Concentration of Measure and High-Dimensional Machine Learning

Aurélien Garivier

October 15, 2026

Preface

This book grew out of a Master's course taught at the École normale supérieure de Lyon, *Concentration of measure and high-dimensional statistical learning*, a course I have the pleasure of teaching together with Guillaume Aubrun, Rémi Gribonval and Jean-Christophe Mourrat. Each chapter corresponds to one lecture. The audience is made of mathematicians, and the text is written for them: every statement is proved, or else explicitly labelled as admitted and given a precise reference.

Three principles have guided the writing.

Self-containedness. The prerequisites are a solid first course in measure-theoretic probability – σ-fields, expectation, independence, conditional expectation, the classical limit theorems – and nothing else. In particular, no previous exposure to statistical learning is assumed; the learning problems that motivate the probabilistic material are set up from scratch in Chapter 1.

Rigour. Concentration inequalities are quantitative statements, and their interest lies in their constants as much as in their shape. I have therefore avoided the "$\lesssim$" style: constants are explicit, hypotheses are minimal and stated in full, and measurability issues – which are real, and which the subject tends to sweep under the carpet – are discussed where they arise rather than ignored.

Remarks. A theorem is rarely interesting by itself; what is interesting is what it says that its neighbours do not, where it is sharp, where it is wasteful, who proved it and what they were trying to do. A substantial part of each chapter consists of such remarks. The reader in a hurry may skip them; the reader who wants to *use* these inequalities should not.

Each chapter ends with a set of exercises, ordered roughly by difficulty. Complete solutions are collected in Appendix A. Exercises marked with a star are used later in the text.

Lyon, October 15, 2026 A. G.

Contents

Preface iii

List of proposed revisions ix

1 From Statistical Learning to Deviation Inequalities 1
- 1.1 What machine learning is about . . . 1
- 1.2 Probabilistic preliminaries and notation . . . 2
- 1.3 The statistical learning framework . . . 3
  - 1.3.1 Data, rules, risk . . . 3
  - 1.3.2 Learning algorithms . . . 4
  - 1.3.3 Two ways of factorising $\mathcal{D}$ . . . 5
  - 1.3.4 General losses, and the regression case . . . 6
- 1.4 The Bayes classifier and the noise of a problem . . . 6
- 1.5 The nearest-neighbour classifier . . . 8
  - 1.5.1 What the analysis says . . . 11
  - 1.5.2 Beyond the cube: covering numbers and intrinsic dimension . . . 14
- 1.6 Limit theorems, and why they are not enough . . . 15
- 1.7 Markov's inequality and its descendants . . . 17
  - 1.7.1 The laws of large numbers, proved . . . 19
  - 1.7.2 The other direction: anti-concentration . . . 20
- 1.8 The Chernoff method . . . 20
- 1.9 Hoeffding's lemma and Hoeffding's inequality . . . 21
- 1.10 First consequences for learning . . . 25
  - 1.10.1 Back to nearest neighbours: the k-NN rule . . . 26
  - 1.10.2 From deviation bounds to almost-sure statements . . . 29
  - 1.10.3 Application: comparing two treatments . . . 30
  - 1.10.4 How many samples? A summary . . . 31
- 1.11 Where we are going . . . 31
- 1.12 Notes and references . . . 32
- 1.13 Exercises . . . 33

2 The Cramér–Chernoff Method 37
- 2.1 What Hoeffding's inequality misses . . . 37
- 2.2 The cumulant generating function as a free energy . . . 37
- 2.3 The Cramér transform and the Cramér–Chernoff inequality . . . 40
  - 2.3.1 The basic Cramér transforms . . . 44
- 2.4 Sub-Gaussian random variables . . . 45
  - 2.4.1 Equivalent characterisations . . . 47
- 2.5 Sub-gamma variables, Bennett and Bernstein . . . 49

2.5.1 The model computation: Poisson deviations . . . 49
2.5.2 Bennett's inequality . . . 53
2.5.3 Bernstein's inequality . . . 56
2.5.4 Four bounds, compared . . . 57
2.6 Beyond sums: the bounded-difference inequality . . . 59
2.6.1 Bernstein for martingales: Freedman's inequality . . . 62
2.7 Applications . . . 63
2.7.1 Finite hypothesis classes: the first PAC bounds . . . 63
2.7.2 Fast rates under Massart's margin condition . . . 66
2.7.3 Rare events: estimating a small probability . . . 68
2.7.4 A preview: random projections and the Johnson–Lindenstrauss lemma 68
2.7.5 Bounded differences at work . . . 69
2.7.6 Back to nearest neighbours: concentration of the leave-one-out error 70
2.7.7 Sums of random vectors . . . 72
2.8 Notes and references . . . 73
2.9 Exercises . . . 74

3 Lower Bounds: Relative Entropy, Large Deviations and No Free Lunch 77
3.1 Why lower bounds . . . 77
3.2 Bernoulli deviations revisited . . . 78
3.2.1 The Chernoff–Hoeffding bound . . . 78
3.2.2 One identity, and Pollard's lemma . . . 79
3.2.3 Worked comparisons . . . 81
3.2.4 A divergence on the set of means . . . 83
3.3 Relative entropy . . . 84
3.3.1 Definition and examples . . . 84
3.3.2 Positivity, convexity, tensorisation . . . 86
3.3.3 Data processing . . . 88
3.3.4 From relative entropy to probabilities of events . . . 89
3.3.5 The variational formula . . . 91
3.3.6 The testing interpretation . . . 92
3.4 Lower bounds for deviations . . . 95
3.4.1 Change of measure: the asymptotic lower bound . . . 95
3.4.2 The entropic way: a non-asymptotic lower bound . . . 97
3.4.3 An alternative route through Bretagnolle–Huber . . . 99
3.4.4 What each level of precision costs . . . 99
3.5 The minimax toolkit . . . 100
3.5.1 Minimax risk and the reduction to testing . . . 101
3.5.2 Le Cam's two-point method . . . 101
3.5.3 Adaptive sampling: comparing two treatments . . . 103
3.5.4 Fano's inequality . . . 104
3.5.5 Assouad's lemma . . . 105
3.6 No free lunch . . . 107
3.6.1 The curse of dimensionality . . . 110
3.7 Notes and references . . . 112
3.8 Exercises . . . 113

4 PAC Learning and Vapnik–Chervonenkis Theory 115
4.1 Introduction . . . 115
4.2 PAC learning and uniform convergence . . . 116

4.2.1 Setting and standing assumptions . . . 116
4.2.2 PAC learnability, revisited . . . 117
4.2.3 Uniform convergence . . . 119
4.2.4 Finite classes . . . 120
4.3 Shattering and the Vapnik–Chervonenkis dimension . . . 120
4.3.1 Definitions . . . 120
4.3.2 Examples . . . 121
4.3.3 Infinite VC dimension forbids learning . . . 125
4.4 The Sauer–Shelah lemma . . . 126
4.4.1 The lemma . . . 126
4.4.2 Two proofs of the polynomial bound . . . 127
4.4.3 First consequences: combining classes . . . 128
4.5 Uniform convergence through Rademacher complexity . . . 130
4.5.1 Rademacher complexity . . . 130
4.5.2 Symmetrisation . . . 131
4.5.3 Massart's lemma . . . 133
4.5.4 Uniform convergence with high probability . . . 134
4.5.5 How sharp is this? The logarithmic factor and the rate . . . 136
4.5.6 The realizable case: a fast rate through the double sample . . . 138
4.6 The fundamental theorem of statistical learning . . . 140
4.6.1 The qualitative statement . . . 140
4.6.2 Lower bounds . . . 140
4.6.3 The quantitative statement . . . 144
4.7 Complements . . . 145
4.7.1 A fixed distribution, almost sure convergence . . . 145
4.7.2 Choosing the class: structural risk minimisation . . . 145
4.8 Notes and references . . . 146
4.9 Exercises . . . 147

5 Beyond Sums of Independent Variables **151**
5.1 Introduction . . . 151
5.2 The missing mass and the Good–Turing estimator . . . 152
5.2.1 Setting and notation . . . 152
5.2.2 Good's lemma: the bias of the estimator . . . 154
5.3 The variance first . . . 155
5.3.1 The variance of the missing mass, by hand . . . 155
5.3.2 The Efron–Stein inequality . . . 156
5.3.3 First consequences . . . 158
5.4 Negative association . . . 160
5.4.1 Definition and first properties . . . 160
5.4.2 Two closure properties . . . 161
5.4.3 The basic examples . . . 162
5.4.4 Chernoff's method for negatively associated variables . . . 164
5.4.5 Sampling without replacement: Hoeffding's comparison . . . 165
5.5 The bounded-difference inequality, revisited . . . 166
5.6 The entropy method, in its simplest form . . . 169
5.6.1 Entropy of a non-negative variable and its two variational formulas . . . 169
5.6.2 Tensorisation of the entropy . . . 169
5.6.3 From entropy to cumulant generating functions . . . 170

5.6.4 Self-bounding functions behave like Poisson variables 171
5.7 Concentration of the missing mass 173
5.7.1 The Good–Turing estimator: bounded differences 173
5.7.2 The missing mass: negative association and tilted variances 173
5.7.3 The confidence region 176
5.7.4 How sharp is this? 177
5.7.5 Poissonisation 178
5.7.6 Histogram classification rules 181
5.8 Notes and references 184
5.9 Exercises 185

6 Beyond Binary-Valued Classes (Under Construction) 189
6.1 Contraction and margins 189

A Solutions to the exercises 191
Chapter 1 191
Chapter 2 195
Chapter 3 200
Chapter 4 202
Chapter 5 206

Bibliography 211

Index 221

List of proposed revisions

Everything marked in red in the text is a proposal, not a correction already made: the original version is left standing next to it, and *both versions are written out in full*, so that adopting one means deleting the other. Boxed blocks, listed below, carry a complete replacement statement or proof; shorter inline comments quote the exact sentence they would substitute, and are not listed. Where a decision was taken in the review of 18 September, the block says so at the end. The identifiers V1, V2, ... are *frozen*: adopting or dropping one revision never renumbers the others. Revisions V12–V57, decided on 2 and 5 October, have been applied and are no longer marked; they are recorded in `goals.txt`, Section 12. Revisions V58–V60, proposed on 15 October with Chapter 5, are pointers from Chapters 1–2 to the new chapter and are not yet discussed.

V58 Pointer to Chapter 5 (sampling without replacement, negative association) . 25
V1 shorter and sharper: go from (a) to (c) directly 48
V2 the two probability bounds need no calculus at all 51
V3 prove (2.21) directly, with the natural λ 52
V4 the optimality claim is not justified as written 54
V5 one constant more, and Remark 2.43 becomes a corollary 54
V6 Step 2 can be deleted: freeze the past and use the ordinary Hoeffding lemma 61
V59 Pointer to Chapter 5 (Efron–Stein, entropy method) 62
V7 Step 4 in one line, by arithmetic–geometric mean 67
V60 Pointer to the self-bounding inequality of Chapter 5 73
V8 the "orthant" argument needs $r \le \frac{1}{2}$ 193
V9 the inequality is invoked in the wrong direction 193
V10 this last sentence is wrong 197
V11 the accuracy claims are too optimistic; here are the real numbers 198

Chapter 1

From Statistical Learning to Deviation Inequalities

Summary. We set up the statistical learning framework, identify the optimal classifier (the Bayes rule) and analyse the simplest learning algorithm of all, the nearest-neighbour rule. That analysis produces a bound on the *expected* risk, and immediately raises the question that occupies the rest of the book: how far can a random quantity be from its expectation? We answer it here in its most elementary form – Markov, Chebyshev, the Chernoff method and Hoeffding's inequality – and we measure precisely what each of these tools costs and what it buys.

1.1 What machine learning is about

Machine learning is the art of building algorithms that construct a model from examples in order to make predictions or take decisions, rather than following instructions written in advance by a programmer. The classical formulation is due to Samuel [Sam59], who described the field as the one that *gives computers the ability to learn without being explicitly programmed*, and the operational one to Mitchell [Mit97]:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Mitchell's sentence is worth reading as a mathematical programme. It asks for three objects – a class of tasks, a performance measure, a notion of experience – and for a statement about a limit: performance improves with experience. Everything in this book is an attempt to make such statements quantitative. We shall not be content with "performance improves"; we want to know *how fast*, *with what confidence*, and *at what price in the dimension of the data*.

The three standard learning models differ by the nature of the experience.

- In *supervised learning*, the experience is a finite list of pairs (x_i, y_i), where x_i describes an individual and y_i is the answer we would like to predict for it. The task is to produce a function $x \mapsto y$. When the answers are discrete one speaks of *classification*, when they are real-valued of *regression*.
- In *unsupervised learning*, the experience is a list of individuals $x_1, \dots, x_n$ with no answers attached, and the task is to exhibit structure: clusters, a low-dimensional summary, a density, a list of anomalies.

- In *reinforcement learning*, the experience is not given in advance at all: an agent interacts with an environment, chooses actions, receives rewards, and must maximise the reward it accumulates. The data it sees are a consequence of the decisions it takes, which makes the statistical problem fundamentally different, and considerably harder.

This book is concerned with the mathematics that underlies all three, and it uses supervised classification as its running example because it is the setting in which the probabilistic phenomena we study appear in their purest form.

Remark 1.1 (Machine learning, statistics, and the object of study). The mathematical content of supervised learning overlaps heavily with mathematical statistics, and it is a fair question to ask what distinguishes them. The honest answer is: the object one is willing to call a solution. Classical statistics is interested in a *model* – a family of distributions indexed by a parameter – and in the estimation of that parameter, because the parameter is meaningful: it is a physical constant, an elasticity, a treatment effect. Statistical learning is interested in a *decision rule*, and it is prepared to accept a rule that predicts well while having no interpretation whatsoever, and to work without a model at all: the distribution of the data is arbitrary and unknown, and the guarantees must hold uniformly over all of them. This is why *distribution-free* and *non-asymptotic* results are the currency of the field, and why concentration inequalities – which are exactly the distribution-free, non-asymptotic replacements for the central limit theorem – occupy the place they do.

1.2 Probabilistic preliminaries and notation

We collect here, without proofs, the facts from measure-theoretic probability that will be used constantly, and we fix the notation. The reader who is comfortable with conditional expectation may skip to Section 1.3 and return here when needed; the references are [Wil91] and [Dur19].

All random objects live on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, rich enough to carry all the variables we introduce. A *random element* of a measurable space $(E, \mathcal{E})$ is a measurable map $\Omega \to E$; its *law* is the image measure. We write $X \sim P$ to mean that X has law P, and $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$ for independent random elements with common law P. Expectation is $\mathbb{E}$, variance $\mathbb{V}\mathrm{ar}$, covariance $\mathbb{C}\mathrm{ov}$; $\mathbb{1}_A$ or $\mathbb{1}\{A\}$ is the indicator of an event, and $a \vee b$, $a \wedge b$, $a_+ = a^+ = a \vee 0$ have their usual meanings; the second form of the positive part is used when a subscript $+$ is already taken, as in λ_+ and x_+ of Chapter 2.

Conditional expectation. Let Z be integrable and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-field. There is a $\mathcal{G}$-measurable integrable random variable, unique up to almost-sure equality and denoted $\mathbb{E}[Z \mid \mathcal{G}]$, such that $\mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{G}]\mathbb{1}_G\big] = \mathbb{E}[Z\mathbb{1}_G]$ for every $G \in \mathcal{G}$. We use without comment:

(i) *linearity* and *positivity*;

(ii) the *tower property*: $\mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{G}]\big] = \mathbb{E}[Z]$, and more generally $\mathbb{E}\big[\mathbb{E}[Z \mid \mathcal{F}'] \mid \mathcal{G}\big] = \mathbb{E}[Z \mid \mathcal{G}]$ for $\mathcal{G} \subseteq \mathcal{F}'$;

(iii) *taking out what is known*: $\mathbb{E}[WZ \mid \mathcal{G}] = W\,\mathbb{E}[Z \mid \mathcal{G}]$ when W is $\mathcal{G}$-measurable and WZ integrable;

(iv) *independence*: if Z is independent of $\mathcal{G}$ then $\mathbb{E}[Z \mid \mathcal{G}] = \mathbb{E}[Z]$;

(v) *conditional Jensen*: $f\big(\mathbb{E}[Z \mid \mathcal{G}]\big) \le \mathbb{E}[f(Z) \mid \mathcal{G}]$ for f convex, whenever both sides are defined.

We write $\mathbb{E}[Z \mid X]$ for $\mathbb{E}[Z \mid \sigma(X)]$ and $\mathbb{P}(A \mid \mathcal{G})$ for $\mathbb{E}[\mathbb{1}_A \mid \mathcal{G}]$.

Regular conditional distributions. If X takes values in a measurable space and Y in a *Polish* space $\mathcal{Y}$ with its Borel σ-field, there exists a *regular conditional distribution* of Y given X: a map $(x, B) \mapsto \kappa(B \mid x)$ such that $\kappa(\cdot \mid x)$ is a probability measure for each x, $\kappa(B \mid \cdot)$ is measurable for each B, and $\mathbb{P}(Y \in B \mid X) = \kappa(B \mid X)$ almost surely [Dud02, Chap. 10]. All the label spaces met in this book – finite sets and $\mathbb{R}^d$ – are Polish, so we use this freely; it is what allows us to speak of "the conditional law of Y given $X = x$" and of the function η of (1.5).

Independence and Fubini. If X and Y are independent with laws P and Q and $f \geq 0$ is measurable, then

$$\mathbb{E}\big[f(X,Y)\big] = \int \Big(\int f(x,y)\, \mathrm{d}Q(y) \Big)\, \mathrm{d}P(x) = \mathbb{E}\big[g(X)\big], \qquad g(x) := \mathbb{E}\big[f(x,Y)\big], \tag{1.1}$$

and g is measurable. Identity (1.1) – "freeze the first variable, average over the second" – is used in almost every proof of this book, most visibly in Lemma 1.23 and in the proof of Theorem 2.55. It also shows that $\mathbb{E}[f(X,Y) \mid X] = g(X)$ almost surely.

Borel–Cantelli. For events $(A_n)_{n \geq 1}$, write $\{A_n \text{ i.o.}\} = \bigcap_N \bigcup_{n \geq N} A_n$. If $\sum_n \mathbb{P}(A_n) < \infty$ then $\mathbb{P}(A_n \text{ i.o.}) = 0$; if moreover the A_n are independent and $\sum_n \mathbb{P}(A_n) = \infty$ then $\mathbb{P}(A_n \text{ i.o.}) = 1$. Only the first half is used below, and it is the standard bridge from a sequence of deviation bounds to an almost-sure statement (Section 1.10.2).

1.3 The statistical learning framework

1.3.1 Data, rules, risk

We fix throughout a measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, the *domain* or *feature space*, whose points describe the individuals we observe, and a measurable space $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, the *label space*, whose points are the answers we want to predict. In binary classification $\mathcal{Y} = \{0, 1\}$; in multiclass classification $\mathcal{Y}$ is finite; in regression $\mathcal{Y} = \mathbb{R}$. Typically $\mathcal{X} = \mathbb{R}^d$ with its Borel σ-field, and d is large: a 28×28 grey-level image is a point of $\mathbb{R}^{784}$.

Definition 1.2 (Statistical model of supervised learning)**.** A *learning problem* is a probability measure $\mathcal{D}$ on the product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{Y}})$. A *training sample of size n* is a sequence

$$\mathcal{S}_n = \big((X_1, Y_1), \ldots, (X_n, Y_n)\big)$$

of independent random pairs, each distributed according to $\mathcal{D}$; we write $\mathcal{S}_n \sim \mathcal{D}^{\otimes n}$. A *decision rule*, or *classifier*, is a measurable map $h : \mathcal{X} \to \mathcal{Y}$, and its *risk* is

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) \neq Y\big) = \mathcal{D}\big(\{(x,y) : h(x) \neq y\}\big). \tag{1.2}$$

The set $\{(x, y) : h(x) \neq y\}$ is measurable as soon as h is measurable and the diagonal of $\mathcal{Y} \times \mathcal{Y}$ is, which we assume throughout (it holds for $\mathcal{Y}$ finite, and more generally for $\mathcal{Y}$ a separable metric space with its Borel σ-field). We write $\mathcal{D}_{\mathcal{X}}$ for the first marginal of $\mathcal{D}$, i.e. the law of X, and we use $\mathbb{P}$ and $\mathbb{E}$ for probability and expectation with respect to whatever randomness is present, indicating the relevant variables when there is a risk of confusion.

Two remarks on Definition 1.2 are in order, and they are not pedantic: they delimit exactly what the theory can and cannot claim.

Remark 1.3 (The i.i.d. assumption). Everything rests on the hypothesis that the examples are independent draws from $\mathcal{D}$, *and that the points we shall have to classify later are drawn from the same* $\mathcal{D}$. This is a strong hypothesis, and it fails in most applications one can think of: medical data collected in one hospital are used in another, a spam filter trained in 2019 faces the spam of 2026, a recommendation system modifies the behaviour of the users whose behaviour it learns from. The mathematics below is exact; its applicability is a modelling question, and the honest position is to regard the i.i.d. framework as the simplest non-trivial idealisation, on which everything else is built.

Remark 1.4 (There is no "true function"). Nothing in Definition 1.2 says that Y is a function of X. In general it is not: two individuals with the same measured features may have different labels, either because the features do not determine the answer or because the labels are noisy. One distinguishes the *realizable* case, in which there exists a measurable h with $\mathbb{P}(h(X) = Y) = 1$ – or, relative to a class $\mathcal{H}$ of classifiers, $\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$ (Definition 2.63) – from the *agnostic* case, in which no such assumption is made. Realizability is a very strong assumption – it says the problem is noiseless – and we shall not make it unless explicitly stated. Section 1.4 quantifies exactly how much of the risk is irreducible noise.

1.3.2 Learning algorithms

A learning algorithm does not produce a classifier: it produces a *classifier-valued function of the data*, and this is where randomness enters.

Definition 1.5 (Learning algorithm)**.** A *learning algorithm* is a sequence $\mathcal{A} = (\mathcal{A}_n)_{n\geq 1}$ of maps

$$\mathcal{A}_n : (\mathcal{X}\times\mathcal{Y})^n \longrightarrow \mathcal{M}(\mathcal{X},\mathcal{Y}), \qquad s \longmapsto \mathcal{A}_n(s),$$

where $\mathcal{M}(\mathcal{X},\mathcal{Y})$ denotes the set of measurable maps $\mathcal{X}\to\mathcal{Y}$, such that the evaluation map $(s,x)\mapsto \mathcal{A}_n(s)(x)$ is measurable from $(\mathcal{X}\times\mathcal{Y})^n\times\mathcal{X}$ to $\mathcal{Y}$. Given a training sample $\mathcal{S}_n$ we write $\widehat{h}_n := \mathcal{A}_n(\mathcal{S}_n)$ for the classifier it returns, and

$$R_n(\mathcal{A}) := \mathbb{E}_{\mathcal{S}_n\sim\mathcal{D}^{\otimes n}}\left[L_{\mathcal{D}}(\widehat{h}_n)\right] \tag{1.3}$$

for its *expected risk* at sample size n.

The joint measurability required in Definition 1.5 is exactly what makes $L_{\mathcal{D}}(\widehat{h}_n)$ a random variable: by Fubini's theorem,

$$L_{\mathcal{D}}(\widehat{h}_n) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}\left\{\mathcal{A}_n(\mathcal{S}_n)(x)\neq y\right\}\,\mathrm{d}\mathcal{D}(x,y)$$

is a measurable function of $\mathcal{S}_n$, so that (1.3) makes sense. Note the structure: $L_{\mathcal{D}}(\widehat{h}_n)$ is a random variable, being a function of the random sample, and $R_n(\mathcal{A})$ is its expectation. A bound on $R_n(\mathcal{A})$ says what happens *on average over training samples*; it says nothing about the probability that the particular sample we were given produced a bad classifier. Turning the first kind of statement into the second is precisely what concentration inequalities do, and we shall carry out the operation explicitly in the next chapter.

Remark 1.6 (Does the empirical risk minimiser exist, and is it measurable?). The learning algorithm most often considered is *empirical risk minimisation* over a class $\mathcal{H}$: $\widehat{h}_n \in \arg\min_{h\in\mathcal{H}} L_{\mathcal{S}_n}(h)$. Two difficulties hide behind this formula, and it is worth naming them once.

Existence. For the 0–1 loss the minimum is always attained, since $L_{\mathcal{S}_n}(h)$ takes only the $n+1$ values $0, \frac{1}{n}, \dots, 1$. For the general losses of Section 1.3.4 the infimum need not be attained: it is for $\mathcal{H}$ finite, or for $\mathcal{H}$ compact and $h \mapsto L_{\mathcal{S}_n}(h)$ lower semicontinuous; in general one works with an *approximate* minimiser, $L_{\mathcal{S}_n}(\widehat{h}_n) \le \inf_{h\in\mathcal{H}} L_{\mathcal{S}_n}(h) + \alpha_n$ with $\alpha_n \to 0$, and the bounds pick up an additive α_n.

Measurability. Even when the minimum is attained, the map $s \mapsto \arg\min_{h\in\mathcal{H}} L_s(h)$ need not admit a measurable selection, so Definition 1.5 may fail; and the quantity $\sup_{h\in\mathcal{H}} \left|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\right|$, which we shall want to bound, is a supremum of uncountably many random variables and need not be a random variable at all. There are two standard remedies. The first is to assume that $\mathcal{H}$ contains a countable subclass $\mathcal{H}_0$ such that every $h \in \mathcal{H}$ is a pointwise limit of elements of $\mathcal{H}_0$; the supremum over $\mathcal{H}$ then equals the supremum over $\mathcal{H}_0$ and is measurable. The second is to assume that $\mathcal{H}$ is *image-admissible Suslin*, a condition under which measurable selections exist [Dud99, Chap. 5]. All the classes met in practice satisfy the first condition, and we shall assume it whenever a supremum over a class appears; in Chapter 4, where infinite classes are the main object, it is stated explicitly as Assumption 4.1. The point of saying so is not pedantry: the failure is real, and counterexamples exist [DGL96, App. C].

Definition 1.7 (Empirical risk)**.** The *empirical risk* of a classifier h on the sample $\mathcal{S}_n$ is

$$L_{\mathcal{S}_n}(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}\,.$$

For a *fixed* h, the empirical risk is an average of n i.i.d. Bernoulli variables with mean $L_{\mathcal{D}}(h)$, so the law of large numbers gives $L_{\mathcal{S}_n}(h) \to L_{\mathcal{D}}(h)$ almost surely. The whole difficulty of learning theory is contained in the word *fixed*: the classifier one actually uses has been chosen using the sample, hence is not independent of it, and for such a classifier $L_{\mathcal{S}_n}(\widehat{h}_n)$ may be a catastrophically optimistic estimate of $L_{\mathcal{D}}(\widehat{h}_n)$. The nearest-neighbour rule of Section 1.5 is the extreme example: its empirical risk is 0 on any sample with distinct feature vectors, whatever the problem.

Example 1.8 (Character recognition)**.** The MNIST data set consists of 70 000 grey-level images of handwritten digits, of size 28×28. In the notation above:

domain set $\mathcal{X}$	$[0,1]^{784}$ (or $\{0,\dots,255\}^{784}$),
label set $\mathcal{Y}$	$\{0,1,\dots,9\}$,
distribution $\mathcal{D}$	unknown, and not a member of any natural parametric family,
sample $\mathcal{S}_n$	60 000 labelled images,
risk $L_{\mathcal{D}}(h)$	probability of misreading a new digit.

The dimension $d = 784$ is large, the sample size $n = 6 \cdot 10^4$ is not, and the ratio $n/d \approx 77$ is the kind of number that makes classical asymptotic statistics uncomfortable. This is the regime the book is about.

1.3.3 Two ways of factorising $\mathcal{D}$

A joint law can be sliced in two ways, and both are used constantly.

Assume $\mathcal{X}$ is a Polish space equipped with its Borel σ-field (this covers $\mathbb{R}^d$ and all spaces met in practice), so that regular conditional distributions exist [Dud02, Chap. 10]. Then

- the pair $(\mathcal{D}_X, \kappa)$, where $\mathcal{D}_X(A) = \mathcal{D}(A \times \mathcal{Y})$ is the law of X and $\kappa(\cdot \mid x)$ is a regular version of the conditional law of Y given $X = x$, determines $\mathcal{D}$ by $\mathcal{D}(A \times B) = \int_A \kappa(B \mid x)\, \mathrm{d}\mathcal{D}_X(x)$. This is the *discriminative* factorisation: it describes how the answer is produced from the features;
- the pair $\big(\mathcal{D}_\mathcal{Y}, (\mathcal{D}(\cdot \mid y))_{y \in \mathcal{Y}}\big)$, where $\mathcal{D}_\mathcal{Y}$ is the law of Y and $\mathcal{D}(\cdot \mid y)$ the conditional law of X given $Y = y$, determines $\mathcal{D}$ as well. This is the *generative* factorisation: it describes how each class generates its individuals.

Classification algorithms are traditionally sorted according to which of the two they attempt to estimate. We shall use the first almost exclusively, through the function η introduced next.

1.3.4 General losses, and the regression case

The 0–1 loss of (1.2) is the natural measure of success for classification, but it is neither the only one nor the most convenient.

Definition 1.9 (Loss and risk)**.** A *loss function* is a measurable map $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ with $\ell(y, y) = 0$. The risk of a decision rule $h : \mathcal{X} \to \mathcal{Y}$ is $L_\mathcal{D}(h) := \mathbb{E}\big[\ell\big(h(X), Y\big)\big]$ and its empirical risk is $L_{\mathcal{S}_n}(h) = \frac{1}{n} \sum_i \ell\big(h(X_i), Y_i\big)$.

