E\|\sum_{i=1}^{n} f(X_i)\|_{\mathcal{F}}$$

where in the last inequality we have used Jensen's inequality and convexity of the norm. To prove the second inequality, let $X_{n+i}, i = 1, \cdots, n$ be an independent copy of $X_1, \cdots, X_n$. Then proceeding as above,
$E\|\sum_{i=1}^{n} f(X_i)\|_{\mathcal{F}} = E\|\sum_{i=1}^{n}(f(X_i) - Ef(X_{n+i})\|_{\mathcal{F}} \le$
$E\|\sum_{i=1}^{n}(f(X_i + a_i) - \sum_{i=1}^{n}(f(X_{n+i} + a_i)\|_{\mathcal{F}}$
which clearly equals

$$E_\epsilon E_X\|\sum_{i:\epsilon_i=1}\epsilon_i(f(X_i)+a_i-f(X_{n+i})-a_i-\sum_{i:\epsilon_i=-1}\epsilon_i(f(X_i)+a_i-f(X_{n+i})-a_i\|_{\mathcal{F}}$$

Now Pr being a product probability measure with identical coordinates, it is invariant by permutations of the coordinates, so that we may exchange $f(X_i)$ and $f(X_{n+i})$ for the i's where $\epsilon_i = -1$ in the last expectation. This gives that the quantity in the last equation equals

$$E_\epsilon E_X\|\sum_{i=1}^{n}\epsilon_i(f(X_i)) + a_i - f(X_{n+i}) - a_i)\|_{\mathcal{F}} \le 2E\|\sum_{i=1}^{n}\epsilon_i(f(X_i)+a_i)\|_{\mathcal{F}}$$

which completes the proof. ■

This simple but very useful result says that we can always compare the size of the expectation of the supremum of an empirical process to a symmetrized process. The idea usual is that the symmetrized *Rademacher* process has conditional on the $X_i's$ a very simple structure. One can then derive results of the Rademacher proecess and integrate the results over the distribution of the $X_i's$ Rademacher averages are quantities that play an important role in empirical process theory[26] and in the theory of Banach spaces[42]. We shall not consider more about Rademacher averages in this paper and suggest [11, 4] as reading material. The concentration results in 6.8 and 6.9 work well for independent data. The first shows how quickly averages concentrate around their expectations: exponentially fast in the size of the data. The second result generalizes the first from a single function to entire function classes. Both results, as stated, depend critically on the independence of the random variables. For time series, we must be able to handle dependent data. In particular, because time-series data are dependent, the length n of a sample path $Y_1, \dots, Y_n$ exaggerates how much information it contains. Knowing the past allows forecasters to predict future data (at least to some degree), so actually observing those future data points gives less information about the underlying process than in the IID case. Thus, while in 5.8 the probability of large discrepancies between empirical means and their expectations decreases exponentially in n, in the dependent case, the effective sample size may be much less than n resulting in looser bounds.

8 Time series

Before we proceed it is worth including a definition of *time series* and a visual example.

Definition 8.1. *A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.*

We include an example of time series data - plotted using freely available USDA data on meat production.[12]
In moving from the IID setting to time series forecasting, we need a number of modifications to our initial setup. Rather than observing input/output pairs (Y_i, X_i), we observe a single sequence of random variables $Y_{1:n} := (Y_1, \dots, Y_n)$ where each Y_i takes values in $\mathbb{R}^p$.[13] We are interested in

[12] This data is included in the Pandas Python library[1] or at `http://www.ers.usda.gov/data-products/livestock-meat-domestic-data.aspx`

[13] We can easily generalize this to arbitrary measurable spaces.