Taking $\mathcal{Y}$ finite and $\ell(a, b) = \mathbb{1}\{a \neq b\}$ returns Definition 1.2. Taking $\mathcal{Y} = \mathbb{R}$ and $\ell(a, b) = (a - b)^2$ gives *least-squares regression*, for which the analogue of Theorem 1.13 is both simpler and sharper.

Proposition 1.10 (Optimality of the regression function)**.** *Let $\mathcal{Y} = \mathbb{R}$, $\ell(a, b) = (a - b)^2$ and assume $\mathbb{E}[Y^2] < \infty$. Let $\eta(x) = \mathbb{E}[Y \mid X = x]$. Then for every measurable $h : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}[h(X)^2] < \infty$,*

$$L_\mathcal{D}(h) - L_\mathcal{D}(\eta) = \mathbb{E}\Big[\big(h(X) - \eta(X)\big)^2\Big] = \big\|h - \eta\big\|^2_{L^2(\mathcal{D}_X)}. \tag{1.4}$$

In particular η minimises the risk, and the excess risk is the squared L^2-distance to it.

Proof. Write $h(X) - Y = \big(h(X) - \eta(X)\big) + \big(\eta(X) - Y\big)$ and expand. The cross term vanishes:

$$\mathbb{E}\Big[\big(h(X) - \eta(X)\big)\big(\eta(X) - Y\big)\Big] = \mathbb{E}\Big[\big(h(X) - \eta(X)\big)\,\mathbb{E}\big[\eta(X) - Y \mid X\big]\Big] = 0,$$

since $\mathbb{E}[Y \mid X] = \eta(X)$. Hence $L_\mathcal{D}(h) = \mathbb{E}[(h(X) - \eta(X))^2] + L_\mathcal{D}(\eta)$. □

Remark 1.11 (Why classification is harder to analyse than regression). Identity (1.4) says that in least-squares regression the excess risk is a *squared distance*: it is a quadratic, Hilbertian quantity, and the whole apparatus of orthogonal projections applies. Compare with (1.6): in classification the excess risk is $\mathbb{E}\big[\big|2\eta - 1\big|\,\mathbb{1}\{h \neq h^*\}\big]$, which is neither a distance nor a smooth function of h, and which weighs errors by the local noise level. It is this asymmetry that makes classification rates subtler than regression rates, and that makes margin conditions (Remark 1.14) necessary in order to obtain fast rates. Lemma 1.18 is the bridge between the two worlds: it bounds a classification excess risk by an L^1 regression error.

1.4 The Bayes classifier and the noise of a problem

Throughout this section $\mathcal{Y} = \{0, 1\}$. Define the *regression function*

$$\eta(x) := \mathbb{P}(Y = 1 \mid X = x) = \mathbb{E}[Y \mid X = x], \tag{1.5}$$

by which we mean a fixed measurable version of the conditional expectation; all statements below hold for $\mathcal{D}_X$-almost every x and do not depend on the version chosen.

Definition 1.12 (Bayes classifier)**.** The *Bayes classifier* of the problem $\mathcal{D}$ is

$$h^*(x) := \mathbb{1}\left\{\eta(x) \geq \tfrac{1}{2}\right\},$$

and the *Bayes risk* is $L^*_{\mathcal{D}} := L_{\mathcal{D}}(h^*)$.

Theorem 1.13 (Optimality of the Bayes classifier)**.** *For every measurable $h : \mathcal{X} \to \{0,1\}$,*

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\Big[\,\big|2\eta(X) - 1\big|\;\mathbb{1}\{h(X) \neq h^*(X)\}\Big] \geq 0, \tag{1.6}$$

and

$$L^*_{\mathcal{D}} = \mathbb{E}\big[\min\{\eta(X), 1-\eta(X)\}\big]. \tag{1.7}$$

In particular h^* minimises the risk among all classifiers.

Proof. Conditioning on X and using that $h(X)$ is $\sigma(X)$-measurable,

$$\begin{aligned} L_{\mathcal{D}}(h) &= \mathbb{P}\big(h(X) \neq Y\big) = \mathbb{E}\Big[\mathbb{P}\big(Y \neq h(X) \mid X\big)\Big] \\ &= \mathbb{E}\Big[\eta(X)\,\mathbb{1}\{h(X)=0\} + \big(1-\eta(X)\big)\,\mathbb{1}\{h(X)=1\}\Big]. \end{aligned} \tag{1.8}$$

Now observe that for $u \in \{0,1\}$,

$$\eta\,\mathbb{1}\{u=0\} + (1-\eta)\,\mathbb{1}\{u=1\} = \eta + (1-2\eta)\,u,$$

so that (1.8) becomes the *linear* identity

$$L_{\mathcal{D}}(h) = \mathbb{E}\big[\eta(X)\big] + \mathbb{E}\big[\big(1-2\eta(X)\big)h(X)\big]. \tag{1.9}$$

Only the second term depends on h, and it is minimised pointwise by taking $h(x) = 1$ exactly where $1 - 2\eta(x) < 0$; any value is optimal where $\eta(x) = \frac{1}{2}$, and $h^* = \mathbb{1}\{\eta \geq \frac{1}{2}\}$ is one such minimiser. Subtracting,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\Big[\big(1-2\eta(X)\big)\big(h(X) - h^*(X)\big)\Big] = \mathbb{E}\Big[\big|2\eta(X)-1\big|\;\mathbb{1}\{h(X)\neq h^*(X)\}\Big],$$

because the product $(1-2\eta)(h-h^*)$ vanishes when $h = h^*$ and equals $\big|1-2\eta\big|$ otherwise: there $h - h^* = \pm 1$, and $1-2\eta$ carries the same sign by definition of h^*. This proves (1.6), and non-negativity is then obvious. Finally, evaluating (1.9) at h^*,

$$L^*_{\mathcal{D}} = \mathbb{E}\Big[\eta(X) + \big(1-2\eta(X)\big)\mathbb{1}\left\{\eta(X) \geq \tfrac{1}{2}\right\}\Big] = \mathbb{E}\big[\min\{\eta(X), 1-\eta(X)\}\big]. \qquad \square$$

The quantity $L_{\mathcal{D}}(h) - L^*_{\mathcal{D}} \geq 0$ is called the *excess risk* of h. It is the only meaningful measure of the quality of a classifier: the risk itself cannot be made small when the problem is noisy, and $L^*_{\mathcal{D}}$ is the price of that noise. For this reason $L^*_{\mathcal{D}}$ is often called the *noise* of the problem.

Remark 1.14 (What the excess-risk identity really says). Identity (1.6) is more informative than the inequality $L_{\mathcal{D}}(h) \geq L^*_{\mathcal{D}}$ it implies. It says that a mistake of h – that is, a point x where h disagrees with h^* – is penalised by the local amount $\big|2\eta(x)-1\big|$, which measures how far the problem is from a coin toss at x. Disagreeing with the Bayes rule where $\eta(x) \approx \frac{1}{2}$ costs almost nothing: there, even the optimal rule is wrong about half of the time. Consequently, the difficulty of a classification problem is governed not by the noise level alone but by *how much mass $\mathcal{D}_X$ puts near the boundary* $\{\eta = \frac{1}{2}\}$. Hypotheses that quantify this – Massart's condition

$\big|2\eta(X)-1\big| \geq 2\gamma > 0$ almost surely, or Tsybakov's margin condition $\mathbb{P}(\big|2\eta(X)-1\big| \leq s) \leq Cs^{\alpha}$ – are exactly what produces so-called *fast rates* n^{-1} instead of $n^{-1/2}$ in statistical learning [MT99; Tsy04; MN06]; the first of them is Assumption 2.69, and Theorem 2.70 carries out the argument in full. We shall meet the probabilistic counterpart of this phenomenon in the next chapter: it is the difference between Hoeffding's inequality and Bernstein's.

Proposition 1.15 (Multiclass case)**.** *Let $\mathcal{Y} = \{1, \dots, M\}$ and $\eta_y(x) := \mathbb{P}(Y = y \mid X = x)$ for $y \in \mathcal{Y}$. Let*

$$h^*(x) := \min\Big\{ y \in \mathcal{Y} : \ \eta_y(x) = \max_{z \in \mathcal{Y}} \eta_z(x) \Big\}.$$

Then h^* is measurable, and for every measurable $h : \mathcal{X} \to \mathcal{Y}$,

$$L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h^*) = \mathbb{E}\Big[\eta_{h^*(X)}(X) - \eta_{h(X)}(X)\Big] \ \geq \ 0, \qquad L^*_{\mathcal{D}} = \mathbb{E}\Big[1 - \max_{y \in \mathcal{Y}} \eta_y(X)\Big].$$

Proof. Measurability: each η_y is measurable, hence so is $x \mapsto \max_z \eta_z(x)$, and $\{h^* = y\} = \bigcap_{z<y}\{\eta_z < \max_w \eta_w\} \cap \{\eta_y = \max_w \eta_w\}$ is measurable. Conditioning on X as in (1.8),

$$L_{\mathcal{D}}(h) = \mathbb{P}\big(h(X) \neq Y\big) = \mathbb{E}\Big[1 - \mathbb{P}\big(Y = h(X) \mid X\big)\Big] = \mathbb{E}\big[1 - \eta_{h(X)}(X)\big],$$

which gives both identities, the inequality holding pointwise because $h^*(x)$ maximises $y \mapsto \eta_y(x)$. □

Remark 1.16. For $M = 2$ this is Theorem 1.13, since $\eta_{h^*} - \eta_h = \big|2\eta - 1\big| \, \mathbb{1}\{h \neq h^*\}$ when $\eta_1 = \eta$ and $\eta_0 = 1 - \eta$. Note that the multiclass Bayes risk $\mathbb{E}[1 - \max_y \eta_y(X)]$ is at most $1 - 1/M$, and that the measurable-selection device – choosing the smallest maximiser – is genuinely needed: $\arg\max_y \eta_y(x)$ is a set, and an arbitrary choice in it need not be a measurable function of x.

Remark 1.17 (The Bayes rule is not available). h^* depends on $\mathcal{D}$ through η, and $\mathcal{D}$ is unknown; the theorem is a benchmark, not an algorithm. It does suggest one, however: estimate η by some $\hat{\eta}$ and use the *plug-in* rule $h_{\hat{\eta}}(x) = \mathbb{1}\{\hat{\eta}(x) \geq \frac{1}{2}\}$. Lemma 1.18 below bounds the price of this substitution, and the nearest-neighbour rule of the next section is, in disguise, the crudest plug-in rule there is: it estimates $\eta(x)$ by the label of the closest observed point.

Lemma 1.18 (Plug-in rules)**.** *Let $\hat{\eta} : \mathcal{X} \to [0,1]$ be measurable and $h_{\hat{\eta}} = \mathbb{1}\{\hat{\eta} \geq \frac{1}{2}\}$. Then*

$$L_{\mathcal{D}}(h_{\hat{\eta}}) - L^*_{\mathcal{D}} \ \leq \ 2\,\mathbb{E}\big[\,\big|\hat{\eta}(X) - \eta(X)\big|\,\big].$$

Proof. By (1.6) it suffices to show that $\big|2\eta(x) - 1\big| \leq 2\big|\hat{\eta}(x) - \eta(x)\big|$ whenever $h_{\hat{\eta}}(x) \neq h^*(x)$. If $h_{\hat{\eta}}(x) = 1$ and $h^*(x) = 0$ then $\hat{\eta}(x) \geq \frac{1}{2} > \eta(x)$, so $\big|\eta(x) - \frac{1}{2}\big| = \frac{1}{2} - \eta(x) \leq \hat{\eta}(x) - \eta(x) = \big|\hat{\eta}(x) - \eta(x)\big|$; the symmetric case is identical. Multiplying by 2 and taking expectations concludes. □

1.5 The nearest-neighbour classifier

We now analyse completely a learning algorithm. The one we choose is the oldest and the simplest: predict for a new point the label of the closest observed point. It was introduced by Fix and Hodges in 1951 [FH51] and its asymptotic analysis by Cover and Hart [CH67] is one of the founding results of nonparametric classification. Its analysis is instructive for three reasons: it requires no assumption on $\mathcal{D}$ beyond a smoothness condition, it exhibits the curse of dimensionality in the cleanest possible form, and it produces a bound on an *expectation*, thereby raising the question that the rest of this book answers.

Definition 1.19 (Nearest-neighbour rule)**.** Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{S}_n$ a training sample. For $x \in \mathcal{X}$ let

$$I(x) := \min\Big\{ i \in \{1, \dots, n\} : \ \rho(x, X_i) = \min_{1 \le j \le n} \rho(x, X_j) \Big\}$$

(the smallest index among the nearest sample points, a convention that breaks ties). The *nearest-neighbour classifier* is $\widehat{h}_n^{\mathrm{NN}}(x) := Y_{I(x)}$.

The map $(x, x_1, \dots, x_n) \mapsto I(x)$ is measurable, being obtained from the continuous functions $\rho(x, x_i)$ by finitely many comparisons; hence $\widehat{h}_n^{\mathrm{NN}}$ satisfies the measurability requirement of Definition 1.5. We write $\widehat{h}^{\mathrm{NN}}$, without index, for the rule seen as a learning algorithm, so that $R_n(\widehat{h}^{\mathrm{NN}}) = \mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}_n^{\mathrm{NN}})\big]$ is its expected risk (1.3).

Assumption 1.20. We assume:

(A1) $\mathcal{Y} = \{0, 1\}$;

(A2) $\mathcal{X} = [0, 1)^d$, equipped with the Euclidean distance $\rho(x, x') = \|x - x'\|_2$;

(A3) the regression function η is c-Lipschitz: $\big|\eta(x) - \eta(x')\big| \le c\, \rho(x, x')$ for all $x, x' \in \mathcal{X}$.

Theorem 1.21 (Finite-sample risk of the nearest-neighbour rule)**.** *Under Assumption 1.20, for every distribution $\mathcal{D}$ satisfying (A3) and every $n \ge 1$,*

$$R_n(\widehat{h}^{\mathrm{NN}}) \ \le \ 2\, L_{\mathcal{D}}^{*} + \big(2 + e^{-1}\big)\, \frac{c\sqrt{d}}{n^{1/(d+1)}} \ \le \ 2\, L_{\mathcal{D}}^{*} + \frac{3\, c\sqrt{d}}{n^{1/(d+1)}}. \tag{1.10}$$

The proof rests on two lemmas, one probabilistic and one geometric. The first compares two Bernoulli variables, the second controls the distance from a point to the sample.

Lemma 1.22 (Comparison of Bernoulli variables)**.** *Let $p, q \in [0, 1]$ and let $U \sim \mathrm{Ber}(p)$, $V \sim \mathrm{Ber}(q)$ be independent. Then*

$$\mathbb{P}(U \ne V) \ \le \ 2 \min\big\{p, 1 - p\big\} + \big|p - q\big|.$$

Proof. Substituting $q = p + u$ once,

$$\mathbb{P}(U \ne V) = p(1-q) + q(1-p) = p + q - 2pq = p + (p+u) - 2p(p+u) = 2p(1-p) + u(1-2p),$$

so that $\mathbb{P}(U \ne V) \le 2p(1-p) + \big|q - p\big|$ because $\big|1 - 2p\big| \le 1$, and $2p(1-p) \le 2\min\{p, 1-p\}$ because $1 - p \le 1$ and $p \le 1$. □

Lemma 1.23 (Expected distance to the nearest neighbour)**.** *Let $X, X_1, \dots, X_n$ be i.i.d. with values in $[0, 1)^d$ and let $X_{I(X)}$ be the nearest neighbour of X among $X_1, \dots, X_n$. Then*

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \ \le \ \big(2 + e^{-1}\big)\, \sqrt{d}\; n^{-1/(d+1)}.$$

Proof. Fix an integer $T \ge 1$ and partition $[0, 1)^d$ into the T^d cells

$$C = \Big\{ \prod_{k=1}^{d} \Big[\tfrac{j_k - 1}{T}, \tfrac{j_k}{T}\Big) \ : \ 1 \le j_1, \dots, j_d \le T \Big\},$$

each of diameter $\sqrt{d}/T$ (see Figure 1.1). For $C \in \mathcal{C}$ write $p_C := \mathcal{D}_X(C)$, so that $\sum_C p_C = 1$.

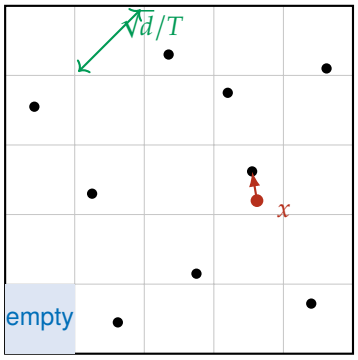


Figure 1.1: The partition used in the proof of Lemma 1.23, here with $d = 2$ and $T = 5$. If the cell of x contains a sample point, the nearest neighbour of x lies within the diameter $\sqrt{d}/T$ of the cell; the cells that contain no sample point – of total probability at most $T^d/(en)$ by (1.11) – are the only ones for which we must resort to the trivial bound $\sqrt{d}$. Refining the grid improves the first term and degrades the second: the optimal trade-off is $T \asymp n^{1/(d+1)}$.

If the cell containing X contains at least one sample point X_i, then $\rho(X, X_{I(X)}) \le \rho(X, X_i) \le \sqrt{d}/T$. Otherwise we use the trivial bound $\rho(X, X_{I(X)}) \le \operatorname{diam}([0,1)^d) = \sqrt{d}$. Hence, denoting by $C(X)$ the cell of X,

$$\rho(X, X_{I(X)}) \;\le\; \frac{\sqrt{d}}{T} + \sqrt{d}\,\mathbb{1}\Big\{\,\forall i \le n,\; X_i \notin C(X)\,\Big\}.$$

Since X is independent of $X_1, \dots, X_n$,

$$\mathbb{P}\big(\forall i \le n,\; X_i \notin C(X)\big) = \sum_{C \in \mathcal{C}} \mathbb{P}\big(X \in C\big)\,\mathbb{P}\big(X_1 \notin C\big)^n = \sum_{C \in \mathcal{C}} p_C\,(1 - p_C)^n.$$

Using $1 - u \le e^{-u}$ and $\max_{u \ge 0} u e^{-nu} = (en)^{-1}$,

$$\sum_{C \in \mathcal{C}} p_C (1 - p_C)^n \;\le\; \sum_{C \in \mathcal{C}} p_C e^{-n p_C} \;\le\; \frac{|\mathcal{C}|}{en} = \frac{T^d}{en}, \tag{1.11}$$

whence

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\le\; \frac{\sqrt{d}}{T} + \frac{\sqrt{d}\,T^d}{e\,n}.$$

It remains to choose T. Take $T = \big\lfloor n^{1/(d+1)} \big\rfloor \ge 1$. Then $T^d \le n^{d/(d+1)}$, so the second term is at most $\sqrt{d}\,n^{-1/(d+1)}/e$; and since $\lfloor u \rfloor \ge u/2$ for every $u \ge 1$, the first term is at most $2\sqrt{d}\,n^{-1/(d+1)}$. Adding the two bounds concludes. □

Proof of Theorem 1.21. Let $(X, Y) \sim \mathcal{D}$ be independent of $\mathcal{S}_n$. By definition,

$$R_n(\widehat{h}^{\mathrm{NN}}) = \mathbb{P}\big(Y \ne Y_{I(X)}\big) = \mathbb{E}\Big[\mathbb{P}\big(Y \ne Y_{I(X)} \;\big|\; X, X_1, \dots, X_n\big)\Big].$$

Condition on $\mathcal{G} := \sigma(X, X_1, \dots, X_n)$. The index $I(X)$ is $\mathcal{G}$-measurable, and conditionally on $\mathcal{G}$ the variables $Y, Y_1, \dots, Y_n$ are independent with $Y \sim \mathrm{Ber}(\eta(X))$ and $Y_i \sim \mathrm{Ber}(\eta(X_i))$; in particular Y and $Y_{I(X)}$ are conditionally independent Bernoulli variables with parameters $\eta(X)$ and $\eta(X_{I(X)})$. Lemma 1.22 gives

$$\mathbb{P}\big(Y \ne Y_{I(X)} \mid \mathcal{G}\big) \;\le\; 2\min\big\{\eta(X), 1 - \eta(X)\big\} + \big|\eta(X) - \eta(X_{I(X)})\big|.$$

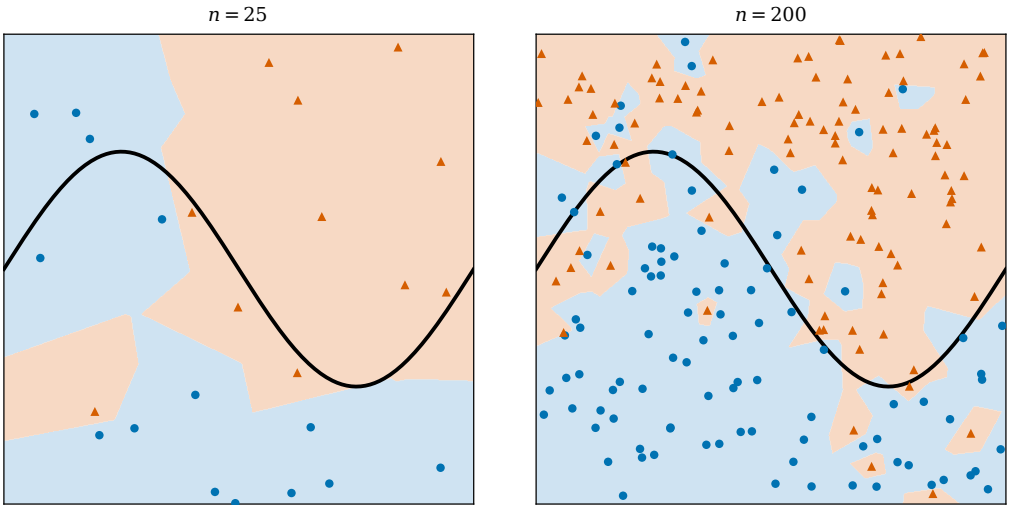


Figure 1.2: The nearest-neighbour rule in dimension $d = 2$, for a problem with a c-Lipschitz regression function η; the black curve is the Bayes boundary $\{\eta = \frac{1}{2}\}$, the shaded regions are the two decision regions of $\widehat{h}_n^{\mathrm{NN}}$, and the markers are the labelled sample points. The decision regions are unions of Voronoi cells, and the rule reproduces the labels of the sample exactly – including its noise, which is the content of Remark 1.71. As n grows, the boundary of the decision region converges to the Bayes boundary, at the speed quantified by Theorem 1.21. Generated by `figures/fig_nn_boundary.py`.

By (A3) the last term is at most $c\rho(X, X_{I(X)})$. Taking expectations and using (1.7) and Lemma 1.23,

$$R_n(\widehat{h}^{\mathrm{NN}}) \;\le\; 2\,\mathbb{E}\big[\min\{\eta(X), 1-\eta(X)\}\big] + c\,\mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\le\; 2L^*_{\mathcal{D}} + \big(2 + e^{-1}\big)\frac{c\sqrt{d}}{n^{1/(d+1)}}. \qquad \square$$

1.5.1 What the analysis says

Theorem 1.21 deserves a long comment; each of the following points is a lesson that recurs throughout the book.

Remark 1.24 (The bound is non-asymptotic and distribution-free). It holds for every $n \ge 1$ and every distribution $\mathcal{D}$ satisfying (A3) – not in the limit, not for a parametric family. The price is that the second term does not depend on $\mathcal{D}$ at all: it is a worst-case term, and it cannot be improved by assuming that the problem at hand is a nice one, unless one adds hypotheses.

Remark 1.25 (Is the factor 2 an artefact?). It is not. Take $\eta \equiv p$ constant with $p < \frac{1}{2}$ (so $c = 0$ and the second term vanishes). Then $L^*_{\mathcal{D}} = p$, whereas the nearest-neighbour rule predicts an independent $\mathrm{Ber}(p)$ label, so its risk is exactly $2p(1-p)$ for every n. As $p \to 0$ the ratio $2p(1-p)/p \to 2$: asymptotically the rule really does lose a factor 2. Cover and Hart [CH67] proved the exact asymptotic statement: under mild conditions the risk of the nearest-neighbour rule converges to a limit L^{NN} satisfying

$$L^*_{\mathcal{D}} \;\le\; L^{\mathrm{NN}} \;\le\; 2\,L^*_{\mathcal{D}}\big(1 - L^*_{\mathcal{D}}\big) \;\le\; 2L^*_{\mathcal{D}},$$

so the constant 2 in (1.10) is not an artefact of the proof. Moreover the example above *attains* the Cover–Hart bound: with $\eta \equiv p$ one has $L^*_{\mathcal{D}} = p$ and nearest-neighbour risk exactly $2p(1-p) = 2L^*_{\mathcal{D}}(1 - L^*_{\mathcal{D}})$, so the inequality is sharp and constant regression functions are its

extremal case. The reason is transparent in Lemma 1.22: we compare the true label with *one* noisy label instead of averaging out the noise. Averaging the labels of the k nearest neighbours with $k \to \infty$ and $k/n \to 0$ removes the factor 2 and yields universal consistency, $L_{\mathcal{D}}(\widehat{h}_n^{k\text{-NN}}) \to L_{\mathcal{D}}^*$ for *every* distribution $\mathcal{D}$; this is Stone's theorem [Sto77], one of the most remarkable results of nonparametric statistics.

Remark 1.26 (The curse of dimensionality). The second term of (1.10) decreases like $n^{-1/(d+1)}$: to make it smaller than ε one needs

$$n \;\geq\; \left(\frac{3c\sqrt{d}}{\varepsilon}\right)^{d+1},$$

a sample size exponential in the dimension. With $c = 1$ and $\varepsilon = 0.1$ this reads $n \geq 7.6 \cdot 10^4$ for $d = 2$, $n \geq 9.1 \cdot 10^{10}$ for $d = 5$, $n \geq 5.6 \cdot 10^{21}$ for $d = 10$ and $n \geq 4.8 \cdot 10^{44}$ for $d = 20$. For the MNIST problem of Example 1.8, where $d = 784$, the bound is vacuous by an astronomical margin. This is not a weakness of the proof. Over the class of all distributions with c-Lipschitz regression function, no learning algorithm can escape a sample size exponential in d: by Theorem 3.78, proved in Chapter 3, every algorithm needs more than $\frac{1}{2}\big(c\sqrt{d/(2\pi e)}\big)^d$ samples to reach an expected risk below $\frac{1}{4}$ uniformly over noiseless c-Lipschitz problems. The nearest-neighbour rule matches this threshold up to the constant in the base (Remark 3.79); see also [DGL96, Chap. 7] and [BD15]. The geometric reason is visible in Lemma 1.23: covering $[0,1]^d$ at resolution $1/T$ requires T^d cells, and a sample can only populate n of them. High-dimensional data are *sparse in space*, always.

That real algorithms nevertheless succeed on MNIST means that the Lipschitz class is not the right class: images live near a low-dimensional subset of $\mathbb{R}^{784}$, and the relevant dimension is not d but some intrinsic dimension. Making this precise is one of the recurring goals of high-dimensional statistics, and one of the motivations for the dimension-reduction results we shall meet later.

Proposition 1.27 (The nearest neighbour really is far, in high dimension)**.** *Let $\mathcal{D}_X$ be the uniform distribution on $[0,1)^d$ and let V_d denote the volume of the Euclidean unit ball of $\mathbb{R}^d$. Then, for every $n \geq 1$,*

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\geq\; \frac{1}{2}\,(nV_d)^{-1/d} \;\geq\; \frac{1}{2}\sqrt{\frac{d}{2\pi e}}\; n^{-1/d}.$$

Proof. For every x and $r > 0$, the Lebesgue measure of $B(x,r) \cap [0,1)^d$ is at most $V_d r^d$, so $\mathbb{P}\big(\rho(x, X_1) < r\big) \leq V_d r^d$ and, by independence and Bernoulli's inequality $(1-u)^n \geq 1 - nu$,

$$\mathbb{P}\big(\rho(x, X_{I(x)}) \geq r\big) = \big(1 - \mathbb{P}(\rho(x, X_1) < r)\big)^n \;\geq\; 1 - nV_d r^d.$$

Hence, for any $r_0 > 0$ with $nV_d r_0^d \leq 1$,

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] = \int_0^\infty \mathbb{P}\big(\rho(X, X_{I(X)}) \geq r\big)\,\mathrm{d}r \;\geq\; \int_0^{r_0} \big(1 - nV_d r^d\big)\,\mathrm{d}r = r_0 - \frac{nV_d r_0^{d+1}}{d+1}.$$

The choice $r_0 = (nV_d)^{-1/d}$ gives $r_0\big(1 - \frac{1}{d+1}\big) = \frac{d}{d+1} r_0 \geq \frac{r_0}{2}$. Finally $V_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ and $\Gamma(x+1) \geq (x/e)^x$, so $V_d^{-1/d} = \Gamma(\frac{d}{2}+1)^{1/d}/\sqrt{\pi} \geq \sqrt{d/(2\pi e)}$. □

Remark 1.28 (Reading the two bounds together). Lemma 1.23 and Proposition 1.27 say that for the uniform distribution

$$\frac{1}{2}\sqrt{\frac{d}{2\pi e}}\; n^{-1/d} \;\leq\; \mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\leq\; \big(2 + e^{-1}\big)\sqrt{d}\; n^{-1/(d+1)}.$$

The upper bound is not tight – the true order is $n^{-1/d}$, and the loss comes from the crude treatment of the empty cells in (1.11), which Remark 1.29 below removes – but the lower bound already contains the essential message, and it is worth stating in words: *to have a sample point within distance ε of a typical point of the unit cube, one needs $n \gtrsim (\sqrt{d}/\varepsilon)^d$ observations*. For $d = 100$ and $\varepsilon = \frac{1}{2}$ this is more than 10^{100}. The diameter of the cube being $\sqrt{d}$, the relative statement is even starker: with any realistic n, the nearest neighbour of a point of $[0,1]^{100}$ is at a distance comparable to the diameter of the whole space. In high dimension, *everything is far from everything*, and any method based on local averaging is in trouble.

Remark 1.29 (Integrating over the scale: the optimal exponent). The gap just described can in fact be closed. The proof of Lemma 1.23 fixes *one* grid and pays $\operatorname{diam}(\mathcal{X}) = \sqrt{d}$ on the event that the cell of X is empty. Integrating the tail over all scales instead – using a different grid for each radius r – removes that loss and produces the exponent $1/d$ of Proposition 1.27.

For every $d \geq 2$, every distribution on $[0,1)^d$ and every $n \geq 1$,

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\leq\; \frac{2d}{d-1}\,\sqrt{d}\,(en)^{-1/d} \;\leq\; 4\sqrt{d}\;n^{-1/d}. \tag{1.12}$$

Indeed, fix $r \in (0, \sqrt{d}\,]$ and partition $[0,1)^d$ into $N(r) = T^d$ cells of side $1/T$ with $T = \lceil \sqrt{d}/r \rceil$, so that each cell has diameter $\sqrt{d}/T \leq r$. If the cell of X contains a sample point then $\rho(X, X_{I(X)}) \leq r$; hence, by (1.11) and $\lceil u \rceil \leq 2u$ for $u \geq 1$,

$$\mathbb{P}\big(\rho(X, X_{I(X)}) > r\big) \;\leq\; \frac{N(r)}{en} \;\leq\; \frac{1}{en}\Big(\frac{2\sqrt{d}}{r}\Big)^d.$$

Since $\rho(X, X_{I(X)}) \leq \sqrt{d}$, integrating and substituting $r = as$ with $a := 2\sqrt{d}\,(en)^{-1/d}$, chosen so that $(2\sqrt{d}/a)^d = en$,

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] = \int_0^{\sqrt{d}} \mathbb{P}\big(\rho(X, X_{I(X)}) > r\big)\,\mathrm{d}r \;\leq\; a\int_0^{\infty} \min\{1, s^{-d}\}\,\mathrm{d}s = a\Big(1 + \frac{1}{d-1}\Big),$$

which is (1.12). Carried into the proof of Theorem 1.21, it gives $R_n(\widehat{h}^{\mathrm{NN}}) \leq 2L^*_{\mathcal{D}} + 4c\sqrt{d}\,n^{-1/d}$ for $d \geq 2$, and the sandwich of Remark 1.28 becomes a genuine two-sided estimate: upper and lower bounds of the same order $\sqrt{d}\,n^{-1/d}$, differing by an absolute factor. For $d = 1$ the integral $\int_1^{\infty} s^{-1}\,\mathrm{d}s$ diverges and the argument gives instead $\frac{2}{en}\big(1 + \log\frac{en}{2}\big) = O(n^{-1}\log n)$.

We keep the simpler one-grid bound of Lemma 1.23 in what follows: it is the argument the figure illustrates, and the exponential dependence on d – the message of Remark 1.26 – is the same for both.

Remark 1.30 (A bound in expectation is not a guarantee). Theorem 1.21 bounds $R_n(\widehat{h}^{\mathrm{NN}}) = \mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n)\big]$, an average over training samples. A practitioner has one sample, and wants to know that *his* classifier is good. Markov's inequality applied to the non-negative random variable $L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n)$ turns the first statement into the second, at a cost: with probability at least $1 - \delta$,

$$L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n) \;\leq\; \frac{1}{\delta}\Big(2L^*_{\mathcal{D}} + 3c\sqrt{d}\,n^{-1/(d+1)}\Big),$$

which is useless for $\delta = 5\%$. What one would like instead is

$$L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n) \;\leq\; R_n(\widehat{h}^{\mathrm{NN}}) + \text{something like } \sqrt{\tfrac{\log(1/\delta)}{n}},$$

a bound whose dependence in δ is logarithmic rather than polynomial. Producing such statements is exactly what concentration inequalities do. The rest of this chapter develops the elementary tools; the nearest-neighbour rule itself will be revisited in Chapter 2 once we have the bounded-difference inequality.

Remark 1.31 (Computational cost). Predicting a single label with the nearest-neighbour rule costs $O(nd)$ operations if implemented naively: the training phase is free and the *prediction* phase is expensive, the opposite of most methods. Data structures such as k-d trees help in small dimension and degrade to linear search as d grows – yet another face of the curse of dimensionality.

1.5.2 Beyond the cube: covering numbers and intrinsic dimension

The only property of $[0,1)^d$ used in Lemma 1.23 is that it can be covered by T^d sets of diameter $\sqrt{d}/T$. Stating the lemma in those terms costs nothing and says considerably more.

Definition 1.32 (Covering number)**.** Let $(\mathcal{X}, \rho)$ be a metric space of finite diameter Δ. For $r > 0$, the *covering number* $N(r)$ is the smallest number of subsets of diameter at most r needed to cover $\mathcal{X}$ (possibly $+\infty$).

Proposition 1.33 (Nearest-neighbour distance in a metric space)**.** *Let $X, X_1, \dots, X_n$ be i.i.d. with values in $(\mathcal{X}, \rho)$, of diameter $\Delta < \infty$. Then for every $r > 0$,*

$$\mathbb{E}\big[\rho\big(X, X_{I(X)}\big)\big] \;\le\; r + \frac{\Delta\, N(r)}{e\, n}.$$

Proof. Let $C_1, \dots, C_{N(r)}$ cover $\mathcal{X}$ with $\operatorname{diam}(C_m) \le r$; replacing C_m by $C_m \setminus (C_1 \cup \cdots \cup C_{m-1})$ we may assume that they are disjoint. The proof of Lemma 1.23 applies verbatim, with $\sqrt{d}/T$ replaced by r, $\sqrt{d}$ by Δ and T^d by $N(r)$; the occupancy bound (1.11) is unchanged. □

Corollary 1.34 (Only the intrinsic dimension matters)**.** *Suppose that $N(r) \le (A/r)^{d_0}$ for all $r \in (0, \Delta]$ and some $A \ge \Delta$ and $d_0 \ge 1$. Then*

$$\mathbb{E}\big[\rho\big(X, X_{I(X)}\big)\big] \;\le\; \big(1 + e^{-1}\big)\Big(\frac{\Delta A^{d_0}}{n}\Big)^{1/(d_0+1)}.$$

Proof. Set $r := \big(\Delta A^{d_0}/n\big)^{1/(d_0+1)}$. If $r \ge \Delta$ the claim is trivial, since $\rho(X, X_{I(X)}) \le \Delta \le r$ almost surely. Otherwise $r \in (0, \Delta)$, the hypothesis on N applies at r, and Proposition 1.33 gives

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \;\le\; r + \frac{\Delta N(r)}{en} \;\le\; r + \frac{\Delta A^{d_0}}{en\, r^{d_0}} = r + \frac{r}{e},$$

because $r^{d_0+1} = \Delta A^{d_0}/n$. □

Remark 1.35 (The optimal exponent again). Integrating over the scale, as in Remark 1.29, improves the exponent here too: if $N(r) \le (A/r)^{d_0}$ for all $r \in (0, \Delta]$ with $d_0 \ge 2$, then

$$\mathbb{E}\big[\rho\big(X, X_{I(X)}\big)\big] \;\le\; \frac{d_0}{d_0 - 1}\, A\, (en)^{-1/d_0} \;\le\; 2A\, n^{-1/d_0}.$$

Indeed, for $r \in (0, \Delta]$ cover $\mathcal{X}$ by $N(r) \le (A/r)^{d_0}$ sets of diameter at most r, disjointified as in Proposition 1.33; if the piece containing X contains a sample point then $\rho(X, X_{I(X)}) \le r$, so that $\mathbb{P}(\rho(X, X_{I(X)}) > r) \le N(r)/(en)$ by (1.11), and

$$\mathbb{E}\big[\rho(X, X_{I(X)})\big] \le \int_0^\infty \min\Big\{1, \frac{1}{en}\Big(\frac{A}{r}\Big)^{d_0}\Big\}\, \mathrm{d}r = a \int_0^\infty \min\{1, s^{-d_0}\}\, \mathrm{d}s = \frac{d_0}{d_0 - 1}\, a$$

with $a = A(en)^{-1/d_0}$.

Remark 1.36 (This is the answer to Remark 1.26). The exponent that governs the nearest-neighbour rule is d_0, the exponent of the covering numbers of the *support of* $\mathcal{D}_X$ – its "intrinsic dimension" – and not the dimension d of the ambient space. For $\mathcal{X} = [0,1]^d$ one has $N(r) \asymp (\sqrt{d}/r)^d$ and $d_0 = d$, which is Lemma 1.23; but if the data lie on a d_0-dimensional submanifold of $\mathbb{R}^d$, or more generally on a set of covering dimension $d_0 \ll d$, then Corollary 1.34 and the proof of Theorem 1.21 give

$$R_n\big(\widehat{h}^{\mathrm{NN}}\big) \;\le\; 2L^*_{\mathcal{D}} + 2c\Big(\frac{\Delta A^{d_0}}{n}\Big)^{1/(d_0+1)},$$

with no reference to d whatsoever. This is the precise form of the statement made informally in Remark 1.26: high-dimensional data are tractable when they are not really high-dimensional. Covering numbers will reappear in a central role when we measure the size of a class of functions rather than of a set of points.

1.6 Limit theorems, and why they are not enough

We leave learning aside for a while and consider the basic object of the theory: the empirical mean

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

of independent, identically distributed real random variables with common law P, mean $\mu = \mathbb{E}[X_1]$ and variance $\sigma^2 = \mathbb{V}\mathrm{ar}(X_1)$, assumed finite when we use it. The two classical theorems describe the behaviour of $\bar{X}_n$ for large n; our purpose here is to delimit precisely what they do *not* say.

Theorem 1.37 (Laws of large numbers)**.** *If* $\mathbb{E}|X_1| < \infty$ *then* $\bar{X}_n \to \mu$ *almost surely, and in particular in probability: for every* $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\big(\big|\bar{X}_n - \mu\big| > \varepsilon\big) = 0$.

Theorem 1.38 (Central limit theorem)**.** *If* $\sigma^2 < \infty$ *then* $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ *converges in distribution to* $\mathcal{N}(0,1)$*: for all* $a < b$,

$$\lim_{n\to\infty} \mathbb{P}\Big(a \le \tfrac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \le b\Big) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,\mathrm{d}u.$$

Proofs are in any textbook, e.g. [Dur19, Chap. 2 and 3]. Write $\beta(n,\varepsilon) := \mathbb{P}(\big|\bar{X}_n - \mu\big| > \varepsilon)$ for the quantity we wish to understand. The law of large numbers states that $\beta(n,\varepsilon) \to 0$ for each fixed ε; it is a qualitative, asymptotic statement, and it answers none of the three questions a user asks:

(i) given n and a confidence level δ, how accurate is $\bar{X}_n$, i.e. what is the smallest ε with $\beta(n,\varepsilon) \le \delta$?

(ii) given ε and δ, how many samples are needed, i.e. what is the smallest n with $\beta(n,\varepsilon) \le \delta$?

(iii) what is the order of magnitude of $\beta(n,\varepsilon)$?

The central limit theorem *suggests* an answer to (iii). Recall the classical estimates for the Gaussian tail, which we shall use repeatedly.

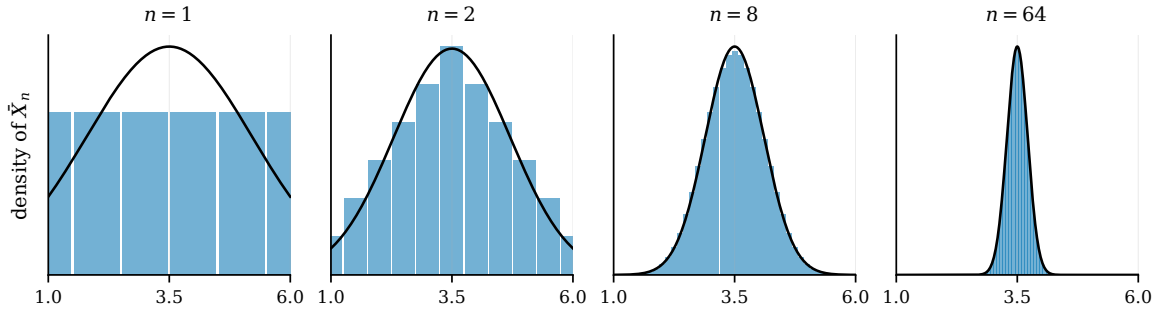


Figure 1.3: The exact law of the empirical mean $\bar{X}_n$ of n fair dice (bars, rescaled as densities) and the Gaussian density with the same mean 3.5 and variance $\frac{35}{12n}$ (solid line), for $n \in \{1, 2, 8, 64\}$. The convergence of Theorem 1.38 is fast in the bulk – already excellent at $n = 8$ – and this is exactly what Remark 1.40 warns against: the picture says nothing about the *tails*, which is where the events of interest live. Generated by `figures/fig_clt_histograms.py`.

Lemma 1.39 (Gaussian tail)**.** *Let $Z \sim \mathcal{N}(0,1)$ and $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$. For every $t > 0$,*

$$\Big(\frac{1}{t} - \frac{1}{t^3}\Big)\varphi(t) \;\le\; \mathbb{P}(Z \ge t) \;\le\; \frac{\varphi(t)}{t}, \qquad \text{and} \qquad \mathbb{P}(Z \ge t) \;\le\; \tfrac{1}{2}e^{-t^2/2}. \tag{1.13}$$

Proof. For the upper bound in the first display, $u \ge t$ implies $1 \le u/t$, so

$$\mathbb{P}(Z \ge t) = \int_t^\infty \varphi(u)\,\mathrm{d}u \le \int_t^\infty \frac{u}{t}\varphi(u)\,\mathrm{d}u = \frac{1}{t}\big[-\varphi(u)\big]_t^\infty = \frac{\varphi(t)}{t},$$

where we used $\varphi'(u) = -u\varphi(u)$. For the lower bound, set $g(u) = (u^{-1} - u^{-3})\varphi(u)$; a direct computation using $\varphi' = -u\varphi$ gives $g'(u) = (3u^{-4} - 1)\varphi(u)$, hence

$$g(t) = -\int_t^\infty g'(u)\,\mathrm{d}u = \int_t^\infty \big(1 - 3u^{-4}\big)\varphi(u)\,\mathrm{d}u \;\le\; \int_t^\infty \varphi(u)\,\mathrm{d}u = \mathbb{P}(Z \ge t).$$

For the last bound, let $f(t) = \frac{1}{2}e^{-t^2/2} - \mathbb{P}(Z \ge t)$. Then $f(0) = 0$, $f(+\infty) = 0$ and $f'(t) = e^{-t^2/2}\big(\frac{1}{\sqrt{2\pi}} - \frac{t}{2}\big)$, which is positive on $[0, 2/\sqrt{2\pi})$ and negative afterwards. Thus f increases from 0 and then decreases to 0, so $f \ge 0$ on $[0, \infty)$. □

Combining Theorem 1.38 with Lemma 1.39 yields the heuristic

$$\beta(n,t) = \mathbb{P}\big(\big|\bar{X}_n - \mu\big| > t\big) \;\approx\; 2\,\mathbb{P}\Big(Z \ge \frac{t\sqrt{n}}{\sigma}\Big) \;\le\; \exp\Big(-\frac{nt^2}{2\sigma^2}\Big), \tag{1.14}$$

an exponential decay in n, governed by the variance. This is the shape of every result in this book. It is, however, only a heuristic, and for two independent reasons.

Remark 1.40 (The central limit theorem says nothing about tails). The convergence in Theorem 1.38 is a convergence of distribution functions, and the best general control on its speed is the Berry–Esseen theorem [Ber41; Ess42]: if $M_3 = \mathbb{E}\,\big|X_1 - \mu\big|^3 < \infty$, then

$$\sup_{x \in \mathbb{R}} \Big|\mathbb{P}\Big(\tfrac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \le x\Big) - \Phi(x)\Big| \;\le\; \frac{C M_3}{\sigma^3\sqrt{n}}, \tag{1.15}$$

where Φ is the standard Gaussian distribution function, with $C \le 0.4690$ [She11], and the rate $n^{-1/2}$ cannot be improved in general. Now compare the two sides of (1.14). With

$\sigma = 1$, $t = 0.1$ and $n = 10^4$, the Gaussian approximation of the right-hand side is about $e^{-50} \approx 2 \cdot 10^{-22}$, whereas the Berry–Esseen error is of order $0.47\, M_3 \cdot 10^{-2}$. The error term is larger than the quantity being estimated by twenty orders of magnitude: at the scale of tail probabilities, (1.15) is *vacuous*. The central limit theorem describes the bulk of the distribution of $\bar{X}_n$, not its tails, and no amount of care in its proof will change that. Refined statements do exist – moderate deviation principles, Cramér's series – but they require exponential moments, which is precisely the hypothesis under which the elementary methods of this chapter work anyway.

Counterexample 1.41 (Finite variance does not imply exponential tails)**.** Let $X_1, X_2, \dots$ be i.i.d. with $\mathbb{P}(X_1 > x) = x^{-4}$ for $x \geq 1$; then $\mu = \mathbb{E}X_1 = 4/3$ and $\sigma^2 < \infty$, so (1.14) applies in the limit sense. Yet for any fixed $t > 0$ and any n,

$$\mathbb{P}\big(\bar{X}_n - \mu > t\big) \;\geq\; \mathbb{P}\Big(X_1 > \mu + nt\Big) \cdot \mathbb{P}\Big(\sum_{i=2}^{n} (X_i - \mu) \geq 0\Big) \;\geq\; \frac{1}{3}\,\big(\mu + nt\big)^{-4}$$

for n large enough, since the second factor converges to $\frac{1}{2}$ by Theorem 1.38. The deviation probability decreases *polynomially* in n, not exponentially: the Gaussian approximation (1.14) is wrong by an arbitrarily large factor in the tails. The mechanism is the one that governs all heavy-tailed sums: a large value of the mean is caused by a single large summand, not by a conspiracy of all of them. Exponential inequalities therefore *must* require more than two moments; boundedness, or the existence of a moment generating function, will be our hypotheses.

The programme for the rest of the chapter is now clear. We want statements of the form

$$\mathbb{P}\big(\bar{X}_n - \mu \geq t\big) \;\leq\; \text{explicit function of } (n, t),$$

valid for every n, under assumptions we can check, and as close as possible to the Gaussian benchmark (1.14).

1.7 Markov's inequality and its descendants

Everything begins with one line.

Theorem 1.42 (Markov's inequality)**.** *Let Z be a non-negative random variable. For every $t > 0$,*

$$\mathbb{P}(Z \geq t) \;\leq\; \frac{\mathbb{E}[Z]}{t}. \tag{1.16}$$

Proof. Since $Z \geq 0$ we have $t\,\mathbb{1}\{Z \geq t\} \leq Z$ pointwise; take expectations. □

Remark 1.43 (Attribution). The inequality is due to Bienaymé [Bie53] and Chebyshev [Tch67]; Markov was Chebyshev's student, and popularised it. The literature is inconsistent, and one finds (1.16) called the Bienaymé–Chebyshev, Chebyshev or Markov inequality depending on the country and the decade. We follow the dominant modern usage: (1.16) is Markov's inequality, and its second-moment corollary below is Chebyshev's.

Remark 1.44 (Markov's inequality is sharp, and that is the problem). Equality holds in (1.16) if and only if $Z \in \{0, t\}$ almost surely. Thus (1.16) cannot be improved without further assumptions – and the extremal variable shows what those assumptions must exclude: a variable that is 0 most of the time and huge occasionally. The decay in $1/t$ is *polynomial*, hopelessly far from the exponential decay (1.14) we are aiming at. All the inequalities of this chapter are obtained by applying (1.16) not to Z itself, but to a well-chosen non-decreasing transformation of Z; the whole art is in the choice.

Proposition 1.45 (Transfer principle)**.** *Let Z be a real random variable.*

(i) If $f : \mathbb{R} \to [0,\infty)$ is non-decreasing, then for every t with $f(t) > 0$,

$$\mathbb{P}(Z \geq t) \;\leq\; \frac{\mathbb{E}[f(Z)]}{f(t)}. \tag{1.17}$$

(ii) If $f : [0,\infty) \to [0,\infty)$ is non-decreasing, then for every $t > 0$ with $f(t) > 0$,

$$\mathbb{P}\big(|Z| \geq t\big) \;\leq\; \frac{\mathbb{E}\big[f(|Z|)\big]}{f(t)}. \tag{1.18}$$

Proof. In both cases the event on the left is contained in $\{f(Z) \geq f(t)\}$, respectively $\{f(|Z|) \geq f(t)\}$, because f is non-decreasing; apply Theorem 1.42 to the non-negative variable $f(Z)$, respectively $f(|Z|)$. □

Three choices of f generate the classical theory. Applied to $|Z - \mathbb{E}Z|$, version (ii) with $f(z) = z^2$ gives Chebyshev's inequality and with $f(z) = z^q$ the moment method; applied to Z itself, version (i) with $f(z) = e^{\lambda z}$ gives the Chernoff method. Only the exponential is admissible in (i) – $z \mapsto (z-\mu)^2$ and $z \mapsto |z-\mu|^q$ are not monotone on $\mathbb{R}$ – and that is precisely why it controls one tail at a time, whereas the first two control $|Z - \mathbb{E}Z|$ and therefore both tails at once.

Corollary 1.46 (Chebyshev's inequality)**.** *If $\mathbb{V}\mathrm{ar}(Z) < \infty$ then for every $t > 0$,*

$$\mathbb{P}\big(|Z - \mathbb{E}Z| \geq t\big) \;\leq\; \frac{\mathbb{V}\mathrm{ar}(Z)}{t^2}. \tag{1.19}$$

Proof. Apply Theorem 1.42 to the non-negative variable $(Z-\mathbb{E}Z)^2$ and the level t^2: $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) = \mathbb{P}((Z - \mathbb{E}Z)^2 \geq t^2) \leq \mathbb{E}[(Z - \mathbb{E}Z)^2]/t^2$. □

Applied to the empirical mean, whose variance is σ^2/n, Chebyshev's inequality gives the first quantitative version of the law of large numbers:

$$\mathbb{P}\big(\big|\bar{X}_n - \mu\big| \geq t\big) \;\leq\; \frac{\sigma^2}{nt^2}, \qquad \text{hence} \qquad n \;\geq\; \frac{\sigma^2}{\delta\,\varepsilon^2} \;\Longrightarrow\; \beta(n,\varepsilon) \leq \delta. \tag{1.20}$$

This is a genuine, non-asymptotic, distribution-free answer to questions (i)–(iii) above, and it is often all one needs. But observe the dependence on the confidence level: to gain one digit of confidence one must multiply the sample size by 10. With $\sigma = 1$, $\varepsilon = 10^{-2}$ and $\delta = 10^{-3}$, (1.20) requires $n \geq 10^7$, whereas the Gaussian heuristic (1.14) suggests that $n \approx 2\log(1/\delta)/\varepsilon^2 \approx 1.4 \cdot 10^5$ suffices – a factor of 72. The entire purpose of exponential inequalities is to replace the factor $1/\delta$ by $\log(1/\delta)$.

Proposition 1.47 (Chebyshev–Cantelli, one-sided)**.** *Let Z have mean μ and finite variance σ^2. For every $t > 0$,*

$$\mathbb{P}\big(Z - \mu \geq t\big) \;\leq\; \frac{\sigma^2}{\sigma^2 + t^2}. \tag{1.21}$$

Proof. Let $a > 0$. Since $\{Z - \mu \geq t\} \subseteq \{Z - \mu + a \geq t + a\}$ and $t + a > 0$, Proposition 1.45 with $f(z) = (z+a)^2\mathbb{1}\{z + a \geq 0\}$, or simply Markov's inequality applied to $(Z - \mu + a)^2$, gives

$$\mathbb{P}(Z - \mu \geq t) \;\leq\; \frac{\mathbb{E}\big[(Z - \mu + a)^2\big]}{(t+a)^2} = \frac{\sigma^2 + a^2}{(t+a)^2}.$$

The right-hand side is minimal at $a = \sigma^2/t$, where it equals $\sigma^2/(\sigma^2 + t^2)$. □

Remark 1.48. Cantelli's bound (1.21) is better than the one-sided use of (1.19), and it is sharp: the two-valued variable taking the value $\mu + t$ with probability $\sigma^2/(\sigma^2 + t^2)$ and the value $\mu - \sigma^2/t$ otherwise achieves equality (Exercise 1.2). Note the gain: for $t \gg \sigma$ the bound is $\approx \sigma^2/t^2$, the same order as Chebyshev's, but the factor 2 coming from the union of the two tails has disappeared.

Remark 1.49 (The moment method). Taking $f(z) = z^q$ in version (ii) of Proposition 1.45, applied to $|Z - \mu|$, yields for every $q \geq 1$

$$\mathbb{P}(|Z - \mu| \geq t) \;\leq\; \frac{\mathbb{E}\,|Z - \mu|^q}{t^q}, \qquad \text{hence} \qquad \mathbb{P}(|Z - \mu| \geq t) \;\leq\; \inf_{q \geq 1} \frac{\mathbb{E}\,|Z - \mu|^q}{t^q}.$$

Optimising over q is a genuine alternative to the exponential method of the next section, and a strictly more powerful one: it can be shown that the optimised moment bound is always at least as good as the optimised Chernoff bound [BLM13, Sec. 2.2], and it applies to variables with no exponential moment at all, for which the Chernoff bound is vacuous (Exercise 1.4). It is also, in general, much harder to compute with – which is why the exponential method dominates practice.

1.7.1 The laws of large numbers, proved

Chebyshev's inequality already delivers the weak law of large numbers, and a fourth moment, through Borel–Cantelli, delivers the strong one.

Corollary 1.50 (Weak law of large numbers)**.** *If* $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2 < \infty$ *then* $\bar{X}_n \to \mu$ *in probability, and quantitatively* $\mathbb{P}(|\bar{X}_n - \mu| \geq t) \leq \sigma^2/(nt^2)$.

Proof. This is (1.20), which tends to 0 for each fixed $t > 0$. □

Proposition 1.51 (Strong law of large numbers under a fourth moment)**.** *If* $\mathbb{E}[X_1^4] < \infty$ *then* $\bar{X}_n \to \mu$ *almost surely.*

Proof. Set $W_i = X_i - \mu$, so that the W_i are i.i.d., centred, with $\mathbb{E}[W_1^4] < \infty$ and $\sigma^2 = \mathbb{E}[W_1^2]$. Expanding $\left(\sum_{i \leq n} W_i\right)^4$ and using independence and $\mathbb{E}[W_i] = 0$, the only surviving terms are those in which every index appears at least twice:

$$\mathbb{E}\Big[\Big(\sum_{i=1}^{n} W_i\Big)^4\Big] = n\,\mathbb{E}[W_1^4] + 3n(n-1)\sigma^4 \;\leq\; Cn^2, \qquad C := \mathbb{E}[W_1^4] + 3\sigma^4.$$

By Markov's inequality applied to the fourth power, for every $t > 0$,

$$\mathbb{P}\big(|\bar{X}_n - \mu| \geq t\big) \leq \frac{\mathbb{E}\big[(\sum_i W_i)^4\big]}{n^4 t^4} \leq \frac{C}{n^2 t^4},$$

which is summable in n. By Borel–Cantelli, for each fixed $t > 0$ almost surely $|\bar{X}_n - \mu| < t$ for all large n; intersecting the countably many almost-sure events obtained with $t = 1/k$, $k \geq 1$, gives $\bar{X}_n \to \mu$ almost surely. □

Remark 1.52. Kolmogorov's theorem asserts the same conclusion under the optimal hypothesis $\mathbb{E}|X_1| < \infty$; the proof is harder, and uses a maximal inequality. The elementary argument above is worth remembering because it displays, in miniature, the mechanism we shall use repeatedly: *a summable sequence of deviation bounds yields an almost-sure statement*. Section 1.10.2 carries it out with an exponential bound instead of a fourth moment, and gets much more.

1.7.2 The other direction: anti-concentration

All the inequalities of this chapter are upper bounds on the probability of a deviation. Lower bounds – statements that a random variable is *not* too concentrated – are equally important: they tell us when an upper bound cannot be improved, and they are the engine of the "second moment method" in probabilistic combinatorics. The most useful elementary one is the following.