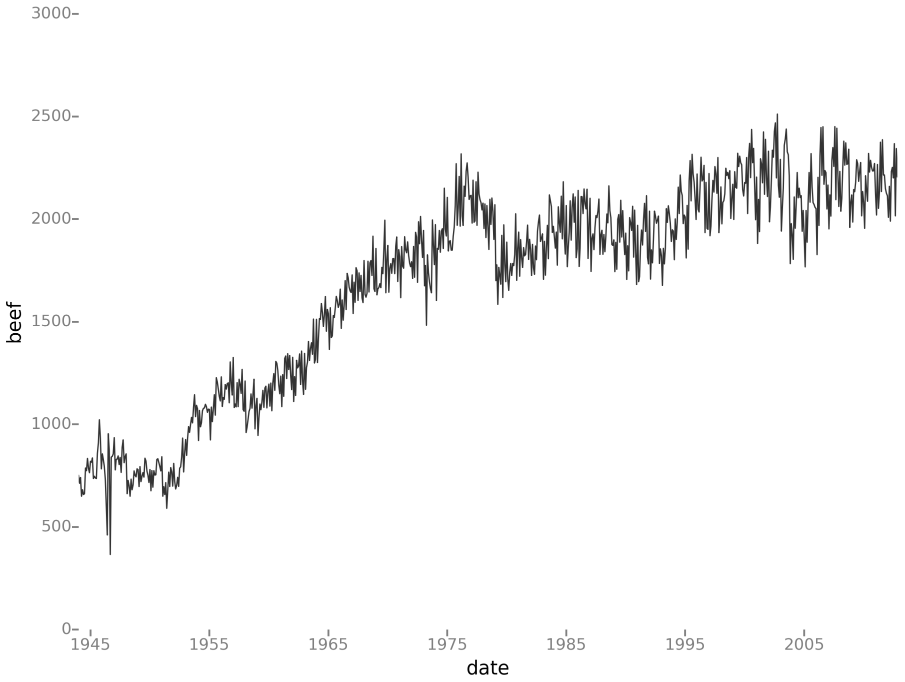

Figure 3: Time Series example showing the changes in Beef consumption over the 20th Century in the US

using functions which take past observations as inputs and predict future values of the process. Specifically, given data from time 1 to time n, we wish to predict time $n+1$.

While we no longer presume IID data, we still need to restrict the sort of dependent process we work with. We first remind the reader of the notion of (strict or strong) stationarity.

Definition 8.2 (Stationarity)**.** *A random sequence Y_∞ is stationary when all its finite-dimensional distributions are time-invariant: for all t and all non-negative integers i and j, the random vectors $Y_{t:t+i}$ and $Y_{t+j:t+i+j}$ have the same distribution.*

Stationarity does not imply that the random variables Y_t are independent across time t, only that the unconditional distribution of Y_t is constant in time. We limit ourselves not just to stationary processes, but also to ones in which widely-separated observations are asymptotically independent. Without this restriction, convergence of the training error to the expected risk could occur arbitrarily slowly, and finite-sample bounds may not exist.[14] The next definition describes the sort of serial dependence which we entertain.

Definition 8.3 (β-Mixing). *Consider a stationary random sequence Y_∞ defined on a probability space $(\Omega, \Sigma, \mathbb{P}_\infty)$. Let $\sigma_{i:j} = \sigma(Y_{i:j})$ be the σ-field of events generated by the appropriate collection of random variables. Let $\mathbb{P}_0$ be the restriction of $\mathbb{P}_\infty$ to $\sigma_{-\infty:0}$, $\mathbb{P}_a$ be the restriction of $\mathbb{P}_\infty$ to $\sigma_{a:\infty}$, and $\mathbb{P}_{0\otimes a}$ be the restriction of $\mathbb{P}_\infty$ to $\sigma(Y_{\infty:0}, Y_{a:\infty})$. The* coefficient of absolute regularity*, or* β-mixing coefficient*, β_a, is given by*

$$\beta_a := ||\mathbb{P}_0 \times \mathbb{P}_a - \mathbb{P}_{0\otimes a}||_{TV}, \tag{54}$$

where $||\cdot||_{TV}$ is the total variation norm. A stochastic process is absolutely regular*, or* β-mixing*, if $\beta_a \to 0$ as $a \to \infty$.*

This is only one of many equivalent characterizations of β-mixing (see [13] for others). This definition makes clear that a process is β-mixing if the joint probability of events which are widely separated in time approaches the product of the individual probabilities, i.e., that Y_∞ is asymptotically independent. Many common time series models are known to be β-mixing, and the rates of decay are known up to constant factors which are functions of the true parameters of the process. Among the processes for which such results are known are ARMA models [43], GARCH models [14], and certain Markov processes — see [24] for an overview. Additionally, functions of β-mixing processes are β-mixing, so if $\mathbb{P}_\infty$ could be specified by a dynamic factor model or DSGE or VAR, the observed data would satisfy this condition.