Proposition 1.53 (Paley–Zygmund inequality)**.** *Let $Z \geq 0$ with $0 < \mathbb{E}[Z^2] < \infty$. For every $\theta \in [0,1]$,*

$$\mathbb{P}\big(Z > \theta\,\mathbb{E}Z\big) \;\geq\; (1-\theta)^2\,\frac{\big(\mathbb{E}Z\big)^2}{\mathbb{E}[Z^2]}.$$

Proof. Split the expectation according to the event $A = \{Z > \theta\mathbb{E}Z\}$:

$$\mathbb{E}[Z] = \mathbb{E}\big[Z\mathbb{1}_A\big] + \mathbb{E}\big[Z\mathbb{1}_{A^c}\big] \leq \mathbb{E}\big[Z\mathbb{1}_A\big] + \theta\,\mathbb{E}[Z],$$

so $\mathbb{E}[Z\mathbb{1}_A] \geq (1-\theta)\mathbb{E}[Z]$. By the Cauchy–Schwarz inequality, $\mathbb{E}[Z\mathbb{1}_A]^2 \leq \mathbb{E}[Z^2]\,\mathbb{P}(A)$, and combining the two gives the claim. □

Corollary 1.54 (Second moment method)**.** *If $Z \geq 0$ then $\mathbb{P}(Z > 0) \geq (\mathbb{E}Z)^2/\mathbb{E}[Z^2]$; in particular, if $\mathbb{V}\mathrm{ar}(Z) = o\big((\mathbb{E}Z)^2\big)$ along a sequence, then $\mathbb{P}(Z > 0) \to 1$.*

Proof. Take $\theta = 0$ in Proposition 1.53, and note that $(\mathbb{E}Z)^2/\mathbb{E}[Z^2] = 1/\big(1 + \mathbb{V}\mathrm{ar}(Z)/(\mathbb{E}Z)^2\big)$. □

Remark 1.55. Corollary 1.54 is the standard tool for proving that a random combinatorial structure exists: one counts the expected number Z of objects of the desired type, checks that the second moment is comparable to the square of the first, and concludes that $Z > 0$ with high probability. We shall meet sharper anti-concentration statements later; for now, note the contrast of philosophies. Concentration says that a random variable is close to its mean and is proved by exponentiating; anti-concentration says that it is not too close and is proved by second moments. The two are not symmetric, and the second is usually the harder. Lower bounds at the exponential scale of large deviations require a different tool, the change of measure, and are the subject of Chapter 3; the table of Section 3.4.4 lines up the lower bounds on deviation probabilities of the book, from Paley–Zygmund to Bahadur–Rao.

1.8 The Chernoff method

Definition 1.56 (Moment and cumulant generating functions)**.** Let Z be a real random variable. Its *moment generating function* is

$$F_Z(\lambda) := \mathbb{E}\big[e^{\lambda Z}\big] \in (0, +\infty], \qquad \lambda \in \mathbb{R},$$

its *cumulant generating function* is $K_Z(\lambda) := \log F_Z(\lambda)$, and, when $\mathbb{E}Z$ exists, its *centred cumulant generating function* is

$$\psi_Z(\lambda) := K_{Z-\mathbb{E}Z}(\lambda) = \log \mathbb{E}\big[e^{\lambda(Z-\mathbb{E}Z)}\big] = K_Z(\lambda) - \lambda\mathbb{E}Z.$$

Theorem 1.57 (Chernoff's bound)**.** *For every real random variable Z and every $t \in \mathbb{R}$,*

$$\mathbb{P}(Z \geq t) \;\leq\; \inf_{\lambda > 0} e^{-\lambda t}\,\mathbb{E}\big[e^{\lambda Z}\big] = \exp\Big(-\sup_{\lambda > 0}\big\{\lambda t - K_Z(\lambda)\big\}\Big). \tag{1.22}$$

Equivalently, for $t \geq 0$, $\mathbb{P}(Z \geq \mathbb{E}Z + t) \leq \exp\big(-\sup_{\lambda>0}\{\lambda t - \psi_Z(\lambda)\}\big)$.

Proof. Fix $\lambda > 0$. The map $z \mapsto e^{\lambda z}$ is non-negative and increasing, so Proposition 1.45 gives $\mathbb{P}(Z \geq t) \leq \mathbb{E}[e^{\lambda Z}]/e^{\lambda t}$. Take the infimum over $\lambda > 0$ and rewrite it as an exponential. □

Three comments explain why this one-line consequence of Markov's inequality is the engine of the whole theory.

Remark 1.58 (Why the exponential). Among all the transfer functions f available in Proposition 1.45, the exponential is the only one that turns a *sum* into a *product*. If $Z_1, \dots, Z_n$ are independent and $S_n = \sum_i Z_i$, then

$$\mathbb{E}\big[e^{\lambda S_n}\big] = \prod_{i=1}^{n} \mathbb{E}\big[e^{\lambda Z_i}\big], \qquad \text{that is} \qquad \psi_{S_n}(\lambda) = \sum_{i=1}^{n} \psi_{Z_i}(\lambda). \tag{1.23}$$

The cumulant generating function is *additive* over independent summands. Independence, which is a statement about products of expectations, meets the exponential, which converts sums into products: this is the entire reason why exponential inequalities for sums of independent variables are easy, and why extending them to non-linear functions of independent variables (Chapter 2 and beyond) is the real work.

Remark 1.59 (The bound is void below the mean). If $t < \mathbb{E}Z$ the right-hand side of (1.22) is at least 1. Indeed Jensen's inequality gives $\mathbb{E}[e^{\lambda Z}] \geq e^{\lambda \mathbb{E}Z} > e^{\lambda t}$ for every $\lambda > 0$. The Chernoff method is a *deviation* method: it says nothing about the bulk, and everything about the tails.

Example 1.60 (The Gaussian case, and what Chernoff costs)**.** Let $Z \sim \mathcal{N}(\mu, \sigma^2)$. A direct computation gives $K_Z(\lambda) = \lambda\mu + \lambda^2\sigma^2/2$, hence $\psi_Z(\lambda) = \lambda^2\sigma^2/2$ and

$$\sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2 \sigma^2}{2} \right\} = \frac{t^2}{2\sigma^2} \qquad \text{(attained at } \lambda = t/\sigma^2\text{)},$$

so that $\mathbb{P}(Z \geq \mu + t) \leq e^{-t^2/2\sigma^2}$ for every $t > 0$. Compare with the exact behaviour given by Lemma 1.39: for $t \geq 2\sigma$,

$$\left(\frac{\sigma}{t} - \frac{\sigma^3}{t^3}\right) \frac{e^{-t^2/2\sigma^2}}{\sqrt{2\pi}} \;\leq\; \mathbb{P}(Z \geq \mu + t) \;\leq\; e^{-t^2/2\sigma^2}.$$

The Chernoff bound gets the exponential rate *exactly* right and loses a factor of order σ/t. This is the general picture, and Chapter 2 will explain why: the exponential rate produced by the Chernoff method is the true large-deviation rate (the matching lower bound is proved in Chapter 3), and the polynomial factor it discards is the object of the Bahadur–Rao refinement.

Remark 1.61 (History). The technique appears in Bernstein's work in the 1920s and in Cramér's 1938 memoir [Cra38], and in a different guise (the "Esscher transform") in actuarial mathematics. It became standard after Chernoff's 1952 paper [Che52] on the asymptotic efficiency of tests, whence the name. It is worth recording that Chernoff himself did not claim it: in a 1996 conversation [Bat96] he attributes the bounding argument to his colleague Herman Rubin and expresses mild embarrassment at the attribution – "I don't think it's fair to call it the Chernoff bound". Stigler's law of eponymy is well represented in this chapter: see also Remark 1.43.

1.9 Hoeffding's lemma and Hoeffding's inequality

To use Theorem 1.57 one needs an upper bound on ψ_Z. The simplest hypothesis under which such a bound is available – and the most useful, since losses and indicator functions are bounded by construction – is boundedness.

Lemma 1.62 (Hoeffding's lemma)**.** *Let Z be a random variable with $\mathbb{E}Z = 0$ and $a \le Z \le b$ almost surely. Then*

$$\psi_Z(\lambda) = \log \mathbb{E}\big[e^{\lambda Z}\big] \;\le\; \frac{\lambda^2 (b-a)^2}{8}, \qquad \text{for all } \lambda \in \mathbb{R}. \tag{1.24}$$

We give two proofs. The first is the natural one, and introduces – in the simplest possible case – the change of measure that will structure the whole of Chapter 2. The second is shorter and purely convexity-based.

First proof: exponential tilting. Since $|Z| \le M := \max(|a|, |b|)$ almost surely, $e^{\lambda Z}$ is bounded by $e^{|\lambda| M}$, so $F_Z(\lambda) = \mathbb{E}e^{\lambda Z}$ is finite for every λ and, by dominated convergence, infinitely differentiable with derivatives obtained by differentiating under the expectation. Set $\psi = \psi_Z = \log F_Z$, well defined since $F_Z > 0$. For $\lambda \in \mathbb{R}$ define a new probability measure $\mathbb{P}_\lambda$ on the underlying space by

$$\frac{\mathrm{d}\mathbb{P}_\lambda}{\mathrm{d}\mathbb{P}} := \frac{e^{\lambda Z}}{\mathbb{E}[e^{\lambda Z}]}, \tag{1.25}$$

the *exponentially tilted* measure. Writing $\mathbb{E}_\lambda$ and $\mathbb{V}\mathrm{ar}_\lambda$ for expectation and variance under $\mathbb{P}_\lambda$, we obtain

$$\psi'(\lambda) = \frac{\mathbb{E}[Z e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} = \mathbb{E}_\lambda[Z], \qquad \psi''(\lambda) = \frac{\mathbb{E}[Z^2 e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \left(\frac{\mathbb{E}[Z e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}\right)^2 = \mathbb{V}\mathrm{ar}_\lambda(Z).$$

Since $\mathbb{P}_\lambda$ is absolutely continuous with respect to $\mathbb{P}$, we still have $a \le Z \le b$ $\mathbb{P}_\lambda$-almost surely, and therefore

$$\mathbb{V}\mathrm{ar}_\lambda(Z) \;\le\; \mathbb{E}_\lambda\left[\left(Z - \frac{a+b}{2}\right)^2\right] \;\le\; \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{4},$$

the first inequality because the variance minimises the quadratic deviation about any point, the second because $|Z - (a+b)/2| \le (b-a)/2$ pointwise. (This bound on the variance of a bounded variable is Popoviciu's inequality [Pop35].) Finally $\psi(0) = 0$ and $\psi'(0) = \mathbb{E}Z = 0$, so Taylor's formula with Lagrange remainder gives, for some λ_0 between 0 and λ,

$$\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\lambda_0) = \frac{\lambda^2}{2}\,\mathbb{V}\mathrm{ar}_{\lambda_0}(Z) \;\le\; \frac{\lambda^2 (b-a)^2}{8}. \qquad \square$$

Second proof: reduction to a Bernoulli variable. We may assume $a < b$, and $a \le 0 \le b$ since $\mathbb{E}Z = 0$. Let $p := \frac{-a}{b-a} \in [0,1]$, let $B \sim \mathrm{Ber}(p)$ and put $W := a + (b-a)B$, the two-point variable on $\{a, b\}$ with $\mathbb{E}W = 0$. By convexity of $z \mapsto e^{\lambda z}$ on $[a, b]$,

$$e^{\lambda z} \;\le\; \frac{b-z}{b-a}\, e^{\lambda a} + \frac{z-a}{b-a}\, e^{\lambda b} \qquad (a \le z \le b),$$

and taking expectations with $\mathbb{E}Z = 0$ gives $\mathbb{E}[e^{\lambda Z}] \le \mathbb{E}[e^{\lambda W}]$, that is $\psi_Z \le \psi_W$: among centred variables with values in $[a, b]$, the two-point law is extremal.

It remains to bound ψ_W, and here the first proof applies again. Since $W = a + (b-a)B$ we have $\psi_W(\lambda) = \psi_{\mathrm{Ber}(p)}\big(\lambda(b-a)\big)$, where

$$\psi_{\mathrm{Ber}(p)}(u) = -pu + \log\big(1 - p + pe^u\big) \tag{1.26}$$

is the centred cumulant generating function of a Bernoulli variable. The first proof, applied to B, gives $\psi''_{\mathrm{Ber}(p)}(u) = \mathbb{V}\mathrm{ar}_u(B) = p_u(1-p_u) \le \frac{1}{4}$, where $p_u = pe^u/(1-p+pe^u)$ is the parameter of the u-tilted Bernoulli law; and $\psi_{\mathrm{Ber}(p)}(0) = \psi'_{\mathrm{Ber}(p)}(0) = 0$. Taylor's formula therefore gives $\psi_{\mathrm{Ber}(p)}(u) \le u^2/8$ for every $u \in \mathbb{R}$ and every $p \in [0,1]$, whence

$$\psi_Z(\lambda) \;\le\; \psi_{\mathrm{Ber}(p)}\big(\lambda(b-a)\big) \;\le\; \frac{\lambda^2 (b-a)^2}{8}. \qquad \square$$

The two proofs are one. The first bounds the variance of the tilted law of Z; the second first replaces Z by the extremal two-point law and then bounds the variance of *its* tilted law. In both, the constant $\frac{1}{8}$ is the worst-case variance $(b-a)^2/4$, halved by Taylor's formula.

Remark 1.63 (The constant $1/8$ cannot be improved). Fix $w > 0$ and let Z take the values $\pm w/2$ with probability $1/2$ each, so that $\mathbb{E}Z = 0$, $a \le Z \le b$ with $a = -w/2$, $b = w/2$, $b - a = w$, and $\mathbb{V}\mathrm{ar}(Z) = w^2/4$. As $\lambda \to 0$,

$$\psi_Z(\lambda) = \frac{\lambda^2}{2}\,\mathbb{V}\mathrm{ar}(Z) + O(\lambda^4) = \frac{\lambda^2 w^2}{8} + O(\lambda^4),$$

so no inequality $\psi_Z(\lambda) \le C\,\lambda^2(b-a)^2$ valid for all bounded Z and all λ can hold with $C < 1/8$. The first proof explains where the constant comes from: it is *exactly* the worst-case variance of a variable confined to an interval of length $b - a$. Lemma 1.62 is therefore best possible *as a statement about bounded variables* – and, for that very reason, wasteful for any particular variable whose variance is smaller than the worst case. Repairing this is the subject of Chapter 2.

Theorem 1.64 (Hoeffding's inequality). *Let $Z_1, \dots, Z_n$ be independent random variables with $a_i \le Z_i \le b_i$ almost surely, and let $S_n = \sum_{i=1}^n Z_i$. Then for every $t > 0$,*

$$\mathbb{P}\big(S_n - \mathbb{E}S_n \ge t\big) \;\le\; \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big), \tag{1.27}$$

and the same bound holds for $\mathbb{P}(S_n - \mathbb{E}S_n \le -t)$; consequently

$$\mathbb{P}\big(|S_n - \mathbb{E}S_n| \ge t\big) \;\le\; 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Proof. Set $v := \frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2$ and $W_i := Z_i - \mathbb{E}Z_i$, so that the W_i are independent, centred, and $W_i \in [a_i - \mathbb{E}Z_i,\, b_i - \mathbb{E}Z_i]$, an interval of length $b_i - a_i$. By (1.23) and Lemma 1.62, for every $\lambda > 0$,

$$\psi_{S_n}(\lambda) = \sum_{i=1}^n \psi_{Z_i}(\lambda) = \sum_{i=1}^n \log \mathbb{E}\big[e^{\lambda W_i}\big] \;\le\; \sum_{i=1}^n \frac{\lambda^2 (b_i - a_i)^2}{8} = \frac{\lambda^2 v}{2}.$$

Theorem 1.57 then yields $\mathbb{P}(S_n - \mathbb{E}S_n \ge t) \le \exp\big(-\lambda t + \lambda^2 v/2\big)$ for every $\lambda > 0$. The quadratic $\lambda \mapsto -\lambda t + \lambda^2 v/2$ is minimal at $\lambda = t/v > 0$, where its value is $-t^2/(2v) = -2t^2/\sum_{i=1}^n (b_i - a_i)^2$; this proves (1.27). Applying it to the variables $-Z_i$, which satisfy $-b_i \le -Z_i \le -a_i$ with intervals of the same lengths, gives the left tail, and the union bound gives the two-sided statement. □

Corollary 1.65 (Empirical mean of bounded variables). *Let $X_1, \dots, X_n$ be i.i.d. with $a \le X_1 \le b$ almost surely and $\mu = \mathbb{E}X_1$. Then for every $t > 0$,*

$$\mathbb{P}\big(\bar{X}_n - \mu \ge t\big) \le e^{-2nt^2/(b-a)^2}, \qquad \mathbb{P}\big(\big|\bar{X}_n - \mu\big| \ge t\big) \le 2e^{-2nt^2/(b-a)^2}. \tag{1.28}$$

Equivalently, for every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\big|\bar{X}_n - \mu\big| \;\le\; (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{1.29}$$

and $n \ge (b-a)^2 \log(2/\delta)/(2\varepsilon^2)$ samples guarantee an accuracy ε at confidence $1 - \delta$.

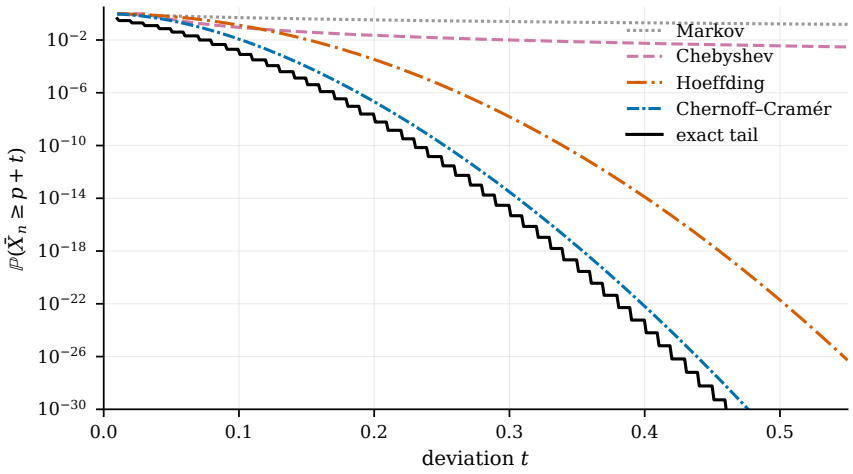


Figure 1.4: The exact tail $\mathbb{P}(\bar{X}_n \geq p + t)$ for $n = 100$ i.i.d. Ber(0.1) variables, together with the bounds of this chapter. Markov's and Chebyshev's inequalities decay polynomially in t and are hopeless beyond $t \approx 0.1$; Hoeffding's is exponential and already usable; the exact Cramér–Chernoff bound $e^{-n\,\mathrm{kl}(p+t,p)}$ of Chapter 2 tracks the true tail over thirty orders of magnitude. Note the logarithmic vertical scale: the visible gap between Hoeffding and the truth is a gap in the *exponent*. Generated by `figures/fig_tails_bernoulli.py`.

Proof. Apply Theorem 1.64 to $Z_i = X_i/n$, for which $b_i - a_i = (b-a)/n$ and $\sum_i (b_i - a_i)^2 = (b-a)^2/n$. Solving $2e^{-2nt^2/(b-a)^2} = \delta$ for t gives (1.29), and solving (1.29) for n gives the sample size. □

Compare (1.29) with the Chebyshev bound (1.20), which gives $\left|\bar{X}_n - \mu\right| \leq \sigma/\sqrt{n\delta}$: the dependence on the confidence level has been improved from $\delta^{-1/2}$ to $\sqrt{\log(1/\delta)}$, which is the whole point. The price is the boundedness assumption, and a variance that has been replaced by its worst-case value, as we now discuss.

Remark 1.66 (Hoeffding versus the central limit theorem). Take $X_i \sim \mathrm{Ber}(p)$, so $a = 0$, $b = 1$ and $\sigma^2 = p(1-p)$. The exponents in (1.28) and in the Gaussian heuristic (1.14) are respectively

$$2nt^2 \qquad \text{and} \qquad \frac{nt^2}{2p(1-p)}.$$

Their ratio is $4p(1-p) \leq 1$, with equality if and only if $p = \frac{1}{2}$. For $p = \frac{1}{2}$ Hoeffding's inequality is thus exactly as strong as the central limit theorem suggests – while being valid for every n. For $p = 10^{-3}$, on the other hand, the true exponent is 250 times larger than Hoeffding's: the bound is correct but massively pessimistic. The reason is visible in Remark 1.63: Lemma 1.62 replaces the variance of Z by the largest variance compatible with its range. Recovering the true variance is exactly what Bennett's and Bernstein's inequalities achieve (Chapter 2).

Remark 1.67 (Hoeffding = Chernoff + Pinsker). For $X_i \sim \mathrm{Ber}(p)$ one can compute the Chernoff bound exactly. With $\mathrm{kl}(q,p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$ the relative entropy between two Bernoulli laws, Section 2.3.1 will show that

$$\mathbb{P}\big(\bar{X}_n \geq p + t\big) \;\leq\; e^{-n\,\mathrm{kl}(p+t,\,p)}.$$

Pinsker's inequality $\mathrm{kl}(q,p) \geq 2(q-p)^2$, proved in Corollary 2.12, then gives back exactly (1.28). In other words, Hoeffding's inequality is the combination of the exact Chernoff bound

with one specific, lossy lower bound on the relative entropy. Everything that is gained in Chapter 2 consists in using better lower bounds on kl – or in not leaving kl at all.

Remark 1.68 (Sampling without replacement). Hoeffding proved considerably more than Theorem 1.64 in his 1963 paper. In particular [Hoe63, Sec. 6], if $\bar{X}_n$ denotes the mean of a sample of size n drawn *without replacement* from a finite population contained in $[a, b]$ with mean μ, then (1.28) still holds verbatim. The proof is not a consequence of Theorem 1.64 – the summands are no longer independent – but of the fact that sampling without replacement produces a smaller moment generating function than sampling with replacement. This is the first instance of a phenomenon we shall meet again: negative dependence helps concentration.

[V58 – Pointer to Chapter 5 (sampling without replacement, negative association)] *Replacement of the last sentence:* "This is the first instance of a phenomenon we shall meet again in Chapter 5: negative dependence helps concentration. Hoeffding's comparison is proved there (Theorem 5.28), in two ways – by his original coupling argument and as a consequence of the *negative association* of samples drawn without replacement (Corollary 5.23)." *Proposed on 15 October 2026 (chapter 5 written): not yet discussed.*

Remark 1.69 (What boundedness cannot be weakened to). Counterexample 1.41 shows that finite variance alone gives no exponential bound. One may ask whether the two-sided bound $a \le Z \le b$ can be replaced by a one-sided bound, $Z \le b$. It cannot, without a further hypothesis: a variable that is bounded above but has a heavy left tail has no finite moment generating function for $\lambda < 0$, and no exponential bound on its *lower* tail. For the *upper* tail, however, a one-sided bound together with a variance bound is enough, and this is precisely the hypothesis of Bennett's inequality.

1.10 First consequences for learning

We can now return to classification and cash in. The following is immediate but deserves to be isolated, because it is the elementary brick from which the whole theory of generalisation is built.

Corollary 1.70 (Deviation of the empirical risk of a fixed classifier)**.** *Let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier chosen* independently of the sample $\mathcal{S}_n$. *Then for every $t > 0$,*

$$\mathbb{P}\Big(\big|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\big| \ge t\Big) \;\le\; 2e^{-2nt^2},$$

so that with probability at least $1 - \delta$, $\big|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\big| \le \sqrt{\log(2/\delta)/(2n)}$.

Proof. The variables $\mathbb{1}\{h(X_i) \ne Y_i\}$, $1 \le i \le n$, are i.i.d., take values in $[0, 1]$ and have expectation $L_{\mathcal{D}}(h)$; apply Corollary 1.65 with $b - a = 1$. □

Remark 1.71 (Why this is not yet a learning guarantee). The hypothesis "chosen independently of the sample" is essential and is violated by every learning algorithm, since the whole point of an algorithm is to choose h using the data. The failure can be total. Consider the nearest-neighbour rule of Section 1.5: if the feature vectors $X_1, \dots, X_n$ are almost surely distinct, then $\widehat{h}_n^{\mathrm{NN}}(X_i) = Y_i$ for every i, hence

$$L_{\mathcal{S}_n}\big(\widehat{h}_n^{\mathrm{NN}}\big) = 0 \qquad \text{almost surely,}$$

while $L_{\mathcal{D}}(\widehat{h}_n^{\mathrm{NN}}) \ge L_{\mathcal{D}}^*$ can be arbitrarily close to $\frac{1}{2}$. The empirical risk of the classifier one has selected is not an estimate of its risk; it is a number that the algorithm has been designed

to make small. Two ways out exist: evaluate the classifier on fresh data (Proposition 1.72), or control the empirical risk *simultaneously for all* h in the class from which the algorithm chooses. The second route is the subject of Section 2.7.1, of Chapter 4 and of later chapters.

Proposition 1.72 (Hold-out estimate of the risk)**.** *Let* $\widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n)$ *be produced by any learning algorithm from a sample* $\mathcal{S}_n$*, and let* $\mathcal{T} = \big((X'_1, Y'_1), \dots, (X'_m, Y'_m)\big)$ *be i.i.d. with law* $\mathcal{D}$ *and independent of* $\mathcal{S}_n$*. Let* $L_{\mathcal{T}}(\widehat{h}_n) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}\{\widehat{h}_n(X'_j) \neq Y'_j\}$ *be the* test error*. Then for every* $t > 0$*,*

$$\mathbb{P}\Big(\big|L_{\mathcal{T}}(\widehat{h}_n) - L_{\mathcal{D}}(\widehat{h}_n)\big| \geq t\Big) \;\leq\; 2e^{-2mt^2}.$$

Proof. Condition on $\mathcal{S}_n$. Given $\mathcal{S}_n$, the classifier $\widehat{h}_n$ is a fixed measurable function, and the variables $\mathbb{1}\{\widehat{h}_n(X'_j) \neq Y'_j\}$, $1 \leq j \leq m$, are i.i.d. Bernoulli with parameter $L_{\mathcal{D}}(\widehat{h}_n)$, because $\mathcal{T}$ is independent of $\mathcal{S}_n$. Corollary 1.65 therefore gives

$$\mathbb{P}\Big(\big|L_{\mathcal{T}}(\widehat{h}_n) - L_{\mathcal{D}}(\widehat{h}_n)\big| \geq t \;\Big|\; \mathcal{S}_n\Big) \;\leq\; 2e^{-2mt^2} \qquad \text{almost surely.}$$

The right-hand side does not depend on $\mathcal{S}_n$; taking expectations concludes. □

This is the mathematical justification of the universal practice of splitting data into a training set and a test set. With $m = 10^4$ test points and $\delta = 5\%$, Proposition 1.72 certifies the test error to within $\pm 1.36 \cdot 10^{-2}$, whatever the algorithm, whatever the distribution, and whatever the dimension. It is worth appreciating how strong – and how cheap – this statement is.

Remark 1.73 (Reusing a test set). Proposition 1.72 requires $\mathcal{T}$ to be independent of $\widehat{h}_n$. A test set used once is; a test set used to choose between N candidate classifiers is not, and the naive confidence interval becomes invalid. The remedy is again a union bound, which costs a factor $\sqrt{\log N}$ (Exercise 1.13). Public benchmarks on which thousands of researchers evaluate thousands of models are, from this point of view, a large-scale experiment in test-set overfitting.

Example 1.74 (How many people to poll)**.** To estimate the proportion p of a population holding an opinion, within $\varepsilon = 3\%$ and with confidence 95%, Corollary 1.65 requires

$$n \;\geq\; \frac{\log(2/\delta)}{2\varepsilon^2} = \frac{\log 40}{2 \cdot 9 \cdot 10^{-4}} \approx 2050.$$

The Gaussian answer, obtained from (1.14) with the worst-case variance $\sigma^2 = \frac{1}{4}$, is $n \geq z_\delta^2/(4\varepsilon^2) \approx 1067$ with $z_\delta = 1.96$. Hoeffding's inequality costs here a factor 1.92 in sample size – and, unlike the Gaussian computation, it is a theorem valid at this very value of n. Moreover the factor is $2\log(2/\delta)/z_\delta^2$, which tends to 1 as $\delta \to 0$ since $z_\delta \sim \sqrt{2\log(1/\delta)}$: at high confidence, Hoeffding's inequality is asymptotically *free*. This is Remark 1.66 seen from the other side: for $p = \frac{1}{2}$ the worst-case variance is the true variance, and nothing is lost.

1.10.1 Back to nearest neighbours: the k-NN rule

We can now improve Theorem 1.21 substantially, and remove the factor 2 that Remark 1.25 showed to be unavoidable for the 1-nearest-neighbour rule. The idea is the one that underlies all of statistics: to reduce noise, average.

Definition 1.75 (k-nearest-neighbour rule)**.** Let $1 \le k \le n$. For $x \in \mathcal{X}$ let $X_{(1)}(x), \dots, X_{(n)}(x)$ be the sample points reordered by increasing distance to x (ties broken by index, as in Definition 1.19), and let $Y_{(1)}(x), \dots, Y_{(n)}(x)$ be the corresponding labels. The *k-nearest-neighbour rule* is

$$\widehat{h}_n^{k\text{-NN}}(x) := \mathbb{1}\Big\{ \frac{1}{k} \sum_{j=1}^{k} Y_{(j)}(x) \;\ge\; \frac{1}{2} \Big\} :$$

it takes a majority vote among the labels of the k closest observations.

As for Definition 1.19, $\widehat{h}^{k\text{-NN}}$ denotes the rule as a learning algorithm and $R_n(\widehat{h}^{k\text{-NN}}) = \mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}_n^{k\text{-NN}})\big]$ its expected risk.

The analysis needs a companion to Lemma 1.23 for the k-th nearest neighbour. Remarkably, the occupancy bound (1.11) has an exact analogue that requires no Chernoff bound at all, only an algebraic identity.

Lemma 1.76 (An exact binomial identity)**.** *For every $n \ge 1$, $k \ge 1$ and $p \in [0,1]$,*

$$p\, \mathbb{P}\big(\mathrm{Bin}(n,p) \le k-1\big) \;\le\; \frac{k}{n+1}.$$

Proof. For $0 \le j \le n$, a direct computation with factorials gives

$$(n+1)p\binom{n}{j}p^j(1-p)^{n-j} = \frac{(n+1)!}{j!\,(n-j)!}p^{j+1}(1-p)^{n-j} = (j+1)\binom{n+1}{j+1}p^{j+1}(1-p)^{n-j},$$

that is, $(n+1)p\,\mathbb{P}\big(\mathrm{Bin}(n,p)=j\big) = (j+1)\mathbb{P}\big(\mathrm{Bin}(n+1,p)=j+1\big)$. Summing over $0 \le j \le k-1$,

$$(n+1)p\,\mathbb{P}\big(\mathrm{Bin}(n,p) \le k-1\big) = \sum_{j=0}^{k-1}(j+1)\,\mathbb{P}\big(\mathrm{Bin}(n+1,p)=j+1\big) \le k\sum_{i=1}^{k}\mathbb{P}\big(\mathrm{Bin}(n+1,p)=i\big) \le k.$$

□

Lemma 1.77 (Distance to the k-th nearest neighbour)**.** *Let $X, X_1, \dots, X_n$ be i.i.d. with values in $[0,1)^d$ and let $1 \le k \le n$. Then*

$$\mathbb{E}\big[\rho(X, X_{(k)}(X))\big] \;\le\; 3\sqrt{d}\left(\frac{k}{n}\right)^{1/(d+1)}.$$

Proof. Partition $[0,1)^d$ into the T^d cells C of side $1/T$ as in the proof of Lemma 1.23. If the cell $C(X)$ containing X contains at least k sample points, then at least k sample points lie within distance $\operatorname{diam} C(X)$ of X, so that $\rho(X, X_{(k)}(X)) \le \sqrt{d}/T$; otherwise we use $\rho(X, X_{(k)}) \le \sqrt{d}$. Since X is independent of the sample, Lemma 1.76 gives

$$\begin{aligned}\mathbb{P}\big(C(X) \text{ contains at most } k-1 \text{ sample points}\big) &= \sum_{C \in \mathcal{C}} p_C\, \mathbb{P}\big(\mathrm{Bin}(n,p_C) \le k-1\big) \\ &\le \frac{k\,|\mathcal{C}|}{n+1} = \frac{k\,T^d}{n+1},\end{aligned}$$

so that $\mathbb{E}\,\rho(X, X_{(k)}) \le \sqrt{d}/T + \sqrt{d}\,kT^d/(n+1)$. Choosing $T = \big\lfloor (n/k)^{1/(d+1)} \big\rfloor \ge 1$ gives $kT^d/(n+1) \le k(n/k)^{d/(d+1)}/n = (k/n)^{1/(d+1)}$ and, as $\lfloor u \rfloor \ge u/2$ for $u \ge 1$, $1/T \le 2(k/n)^{1/(d+1)}$. □

Remark 1.78 (The optimal rate for the k-NN rule). Scale integration improves Lemma 1.77 exactly as it improved Lemma 1.23 in Remark 1.29: for $d \geq 2$ and $1 \leq k \leq n$,

$$\mathbb{E}\big[\rho(X, X_{(k)}(X))\big] \;\leq\; \frac{2d}{d-1}\sqrt{d}\Big(\frac{k}{n+1}\Big)^{1/d} \;\leq\; 4\sqrt{d}\Big(\frac{k}{n}\Big)^{1/d}. \tag{1.30}$$

Indeed, for $r \in (0, \sqrt{d}\,]$ partition $[0,1)^d$ into $N(r) = T^d$ cells of side $1/T$ with $T = \lceil \sqrt{d}/r \rceil \leq 2\sqrt{d}/r$; if the cell of X contains at least k sample points then $\rho(X, X_{(k)}(X)) \leq r$, so that Lemma 1.76 gives $\mathbb{P}(\rho(X, X_{(k)}) > r) \leq kN(r)/(n+1)$, and one integrates and substitutes $r = as$ with $a = 2\sqrt{d}\,(k/(n+1))^{1/d}$ as before.

Carried into the proof of Theorem 1.79, (1.30) gives

$$R_n\big(\widehat{h}^{k\text{-NN}}\big) \;\leq\; L^*_{\mathcal{D}} + \frac{1}{\sqrt{e\,k}} + 8c\sqrt{d}\Big(\frac{k}{n}\Big)^{1/d},$$

and balancing the two terms with $k = \lceil n^{2/(d+2)} \rceil$ yields

$$R_n\big(\widehat{h}^{k\text{-NN}}\big) - L^*_{\mathcal{D}} \;\leq\; \frac{1 + 8c\sqrt{d}\,2^{1/d}}{n^{1/(d+2)}}.$$

The exponent $1/(d+2)$ is worth noting: it is the optimal rate for classification with a c-Lipschitz regression function and no margin assumption, so that the k-nearest-neighbour rule, tuned in this way, is rate-optimal over that class. The statement proved above, with the simpler one-grid lemma, gives the slightly weaker $n^{-1/(d+3)}$.

Theorem 1.79 (Finite-sample risk of the k-NN rule)**.** *Under Assumption 1.20, for every* $1 \leq k \leq n$,

$$R_n\big(\widehat{h}^{k\text{-NN}}\big) \;\leq\; L^*_{\mathcal{D}} + \frac{1}{\sqrt{e\,k}} + 6c\sqrt{d}\,\Big(\frac{k}{n}\Big)^{1/(d+1)}. \tag{1.31}$$

In particular, choosing $k = \lceil n^{2/(d+3)} \rceil$,

$$R_n\big(\widehat{h}^{k\text{-NN}}\big) - L^*_{\mathcal{D}} \;\leq\; \frac{1 + 6c\sqrt{d}\,2^{1/(d+1)}}{n^{1/(d+3)}}, \tag{1.32}$$

so that the k-nearest-neighbour rule is consistent*: its risk converges to the Bayes risk, for every distribution satisfying (A3).*

Proof. Let $(X, Y) \sim \mathcal{D}$ be independent of $\mathcal{S}_n$, let $\mathcal{G} = \sigma(X, X_1, \ldots, X_n)$, and write $p = \eta(X)$, $\hat{p} = \frac{1}{k}\sum_{j \leq k} Y_{(j)}(X)$ and $\bar{p} = \mathbb{E}[\hat{p} \mid \mathcal{G}] = \frac{1}{k}\sum_{j \leq k} \eta\big(X_{(j)}(X)\big)$. Conditionally on $\mathcal{G}$, the variables $Y, Y_{(1)}, \ldots, Y_{(k)}$ are independent Bernoulli with parameters p and $\eta(X_{(j)}(X))$.

Step 1: the conditional excess risk. Suppose first $p \geq \frac{1}{2}$, so that $h^*(X) = 1$ and $\min\{p, 1-p\} = 1 - p$. Then

$$\mathbb{P}\big(\widehat{h}_n^{k\text{-NN}}(X) \neq Y \mid \mathcal{G}\big) = p\,\mathbb{P}\big(\hat{p} < \tfrac{1}{2} \mid \mathcal{G}\big) + (1-p)\big(1 - \mathbb{P}(\hat{p} < \tfrac{1}{2} \mid \mathcal{G})\big) = (1-p) + (2p-1)\mathbb{P}\big(\hat{p} < \tfrac{1}{2} \mid \mathcal{G}\big).$$

The case $p < \frac{1}{2}$ is symmetric, and in both cases

$$\mathbb{P}\big(\widehat{h}_n^{k\text{-NN}}(X) \neq Y \mid \mathcal{G}\big) = \min\big\{p, 1-p\big\} + \big|2p - 1\big|\;\mathbb{P}\big(V \mid \mathcal{G}\big), \tag{1.33}$$

where V is the event that the majority vote disagrees with $h^*(X)$.

Step 2: Hoeffding controls the vote. Since η is c-Lipschitz, $\big|\bar{p} - p\big| \leq \frac{c}{k}\sum_{j \leq k} \rho(X, X_{(j)}(X)) \leq c\rho_k$ with $\rho_k := \rho(X, X_{(k)}(X))$. Assume again $p \geq \frac{1}{2}$, so $V = \{\hat{p} < \frac{1}{2}\}$ and $\bar{p} - \frac{1}{2} \geq (p - \frac{1}{2}) - c\rho_k$.

If the right-hand side is non-positive the bound below is trivial; otherwise Corollary 1.65, applied conditionally on $\mathcal{G}$ to the k independent $[0,1]$-valued variables $Y_{(j)}$, gives

$$\mathbb{P}\big(V \mid \mathcal{G}\big) \le \mathbb{P}\big(\hat{p} - \bar{p} < \tfrac{1}{2} - \bar{p} \mid \mathcal{G}\big) \le \exp\Big(-2k\big(\bar{p} - \tfrac{1}{2}\big)^2\Big) \le \exp\Big(-2k\,U^2\Big), \quad U := \Big(\big|p - \tfrac{1}{2}\big| - c\rho_k\Big)_+ .$$

The same bound holds when $p < \frac{1}{2}$, by symmetry.

Step 3: integration. Since $\big|2p-1\big| = 2\big|p - \frac{1}{2}\big| \le 2U + 2c\rho_k$, (1.33) and Step 2 give

$$R_n\big(\widehat{h}^{k\text{-NN}}\big) - L^*_{\mathcal{D}} \;\le\; \mathbb{E}\big[(2U + 2c\rho_k)e^{-2kU^2}\big] \;\le\; \sup_{u \ge 0}\big(2u\, e^{-2ku^2}\big) + 2c\,\mathbb{E}[\rho_k],$$

where we used (1.7) for the Bayes risk and $e^{-2kU^2} \le 1$ in the second term. The supremum is attained at $u = 1/(2\sqrt{k})$ and equals $1/\sqrt{ek}$, and $\mathbb{E}[\rho_k] \le 3\sqrt{d}(k/n)^{1/(d+1)}$ by Lemma 1.77; this proves (1.31). For (1.32), take $k = \lceil n^{2/(d+3)} \rceil \le 2n^{2/(d+3)}$ and note that $1/\sqrt{ek} \le n^{-1/(d+3)}$ and $(k/n)^{1/(d+1)} \le 2^{1/(d+1)} n^{-1/(d+3)}$, since $\frac{2}{d+3} - 1 = -\frac{d+1}{d+3}$. □

Remark 1.80 (What averaging bought). Compare (1.32) with (1.10): the factor 2 in front of $L^*_{\mathcal{D}}$ has disappeared, at the price of a slightly slower rate, $n^{-1/(d+3)}$ instead of $n^{-1/(d+1)}$. The trade-off is visible in (1.31): a large k averages away the label noise – the term $1/\sqrt{ek}$, which is a pure concentration term and comes straight from Hoeffding's inequality – but forces the rule to look at points that are far away, whose regression function may differ from $\eta(X)$ – the term $(k/n)^{1/(d+1)}$, which is pure geometry. Choosing k balances a *statistical* and an *approximation* error, an arbitration that the reader will meet in every chapter of this book and in every corner of nonparametric statistics.

Two further comments. First, Theorem 1.79 is a quantitative version of Stone's consistency theorem [Sto77], which holds without any smoothness assumption but gives no rate; here the Lipschitz hypothesis (A3) is what converts consistency into a rate. Second, the exponential dependence on d has not gone away, and cannot: the second term of (1.31) is the geometric term of Remark 1.26 again.

1.10.2 From deviation bounds to almost-sure statements

An exponential bound valid for every n can be summed, and Remark 1.52 then converts it into a trajectorial statement.

Proposition 1.81. *Let $X_1, X_2, \dots$ be i.i.d. with $a \le X_1 \le b$ almost surely and mean μ. Then, almost surely,*

$$\limsup_{n \to \infty} \sqrt{\frac{2n}{\log n}}\,\big|\bar{X}_n - \mu\big| \;\le\; b - a.$$

Proof. Fix $\gamma > 0$ and set $t_n = (b-a)\sqrt{(1+\gamma)\log n/(2n)}$. By (1.28),

$$\mathbb{P}\big(\big|\bar{X}_n - \mu\big| \ge t_n\big) \le 2\exp\Big(-\frac{2nt_n^2}{(b-a)^2}\Big) = 2e^{-(1+\gamma)\log n} = \frac{2}{n^{1+\gamma}},$$

which is summable. By Borel–Cantelli, almost surely $\big|\bar{X}_n - \mu\big| < t_n$ for all large n, i.e. $\limsup_n \sqrt{2n/\log n}\,\big|\bar{X}_n - \mu\big| \le (b-a)\sqrt{1+\gamma}$. Letting $\gamma \downarrow 0$ along a sequence concludes. □

Remark 1.82 (Hoeffding versus the law of the iterated logarithm). The truth is better. The law of the iterated logarithm states that, for i.i.d. variables with variance σ^2,

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\sigma^2 \log\log n}}\,(\bar{X}_n - \mu) = 1 \qquad \text{almost surely},$$

so the correct normalisation involves $\log\log n$, not $\log n$. The loss in Proposition 1.81 is the price of the union bound implicit in Borel–Cantelli: we paid for each n separately, whereas the fluctuations of $\bar{X}_n$ and $\bar{X}_{n+1}$ are almost the same. Recovering $\log\log n$ requires treating blocks of indices simultaneously, which is done with a *maximal* inequality. We shall see such inequalities when we study martingales.

1.10.3 Application: comparing two treatments

Suppose two treatments A and B produce independent outcomes in $[0,1]$ with unknown means μ_A and μ_B, and that we may allocate n subjects to each. Let $\bar{A}_n$ and $\bar{B}_n$ be the two empirical means.

Proposition 1.83. *For every $t > 0$, $\mathbb{P}\big(\big|(\bar{A}_n - \bar{B}_n) - (\mu_A - \mu_B)\big| \ge t\big) \le 2e^{-nt^2/2}$. Consequently, if $|\mu_A - \mu_B| \ge \varepsilon$ then $n \ge 2\log(2/\delta)/\varepsilon^2$ subjects per arm suffice to identify the better treatment with probability at least $1-\delta$.*

Proof. Pair the subjects arbitrarily and write $\bar{A}_n - \bar{B}_n = \frac{1}{n}\sum_{i=1}^n (A_i - B_i)$, an average of n independent variables with values in $[-1,1]$ and mean $\mu_A - \mu_B$. Corollary 1.65 with $b - a = 2$ gives $2\exp\big(-2nt^2/4\big)$. For the second claim, take $t = \varepsilon$: on the complementary event, $\bar{A}_n - \bar{B}_n$ has the same sign as $\mu_A - \mu_B$. □

With $\varepsilon = 10^{-2}$ and $\delta = 5\%$ this requires $n \approx 7.4 \cdot 10^4$ subjects *per arm*: detecting small effects is expensive, and no inequality will change that – the requirement $n \asymp \varepsilon^{-2}$ is a feature of the problem, not of the proof. This is proved in Proposition 3.60: when $\varepsilon \le \frac{1}{4}$ and $\delta < \frac{1}{4}$, even an adaptive allocation of the subjects needs at least $\frac{15}{32}\log(1/(4\delta))/\varepsilon^2$ subjects in total.

Remark 1.84 (Why you may not peek). Proposition 1.83 is a statement about a *fixed* n, decided before the data are seen. A tempting practice is to monitor the experiment and stop as soon as the confidence interval excludes 0. This invalidates the guarantee completely. Indeed, assuming the outcomes are not deterministic, so that $\sigma^2 = \mathbb{V}\mathrm{ar}(A_1 - B_1) > 0$, the law of the iterated logarithm (Remark 1.82) gives, for any constant C,

$$\big|\bar{A}_n - \bar{B}_n - (\mu_A - \mu_B)\big| \;\ge\; \frac{C}{\sqrt{n}} \qquad \text{for infinitely many } n, \quad \text{almost surely},$$

since the left-hand side is of order $\sqrt{\log\log n/n} \gg n^{-1/2}$ along a subsequence. An experimenter willing to wait will therefore *always* eventually observe a "significant" difference, even when $\mu_A = \mu_B$. The cure is to use bounds that hold simultaneously for all n – confidence sequences, of half-width $\asymp \sqrt{\log\log n/n}$ – rather than for one n at a time. This is a recurring theme in sequential analysis and in the theory of bandit algorithms.

Remark 1.85 (Hold-out, cross-validation, and what we can prove). Proposition 1.72 analyses the *hold-out* estimate, in which the data are split once. Practitioners prefer V-fold cross-validation, which splits the data V times and averages the V test errors, thereby using every observation both for training and for testing. The analysis is much harder, because the V estimates are dependent – they share training data – and because the classifier changes from fold to fold. There is no analogue of Proposition 1.72 of comparable simplicity and generality.

The leave-one-out case $V = n$ can nevertheless be handled for specific, *stable* algorithms; we shall do it for the nearest-neighbour rule in Section 2.7.6, once the bounded-difference inequality is available, and the stability required will be a geometric property of the rule.

1.10.4 How many samples? A summary

It is worth tabulating what the three answers of this chapter cost. Take $X_i \in [0,1]$ i.i.d. with variance $\sigma^2 = \frac{1}{4}$ – the worst case for a $[0,1]$-valued variable, attained by $\mathrm{Ber}(\frac{1}{2})$ – and ask for $\mathbb{P}\big(\big|\bar{X}_n - \mu\big| \geq \varepsilon\big) \leq \delta$ with $\varepsilon = 10^{-2}$. Chebyshev's inequality (1.20) requires $n \geq \sigma^2/(\delta\varepsilon^2)$, Hoeffding's (1.28) requires $n \geq \log(2/\delta)/(2\varepsilon^2)$, and the central limit theorem *suggests* $n \geq z_\delta^2 \sigma^2/\varepsilon^2$, where z_δ is the $(1-\delta/2)$-quantile of $\mathcal{N}(0,1)$.

δ	Chebyshev	Hoeffding	Gaussian (not a theorem)
10^{-1}	$2.5 \cdot 10^4$	$1.5 \cdot 10^4$	$6.8 \cdot 10^3$
10^{-2}	$2.5 \cdot 10^5$	$2.6 \cdot 10^4$	$1.7 \cdot 10^4$
10^{-4}	$2.5 \cdot 10^7$	$5.0 \cdot 10^4$	$3.8 \cdot 10^4$
10^{-6}	$2.5 \cdot 10^9$	$7.3 \cdot 10^4$	$6.0 \cdot 10^4$

Chebyshev's requirement is proportional to $1/\delta$ and becomes absurd; Hoeffding's grows like $\log(1/\delta)$ and approaches the Gaussian benchmark as the confidence increases – a factor 2.2 at $\delta = 10^{-1}$, 1.6 at 10^{-2}, and only 1.2 at 10^{-6} (Example 1.74). This table is the whole argument for exponential inequalities, in four lines.

1.11 Where we are going

Hoeffding's inequality is the first of a family, and each of its weaknesses points to a chapter of this book.

It is *variance-blind*: Remark 1.66 showed that it replaces the variance of a bounded variable by the largest value compatible with its range. Recovering the true variance is the object of Bennett's and Bernstein's inequalities, and more generally of the refined analysis of the cumulant generating function, in Chapter 2.

It applies only to *sums*. The quantities of interest in learning – $\sup_{h \in \mathcal{H}} \big|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\big|$, the risk of a rule built from the data, the error of an estimator – are non-linear functions of the sample. The bounded-difference inequality of Chapter 2 is the first extension; the entropy method, the Efron–Stein inequality and Talagrand's inequality go further.

It is an *upper* bound, as is the nearest-neighbour theorem. Apart from the Paley–Zygmund inequality of Section 1.7.2, this chapter has no way of telling whether such bounds can be improved: whether the exponent of a deviation inequality is the right one, whether a cleverer algorithm could escape the curse of dimensionality met by the nearest-neighbour rule (Remark 1.26), whether any algorithm can learn without assumptions on $\mathcal{D}$. These are questions about *lower* bounds, answered in Chapter 3 with relative entropy and the change of measure.

It bounds a deviation *around the mean*, and says nothing about the mean itself. For a sum the mean is trivial; for a supremum it is the whole problem, and it is what Vapnik–Chervonenkis theory, Rademacher complexities and chaining are designed to bound; the first two are the subject of Chapter 4.

Finally, it is an inequality about *independent* variables. Dependence – martingales, Markov chains, negatively associated variables – requires its own tools.

Each of these directions is taken up in the chapters that follow.

1.12 Notes and references

The statistical learning framework of Section 1.3 follows [SB14], whose first chapters are an excellent companion to this one; the measure-theoretic care taken here with learning algorithms is treated at length in [DGL96, Chap. 1]. The Bayes classifier and the excess-risk identity (1.6) go back to the beginnings of decision theory; the form we use, and the systematic exploitation of $|2\eta - 1|$ as a local measure of difficulty, are taken from [DGL96] and [BBL05].

The nearest-neighbour rule was introduced in an unpublished technical report of Fix and Hodges [FH51], reprinted in *International Statistical Review* in 1989. The asymptotic analysis, including the bound $L^{\mathrm{NN}} \le 2L^*_{\mathcal{D}}(1 - L^*_{\mathcal{D}})$, is due to Cover and Hart [CH67]; Stone's universal consistency theorem for k-nearest neighbours is [Sto77]. The finite-sample statement of Theorem 1.21, with its elementary partition argument, is classical; see [DGL96, Chap. 5–6] and the monograph [BD15] entirely devoted to the method.

Markov's inequality and its history are discussed in [BLM13, Chap. 2]. Cantelli's one-sided inequality is [Can28]. Hoeffding's lemma and inequality are [Hoe63], a paper that contains much more than the result that bears its name – in particular the treatment of sampling without replacement and a comparison of the binomial with general bounded variables already present in [Hoe56]. Okamoto [Oka59] obtained related binomial bounds at about the same time. The sharpest known improvements of Hoeffding's inequality for bounded variables, replacing the Gaussian tail by a binomial one, are due to Bentkus [Ben04]. The exponential bounding technique of Section 1.8 is usually attributed to Chernoff [Che52], who attributes it to Herman Rubin [Bat96]; see [Cra38] for the ancestor of the whole theory, to which we return in Chapter 2.

A historical note

The oldest inequality of the chapter is (1.16), which Bienaymé published in 1853 [Bie53] in a defence of Laplace's method of least squares, and which Chebyshev rediscovered and generalised in 1867 [Tch67]; the attribution to Markov, Chebyshev's student, is an instance of Stigler's law. The exponential method has a longer and more tangled history: it appears in Bernstein's work of the 1920s, in Esscher's 1932 paper on the collective theory of risk [Ess32] – where the tilted measure (2.1) is introduced for actuarial purposes – and in Cramér's 1938 memoir [Cra38], which was likewise motivated by insurance mathematics and which contains the large-deviation theorem of Chapter 2. Chernoff's 1952 paper [Che52] made the method standard in statistics, and gave it its name, although Chernoff attributed the argument to Herman Rubin [Bat96].

Hoeffding's 1963 paper [Hoe63] is the direct ancestor of modern concentration theory. It contains, besides Theorem 1.64, the treatment of sampling without replacement (Remark 1.68), a comparison theorem showing that among bounded variables with a given mean the Bernoulli law is extremal, and – in a section that is often overlooked – the martingale argument that Azuma [Azu67] rediscovered four years later and that underlies the bounded-difference inequality of Chapter 2. The combinatorial potential of that argument was realised in the 1980s, in particular by Shamir and Spencer [SS87], and systematised by McDiarmid [McD89].

The phrase *concentration of measure* itself comes from a different tradition – the local theory of Banach spaces, and the work of Milman on Lévy's isoperimetric inequality on the sphere – whose fusion with the probabilistic tradition described above, largely through the work of Talagrand [Tal95] and Ledoux [Led01], produced the subject as it now stands. Both halves are represented in this book.

1.13 Exercises

Exercise 1.1 (Chebyshev is sharp)**.** Show that for every $\sigma > 0$ and every $t \geq \sigma$ there exists a random variable Z with mean 0 and variance σ^2 such that $\mathbb{P}(|Z| \geq t) = \sigma^2/t^2$. What happens for $t < \sigma$?

Exercise 1.2 (⋆ Sharpness of Cantelli's inequality)**.** Let $\sigma, t > 0$. Exhibit a random variable with mean 0 and variance σ^2 for which (1.21) is an equality.

Exercise 1.3 (The second moment method at work)**.** Let $G \sim G(n,p)$ be an Erdős–Rényi random graph and let Z be its number of triangles.

(a) Show that $\mathbb{E}[Z] = \binom{n}{3}p^3$ and that $\mathbb{V}\mathrm{ar}(Z) = \binom{n}{3}p^3(1-p^3) + 3\binom{n}{3}(n-3)\big(p^5 - p^6\big)$. *Hint: two triples of vertices sharing at most one vertex give independent indicators.*

(b) Deduce from Corollary 1.54 that $\mathbb{P}(Z > 0) \to 1$ as soon as $np \to \infty$.

(c) Show conversely, using Markov's inequality, that $\mathbb{P}(Z > 0) \to 0$ when $np \to 0$. (The threshold $p \asymp 1/n$ for the appearance of triangles is one of the first results of random graph theory.)

Exercise 1.4 (⋆ When moments beat Chernoff)**.** Let $Z \geq 1$ with $\mathbb{P}(Z > x) = x^{-\alpha}$ for $x \geq 1$, where $\alpha > 0$.

(a) Show that $\mathbb{E}[e^{\lambda Z}] = +\infty$ for every $\lambda > 0$, so that the Chernoff bound (1.22) is vacuous.

(b) Compute $\mathbb{E}[Z^q]$ for $0 < q < \alpha$ and deduce, by optimising over q, a bound on $\mathbb{P}(Z \geq t)$ of the correct polynomial order.

Exercise 1.5 (Popoviciu's inequality: the equality case)**.** Let $Z \in [a,b]$ almost surely. Show that $\mathbb{V}\mathrm{ar}(Z) \leq (b-a)^2/4$, with equality if and only if Z takes the two values a and b with probability $1/2$ each.

Exercise 1.6 (⋆ Rademacher variables are sub-Gaussian)**.** Let ξ be a Rademacher variable, i.e. $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = \frac{1}{2}$. Show, by expanding the exponential into a power series and comparing with the series of $e^{\lambda^2/2}$, that

$$\mathbb{E}\big[e^{\lambda \xi}\big] \;\leq\; e^{\lambda^2/2}, \qquad \lambda \in \mathbb{R}.$$

Compare with the bound given by Lemma 1.62. Deduce a deviation bound for $\sum_{i=1}^n a_i \xi_i$ with $a \in \mathbb{R}^n$, and check that it is optimal in the exponent when all the a_i are equal.

Exercise 1.7 (The exact Chernoff bound for Bernoulli sums)**.** Let $X_1, \dots, X_n$ be i.i.d. $\mathrm{Ber}(p)$ and let $\mathrm{kl}(q,p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$ for $p, q \in (0,1)$.

(a) Compute $K_{X_1}(\lambda) = \log\big(1-p+pe^{\lambda}\big)$ and show that, for $q \in (p,1)$, $\sup_{\lambda>0}\{\lambda q - K_{X_1}(\lambda)\} = \mathrm{kl}(q,p)$, the supremum being attained at $\lambda^* = \log\frac{q(1-p)}{p(1-q)}$.

(b) Deduce that $\mathbb{P}(\bar{X}_n \geq q) \leq e^{-n\,\mathrm{kl}(q,p)}$.

(c) Verify that $n\bar{X}_n \sim \mathrm{Bin}(n,p)$ and compare, numerically, the bound of (b) with the exact binomial tail for $n = 100$, $p = 0.1$, $q = 0.2$.

Exercise 1.8 (Pinsker's inequality for Bernoulli laws)**.** With the notation of Exercise 1.7, show that $\mathrm{kl}(q,p) \geq 2(q-p)^2$ for all $p, q \in (0,1)$, and deduce Corollary 1.65 for Bernoulli variables. *Hint: fix p and study $g(q) = \mathrm{kl}(q,p) - 2(q-p)^2$; compute g'' and use $u(1-u) \leq \frac{1}{4}$.*

Exercise 1.9 (⋆ Maximum of sub-Gaussian variables)**.** Let $Z_1, \dots, Z_N$ be real random variables, not necessarily independent, such that $\psi_{Z_i}(\lambda) \le \lambda^2 v/2$ for all $\lambda > 0$ and all i, and $\mathbb{E}Z_i = 0$. Show that

$$\mathbb{E}\Big[\max_{1\le i\le N} Z_i\Big] \;\le\; \sqrt{2v\log N}.$$

Hint: $\exp(\lambda\,\mathbb{E}\max_i Z_i) \le \mathbb{E}\exp(\lambda\max_i Z_i) \le \sum_i \mathbb{E}e^{\lambda Z_i}$*; optimise in* λ. Apply this to N Rademacher sums, and comment on the dependence in N.

Exercise 1.10 (Cost-sensitive classification)**.** Let $\mathcal{Y} = \{1, \dots, M\}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ be a loss matrix, the risk being $L_{\mathcal{D}}(h) = \mathbb{E}[\ell(h(X), Y)]$ as in Definition 1.9.

(a) Show that $h^*(x) \in \arg\min_{a\in\mathcal{Y}} \sum_y \ell(a, y)\eta_y(x)$ minimises the risk, and compute the excess risk of an arbitrary h.

(b) Specialise to $M = 2$ with $\ell(1, 0) = c_1$, $\ell(0, 1) = c_0$ and $\ell(y, y) = 0$: show that the optimal rule is $h^*(x) = \mathbb{1}\{\eta(x) \ge c_1/(c_0 + c_1)\}$, and comment on the case where false negatives are much more costly than false positives.