Knowing β_a would let us determine the effective sample size of time series $Y_{1:n}$. In effect, having n dependent-but-mixing data points is like having $\mu < n$ independent ones. Once we determine the correct μ, we can (as we will now show) use concentration results for IID data like those in 5.8 and 6.8 with small corrections.

[14] In fact, [2] demonstrate that for ergodic processes, finite VC dimension is enough to give consistency, but not rates.

9 Risk bounds

With the relevant background in place, we can put the pieces together to derive our results. We use β-mixing to find out how much information is in the data and VC dimension to measure the capacity of the state-space model's prediction functions. The result is a bound on the generalization error of the chosen function $\widehat{f}$. After slightly modifying the definition of "risk" to fit the time-series forecasting scenario, and stating necessary technical assumptions, we derive risk bounds for wide classes of economic forecasting models.

9.1 Setup and assumptions

We observe a finite subsequence of random vectors $Y_{1:n}$ from a process Y_∞ defined on a probability space $(\Omega, \Sigma, \mathbb{P}_\infty)$, with $Y_i \in \mathbb{R}^p$. We make the following assumption on the process.

Assumption 2. $\mathbb{P}_\infty$ is a stationary, β-mixing process with mixing coefficients[15] β_a, $\forall a > 0$.

Under stationarity, the marginal distribution of Y_t is the same for all t. We deal mainly with the joint distribution of $Y_{1:n+1}$, where we observe the first n observations and try predicting Y_{n+1}. For the rest of this paper, we will call this joint distribution $\mathbb{P}$. Our results extend to predicting more than one step ahead, but the notation becomes cumbersome.
We must define generalization error and training error slightly differently for time series than in the IID setting. Using the same notion of loss functions as before, we consider prediction functions $f : \mathbb{R}^{n\times p} \to \mathbb{R}^p$

Definition 9.1 (Time series risk).

$$R_n(f) := \mathbb{E}\Big[\ell\left(Y_{n+1} - f(Y_{1:n})\right)\Big]. \tag{55}$$

The expectation is taken with respect to the joint distribution $\mathbb{P}$ and therefore depends on n. The function f may use some or all of the past to generate predictions. A function using only the most recent d observations

[15] In order to apply the results, one must either know β_a for some a or be able to estimate it with sufficient precision and accuracy. [39] shows how to estimate the mixing coefficients non-parametrically, based on a single sample from the process.

as inputs will be said to have *fixed memory* of length d. Other functions have *growing memory*, i.e., f may use all the previous data to predict the next data point. This incongruity makes the notation for time series training error somewhat problematic.

We will define the training error with a subscript $i \in \mathbb{N}$ on f within the summation. Strictly speaking, there is only one function f which we are using to make forecasts. In typical fixed memory settings — standard VAR forecasting models and so on — $f_i = f_j = f$ for all $i, j \in \mathbb{N}$. But for models with growing memory, a fixed forecasting method — an ARMA model, DSGE,[16] or linear state-space model — will use all of the past to make predictions, so the dimension of the domain changes with i. We write the risk of f as a single function, because, once we parameterize a forecasting method, an entire sequence of forecasting functions $f_1, f_2, \ldots$ is determined.

Definition 9.2 (Time series training error)**.**

$$\widehat{R}_n(f) := \frac{1}{n-d-1}\sum_{i=d}^{n-1} \ell\left(Y_{i+1} - f_i(Y_{1:i})\right). \tag{56}$$

In order to make use of this single definition of training error, we let $d \geq 0$. In fixed memory cases — say an AR(2) — d has an obvious meaning, while with growing memory, $d = 0$ is allowed.

To control the generalization error for time series forecasting, we make one final assumption, about the possible magnitude of the losses. Specifically, we weaken the bounded loss assumption we used in 4.3 to allow for unbounded loss as long as we retain some control on moments of the loss.

Assumption 3. Assume that for all $f \in \mathcal{F}$

$$Q_n(f) := \sqrt{\mathbb{E}_{\mathbb{P}}\left[\ell\left(Y_{n+1} - f(Y_{1:n})\right)^2\right]} \leq M < \infty. \tag{57}$$

3 is still quite general, allowing even some heavy tailed distributions.