Exercise 1.11 (Nearest neighbours: ties and conventions)**.** Show that Theorem 1.21 remains valid if ties in Definition 1.19 are broken by any measurable rule, for instance uniformly at random among the nearest points. Where exactly is the tie-breaking convention used in the proof?

Exercise 1.12 (Is Lemma 1.23 tight?)**.** Let $\mathcal{D}_X$ be uniform on $[0, 1)^d$.

(a) For $d = 1$, compute $\mathbb{E}\big[\rho(X, X_{I(X)})\big]$ exactly, or up to a constant factor, and compare with the bound $\asymp n^{-1/2}$ of Lemma 1.23.

(b) For general d, show that $\mathbb{E}\big[\rho(X, X_{I(X)})\big] \asymp n^{-1/d}$, so that the exponent $1/(d + 1)$ of Lemma 1.23 is not optimal for this particular $\mathcal{D}_X$. Where does the proof lose?

Exercise 1.13 (⋆ Reusing a test set)**.** Let $h_1, \dots, h_N$ be classifiers, all independent of a test sample $\mathcal{T}$ of size m, and let $\hat{k} \in \arg\min_k L_{\mathcal{T}}(h_k)$.

(a) Show that, with probability at least $1 - \delta$, $\big|L_{\mathcal{T}}(h_{\hat{k}}) - L_{\mathcal{D}}(h_{\hat{k}})\big| \le \sqrt{\log(2N/\delta)/(2m)}$.

(b) Show that $\mathbb{E}\big[L_{\mathcal{T}}(h_{\hat{k}})\big]$ can be strictly smaller than $\min_k L_{\mathcal{D}}(h_k)$, and construct an example where the gap is of order $\sqrt{\log N/m}$.

Exercise 1.14 (⋆ Median of means)**.** Let $X_1, \dots, X_n$ be i.i.d. with mean μ and *finite variance* σ^2 – no boundedness, no exponential moment. Let J be an odd integer, split the sample into J blocks of size $m = \lfloor n/J \rfloor$, let $\bar{X}^{(1)}, \dots, \bar{X}^{(J)}$ be the block means and let $\hat{\mu} := \mathrm{median}\big(\bar{X}^{(1)}, \dots, \bar{X}^{(J)}\big)$.

(a) Show that $\mathbb{P}\big(\big|\bar{X}^{(j)} - \mu\big| > 2\sigma/\sqrt{m}\big) \le \frac{1}{4}$.

(b) Show that if $\big|\hat{\mu} - \mu\big| > 2\sigma/\sqrt{m}$ then at least $J/2$ of the block means are at distance more than $2\sigma/\sqrt{m}$ from μ.

(c) Using Theorem 1.64, deduce that $\mathbb{P}\big(\big|\hat{\mu} - \mu\big| > 2\sigma/\sqrt{m}\big) \le e^{-J/8}$, and conclude: for every $\delta \in (0, \frac{1}{2}]$, if J is the smallest odd integer $\ge 8\log(1/\delta)$ and $n \ge J$, then

$$\big|\hat{\mu} - \mu\big| \;\le\; 10\,\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$.

Comment: compare with Counterexample 1.41. Sub-Gaussian deviations are achievable with two moments only – but by a different estimator. It is the empirical mean, not the problem, that requires exponential moments.

Exercise 1.15 (A refinement of Popoviciu's inequality)**.** Let $a \le Z \le b$ almost surely, with $\mu = \mathbb{E}Z$.

(a) Show that $\mathbb{V}\mathrm{ar}(Z) \le (b-\mu)(\mu-a)$. *Hint: expand* $\mathbb{E}[(Z-a)(b-Z)] \ge 0$.

(b) Deduce Exercise 1.5, and check that for $Z \sim \mathrm{Ber}(p)$ the refinement is an equality.

(c) Where could this refinement be inserted in the first proof of Lemma 1.62, and why does it *not* improve (1.24)? *Hint: what is the mean of the tilted law?*

Exercise 1.16 (Allocating a budget between two treatments)**.** In the setting of Section 1.10.3, suppose n_A subjects are allocated to treatment A and n_B to treatment B.

(a) Show that $\mathbb{P}\big(\big|(\bar{A}_{n_A} - \bar{B}_{n_B}) - (\mu_A - \mu_B)\big| \ge t\big) \le 2\exp\big(-2t^2/(n_A^{-1} + n_B^{-1})\big)$.

(b) Under the budget constraint $n_A + n_B = N$, which allocation minimises the bound? Would the answer change if the outcomes of A were known to lie in $[0,1]$ and those of B in $[0, \frac{1}{2}]$?

Exercise 1.17 (Choosing k and n for the k-NN rule)**.** Under Assumption 1.20 with $c = 1$, use (1.31) to determine how many observations are needed so that $R_n(\widehat{h}^{k\text{-NN}}) \le L^*_{\mathcal{D}} + \varepsilon$, and which k to use, as a function of d and ε. Evaluate for $\varepsilon = 0.1$ and $d = 2, 5, 10$, and compare with Remark 1.26.

Exercise 1.18 (A Gaussian bound without Chernoff)**.** Show that $\mathrm{erfc}(z) := \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-u^2}\,\mathrm{d}u \le e^{-z^2}$ for all $z \ge 0$, and recover the last inequality of Lemma 1.39. Compare, for t large and for t small, with the Chernoff bound of Example 1.60.

Chapter 2

The Cramér–Chernoff Method

Summary. Chapter 1 used the Chernoff method with a single, crude bound on the cumulant generating function. We now look at that function properly. It turns out to be a free energy, its derivatives are the mean and the variance of an exponentially tilted – Gibbs – measure, and its Legendre transform is the exact exponential rate of deviation. From this picture we extract the two classes of variables on which everything rests, sub-Gaussian and sub-gamma, we prove Bennett's and Bernstein's inequalities, we go beyond sums with the bounded-difference inequality, and we return to learning with variance-adaptive risk bounds.

2.1 What Hoeffding's inequality misses

Let $X_1, \dots, X_n$ be i.i.d. Ber(p) with p small – think of $p = 10^{-3}$, the probability that a given e-mail is misclassified, or that a given component fails. Hoeffding's inequality (1.28) gives

$$\mathbb{P}\big(\bar{X}_n \geq p + t\big) \;\leq\; e^{-2nt^2},$$

whereas the central limit theorem suggests $e^{-nt^2/(2p(1-p))}$, an exponent 250 times larger for $p = 10^{-3}$. The discrepancy is not a matter of constants. For $t = p$, that is, for the event that the empirical frequency is twice its expectation – a perfectly ordinary event to ask about – Hoeffding's bound is e^{-2np^2}, which requires $n \geq 2.3 \cdot 10^6$ samples to be smaller than 1%, while the exact Chernoff bound (2.12) requires only $n \geq 1.2 \cdot 10^4$: a factor of 193.

The reason was identified in Remark 1.63: Hoeffding's lemma bounds the cumulant generating function of a variable in $[a, b]$ by the one associated with the *largest possible* variance $(b-a)^2/4$. If we want inequalities that see the actual variance, we must look at the cumulant generating function itself. That is the programme of this chapter, and the first step is to understand what kind of object $\lambda \mapsto \log \mathbb{E}e^{\lambda Z}$ is.

2.2 The cumulant generating function as a free energy

Throughout this section Z is a real random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with law P. Recall from Definition 1.56 that $K_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$, and set

$$\mathrm{dom}(K_Z) := \{\lambda \in \mathbb{R} : \; K_Z(\lambda) < \infty\}\,.$$

Proposition 2.1 (Elementary properties)**.** $\mathrm{dom}(K_Z)$ *is an interval containing* 0*, and* K_Z *is convex and lower semicontinuous on* $\mathbb{R}$*. If* $\mathbb{E}Z$ *exists then* $K_Z(\lambda) \geq \lambda \mathbb{E}Z$ *for all* λ*, with equality at* $\lambda = 0$*.*

Proof. Let $\lambda_1, \lambda_2 \in \operatorname{dom}(K_Z)$ and $\theta \in (0,1)$. By Hölder's inequality with exponents $1/\theta$ and $1/(1-\theta)$,

$$\mathbb{E}\big[e^{(\theta\lambda_1+(1-\theta)\lambda_2)Z}\big] = \mathbb{E}\Big[\big(e^{\lambda_1 Z}\big)^{\theta}\big(e^{\lambda_2 Z}\big)^{1-\theta}\Big] \le \big(\mathbb{E}e^{\lambda_1 Z}\big)^{\theta}\big(\mathbb{E}e^{\lambda_2 Z}\big)^{1-\theta},$$

which is finite; taking logarithms gives both the convexity of $\operatorname{dom}(K_Z)$ and that of K_Z. Lower semicontinuity follows from Fatou's lemma. Finally $K_Z(\lambda) \ge \log e^{\lambda \mathbb{E}Z} = \lambda \mathbb{E}Z$ by Jensen's inequality, and $K_Z(0) = 0$. □

The next definition is the heart of the chapter.

Definition 2.2 (Exponential tilting, Gibbs measure)**.** Let $\lambda \in \operatorname{dom}(K_Z)$. The *$\lambda$-tilted* probability measure $\mathbb{P}_\lambda$ is defined by

$$\frac{\mathrm{d}\mathbb{P}_\lambda}{\mathrm{d}\mathbb{P}} := \frac{e^{\lambda Z}}{\mathbb{E}\big[e^{\lambda Z}\big]} = e^{\lambda Z - K_Z(\lambda)}. \tag{2.1}$$

We write $\mathbb{E}_\lambda$ and $\mathbb{V}\mathrm{ar}_\lambda$ for expectation and variance under $\mathbb{P}_\lambda$, and P_λ for the law of Z under $\mathbb{P}_\lambda$, the *tilted law*.

The density (2.1) is a Boltzmann–Gibbs weight: if one regards $-Z$ as an energy and λ as an inverse temperature (or, better, Z as an observable and λ as an external field conjugate to it), then $\mathbb{P}_\lambda$ is the equilibrium measure of the system in that field, $\mathbb{E}[e^{\lambda Z}]$ is its partition function and $K_Z(\lambda)$ its free energy up to sign and normalisation. The following theorem is the mathematical content of that analogy.

Theorem 2.3 (Smoothness and the meaning of the derivatives)**.** *Let λ_0 belong to the interior of* $\operatorname{dom}(K_Z)$*, assumed non-empty. Then:*

(i) Z has moments of every order under $\mathbb{P}_{\lambda_0}$, and K_Z is infinitely differentiable on $\operatorname{int}\operatorname{dom}(K_Z)$*, the derivatives being obtained by differentiating under the expectation sign;*

(ii) $K_Z'(\lambda) = \mathbb{E}_\lambda[Z]$ and $K_Z''(\lambda) = \mathbb{V}\mathrm{ar}_\lambda(Z) \ge 0$ for every $\lambda \in \operatorname{int}\operatorname{dom}(K_Z)$;

(iii) consequently K_Z is convex, $\lambda \mapsto \mathbb{E}_\lambda[Z]$ is non-decreasing, and K_Z is strictly convex on $\operatorname{int}\operatorname{dom}(K_Z)$ *unless Z is almost surely constant;*

(iv) $K_Z'(0) = \mathbb{E}[Z]$ and $K_Z''(0) = \mathbb{V}\mathrm{ar}(Z)$ when 0 is interior to $\operatorname{dom}(K_Z)$*.*

Proof. (i) Choose $r > 0$ with $\lambda_0 \pm 2r \in \operatorname{dom}(K_Z)$. For $|\lambda - \lambda_0| \le r$ and $z \in \mathbb{R}$, splitting according to the sign of z,

$$e^{\lambda z} \;\le\; e^{(\lambda_0+r)z} + e^{(\lambda_0-r)z},$$

and, since $e^u \ge u^k/k!$ for $u \ge 0$ gives $|z|^k \le \frac{k!}{r^k} e^{r|z|} \le \frac{k!}{r^k}\big(e^{rz} + e^{-rz}\big)$,

$$|z|^k\, e^{\lambda z} \;\le\; \frac{k!}{r^k}\big(e^{rz}+e^{-rz}\big)\big(e^{(\lambda_0+r)z}+e^{(\lambda_0-r)z}\big) \;\le\; \frac{k!}{r^k}\big(e^{(\lambda_0+2r)z} + 2e^{\lambda_0 z} + e^{(\lambda_0-2r)z}\big),$$

where λ_0 and $\lambda_0 \pm 2r$ belong to $\operatorname{dom}(K_Z)$. The right-hand side is $\mathbb{P}$-integrable, so the family $\{Z^k e^{\lambda Z}\}_{|\lambda-\lambda_0|\le r}$ is dominated by a fixed integrable variable. Differentiation under the integral sign is therefore legitimate at every order, and $\mathbb{E}_{\lambda_0}|Z|^k < \infty$ for all k.

(ii) Write $F_Z(\lambda) = \mathbb{E}[e^{\lambda Z}] > 0$. By (i), $F_Z'(\lambda) = \mathbb{E}[Z e^{\lambda Z}]$ and $F_Z''(\lambda) = \mathbb{E}[Z^2 e^{\lambda Z}]$, hence

$$K_Z'(\lambda) = \frac{F_Z'(\lambda)}{F_Z(\lambda)} = \mathbb{E}\Big[Z\, e^{\lambda Z - K_Z(\lambda)}\Big] = \mathbb{E}_\lambda[Z],$$

$$K_Z''(\lambda) = \frac{F_Z''(\lambda)}{F_Z(\lambda)} - \left(\frac{F_Z'(\lambda)}{F_Z(\lambda)}\right)^2 = \mathbb{E}_\lambda[Z^2] - \big(\mathbb{E}_\lambda[Z]\big)^2 = \mathbb{V}\mathrm{ar}_\lambda(Z).$$

(iii) A non-negative second derivative means convexity and a non-decreasing first derivative. If $K_Z''(\lambda) = 0$ for some interior λ, then Z is $\mathbb{P}_\lambda$-almost surely constant, hence $\mathbb{P}$-almost surely constant since $\mathbb{P}_\lambda \sim \mathbb{P}$. (iv) is the case $\lambda = 0$ of (ii). □

Remark 2.4 (The physical dictionary). Theorem 2.3 is a theorem of statistical mechanics in disguise, and the dictionary is worth writing down explicitly, because it is a reliable source of intuition – and, in later chapters, of technique.

Probability	**Statistical physics**
observable Z, field λ	magnetisation, external field
$\mathbb{E}[e^{\lambda Z}] = F_Z(\lambda)$ (Definition 1.56)	partition function
$K_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$	free energy (up to sign and temperature)
tilted measure $\mathbb{P}_\lambda$	Gibbs measure in the field λ
$K_Z'(\lambda) = \mathbb{E}_\lambda[Z]$	mean magnetisation
$K_Z''(\lambda) = \mathbb{V}\mathrm{ar}_\lambda(Z) \geq 0$	susceptibility; fluctuation–response
convexity of K_Z	thermodynamic stability
K_Z^* (Section 2.3)	entropy as a function of magnetisation

The identity $K_Z'' = \mathbb{V}\mathrm{ar}_\lambda(Z)$ is the fluctuation–dissipation relation: the response of the mean to an infinitesimal change of the field equals the fluctuation of the observable. Its non-negativity – a triviality here, a physical principle there – is exactly the convexity of the free energy. We shall use it constantly in the form: *to bound a cumulant generating function, it suffices to bound the variance of the tilted law*. This is precisely how we proved Hoeffding's lemma (Lemma 1.62, first proof), and how we shall prove Bennett's.

The identity $K_Z'' = \mathbb{V}\mathrm{ar}_\lambda(Z)$ has a consequence that will organise the whole chapter. Since $\psi_Z(0) = \psi_Z'(0) = 0$ and $\psi_Z'' = K_Z''$, a Taylor expansion with integral remainder expresses the cumulant generating function as an *average of tilted variances*.

Notation 2.5. For λ in the interior of $\mathrm{dom}(K_Z)$ we write

$$m(\lambda) := K_Z'(\lambda) = \mathbb{E}_\lambda[Z], \qquad \varsigma^2(\lambda) := K_Z''(\lambda) = \mathbb{V}\mathrm{ar}_\lambda(Z)$$

for the mean and the variance *under the tilted measure* $\mathbb{P}_\lambda$, keeping $\mu = \mathbb{E}Z$ and $\sigma^2 = \mathbb{V}\mathrm{ar}(Z)$ for the untilted ones. Thus $m(0) = \mu$, $\varsigma^2(0) = \sigma^2$, and m is increasing.

Proposition 2.6 (Variance representation of ψ_Z)**.** *Assume* $0 \in \mathrm{int}\,\mathrm{dom}(K_Z)$. *For every* λ *in the interior of* $\mathrm{dom}(K_Z)$,

$$\psi_Z(\lambda) = \int_0^\lambda (\lambda - \ell)\,\varsigma^2(\ell)\,\mathrm{d}\ell \;=\; \frac{\lambda^2}{2}\int_0^1 \varsigma^2(\lambda t)\,2(1-t)\,\mathrm{d}t. \tag{2.2}$$

Since $\int_0^1 2(1-t)\,\mathrm{d}t = 1$, *the right-hand side exhibits* $2\psi_Z(\lambda)/\lambda^2$ *as a weighted average of the variances of the tilted laws* P_ℓ, ℓ *between* 0 *and* λ. *In particular, if* $\varsigma^2(\ell) \leq v$ *for all such* ℓ, *then* $\psi_Z(\lambda) \leq \lambda^2 v/2$.

Proof. By Theorem 2.3, $\psi_Z = K_Z - \lambda\mathbb{E}Z$ is twice continuously differentiable on $\mathrm{int}\,\mathrm{dom}(K_Z)$, with $\psi_Z(0) = 0$, $\psi_Z'(0) = m(0) - \mu = 0$ and $\psi_Z'' = \varsigma^2$. Taylor's formula with integral remainder gives the first equality, and the substitution $\ell = \lambda t$ the second. □

Remark 2.7. Proposition 2.6 turns every bound on a cumulant generating function into a bound on a variance, which is a far more concrete object. All the inequalities of this chapter are obtained in this way:

$$\begin{aligned} &\varsigma^2(\ell) \le v \text{ for all } \ell && \rightsquigarrow \text{ sub-Gaussian, Hoeffding;} \\ &\varsigma^2(\ell) \le m(\ell) \text{ for all } \ell && \rightsquigarrow \text{ Bennett;} \\ &\varsigma^2(\ell) \le v\big(1 - c\ell\big)^{-3} && \rightsquigarrow \text{ sub-gamma, Bernstein.} \end{aligned}$$

The first proof of Hoeffding's lemma (Lemma 1.62) was exactly the first line, in the special case $v = (b-a)^2/4$ given by Popoviciu's inequality.

Example 2.8 (The basic cumulant generating functions)**.** We record, once and for all, the computations used throughout the chapter.

(a) $Z \sim \mathcal{N}(\mu, \sigma^2)$: $K_Z(\lambda) = \lambda\mu + \lambda^2\sigma^2/2$, $\operatorname{dom}(K_Z) = \mathbb{R}$; $\psi_Z(\lambda) = \lambda^2\sigma^2/2$. The tilted law is $\mathcal{N}(\mu + \lambda\sigma^2, \sigma^2)$: tilting *translates* a Gaussian.

(b) $Z \sim \mathrm{Ber}(p)$: $K_Z(\lambda) = \log\big(1 - p + pe^{\lambda}\big)$, $\operatorname{dom}(K_Z) = \mathbb{R}$. The tilted law is $\mathrm{Ber}(p_\lambda)$ with $p_\lambda = pe^{\lambda}/(1 - p + pe^{\lambda})$: tilting a Bernoulli variable changes its parameter, and $\lambda \mapsto p_\lambda$ is an increasing bijection from $\mathbb{R}$ onto $(0,1)$.

(c) $Z \sim \mathrm{Poi}(\sigma^2)$: $K_Z(\lambda) = \sigma^2(e^{\lambda} - 1)$, $\operatorname{dom}(K_Z) = \mathbb{R}$; the tilted law is $\mathrm{Poi}(\sigma^2 e^{\lambda})$.

(d) $Z \sim \mathrm{Exp}(1)$: $K_Z(\lambda) = -\log(1-\lambda)$ for $\lambda < 1$ and $+\infty$ for $\lambda \ge 1$; the tilted law is $\mathrm{Exp}(1-\lambda)$. Here $\operatorname{dom}(K_Z) = (-\infty, 1)$ is a genuine half-line: the domain of a cumulant generating function records how heavy the tail is.

(e) $Z = X^2$ with $X \sim \mathcal{N}(0,1)$: $K_Z(\lambda) = -\frac{1}{2}\log(1 - 2\lambda)$ for $\lambda < \frac{1}{2}$, $+\infty$ otherwise.

2.3 The Cramér transform and the Cramér–Chernoff inequality

Definition 2.9 (Cramér transform)**.** The *Cramér transform* of Z is the Legendre–Fenchel conjugate of its cumulant generating function,

$$K_Z^*(x) := \sup_{\lambda \in \mathbb{R}} \big\{\lambda x - K_Z(\lambda)\big\}, \qquad x \in \mathbb{R}.$$

When $\mathbb{E}Z$ exists we also write $\psi_Z^*(t) = \sup_{\lambda\in\mathbb{R}}\{\lambda t - \psi_Z(\lambda)\} = K_Z^*(\mathbb{E}Z + t)$ for the transform of the centred variable.

Proposition 2.10 (Properties of the Cramér transform)**.** *Assume $\mathbb{E}Z$ exists and write $\mu = \mathbb{E}Z$. Then*

(i) K_Z^* is convex, lower semicontinuous and non-negative, and $K_Z^*(\mu) = 0$;

(ii) K_Z^* is non-decreasing on $[\mu, +\infty)$ and non-increasing on $(-\infty, \mu]$;

(iii) if $\lambda_+ := \sup \operatorname{dom}(K_Z) > 0$, then for $x > \mu$, $K_Z^*(x) = \sup_{\lambda>0}\{\lambda x - K_Z(\lambda)\}$;

(iv) if $\lambda_x \in \operatorname{int}\operatorname{dom}(K_Z)$ satisfies $K_Z'(\lambda_x) = x$, then the supremum is attained at λ_x and

$$K_Z^*(x) = \lambda_x x - K_Z(\lambda_x), \qquad \text{and} \qquad \mathbb{E}_{\lambda_x}[Z] = x.$$

Proof. (i) A supremum of affine functions of x is convex and lower semicontinuous. Taking $\lambda = 0$ gives $K_Z^* \geq 0$. By Proposition 2.1, $\lambda\mu - K_Z(\lambda) \leq 0$ for every λ, with equality at $\lambda = 0$, so $K_Z^*(\mu) = 0$. (ii) By (i), K_Z^* is a non-negative convex function vanishing at μ, hence non-increasing before μ and non-decreasing after. (iii) For $x > \mu$ and $\lambda \leq 0$, Proposition 2.1 gives $\lambda x - K_Z(\lambda) \leq \lambda x - \lambda\mu = (x-\mu)\lambda \leq 0$. For the reverse inequality, K_Z is convex, finite on $[0, \lambda_+)$ and vanishes at 0, so $K_Z(\lambda) \leq (\lambda/\lambda_1)K_Z(\lambda_1)$ for $0 < \lambda < \lambda_1 < \lambda_+$ and $\limsup_{\lambda \downarrow 0} K_Z(\lambda) \leq 0$; with lower semicontinuity this gives $K_Z(\lambda) \to 0$ as $\lambda \to 0^+$, whence $\sup_{\lambda>0}\{\lambda x - K_Z(\lambda)\} \geq 0$. The hypothesis $\lambda_+ > 0$ is needed: if $K_Z = +\infty$ on $(0,\infty)$ that supremum is $-\infty$, while $K_Z^*(x) = 0$ by (i) and (ii). (iv) The function $\lambda \mapsto \lambda x - K_Z(\lambda)$ is concave and differentiable on $\operatorname{int}\operatorname{dom}(K_Z)$, with derivative $x - K_Z'(\lambda)$ vanishing at λ_x; the last identity is Theorem 2.3(ii). □

The Cramér transform admits a representation dual to (2.2), and just as useful: it is an average of the *inverses* of the tilted variances. In the rest of this section we assume, to avoid degenerate cases, that Z is not almost surely constant and that $(\lambda_-, \lambda_+) := \operatorname{int}\operatorname{dom}(K_Z)$ contains 0. By Theorem 2.3, $m = K_Z'$ is then a C^∞ increasing bijection from (λ_-, λ_+) onto an open interval $(x_-, x_+) \ni \mathbb{E}Z$; we write $\lambda(\cdot) = m^{-1}(\cdot)$ for its inverse, so that $\lambda(x) = \lambda_x$ is the tilt that makes x typical, as in Proposition 2.10(iv).

Proposition 2.11 (Variance representation of the Cramér transform)**.** *For every $x \in (x_-, x_+)$,*

$$\big(K_Z^*\big)'(x) = \lambda(x), \qquad \big(K_Z^*\big)''(x) = \frac{1}{\varsigma^2\big(\lambda(x)\big)}, \tag{2.3}$$

and consequently, with $\mu = \mathbb{E}Z$,

$$K_Z^*(x) = \int_\mu^x \frac{x-u}{\varsigma^2\big(\lambda(u)\big)}\,\mathrm{d}u = \frac{(x-\mu)^2}{2}\int_0^1 \frac{2(1-t)\,\mathrm{d}t}{\varsigma^2\big(\lambda(\mu + t(x-\mu))\big)}. \tag{2.4}$$

In particular, if $\varsigma^2(\ell) \leq v$ for every ℓ between 0 and $\lambda(x)$, then $K_Z^*(x) \geq (x-\mu)^2/(2v)$; and if $\varsigma^2(\ell) \geq w$ on that range, then $K_Z^*(x) \leq (x-\mu)^2/(2w)$.

Proof. By Proposition 2.10(iv), $K_Z^*(x) = \lambda(x)x - K_Z(\lambda(x))$ on (x_-, x_+). Differentiating and using $K_Z'(\lambda(x)) = x$,

$$\big(K_Z^*\big)'(x) = \lambda(x) + x\lambda'(x) - K_Z'\big(\lambda(x)\big)\lambda'(x) = \lambda(x),$$

and differentiating once more, $\big(K_Z^*\big)''(x) = \lambda'(x) = 1/m'(\lambda(x)) = 1/\varsigma^2(\lambda(x))$ by the inverse function theorem. Since $K_Z^*(\mu) = 0$ and $(K_Z^*)'(\mu) = \lambda(\mu) = 0$, Taylor's formula with integral remainder gives the first equality in (2.4), and the substitution $u = \mu + t(x-\mu)$ the second. The two consequences follow because $\int_0^1 2(1-t)\,\mathrm{d}t = 1$. □

Propositions 2.6 and 2.11 reduce the whole Chernoff programme to a single question: *how large can the variance of the tilted law be?* The following corollary answers it in the bounded case and, as a by-product, re-proves Hoeffding's lemma and proves Pinsker's inequality.

Corollary 2.12 (Hoeffding and Pinsker revisited)**.** *Let $a \leq Z \leq b$ almost surely, with $\mu = \mathbb{E}Z$. Then* $\operatorname{dom}(K_Z) = \mathbb{R}$,

$$\psi_Z(\lambda) \leq \frac{\lambda^2(b-a)^2}{8} \quad (\lambda \in \mathbb{R}), \qquad \text{and} \qquad K_Z^*(\mu+\varepsilon) \geq \frac{2\varepsilon^2}{(b-a)^2} \quad (\varepsilon > 0).$$

In particular, taking $Z \sim \mathrm{Ber}(p)$ and $b - a = 1$,

$$\mathrm{kl}(q,p) \geq 2(q-p)^2 \qquad \text{for all } p, q \in (0,1), \tag{2.5}$$

which is Pinsker's inequality for Bernoulli laws.

Proof. Each tilted measure $\mathbb{P}_\lambda$ is absolutely continuous with respect to $\mathbb{P}$, so $a \le Z \le b$ holds $\mathbb{P}_\lambda$-almost surely and Popoviciu's inequality gives $\varsigma^2(\lambda) \le (b-a)^2/4$ for every λ. Now apply Proposition 2.6 and Proposition 2.11 with $v = (b-a)^2/4$; the latter covers $\mu + \varepsilon \in (x_-, x_+) = (\operatorname{ess\,inf} Z, \operatorname{ess\,sup} Z)$, and the remaining values are covered as follows. If $\mu + \varepsilon > \operatorname{ess\,sup} Z$ then $(\mu+\varepsilon)\lambda - K_Z(\lambda) \ge \big(\mu + \varepsilon - \operatorname{ess\,sup} Z\big)\lambda \to +\infty$ as $\lambda \to +\infty$, so $K_Z^*(\mu+\varepsilon) = +\infty$; and the boundary point $\mu+\varepsilon = \operatorname{ess\,sup} Z$ follows by the lower semicontinuity of K_Z^* (Proposition 2.10(i)).

For the last statement, $K^*_{\mathrm{Ber}(p)}(q) = \mathrm{kl}(q,p)$ by Proposition 2.23(b), which gives (2.5) for $q > p$; the case $q < p$ follows from the identity $\mathrm{kl}(q,p) = \mathrm{kl}(1-q, 1-p)$ – apply what precedes to $1 - Z \sim \mathrm{Ber}(1-p)$ and the level $1-q > 1-p$ – and $q = p$ is trivial. □

Remark 2.13. Corollary 2.12 closes a loop opened in Chapter 1: Remark 1.67 asserted that Hoeffding's inequality is the exact Chernoff bound composed with Pinsker's inequality, and we can now see that both statements come from the *same* elementary fact – a variable confined to an interval of length $b-a$ has variance at most $(b-a)^2/4$, under every tilt. By the data-processing inequality, the Bernoulli case (2.5) also implies Pinsker's inequality $\mathrm{TV}(P,Q) \le \sqrt{\mathrm{KL}(P\|Q)/2}$ for arbitrary laws, where TV is the total variation distance (Definition 3.26 and Proposition 3.28). Everything that follows consists in replacing that worst-case bound on $\varsigma^2(\lambda)$ by a better one.

Theorem 2.14 (Cramér–Chernoff inequality)**.** *For every real random variable Z with $\mathbb{E}Z = \mu$ and every $x \ge \mu$,*

$$\mathbb{P}(Z \ge x) \;\le\; e^{-K_Z^*(x)}. \tag{2.6}$$

If moreover $Z = S_n = X_1 + \cdots + X_n$ with X_i i.i.d. copies of X, then $K_{S_n} = nK_X$, hence $K^*_{S_n}(nx) = n\,K_X^*(x)$ and

$$\mathbb{P}\big(\bar{X}_n \ge x\big) \;\le\; e^{-n\,K_X^*(x)}, \qquad x \ge \mathbb{E}X. \tag{2.7}$$

Proof. If $\lambda_+ = \sup \mathrm{dom}(K_Z) > 0$ this is immediate from Theorem 1.57 and Proposition 2.10(iii). If $\lambda_+ = 0$, then $K_Z^*(x) = 0$ for every $x > \mu$ by Proposition 2.10(i)–(ii) and the bound reads $\mathbb{P}(Z \ge x) \le 1$: it is true and empty. The identity $K^*_{S_n}(nx) = \sup_\lambda\{n\lambda x - nK_X(\lambda)\} = nK_X^*(x)$ uses (1.23). □

Remark 2.15 (When the bound says nothing). The second case above is not a curiosity: it is exactly the situation of Counterexample 1.41 and Exercise 1.4, where Z has no exponential moment at all. The Chernoff method is then vacuous, and Remark 1.49 is the right tool instead. One may read $\lambda_+ > 0$ as the precise hypothesis under which the whole of this chapter has content.

Remark 2.16 (Tilt until the deviation becomes typical). Proposition 2.10(iv) is the conceptual core of the method. To bound the probability that Z exceeds x, one changes the measure to $\mathbb{P}_{\lambda_x}$, where λ_x is chosen so that $\mathbb{E}_{\lambda_x}[Z] = x$: under the tilted measure, the deviation we are asking about is no longer a deviation at all – it is the typical behaviour. The price of the change of measure is the likelihood ratio, and the logarithm of its typical value is exactly

$$\mathbb{E}_{\lambda_x}\left[\log \frac{\mathrm{d}\mathbb{P}_{\lambda_x}}{\mathrm{d}\mathbb{P}}\right] = \lambda_x \mathbb{E}_{\lambda_x}[Z] - K_Z(\lambda_x) = \lambda_x x - K_Z(\lambda_x) = K_Z^*(x).$$

The Cramér transform is therefore a *relative entropy*: the Kullback–Leibler divergence of the tilted law with respect to the original one. (For probability measures $Q \ll P$, the relative entropy is $\mathrm{KL}(Q\|P) = \int \log \frac{\mathrm{d}Q}{\mathrm{d}P}\,\mathrm{d}Q \in [0, +\infty]$, and $\mathrm{KL}(Q\|P) = +\infty$ if $Q \not\ll P$; it is non-negative and vanishes only for $Q = P$. These facts are proved in Section 3.3, see

Definition 3.12 and Proposition 3.16.) This is not an accident, and Proposition 2.17 below turns it into an identity: $K_Z^*(x)$ is the *smallest* relative entropy of a law under which the deviation is not a deviation. The picture explains both why the bound (2.6) is natural and why, as Theorem 2.19 says, it is exponentially exact.

Proposition 2.17 (The Cramér transform is a relative entropy)**.** *Let P be the law of Z and, for $x \in (\mathbb{E}Z, x_+)$, let $\lambda(x) > 0$ and $P_{\lambda(x)}$ be as above. Then*

$$\mathrm{KL}\big(P_{\lambda(x)} \,\|\, P\big) = \lambda(x)\, x - K_Z\big(\lambda(x)\big) = K_Z^*(x), \tag{2.8}$$

and

$$K_Z^*(x) = \inf\Big\{\mathrm{KL}(Q\|P) :\ Q \ll P,\ \mathbb{E}_Q[Z] \text{ well defined and } \geq x\Big\}, \tag{2.9}$$

the infimum being attained at $Q = P_{\lambda(x)}$.

Proof. Write $\lambda = \lambda(x)$. By (2.1), $\log \frac{\mathrm{d}P_\lambda}{\mathrm{d}P}(Z) = \lambda Z - K_Z(\lambda)$, so

$$\mathrm{KL}\big(P_\lambda \| P\big) = \mathbb{E}_\lambda\big[\lambda Z - K_Z(\lambda)\big] = \lambda\, m(\lambda) - K_Z(\lambda) = \lambda x - K_Z(\lambda) = K_Z^*(x),$$

using $m(\lambda(x)) = x$ and Proposition 2.10(iv); this is (2.8), and it shows that the infimum in (2.9) is at most $K_Z^*(x)$.

Conversely, let $Q \ll P$ be such that $\mathbb{E}_Q[Z]$ is defined in $(-\infty, +\infty]$ and $\mathbb{E}_Q[Z] \geq x$. Since P_λ and P are equivalent, $Q \ll P_\lambda$, and the non-negativity of $\mathrm{KL}(Q\|P_\lambda)$ (Proposition 3.16) reads

$$\mathrm{KL}(Q\|P) \ \geq\ \lambda\, \mathbb{E}_Q[Z] - K_Z(\lambda), \tag{2.10}$$

the Donsker–Varadhan inequality: when both divergences are finite it is the identity $\mathrm{KL}(Q\|P) = \mathrm{KL}(Q\|P_\lambda) + \lambda\mathbb{E}_Q[Z] - K_Z(\lambda)$, obtained from $\frac{\mathrm{d}Q}{\mathrm{d}P} = \frac{\mathrm{d}Q}{\mathrm{d}P_\lambda}\frac{\mathrm{d}P_\lambda}{\mathrm{d}P}$ and $\log \frac{\mathrm{d}P_\lambda}{\mathrm{d}P} = \lambda Z - K_Z(\lambda)$, and it is trivial otherwise. Since $\lambda > 0$ and $\mathbb{E}_Q[Z] \geq x$, the right-hand side of (2.10) is at least $\lambda x - K_Z(\lambda) = K_Z^*(x)$. □

Remark 2.18 (Why this matters). Identity (2.9) is the simplest instance of the *Gibbs variational principle*, and it is the bridge between the Chernoff method and information theory: the exponential cost of a deviation is the minimal amount of relative entropy needed to make that deviation typical. Read backwards, it is also the germ of the entropy method: bounding a cumulant generating function amounts to bounding a relative entropy, and one may then use the whole arsenal of entropy inequalities – tensorisation, log-Sobolev inequalities – instead of the tilting computations of this chapter. We shall follow that road in a later chapter; for sums of independent variables the elementary route below is shorter.

Theorem 2.19 (Cramér's theorem)**.** *Let $X, X_1, X_2, \dots$ be i.i.d. real random variables with $\mathbb{E}X = \mu$ and let $S_n = \sum_{i\leq n} X_i$. For every $x \geq \mu$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \geq nx\big) \ \leq\ -K_X^*(x),$$

and if x belongs to the interior of $\{K_X^* < \infty\}$ then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \geq nx\big) = -K_X^*(x).$$

Proof of the upper bound, and sketch of the lower bound. The upper bound is (2.7). For the lower bound, which we admit and for which we refer to [DZ98, Thm. 2.2.3] or [Fel68, Chap. 27], the idea is exactly Remark 2.16. Suppose for simplicity that λ_x with $K_X'(\lambda_x) = x$ exists in the

interior of $\mathrm{dom}(K_X)$, and let $\mathbb{P}_{\lambda_x}$ be the corresponding tilted measure, under which X has mean x and finite variance $\varsigma^2(\lambda_x)$. Writing the event in terms of the likelihood ratio and restricting to $S_n \in [nx, nx + a\sqrt{n}]$,

$$\begin{aligned}\mathbb{P}(S_n \geq nx) &\geq \mathbb{E}_{\lambda_x}\Big[e^{-\lambda_x S_n + nK_X(\lambda_x)}\mathbb{1}\left\{S_n \in [nx, nx + a\sqrt{n}]\right\}\Big] \\ &\geq e^{-nK_X^*(x) - \lambda_x a\sqrt{n}}\,\mathbb{P}_{\lambda_x}\big(S_n \in [nx, nx + a\sqrt{n}]\big),\end{aligned}$$

and the last probability converges, by the central limit theorem under $\mathbb{P}_{\lambda_x}$, to a positive constant depending on $a/\varsigma(\lambda_x)$. Taking logarithms, dividing by n and letting $n \to \infty$ gives the lower bound. □

Remark 2.20 (What is proved, and where). The lower bound sketched above is proved in full in Chapter 3: Theorem 3.42 and Corollary 3.44 treat Bernoulli variables, and Proposition 3.45 turns the sketch into a proof for every $x \in [\mu, x_+)$, where $x_+ = \lim_{\lambda \uparrow \lambda_+} K_X'(\lambda)$, as soon as $K_X(\lambda) < \infty$ for some $\lambda > 0$. This covers the whole statement when $\mathrm{dom}(K_X) \supset [0, \infty)$ – bounded, finitely-valued, Gaussian, Poisson and all sub-Gaussian variables – and more generally when K_X is steep, i.e. $K_X'(\lambda) \to \sup\{K_X^* < \infty\}$ as $\lambda \uparrow \lambda_+$ (Remark 3.46). Only the residual non-steep and heavy-tailed cases remain admitted, with the reference [DZ98, Thm. 2.2.3]. A non-asymptotic version is Theorem 3.47 for Bernoulli variables, with the explicit second-order term of Corollary 3.48, and Remark 3.50 for general laws.

Remark 2.21 (Exactly the right exponent, and not more). Theorem 2.19 says that the Chernoff method loses *nothing* at the exponential scale: $e^{-nK^*(x)}$ is the truth. It also says nothing about subexponential factors, and there the bound does lose: the sharp asymptotics, due to Bahadur and Rao [BR60], read

$$\mathbb{P}\big(S_n \geq nx\big) \;\sim\; \frac{e^{-nK_X^*(x)}}{\lambda_x\,\varsigma(\lambda_x)\sqrt{2\pi n}} \qquad (n \to \infty)$$

in the non-lattice case, with λ_x and $\varsigma^2(\lambda_x)$ as in the proof above. The missing factor $\asymp n^{-1/2}$ is exactly the one we already saw in the Gaussian case (Example 1.60). We shall not use Bahadur–Rao asymptotics, but it is useful to know what the Chernoff bound costs: a polynomial factor, never an exponential one.

2.3.1 The basic Cramér transforms

We now compute K^* in the cases that matter. The Bernoulli computation deserves a name of its own, since the function it produces will reappear in every comparison of this chapter.

Notation 2.22 (Bernoulli relative entropy). For $p, q \in (0,1)$ set

$$\mathrm{kl}(q,p) := q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p},$$

extended by $\mathrm{kl}(0,p) = \log\frac{1}{1-p}$ and $\mathrm{kl}(1,p) = \log\frac{1}{p}$. This is the Kullback–Leibler divergence $\mathrm{KL}\big(\mathrm{Ber}(q)\|\mathrm{Ber}(p)\big)$; it is non-negative, vanishes only at $q = p$, and is strictly convex in q.

Proposition 2.23 (Cramér transforms of the basic laws). *(a)* $X \sim \mathcal{N}(\mu, \sigma^2)$: $K_X^*(x) = (x-\mu)^2/(2\sigma^2)$ *for all* $x \in \mathbb{R}$.

(b) $X \sim \mathrm{Ber}(p)$: $K_X^*(q) = \mathrm{kl}(q,p)$ *for* $q \in [0,1]$, *and* $+\infty$ *outside; the optimal tilt is* $\lambda_q = \log\frac{q(1-p)}{p(1-q)}$.

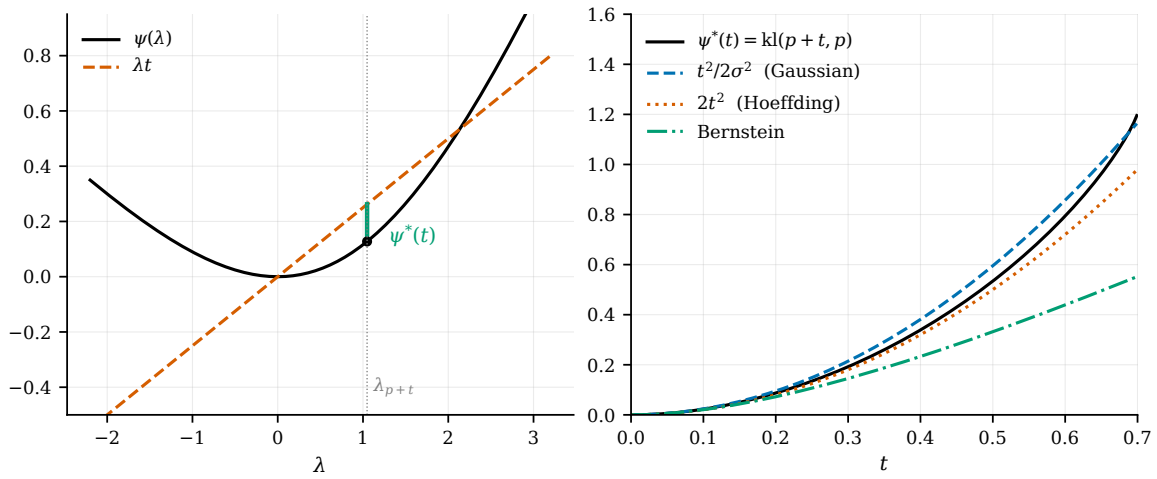


Figure 2.1: The Cramér transform for $X \sim \mathrm{Ber}(0.3)$. *Left:* the centred cumulant generating function ψ (a convex function vanishing to second order at the origin, by Proposition 2.6) and the line $\lambda \mapsto \lambda t$; $\psi^*(t)$ is the largest vertical gap between the two, attained at the tilt λ_{p+t} for which the tilted mean equals $p + t$ (Proposition 2.10(iv)). *Right:* $\psi^*(t) = \mathrm{kl}(p + t, p)$ together with the three lower bounds that produce, respectively, the Gaussian heuristic, Bernstein's inequality and Hoeffding's. Near $t = 0$ all four agree to second order, with curvature $1/\sigma^2$; far from 0 only the exact transform captures the true growth. Generated by `figures/fig_legendre.py`.

(c) $X \sim \mathrm{Poi}(\sigma^2)$: $K_X^*(x) = x \log \frac{x}{\sigma^2} - x + \sigma^2 = \sigma^2\, h\big(\frac{x}{\sigma^2} - 1\big)$ *for* $x > 0$, *where*

$$h(u) := (1+u)\log(1+u) - u, \qquad u \geq -1. \tag{2.11}$$

(d) $X \sim \mathrm{Exp}(1)$: $K_X^*(x) = x - 1 - \log x$ *for* $x > 0$.

Proof. (a) $\lambda x - \lambda\mu - \lambda^2\sigma^2/2$ is maximal at $\lambda = (x-\mu)/\sigma^2$, with value $(x-\mu)^2/(2\sigma^2)$. (b) With $K_X(\lambda) = \log(1-p+pe^\lambda)$, the derivative of $\lambda q - K_X(\lambda)$ is $q - pe^\lambda/(1-p+pe^\lambda)$, which vanishes when $pe^\lambda/(1-p+pe^\lambda) = q$, that is $e^\lambda = \frac{q(1-p)}{p(1-q)}$. Substituting, $1-p+pe^{\lambda_q} = 1-p+\frac{q(1-p)}{1-q} = \frac{1-p}{1-q}$, so

$$K_X^*(q) = q\log\frac{q(1-p)}{p(1-q)} - \log\frac{1-p}{1-q} = q\log\frac{q}{p} + q\log\frac{1-p}{1-q} + \log\frac{1-q}{1-p} = \mathrm{kl}(q,p).$$

(c) $\lambda x - \sigma^2(e^\lambda - 1)$ is maximal at $e^\lambda = x/\sigma^2$, with value $x\log(x/\sigma^2) - x + \sigma^2$; the second form follows by substituting $u = x/\sigma^2 - 1$ in (2.11). (d) $\lambda x + \log(1-\lambda)$ is maximal at $\lambda = 1 - 1/x$ (which is < 1), with value $x - 1 - \log x$. □

Combining (b) with (2.7) gives the exact Chernoff bound for Bernoulli sums, which will serve as our benchmark throughout:

$$X_1, \dots, X_n \text{ i.i.d. } \mathrm{Ber}(p), \qquad \mathbb{P}\big(\bar{X}_n \geq q\big) \;\leq\; e^{-n\,\mathrm{kl}(q,p)} \quad \text{for } q \in (p, 1). \tag{2.12}$$

2.4 Sub-Gaussian random variables

The Gaussian is the benchmark: its centred cumulant generating function is exactly $\lambda^2\sigma^2/2$ (Example 2.8). The class of variables that do no worse deserves a name.

Definition 2.24 (Sub-Gaussian variable)**.** Let $v > 0$. A random variable Z with $\mathbb{E}Z$ finite is *sub-Gaussian with variance proxy* v, written $Z \in \mathrm{subG}(v)$, if

$$\psi_Z(\lambda) = \log \mathbb{E}\big[e^{\lambda(Z-\mathbb{E}Z)}\big] \;\le\; \frac{\lambda^2 v}{2} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{2.13}$$

Proposition 2.25 (Basic properties)**.** *Let $Z \in \mathrm{subG}(v)$ with $\mathbb{E}Z = \mu$.*

(i) (Tails.) *For every $t > 0$,*

$$\mathbb{P}(Z - \mu \ge t) \le e^{-t^2/(2v)}, \qquad \mathbb{P}\big(|Z - \mu| \ge t\big) \le 2e^{-t^2/(2v)}.$$

Equivalently, with probability at least $1 - \delta$, $|Z - \mu| \le \sqrt{2v \log(2/\delta)}$.

(ii) (The proxy dominates the variance.) $\mathbb{V}\mathrm{ar}(Z) \le v$.

(iii) (Scaling and sums.) *$aZ + b \in \mathrm{subG}(a^2 v)$ for $a, b \in \mathbb{R}$; and if $Z_1, \dots, Z_n$ are* independent *with $Z_i \in \mathrm{subG}(v_i)$, then $\sum_i Z_i \in \mathrm{subG}\big(\sum_i v_i\big)$. In particular, if the Z_i are i.i.d. in $\mathrm{subG}(v)$ then $\bar{Z}_n \in \mathrm{subG}(v/n)$ and*

$$\mathbb{P}\big(|\bar{Z}_n - \mu| \ge t\big) \;\le\; 2e^{-nt^2/(2v)}.$$

Proof. (i) By Theorem 1.57, $\mathbb{P}(Z - \mu \ge t) \le \exp(-\lambda t + \lambda^2 v/2)$ for every $\lambda > 0$; the exponent is minimal at $\lambda = t/v$, with value $-t^2/(2v)$. Applying this to $-Z$, which is in $\mathrm{subG}(v)$ because (2.13) holds for all $\lambda \in \mathbb{R}$, gives the lower tail. (ii) Since (2.13) holds for every real λ, we have $\mathrm{dom}(K_Z) = \mathbb{R}$ and Theorem 2.3 applies at $\lambda = 0$: $\psi_Z(\lambda) = \frac{\lambda^2}{2}\,\mathbb{V}\mathrm{ar}(Z) + o(\lambda^2)$ as $\lambda \to 0$, so $\mathbb{V}\mathrm{ar}(Z) = \lim_{\lambda \to 0} 2\psi_Z(\lambda)/\lambda^2 \le v$. (iii) $\psi_{aZ+b}(\lambda) = \psi_Z(a\lambda) \le a^2\lambda^2 v/2$, and additivity is (1.23). □

Example 2.26 (The basic sub-Gaussian variables)**.** (a) $Z \sim \mathcal{N}(\mu, \sigma^2)$ is in $\mathrm{subG}(\sigma^2)$, with equality in (2.13) for every λ: the variance proxy is the variance.

(b) Any Z with $a \le Z \le b$ almost surely is in $\mathrm{subG}\big((b-a)^2/4\big)$: this is Hoeffding's lemma (Lemma 1.62). Theorem 1.64 is therefore the special case of Proposition 2.25(iii) obtained by using this proxy.

(c) A Rademacher variable is in $\mathrm{subG}(1)$, and this is the exact variance proxy (Exercise 1.6). Note that (b) gives the same value $(1 - (-1))^2/4 = 1$: for symmetric two-point variables Hoeffding's lemma is exact.

(d) If $\xi_1, \dots, \xi_n$ are independent Rademacher variables and $a \in \mathbb{R}^n$, then $\sum_i a_i \xi_i \in \mathrm{subG}(\|a\|_2^2)$, whence $\mathbb{P}\big(|\sum_i a_i \xi_i| \ge t\big) \le 2e^{-t^2/(2\|a\|_2^2)}$.

Counterexample 2.27 (Squares of Gaussians are not sub-Gaussian)**.** Let $X \sim \mathcal{N}(0,1)$ and $Z = X^2$. By Example 2.8(e), $\mathbb{E}[e^{\lambda Z}] = +\infty$ for $\lambda \ge \frac{1}{2}$, so (2.13) fails for every v: a sub-Gaussian variable has a cumulant generating function finite on the whole line. The tail of Z is genuinely exponential rather than Gaussian, since $\mathbb{P}(X^2 > t) = 2\mathbb{P}(X > \sqrt{t}) \asymp t^{-1/2}e^{-t/2}$. Such variables are the object of Section 2.5.

The converse implication fails too: a finite cumulant generating function does not make a variable sub-Gaussian. If $Z \sim \mathrm{Poi}(\sigma^2)$ then $\psi_Z(\lambda) = \sigma^2(e^\lambda - 1 - \lambda)$ is finite for all λ but grows faster than any quadratic as $\lambda \to +\infty$. The right-hand tail of a Poisson variable is lighter than Gaussian close to the mean and heavier far from it – which is exactly the content of Bennett's inequality below.

Proposition 2.28 (Maximum of sub-Gaussian variables)**.** *Let $Z_1, \dots, Z_N$ be centred random variables,* not assumed independent, *with $Z_i \in \mathrm{subG}(v)$ for every i. Then*

$$\mathbb{E}\Big[\max_{1\le i\le N} Z_i\Big] \;\le\; \sqrt{2v\log N}, \qquad \text{and} \qquad \mathbb{E}\Big[\max_{1\le i\le N} |Z_i|\Big] \;\le\; \sqrt{2v\log(2N)}.$$

Proof. For $\lambda > 0$, Jensen's inequality gives

$$\exp\Big(\lambda\,\mathbb{E}\max_i Z_i\Big) \le \mathbb{E}\Big[e^{\lambda \max_i Z_i}\Big] = \mathbb{E}\Big[\max_i e^{\lambda Z_i}\Big] \le \sum_{i=1}^{N} \mathbb{E}\big[e^{\lambda Z_i}\big] \le N e^{\lambda^2 v/2},$$

hence $\mathbb{E}\max_i Z_i \le \frac{\log N}{\lambda} + \frac{\lambda v}{2}$ for every $\lambda > 0$. Choosing $\lambda = \sqrt{2\log N / v}$ gives the first bound (this was Exercise 1.9). The second follows by applying the first to the $2N$ variables $\pm Z_1, \dots, \pm Z_N$. □

Remark 2.29 ($\sqrt{\log N}$ is the truth). The bound is of the correct order: if the Z_i are i.i.d. $\mathcal{N}(0, v)$ then $\mathbb{E}\max_i Z_i \sim \sqrt{2v\log N}$ as $N \to \infty$. Proposition 2.28 is the prototype of all the "union bound at the level of expectations" arguments of statistical learning: the price of taking a maximum over N objects is $\sqrt{\log N}$, which is why bounds of the form $\sqrt{\log|\mathcal{H}|\,/n}$ appear everywhere in Section 2.7.1, and why controlling an *infinite* class requires replacing $\log N$ by an entropy or a chaining argument. Chapter 4 develops the first: there Proposition 2.28 reappears as Massart's lemma (Lemma 4.44), and $\log N$ becomes the logarithm of the number of patterns of the class on the sample, at most $\mathrm{VCdim}(\mathcal{H})\log(en/\mathrm{VCdim}(\mathcal{H}))$ for a class $\mathcal{H}$ with $\mathrm{VCdim}(\mathcal{H}) \le n$ (Corollary 4.45).

2.4.1 Equivalent characterisations

In practice one rarely verifies (2.13) directly. The following theorem says that sub-Gaussianity can equivalently be read on the tails, on the moments, or on the exponential moment of the square – at the price of absolute constants, which is the one place in this chapter where we cannot keep track of the exact numerology and should not pretend to.

Theorem 2.30 (Characterisations of sub-Gaussianity)**.** *Let Z be a real random variable. Consider the following properties.*

(a) (Tails.) *There is $C_1 > 0$ with $\mathbb{P}(|Z| \ge t) \le 2e^{-t^2/C_1^2}$ for all $t \ge 0$.*

(b) (Moments.) *There is $C_2 > 0$ with $\big(\mathbb{E}|Z|^q\big)^{1/q} \le C_2\sqrt{q}$ for all $q \ge 1$.*

(c) (Exponential moment of the square.) *There is $C_3 > 0$ with $\mathbb{E}\big[e^{Z^2/C_3^2}\big] \le 2$.*

(d) (Moment generating function.) *$\mathbb{E}Z = 0$ and there is $C_4 > 0$ with $\mathbb{E}\big[e^{\lambda Z}\big] \le e^{C_4^2\lambda^2}$ for all $\lambda \in \mathbb{R}$ – that is, $Z \in \mathrm{subG}(2C_4^2)$.*

Then (a) ⇒ (b) ⇒ (c) *with $C_2 \le 2C_1$ and $C_3 \le 2\sqrt{e}\,C_2$; if in addition $\mathbb{E}Z = 0$ then* (c) ⇒ (d) *with $C_4 \le C_3$; and* (d) ⇒ (a) *with $C_1 \le 2C_4$. In particular, for centred variables the four properties are equivalent, and the four parameters are equivalent up to absolute multiplicative constants.*

Proof. **(a) ⇒ (b).** Using $\mathbb{E}|Z|^q = \int_0^\infty q t^{q-1}\mathbb{P}(|Z| > t)\,\mathrm{d}t$ and the substitution $u = t^2/C_1^2$,

$$\mathbb{E}|Z|^q \;\le\; 2q\int_0^\infty t^{q-1}e^{-t^2/C_1^2}\,\mathrm{d}t = q\,C_1^q\int_0^\infty u^{q/2-1}e^{-u}\,\mathrm{d}u = q\,C_1^q\,\Gamma\big(q/2\big).$$

If $q \geq 2$ then $\Gamma(q/2) \leq (q/2)^{q/2}$, because $\Gamma(x) \leq x^x$ for $x \geq 1$: indeed $\Gamma(x) \leq \Gamma(\lceil x \rceil) = (\lceil x \rceil - 1)! \leq j^j \leq x^x$ with $j = \lceil x \rceil - 1 \leq x$, using that $\Gamma \leq 1$ on $[1,2]$ and that Γ is increasing on $[2,\infty)$. Hence

$$\left(\mathbb{E}|Z|^q\right)^{1/q} \leq C_1 q^{1/q}\sqrt{q/2} \leq C_1 e^{1/e}\sqrt{q/2} \leq 2C_1\sqrt{q}.$$

If $1 \leq q < 2$ then $\Gamma(q/2) = \Gamma(q/2+1)/(q/2) \leq 2/q$ since $\Gamma \leq 1$ on $[1,2]$, so $q\Gamma(q/2) \leq 2$ and $\left(\mathbb{E}|Z|^q\right)^{1/q} \leq C_1 2^{1/q} \leq 2C_1 \leq 2C_1\sqrt{q}$.

(b) $\Rightarrow$ (c). Expanding the exponential and using (b) with $q = 2k$, together with $k! \geq (k/e)^k$,

$$\mathbb{E}\left[e^{Z^2/C_3^2}\right] = 1 + \sum_{k\geq 1} \frac{\mathbb{E}[Z^{2k}]}{k!\, C_3^{2k}} \leq 1 + \sum_{k\geq 1} \frac{C_2^{2k}(2k)^k}{k!\, C_3^{2k}} \leq 1 + \sum_{k\geq 1} \left(\frac{2eC_2^2}{C_3^2}\right)^k.$$

Choosing $C_3^2 = 4eC_2^2$ makes the ratio equal to $\frac{1}{2}$ and the sum equal to 1, so that the left-hand side is at most 2.