[16] A DSGE is a nonlinear system of expectational difference equations, so estimating the parameters is nontrivial. Likelihood methods typically work by finding a linear approximation using Taylor expansions and the Kalman filter, though increasingly complex nonlinear methods are now intensely studied. See for instance [23] or [22]

9.2 Fixed memory

We can now state our results giving finite sample risk bounds for the problem of time series forecasting. We will only consider the fixed-memory situation, even though most DSGE[23] models having growing memory.

Theorem 9.3. *Suppose that 2 and 3 hold, that the model class $\mathcal{F}$ has a fixed memory length $d < n$, and that we have a sample $\mathbf{Y}_1^n$. Let μ and a be integers such that $2\mu a + d \leq n$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} > \epsilon\right) \tag{58}$$

$$\leq 8(2\mu+1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu \exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right)}{4}\right\} + 2\mu\beta_{a-d},$$

where $W(\cdot)$ is the Lambert W function.

The implications of this theorem are considerable. Given a finite sample of length n, we can say that with high probability, future prediction errors will not be much larger than our observed training errors. It makes no difference whether the model is correctly specified. This stands in stark contrast to model selection tools like AIC or BIC which appeal to asymptotics. Moreover, given a model class $\mathcal{F}$, we can say exactly how much data we need to have good control of the prediction risk. As the effective data size increases, the training error is a better and better estimate of the generalization error, uniformly over all of $\mathcal{F}$.

The Lambert W function in the exponential term deserves some explanation. The Lambert W function is defined as the inverse of $f(w) = w \exp w$ (cf. [18]). A strictly, but only slightly, worse bound can be achieved by noting that

$$\exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right) \leq \frac{\epsilon^{8/3}}{4^{2/3}} \tag{59}$$

for all $\epsilon \in [0,1]$.

The difference between expected and empirical risk is only interesting when $R_n(f)$ exceeds $\widehat{R}_n(f)$. Due to the supremum, events where the training error exceeds the expected risk are irrelevant. Therefore, we are only concerned with $0 \leq \widehat{R}_n(f) \leq R_n(f)$. Of course, as discussed in 4.3, for most estimation procedures, f is chosen to make $\widehat{R}_n(f)$ as small as possible.

One way to understand this theorem is to visualize the tradeoff between confidence ϵ and effective data μ. Consider, by way of illustration, what happens when $\textsc{vcd}(\mathcal{F}) = 1$, $\beta_a = 0$, and $M = 1$. Then 9.3 and 59 become

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} R_n(f) - \widehat{R}_n(f) > \epsilon\right) \leq 8\exp\left\{\log(2\mu+1) - \frac{\mu\epsilon^{8/3}}{4^{5/3}}\right\} \tag{60}$$

Our goal is to minimize ϵ, thereby ensuring that the relative difference between the expected risk and the training risk is small. At the same time we want to minimize the right side of the bound so that the probability of "bad" outcomes — samples where the difference in risks exceeds ϵ — is small. Of course we want to do this with as little data as possible, but the smaller we take ϵ, the larger we must take μ to compensate. We depict this tradeoff in 4.

The figure is structured so that movement toward the origin is preferable. We have tighter control on the difference in risks with less data. But moving in that direction leads to an increased probability of the bad event — that the difference in risks exceeds ϵ. The bound becomes trivial below the solid black line (the bad event occurs with probability no larger than one). The desire for the bad event to occur with low probability forces the decision boundary to the upper right.

Another way to interpret the plot is as a set of indifference curves. Anywhere in the same color region is equally desirable in the sense that the probability of equally bad events is the same. So if we had a budget constraint trading ϵ and data (i.e. a line with negative slope), we could optimize within the budget set to find the lowest probability allowable.

Before we prove 9.3, we will state a corollary which puts the same result in a form that is sometimes easier to use.

Corollary 9.4. *Under the conditions of 9.3, for any $f \in \mathcal{F}$, the following bound holds with probability at least $1-\eta$, for all $\eta > 2\mu\beta_{a-d}$:*

$$R_n(f) \leq \widehat{R}_n(f) + M\sqrt{\frac{\mathcal{E}(4-\log\mathcal{E})}{2}}, \tag{61}$$

with

$$\mathcal{E} = \frac{4\,\textsc{vcd}(\mathcal{F})\log(2\mu+1) + \log 8/\eta'}{\mu}, \tag{62}$$

and $\eta' = \eta - 2\mu\beta_{a-d}$.

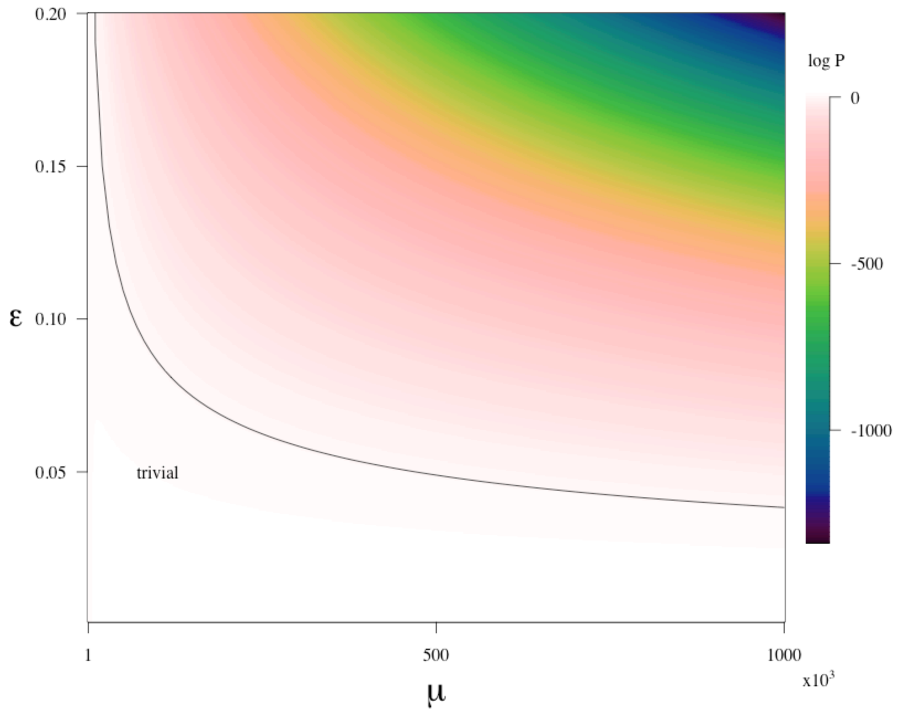

Figure 4: Visualizing the tradeoff between confidence (ϵ, y-axis) and effective data (μ, x-axis). The black curve indicates the region where the bound becomes trivial. Below this line, the probability is bounded by 1. Darker colors indicate lower probability of the "bad" event — that the difference in risks exceeds ϵ. The colors correspond to the natural logarithm of the bound on this probability.

We now prove both 9.3 and 9.4 to provide the reader with some intuition for the types of arguments necessary. We defer proof of the remainder of the theorems in this section to the appendix.

Proof of 9.3 and 9.4. The first step is to move from the actual sample size n to the effective sample size μ which depends on the β-mixing behavior. Let a and μ be non-negative integers such that $2a\mu + d \leq n$. Now divide

$\mathbf{Y}_1^n$ into 2μ blocks, each of length a, ignoring the remainder. Identify the blocks as follows:

$$U_j = \{Y_i : 2(j-1)a + 1 \leq i \leq (2j-1)a\}, \tag{63}$$

$$V_j = \{Y_i : (2j-1)a + 1 \leq i \leq 2ja\}. \tag{64}$$

Let $\mathbf{U}$ be the sequence of odd blocks U_j, and let $\mathbf{V}$ be the sequence of even blocks V_j. Finally, let $\mathbf{U}'$ be a sequence of blocks which are mutually independent and such that each block has the same distribution as a block from the original sequence. That is construct U_j' such that

$$\mathcal{L}(U_j') = \mathcal{L}(U_j) = \mathcal{L}(U_1), \tag{65}$$

where $\mathcal{L}(\cdot)$ means the probability law of the argument.

Let $\widehat{R}_{\mathbf{U}}(f)$, $\widehat{R}_{\mathbf{U}'}(f)$, and $\widehat{R}_{\mathbf{V}}(f)$ be the empirical risk of f based on the block sequences $\mathbf{U}$, $\mathbf{U}'$, and $\mathbf{V}$ respectively. Clearly $\widehat{R}_n(f) = \frac{1}{2}(\widehat{R}_{\mathbf{U}}(f) + \widehat{R}_{\mathbf{V}}(f))$. Then,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} > \epsilon\right) \tag{66}$$

$$= \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left[\frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{2Q_n(f)} + \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{2Q_n(f)}\right] > \epsilon\right)$$

$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{Q_n(f)} + \sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{Q_n(f)} > 2\epsilon\right) \tag{67}$$

$$\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon\right) + \mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{V}}(f)}{Q_n(f)} > \epsilon\right) \tag{68}$$

$$= 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon\right). \tag{69}$$

Now, apply Lemma 4.1 in [58] (reproduced as B.1 in B) to the of the event $\left\{\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon\right\}$. This allows us to move from statements about dependent blocks to statements about independent blocks with a slight correction. Therefore we have,

$$\begin{aligned}
&2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}}(f)}{Q_n(f)} > \epsilon\right) \\
&\quad \leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{Q_n(f)} > \epsilon\right) + 2(\mu - 1)\beta_{a-d}, \qquad (70)
\end{aligned}$$

where the probability on the right is for the σ-field generated by the independent block sequence $\mathbf{U}'$. Therefore,

$$\begin{aligned}
&2\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_{\mathbf{U}'}(f)}{Q_n(f)} > \epsilon\right) \\
&\quad \leq 8(2\mu+1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{\mu \exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right)}{4}\right\} \qquad (71)
\end{aligned}$$

where we have applied Theorem 7 in [19] (reproduced as B.2) to bound the independent blocks $\mathbf{U}'$.

To prove the corollary, set the right hand side of 71 to η, take $\eta' = \eta - 2(\mu - 1)\beta_{a-d}$, and solve for ϵ. We get that for all $f \in \mathcal{F}$, with probability at least $1 - \eta$,

$$\frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} \leq \epsilon. \qquad (72)$$

Solving the equation

$$\eta' = 8(2\mu+1)^h \exp\left\{-\frac{\mu \exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right)}{4}\right\} \qquad (73)$$

implies

$$\epsilon = M\sqrt{\frac{\mathcal{E}(4 - \log \mathcal{E})}{2}} \qquad (74)$$

with

$$\mathcal{E} = \frac{4\,\textsc{vcd}(\mathcal{F})\log(2\mu+1)+\log 8/\eta'}{\mu}. \tag{75}$$

■

The only obstacle to the use of 9.3 is knowledge of $\textsc{vcd}(\mathcal{F})$. For some models, the VC dimension can be calculated explicitly.

Theorem 9.5. *For the class of AR(d) models, $\mathcal{F}_{AR}(d)$,*

$$\textsc{vcd}(\mathcal{F}_{AR}(d)) = d+1. \tag{76}$$

For the class of VAR(d) models with k time series, $\mathcal{F}_{VAR}(k,d)$,

$$\textsc{vcd}(\mathcal{F}_{VAR}(k,d)) = kd+1. \tag{77}$$

9.5 applies equally to Bayesian VARs. However, this is likely too conservative as the prior tends to restrict the effective complexity of the function class.[17] We will not consider real-world examples here but we refer you to [39] for further details, in fact a lot of the above is liberally taken from that paper.

10 Discussion

The classical results in Statistical Learning theory are only concerned with independent data, we introduced some of these and VC theory[55] and we introduced some time series results in 8. While explaining how to compute risk bounds for Time series we introduced the class of examples called mixing. We refer the curious reader, intrigued by mixing theory to look at

[17] Here we should mention that these risk bounds are frequentist in nature. We mean that if one treats Bayesian methods as a regularization technique and predicts with the posterior mean or mode, then our results hold. However, from a subjective Bayesian perspective, our results add nothing since all inference can be derived from the posterior. For further discussion of the frequentist risk properties of Bayesian methods under misspecification, see for example [49]

the following references [6, 33, 39] and to [13] for a survey article.. We only briefly spoke about regression and time-series, this is a vast topic all of its' own and we recommend [40] and the references in there. We unfortunately didn't have time to discuss the beautiful and technically difficult calculations for computing Rademacher Averages, in [11] this is covered and in other textbooks such as [10, 45]. As noted in [4] these have been calculated for many of the standard Machine Learning applications such as Neural Networks and Support Vector Machines. Anyone interested in the exciting world of data-mining and machine learning can of course learn from [29] or your favourite Machine Learning textbook. There are many other sorts of bounds which can be used from Empirical Process theory or Statistical Learning Theory to help compute Rademacher Averages, such as the Haussler bound [27] or the Dudley Bound[25]. These bounds are worth a chapter all of their own, but the reader has already mastered the necessary pre-requisites.

There are other kinds of concentration inequalities of Bernstein type - and these are examined in [45, 51]. They are basically concentration inequalities that include variance and other moments - we only considered expectation. For those interested in the links between statistical learning theory and stochastic optimization they can read [15] and a free pdf is available (at the time of writing) on Prof Catoni's website. This thesis takes results from many different fields, and indeed it is hard to tell when one field ends, and another begins - for example we did not examine the Information Theoretic view of Concentration of Measure in much detail, although we did briefly consider some examples - the interested reader who cares about coding theory and such can look at the excellent monograph by Raginsky[47].

A Machine Learning - Computational Techniques

It is worth mentioning for all Statisticians - that it is useful to generate simulations, especially in applied problems from some of the excellent Machine Learning libraries out there. I cite two for your research, there is the excellent Theano, which is new as of the time of writing. We refer to [7] for further details. Most diagrams and plots in this publication are produced by [44] another excellent Python library.

Readers interested in Statistical Learning and with other programming languages can find advice on R in [29] and sample code is included in [39]

B Auxiliary results

Lemma B.1 (Lemma 4.1 in [58]). *Let Z be an event with respect to the block sequence* $\mathbf{U}$. *Then,*

$$|\mathbb{P}(Z) - \widetilde{\mathbb{P}}(Z)| \leq \beta_a(\mu - 1), \tag{78}$$

where the first probability is with respect to the dependent block sequence, $\mathbf{U}$, *and* $\widetilde{\mathbb{P}}$ *is with respect to the independent sequence,* $\mathbf{U}'$.

This lemma essentially gives a method of applying IID results to β-mixing data. Because the dependence decays as we increase the separation between blocks, widely spaced blocks are nearly independent of each other. In particular, the difference between expectations over these nearly independent blocks and expectations over blocks which are actually independent can be controlled by the β-mixing coefficient.

Lemma B.2 (Theorem 7 in [19]). *Under 3,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} > \epsilon\sqrt{2 + \log\frac{1}{\epsilon}}\right) \leq 4(2n+1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{n\epsilon^2}{4}\right\} \tag{79}$$

Corollary B.3.

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \frac{R_n(f) - \widehat{R}_n(f)}{Q_n(f)} > \epsilon\right)$$
$$\leq 4(2n+1)^{\text{VCD}(\mathcal{F})} \exp\left\{-\frac{n\exp\left(W\left(-\frac{2\epsilon^2}{e^4}\right)+4\right)}{4}\right\}. \tag{80}$$

C Proofs of selected results

Proof of 9.5. The VC dimension of a linear classifier $f : \mathbb{R}^d \to \{0,1\}$ is d (cf. [55]). Real valued predictions have an extra degree of freedom.

For the VAR case, we are interested in the VC dimension of a multivariate linear classifier. Thus, one must be able to shatter collections of vectors where each vector is a binary sequence of length k. For a VAR, each coordinate is independent, thus, one can shatter a collection of vectors if one can shatter each coordinate projection. The result then follows from the AR case. ∎

References

[1] pandas: Python Data Analysis Library. Online, 2012.

[2] Terrence M. Adams and Andrew B. Nobel. Uniform convergence of vapnikchervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 07 2010.

[3] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.

[4] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003.

[5] Alexander Barvinok. Math 710: Measure concentration. *Lecture Notes*, 2005.

[6] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013.

[7] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

[8] S.N Bernstein. Probability theory (4th ed.)(in russian), 1946.

[9] Sergej G Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[10] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[11] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In Olivier Bousquet, Ulrike Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer Berlin Heidelberg, 2004.

[12] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[13] Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[14] Marine Carrasco and Xiaohong Chen. Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory*, null:17–39, 2 2002.

[15] O. Catoni and J. Picard. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI - 2001*. Lecture notes in mathematics. Springer, 2004.

[16] S. Chatterjee. An error bound in the Sudakov-Fernique inequality. *ArXiv Mathematics e-prints*, October 2005.

[17] S. Chatterjee and P. S. Dey. Applications of Stein's method for concentration inequalities. *ArXiv e-prints*, June 2009.

[18] Gonnet G. Hare D. Jeffrey D. Corless, R. and D. Knuth. On the lambert w function. *Advances in Computational Mathematics*, 5(1).

[19] C. Cortes, Y. Mansour, and M. Mohri. *Learning bounds for importance weighting," in Advances in Neural Information Processing Systems 23*, volume 23. MIT Press, 2010.

[20] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[21] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[22] David N. DeJong, Roman Liesenfeld, Guilherme V. Moura, Hariharan Dharmarajan, and Jean-François Richard. Efficient likelihood evaluation of state-space representations, 2009.

[23] D.N. DeJong and C. Dave. *Structural Macroeconometrics: (Second Edition)*. Princeton University Press, 2011.

[24] P. Doukhan. *Mixing: properties and examples*. Lecture notes in statistics. Springer-Verlag, 1994.

[25] R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

[26] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999.

[27] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247 – 261, 1989.

[28] Evarist Giné and Joel Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, pages 929–989, 1984.

[29] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2001.

[30] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[31] WilliamB. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, 1986.

[32] M. J. Kearns and L. K. Saul. Large deviation methods for approximate probabilistic inference in *Proceedings of the 14th Conference on Uncertaintly in Artifical Intelligence*, 1998.

[33] L. Kontorovich. Metric and Mixing Sufficient Conditions for Concentration of Measure. *ArXiv Mathematics e-prints*, October 2006.

[34] M. Ledoux. Concentration of measure and logarithmic sobolev inequalities in *Séminaire de Probabilités XXXIII* ser. lecture notes in math, 1999.

[35] G. Lugosi. Concentration of measure inequalities - lecture notes, 2009.

[36] Katalin Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.

[37] C. McDiarmid. Centering sequences with bounded differences. *Combinatorics, Probability and Computing*, 6, 1997.

[38] C. McDiarmid. *Concentration,*. Probabilistic Methods for Algorithmic Discrete Mathematics. Springer, 1998.

[39] D. J. McDonald, C. Rohilla Shalizi, and M. Schervish. Time series forecasting: model evaluation and selection using nonparametric risk bounds. *ArXiv e-prints*, December 2012.

[40] A.D.R. McQuarrie and C.L. Tsai. *Regression and Time Series Model Selection.* World Scientific, 1998.

[41] V.D. Milman. A new proof of a. dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Prilozhen. (in Russian)*, 5 (4):2837, 1971.

[42] M.Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs.* American Mathematical Society, 2001.

[43] Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic Processes and their Applications*, 29(2):309–315, September 1988.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] P.Massart. *The Concentration of Measure Phenomenon*. Lecture Notes in Mathematics. Springer, 2007.

[46] D. Pollard. *Empirical Processes: Theory and Applications*. Conference Board of the Mathematical Science: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1990.

[47] Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding. *CoRR*, abs/1212.4663, 2012.

[48] D. Romer. *Advanced Macroeconomics*. The McGraw-Hill series in economics. McGraw-Hill Education, 2011.

[49] Cosma Rohilla Shalizi. Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, 3:1039–1074, 2009.

[50] M. Talagrand. Concentration of measure and isoperimteric inequalities in product space. *Publications Mathématiques de l'I.H.E.S*, 81:73–205, 1995.

[51] Michel Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[52] T. Tao. *Topics in Random Matrix Theory*. Graduate studies in mathematics. American Mathematical Society, 2012.

[53] Hugo Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478(13):1 – 69, 2009.

[54] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4), August 2012.

[55] V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer, 2000.

[56] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[57] S. R. S. Varadhan. Large deviations. *The Annals of Probability*, 36(2):397–419, 03 2008.

[58] V.Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

[59] Peter Whittle. Large-deviation theory. In *Probability via Expectation*, Springer Texts in Statistics, pages 306–316. Springer New York, 2000.