(c) $\Rightarrow$ (d), assuming $\mathbb{E}Z = 0$. We use the elementary inequality

$$e^u \;\leq\; u + e^{u^2}, \qquad u \in \mathbb{R}, \tag{2.14}$$

proved below. If $|\lambda| \leq 1/C_3$, then $a := \lambda^2 C_3^2 \leq 1$ and, by (2.14) with $u = \lambda Z$, by $\mathbb{E}Z = 0$ and by Jensen's inequality applied to the concave function $x \mapsto x^a$,

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq \mathbb{E}\left[e^{\lambda^2 Z^2}\right] = \mathbb{E}\left[\left(e^{Z^2/C_3^2}\right)^a\right] \leq \left(\mathbb{E}\left[e^{Z^2/C_3^2}\right]\right)^a \leq 2^a = e^{\lambda^2 C_3^2 \log 2}.$$

If $|\lambda| \geq 1/C_3$, we use $\lambda z \leq \lambda^2 C_3^2/4 + z^2/C_3^2$ (which is $0 \leq (\lambda C_3/2 - z/C_3)^2$) and $2 \leq e^{\lambda^2 C_3^2 \log 2}$:

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq e^{\lambda^2 C_3^2/4}\mathbb{E}\left[e^{Z^2/C_3^2}\right] \leq 2e^{\lambda^2 C_3^2/4} \leq e^{\lambda^2 C_3^2(\frac{1}{4}+\log 2)}.$$

In both cases $\mathbb{E}[e^{\lambda Z}] \leq e^{C_4^2\lambda^2}$ with $C_4^2 = (\frac{1}{4} + \log 2)C_3^2 \leq C_3^2$.

(d) $\Rightarrow$ (a). By Theorem 1.57, $\mathbb{P}(Z \geq t) \leq \exp(-\lambda t + C_4^2\lambda^2)$, minimal at $\lambda = t/(2C_4^2)$ with value $e^{-t^2/(4C_4^2)}$; the same holds for $-Z$ since (d) is invariant under $\lambda \mapsto -\lambda$. Adding the two tails gives (a) with $C_1 = 2C_4$.

Proof of (2.14). Let $f(u) = e^{u^2} + u - e^u$, so $f(0) = 0$ and $f'(0) = 0$. Then $f''(u) = (2+4u^2)e^{u^2} - e^u$. For $u \leq 0$, $e^u \leq 1 \leq 2e^{u^2}$. For $0 \leq u \leq \log 2$, $e^u \leq 2 \leq (2+4u^2)e^{u^2}$. For $u \geq 1$, $u^2 \geq u$ so $e^u \leq e^{u^2} \leq (2+4u^2)e^{u^2}$. For $\log 2 \leq u \leq 1$, $(2+4u^2)e^{u^2} \geq (2+4\log^2 2)e^{\log^2 2} \geq 3.9 > e \geq e^u$. Hence $f'' \geq 0$ everywhere, f is convex with a critical point at 0, so $f \geq f(0) = 0$. □

[V1 – shorter and sharper: go from (a) to (c) directly] The detour through the moments costs the Γ-function estimates and loses a factor ≈ 6.6. Here is the restructured theorem, written out.

Theorem 2.30′. *With (a), (c), (d) as in the text,* (a) $\Rightarrow$ (c) *with* $C_3 \leq \sqrt{3}C_1$; (c) $\Rightarrow$ (d) *with* $C_4 \leq C_3$ *if* $\mathbb{E}Z = 0$; *and* (d) $\Rightarrow$ (a) *with* $C_1 \leq 2C_4$. *Moreover* (c) $\Rightarrow$ (b) *with* $C_2 \leq 2C_3$.

Proof of (a) $\Rightarrow$ (c). Writing $e^{Z^2/s^2} - 1 = \int_0^{|Z|} \frac{2t}{s^2} e^{t^2/s^2}\,\mathrm{d}t$ and applying Fubini,

$$\mathbb{E}\left[e^{Z^2/s^2}\right] = 1 + \int_0^\infty \frac{2t}{s^2} e^{t^2/s^2}\,\mathbb{P}(|Z| > t)\,\mathrm{d}t \;\leq\; 1 + \frac{4}{s^2}\int_0^\infty t\, e^{-t^2\left(\frac{1}{C_1^2} - \frac{1}{s^2}\right)}\,\mathrm{d}t = 1 + \frac{2}{\frac{s^2}{C_1^2} - 1}$$

for every $s > C_1$. Taking $s^2 = 3C_1^2$ makes the right-hand side equal to 2. □

Proof of (c) $\Rightarrow$ (b). From $e^{Z^2/C_3^2} \geq \frac{1}{k!}\left(Z^2/C_3^2\right)^k$ we get $\mathbb{E}|Z|^{2k} \leq 2\,k!\,C_3^{2k} \leq 2k^k C_3^{2k}$, so that for

even $q = 2k$,

$$\big(\mathbb{E}\,|Z|^q\big)^{1/q} \le 2^{1/q} C_3 \sqrt{q/2} \le 2C_3\sqrt{q},$$

and for general $q \ge 1$ by monotonicity of the L^q norms. □

The implications (c) ⇒ (d) and (d) ⇒ (a), and the proof of (2.14), are unchanged. What disappears is the whole of the present (a) ⇒ (b) – the bound $\mathbb{E}\,|Z|^q \le q C_1^q \Gamma(q/2)$, the estimate $\Gamma(x) \le x^x$, and the separate case $1 \le q < 2$ – and the constants improve from $C_3 \le 2\sqrt{e}\, C_2 \le 4\sqrt{e}\, C_1 \approx 6.6 C_1$ to $C_3 \le \sqrt{3} C_1$.

Review of 18 September: not yet discussed.

Definition 2.31 (The sub-Gaussian norm)**.** The quantity $\|Z\|_{\mathrm{sG}} := \inf\{s > 0 : \mathbb{E}[e^{Z^2/s^2}] \le 2\}$ is called the *sub-Gaussian norm* of Z; it is the Orlicz norm associated with $x \mapsto e^{x^2} - 1$, usually written $\|Z\|_{\psi_2}$, a notation we avoid since ψ is our centred cumulant generating function. It is indeed a norm on the space of sub-Gaussian variables (Exercise 2.3), and Theorem 2.30 says that all the parameters $C_1, \dots, C_4$ are equivalent to it up to absolute constants.

Remark 2.32 (Two cultures). The reader will meet two styles in the literature. The *variance proxy* style of Definition 2.24 keeps explicit constants and gives bounds such as $e^{-t^2/2v}$ that are sharp for Gaussian variables; it is the style of [BLM13] and of this book. The *Orlicz norm* style of Definition 2.31 is insensitive to constants, writes $a \lesssim b$ throughout and gains enormously in flexibility – it is the style of [Ver18; Ver12], and it is the right one when one studies random matrices or random projections, where the constants are hopeless anyway. Neither is better; but mixing them silently is a reliable source of error, and we shall always say which one we use.

2.5 Sub-gamma variables, Bennett and Bernstein

Sub-Gaussianity is the right notion for bounded or Gaussian variables, but Counterexample 2.27 shows that many natural variables have a cumulant generating function that blows up at a finite λ. The following class, modelled on the gamma distribution, is the correct extension, and it is also – this is the point of the section – the class in which *bounded variables with small variance* naturally live.

2.5.1 The model computation: Poisson deviations

Before introducing the general class, we carry out completely the computation that it imitates. Let $X \sim \mathrm{Poi}(\sigma^2)$; we name the parameter σ^2, as in Example 2.8, because it is simultaneously the mean and the *variance* of X, and it is as a variance that it appears in every statement below.

Proposition 2.33 (Poisson tails)**.** *Let $X \sim \mathrm{Poi}(\sigma^2)$ and let $h(u) = (1+u)\log(1+u) - u$ as in (2.11). Then $\psi_X(\lambda) = \sigma^2 \varphi(\lambda)$ with $\varphi(\lambda) = e^\lambda - \lambda - 1$, and $K_X^*(\sigma^2 + \varepsilon) = \sigma^2 h\big(\varepsilon/\sigma^2\big)$, so that*

$$\mathbb{P}\big(X \ge \sigma^2 + \varepsilon\big) \le \exp\Big(-\sigma^2 h\Big(\frac{\varepsilon}{\sigma^2}\Big)\Big), \qquad \mathbb{P}\big(X \le \sigma^2 - \varepsilon\big) \le \exp\Big(-\sigma^2 h\Big(\frac{-\varepsilon}{\sigma^2}\Big)\Big) \le \exp\Big(-\frac{\varepsilon^2}{2\sigma^2}\Big) \tag{2.15}$$

for $\varepsilon \ge 0$ (and $\varepsilon \le \sigma^2$ on the left). Moreover h admits the integral representation

$$h(u) = \frac{u^2}{2} \int_0^1 \frac{2(1-t)\,\mathrm{d}t}{1+tu}, \qquad u > -1, \tag{2.16}$$

from which follow

$$h(u) \;\ge\; \frac{u^2}{2} \;\; (u \le 0), \qquad \frac{u^2}{2\bigl(1+\frac{u}{3}\bigr)} \;\le\; h(u) \;\le\; \frac{u^2}{2} \;\; (u \ge 0). \tag{2.17}$$

Proof. The cumulant generating function is $K_X(\lambda) = \sigma^2(e^\lambda - 1)$ (Example 2.8), so $\psi_X(\lambda) = \sigma^2(e^\lambda - 1 - \lambda)$ and, by Proposition 2.23(c) with $x = \sigma^2 + \varepsilon$,

$$K_X^*(x) = x \log \frac{x}{\sigma^2} - x + \sigma^2 = \sigma^2\Bigl[(1+u)\log(1+u) - u\Bigr] = \sigma^2 h(u), \qquad u = \frac{\varepsilon}{\sigma^2},$$

whence (2.15) by Theorem 2.14 (applied to X and to $-X$; note K^* is the same function on both sides of the mean, so the left tail is governed by $h(-\varepsilon/\sigma^2)$).

For (2.16) we use Proposition 2.11. Under $\mathbb{P}_\lambda$ the law of X is $\mathrm{Poi}(\sigma^2 e^\lambda)$, so $m(\lambda) = \varsigma^2(\lambda) = \sigma^2 e^\lambda$: the tilted variance equals the tilted mean. Hence $\varsigma^2(\lambda(z)) = z$ for every $z > 0$, and (2.4) reads, with $\mu = \sigma^2$ and $x = \sigma^2(1+u)$,

$$\sigma^2 h(u) = K_X^*(x) = \frac{(x-\sigma^2)^2}{2}\int_0^1 \frac{2(1-t)\,\mathrm{d}t}{\sigma^2 + t(x-\sigma^2)} = \frac{\sigma^4 u^2}{2}\int_0^1 \frac{2(1-t)\,\mathrm{d}t}{\sigma^2(1+tu)},$$

which is (2.16).

The bounds (2.17) follow at once, since $\mathrm{d}\nu(t) = 2(1-t)\,\mathrm{d}t$ is a probability measure on $[0,1]$: for $u \le 0$ we have $1 + tu \le 1$ and the integral is ≥ 1; for $u \ge 0$ it is ≤ 1; and by Jensen's inequality applied to the convex function $y \mapsto 1/y$,

$$\int_0^1 \frac{\mathrm{d}\nu(t)}{1+tu} \;\ge\; \frac{1}{\int_0^1 (1+tu)\,\mathrm{d}\nu(t)} = \frac{1}{1+\frac{u}{3}},$$

because $\int_0^1 t\;\mathrm{d}\nu(t) = \int_0^1 2t(1-t)\,\mathrm{d}t = \frac{1}{3}$. □

Two features of (2.15) deserve to be retained, because they recur in every statement of this section. The *left* tail of a Poisson variable is sub-Gaussian with the exact variance, $e^{-\varepsilon^2/2\sigma^2}$: deviations *below* the mean are never worse than Gaussian. The right tail is heavier, and (2.17) quantifies the loss: the Gaussian exponent $\varepsilon^2/2\sigma^2$ must be divided by $1 + \varepsilon/(3\sigma^2)$.

Corollary 2.34 (Inverting Bennett's function)**.** *The function h is an increasing bijection from $[0,\infty)$ onto $[0,\infty)$, and its inverse satisfies*

$$h^{-1}(x) \;\le\; \sqrt{2x} + \frac{x}{3}, \qquad x \ge 0. \tag{2.18}$$

Consequently, if $X \sim \mathrm{Poi}(\sigma^2)$, then for every $\delta \in (0,1)$,

$$\mathbb{P}\Bigl(X \;\ge\; \sigma^2 + \sqrt{2\sigma^2 \log\tfrac{1}{\delta}} + \tfrac{1}{3}\log\tfrac{1}{\delta}\Bigr) \;\le\; \delta, \qquad \mathbb{P}\Bigl(X \;\le\; \sigma^2 - \sqrt{2\sigma^2 \log\tfrac{1}{\delta}}\Bigr) \;\le\; \delta.$$

Proof. $h(0) = 0$, $h'(u) = \log(1+u) > 0$ for $u > 0$ and $h(u) \to \infty$, so h is an increasing bijection of $[0,\infty)$. Inequality (2.18) is equivalent to $h\bigl(\sqrt{2x} + x/3\bigr) \ge x$. Substituting $s = \sqrt{2x}$, i.e. $x = s^2/2$, and writing $u = s + s^2/6$, we must show $F(s) := h(u) - s^2/2 \ge 0$ for $s \ge 0$. Now $F(0) = 0$ and

$$F'(s) = h'(u)\Bigl(1 + \frac{s}{3}\Bigr) - s = \log\Bigl(1 + s + \frac{s^2}{6}\Bigr)\cdot\frac{3+s}{3} - s,$$

so $F'(s) \geq 0$ if and only if $G(s) := \log\bigl(1 + s + \frac{s^2}{6}\bigr) - \frac{3s}{3+s} \geq 0$. We have $G(0) = 0$ and

$$G'(s) = \frac{2(3+s)}{6+6s+s^2} - \frac{9}{(3+s)^2} = \frac{2(3+s)^3 - 9(6+6s+s^2)}{(6+6s+s^2)(3+s)^2} = \frac{9s^2+2s^3}{(6+6s+s^2)(3+s)^2} \;\geq\; 0,$$

so $G \geq 0$, hence $F' \geq 0$ and $F \geq 0$. The two probability bounds follow from (2.15): for the right tail, $\sigma^2 h(\varepsilon/\sigma^2) \geq \log(1/\delta)$ as soon as $\varepsilon/\sigma^2 \geq h^{-1}\bigl(\log(1/\delta)/\sigma^2\bigr)$, which by (2.18) holds for $\varepsilon \geq \sqrt{2\sigma^2\log(1/\delta)} + \frac{1}{3}\log(1/\delta)$; for the left tail use $h(-u) \geq u^2/2$. □

[V2 – the two probability bounds need no calculus at all] Lemma 2.46 – the elementary inequality $e^x - 1 - x \leq x^2/(2(1-x/3))$, which depends on nothing in this chapter – makes the calculus with F and G unnecessary, provided it is moved before this subsection. Replacement for Corollary 2.34 and its proof:

Corollary 2.34′. *Let $X \sim \mathrm{Poi}(\sigma^2)$. For every $\delta \in (0,1)$,*

$$\mathbb{P}\Bigl(X \;\geq\; \sigma^2 + \sqrt{2\sigma^2\log\tfrac{1}{\delta}} + \tfrac{1}{3}\log\tfrac{1}{\delta}\Bigr) \leq \delta, \qquad \mathbb{P}\Bigl(X \;\leq\; \sigma^2 - \sqrt{2\sigma^2\log\tfrac{1}{\delta}}\Bigr) \leq \delta.$$

Proof. Put $Z = X - \sigma^2$, so that $\psi_Z(\lambda) = \sigma^2(e^\lambda - 1 - \lambda)$ by Proposition 2.33. By Lemma 2.46 with $x = \lambda \in [0,3)$,

$$\psi_Z(\lambda) \;\leq\; \frac{\lambda^2\sigma^2}{2\bigl(1-\frac{\lambda}{3}\bigr)}, \qquad 0 < \lambda < 3,$$

that is, $Z \in \Gamma_+(\sigma^2, \frac{1}{3})$, and (2.20) with $y = \log(1/\delta)$ gives the first bound. The second is the left-hand half of (2.15) together with $h(-u) \geq u^2/2$. □

The analytic statement $h^{-1}(x) \leq \sqrt{2x} + x/3$ is then a corollary – it is *equivalent* to the first display, by (2.15) – and may be kept, with its present calculus proof, as a remark or an exercise.

Review of 18 September: not yet discussed.

Remark 2.35 (The shape of all the bounds to come). Corollary 2.34 is the prototype of the two-regime bound announced in Remark 2.38: a Gaussian term $\sqrt{2\sigma^2\log(1/\delta)}$ driven by the variance, plus a term $\frac{1}{3}\log(1/\delta)$ driven by the size of the jumps – here 1, the spacing of the Poisson lattice. Bennett's and Bernstein's inequalities say that *no* centred variable with variance σ^2 and increments bounded by b does worse than the Poisson variable with the same parameters, and the general definition below is designed exactly so that this computation goes through verbatim.

Definition 2.36 (Sub-gamma variable)**.** Let $v > 0$ and $c \geq 0$. A centred random variable Z is *sub-gamma on the right tail with variance factor v and scale parameter c*, written $Z \in \Gamma_+(v,c)$, if

$$\psi_Z(\lambda) = \log\mathbb{E}\bigl[e^{\lambda Z}\bigr] \;\leq\; \frac{\lambda^2 v}{2(1-c\lambda)} \qquad \text{for all } \lambda \in \bigl(0, \tfrac{1}{c}\bigr). \tag{2.19}$$

It is sub-gamma on the left tail if $-Z$ is, and sub-gamma if both hold.

For $c = 0$ this is sub-Gaussianity with proxy v (restricted to $\lambda > 0$); for $c > 0$ the bound degrades as $\lambda \uparrow 1/c$, which is exactly what a variable with an exponential tail requires.

Theorem 2.37 (Tails of sub-gamma variables)**.** *Let $Z \in \Gamma_+(v,c)$. Then for every $y > 0$,*

$$\mathbb{P}\Bigl(Z \;\geq\; \sqrt{2vy} + cy\Bigr) \;\leq\; e^{-y}, \tag{2.20}$$

and consequently, for every $t > 0$,

$$\mathbb{P}(Z \geq t) \;\leq\; \exp\Bigl(-\frac{t^2}{2(v+ct)}\Bigr). \tag{2.21}$$

Proof. Fix $y > 0$, put $a := \sqrt{v/(2y)}$ and $\lambda := 1/(c+a) \in (0, 1/c)$, and let $t = \sqrt{2vy} + cy$. Note that $v = 2ya^2$ and $\sqrt{2vy} = 2ay$. Then

$$1 - c\lambda = \frac{a}{c+a}, \qquad \text{so} \qquad \frac{\lambda^2 v}{2(1-c\lambda)} = \frac{1}{(c+a)^2} \cdot \frac{v(c+a)}{2a} = \frac{v}{2a(c+a)} = \frac{ya}{c+a},$$

while $\lambda t = \dfrac{2ay + cy}{c+a}$. Therefore

$$\lambda t - \frac{\lambda^2 v}{2(1-c\lambda)} = \frac{y(2a+c) - ya}{c+a} = \frac{y(a+c)}{a+c} = y,$$

and Theorem 1.57 gives $\mathbb{P}(Z \geq t) \leq \exp\bigl(-\lambda t + \psi_Z(\lambda)\bigr) \leq e^{-y}$, which is (2.20).

For (2.21), fix $t > 0$ and set $y = t^2/\bigl(2(v+ct)\bigr)$. We claim that $\sqrt{2vy} + cy \leq t$, which by (2.20) gives the result. Writing $w = ct/v \geq 0$ and $r = \sqrt{1+w} \geq 1$,

$$\frac{\sqrt{2vy} + cy}{t} = \sqrt{\frac{v}{v+ct}} + \frac{ct}{2(v+ct)} = \frac{1}{r} + \frac{r^2-1}{2r^2},$$

and $\frac{1}{r} + \frac{r^2-1}{2r^2} \leq 1$ is equivalent, after multiplication by $2r^2 > 0$, to $2r + r^2 - 1 \leq 2r^2$, i.e. to $0 \leq (r-1)^2$. □

[V3 – prove (2.21) directly, with the natural λ] The value $\lambda = 1/(c+a)$ is needed for (2.20); for (2.21) alone, the tilt of Bernstein's classical proof makes the computation transparent. Alternative proof of (2.21) (the proof of (2.20) is unchanged):

Proof (replacement). Let $t > 0$ and set $\lambda := \dfrac{t}{v+ct}$, which lies in $(0, 1/c)$ since $ct < v + ct$. Then $1 - c\lambda = \dfrac{v}{v+ct}$, so that

$$\frac{\lambda^2 v}{2(1-c\lambda)} = \frac{t^2}{(v+ct)^2} \cdot \frac{v+ct}{2} = \frac{t^2}{2(v+ct)} = \frac{\lambda t}{2},$$

whence $\lambda t - \frac{\lambda^2 v}{2(1-c\lambda)} = \frac{t^2}{2(v+ct)}$ and Theorem 1.57 gives (2.21). □

Amended 5 October 2026. This block first went on to derive (2.20) from (2.21). That derivation was invalid: (2.21) is strictly weaker when $c > 0$ (for $v = c = y = 1$ it gives only $e^{-0.854}$ at the level $\sqrt{2vy} + cy$), and the implication goes the other way, as in the proof above. The block now offers only a proof of (2.21).

Two benefits: the tilt is the one every reader has already seen in Bernstein's proof, and the identity "*the optimised exponent equals half of λt*" is worth remembering.

Review of 18 September: not yet discussed.

Remark 2.38 (The two regimes). Inequality (2.20) is the form to remember: with probability at least $1 - \delta$,

$$Z \;<\; \underbrace{\sqrt{2v\log(1/\delta)}}_{\text{Gaussian regime}} + \underbrace{c\log(1/\delta)}_{\text{Poisson/exponential regime}}.$$

For moderate deviations ($y \ll v/c^2$) the first term dominates and the variable behaves as if it were Gaussian with variance v; for large deviations the second takes over and the tail is exponential with scale c. Every inequality in this section is a version of this dichotomy, and the whole art of using them consists in knowing which of the two terms one is in.

2.5.2 Bennett's inequality

Lemma 2.39 (Bennett's lemma)**.** *Let Z be a centred random variable with $Z \le b$ almost surely for some $b > 0$, and $\mathbb{E}[Z^2] \le \sigma^2$. Then for every $\lambda > 0$,*

$$\psi_Z(\lambda) \;\le\; \frac{\sigma^2}{b^2}\big(e^{\lambda b} - 1 - \lambda b\big). \tag{2.22}$$

Proof. Let $\varphi_2(u) := \varphi(u)/u^2 = (e^u - 1 - u)/u^2$ for $u \neq 0$ and $\varphi_2(0) = \frac{1}{2}$, with φ as in Proposition 2.33. We first check the integral representation

$$\varphi_2(u) = \int_0^1 (1-s)e^{su}\,\mathrm{d}s, \qquad u \in \mathbb{R}. \tag{2.23}$$

Indeed, for $u \neq 0$, $\int_0^1 e^{su}\,\mathrm{d}s = (e^u - 1)/u$ and, integrating by parts, $\int_0^1 s e^{su}\,\mathrm{d}s = e^u/u - (e^u - 1)/u^2$; subtracting gives $(e^u - 1 - u)/u^2$, and the case $u = 0$ is immediate. Differentiating (2.23) under the integral sign yields $\varphi_2'(u) = \int_0^1 s(1-s)e^{su}\,\mathrm{d}s > 0$, so φ_2 is (strictly) increasing on $\mathbb{R}$.

Now write $e^x = 1 + x + x^2\varphi_2(x)$, so that, using $\mathbb{E}Z = 0$,

$$\mathbb{E}\big[e^{\lambda Z}\big] = 1 + \lambda^2\,\mathbb{E}\big[Z^2\varphi_2(\lambda Z)\big].$$

Since $\lambda > 0$ and $Z \le b$, monotonicity of φ_2 gives $\varphi_2(\lambda Z) \le \varphi_2(\lambda b)$ pointwise, hence

$$\mathbb{E}\big[e^{\lambda Z}\big] \le 1 + \lambda^2\sigma^2\varphi_2(\lambda b) = 1 + \frac{\sigma^2}{b^2}\big(e^{\lambda b} - 1 - \lambda b\big) \le \exp\Big(\frac{\sigma^2}{b^2}\big(e^{\lambda b} - 1 - \lambda b\big)\Big),$$

using $1 + x \le e^x$ in the last step. □

Theorem 2.40 (Bennett's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent centred random variables with $Z_i \le b$ almost surely, and set $v := \sum_{i=1}^n \mathbb{E}[Z_i^2]$ and $S_n = \sum_{i=1}^n Z_i$. Then for every $t > 0$,*

$$\mathbb{P}\big(S_n \ge t\big) \;\le\; \exp\Big(-\frac{v}{b^2}\,h\Big(\frac{bt}{v}\Big)\Big), \qquad h(u) = (1+u)\log(1+u) - u. \tag{2.24}$$

Proof. By (1.23) and Lemma 2.39, $\psi_{S_n}(\lambda) \le \frac{v}{b^2}(e^{\lambda b} - 1 - \lambda b)$ for $\lambda > 0$. Theorem 1.57 gives $\mathbb{P}(S_n \ge t) \le \exp\big(-g(\lambda)\big)$ with $g(\lambda) = \lambda t - \frac{v}{b^2}(e^{\lambda b} - 1 - \lambda b)$. Now $g'(\lambda) = t - \frac{v}{b}(e^{\lambda b} - 1)$ vanishes at $\lambda^* = \frac{1}{b}\log\big(1 + \frac{bt}{v}\big) > 0$, and g is concave, so the maximum is attained there. Writing $u = bt/v$, so that $t = uv/b$ and $e^{\lambda^* b} - 1 = u$,

$$g(\lambda^*) = \frac{uv}{b^2}\log(1+u) - \frac{v}{b^2}\big(u - \log(1+u)\big) = \frac{v}{b^2}\Big[(1+u)\log(1+u) - u\Big] = \frac{v}{b^2}h(u). \qquad \square$$

Remark 2.41 (Bennett's bound is the Poisson bound). The function h of (2.11) is exactly the Cramér transform of the Poisson law (Proposition 2.23(c)), and this is no coincidence. Let $N \sim \mathrm{Poi}(\sigma^2/b^2)$ and let $Z = b(N - \sigma^2/b^2)$. Then $\mathbb{E}Z = 0$, $\mathbb{V}\mathrm{ar}(Z) = b^2\,\mathbb{V}\mathrm{ar}(N) = \sigma^2$, and

$$\psi_Z(\lambda) = \frac{\sigma^2}{b^2}\big(e^{\lambda b} - 1\big) - \lambda b\,\frac{\sigma^2}{b^2} = \frac{\sigma^2}{b^2}\big(e^{\lambda b} - 1 - \lambda b\big)\,:$$

Bennett's lemma is an equality for the centred, scaled Poisson variable. In other words, (2.22) says that among centred variables with variance σ^2 and jumps bounded above by b, none has a cumulant generating function larger than the Poisson one – and (2.24) is precisely the Chernoff bound for a Poisson sum. This explains the shape of h, and it says that Theorem 2.40 cannot be improved as a function of (v, b, t) alone at the exponential scale.

[V4 – the optimality claim is not justified as written] The computation is right – the right-hand side of (2.22) *is* the centred cumulant generating function of $Z = b(N - \sigma^2/b^2)$, $N \sim \mathrm{Poi}(\sigma^2/b^2)$ – but that variable does *not* satisfy the hypothesis of Lemma 2.39: it is unbounded above, only its jumps have size b. So (2.22) is an equality for no member of the class $\{Z \le b,\ \mathbb{E}Z^2 \le \sigma^2\}$; the extremal two-point member of that class, taking the values b and $-\sigma^2/b$ with probabilities $\frac{\sigma^2}{b^2+\sigma^2}$ and $\frac{b^2}{b^2+\sigma^2}$, has ψ strictly below the bound. The optimality is nevertheless true, by a limit *inside* the class. Replacement for the last two sentences of the remark:

In other words, (2.22) compares a centred variable with jumps bounded above by b to the Poisson variable with the same variance, and (2.24) is precisely the Chernoff bound for a Poisson sum. This also shows that Theorem 2.40 cannot be improved as a function of (v, b, t) alone at the exponential scale, by a limit inside the class: take $Z_i = b(B_i - p)$ with $B_i \sim \mathrm{Ber}(p)$ independent, and let $n \to \infty$, $p \to 0$ with $nb^2p \to v$. Then $Z_i \le b$ and $\sum_i \mathbb{E}Z_i^2 \to v$, while $S_n = b\big(\mathrm{Bin}(n,p) - np\big)$ converges in law to $b\big(\mathrm{Poi}(v/b^2) - v/b^2\big)$, whose tail is, by Proposition 2.33 and Theorem 2.19, exactly $\exp\big(-\frac{v}{b^2}h(\frac{bt}{v})\big)$ at the exponential scale. Many rare jumps, not one bounded jump, is the extremal configuration – which is why Bennett's function is the Poisson rate function.

Review of 18 September: not yet discussed.

Lemma 2.39 bounds the cumulant generating function; the variance representation of Section 2.3 gives the same result by bounding the tilted variance directly, and in a form that is often sharper and always more transparent.

Proposition 2.42 (Bennett through tilted variances)**.** *Let X be a random variable with $0 \in \operatorname{int}\operatorname{dom}(K_X)$ and $\mu = \mathbb{E}X > 0$, and suppose that*

$$\varsigma^2(\lambda) \ \le\ m(\lambda) \qquad \text{for every } \lambda \ge 0 \text{ in } \operatorname{dom}(K_X). \tag{2.25}$$

Then, for every $x > \mu$,

$$K_X^*(x) \ \ge\ x \log \frac{x}{\mu} - (x - \mu) = \mu\, h\Big(\frac{x-\mu}{\mu}\Big),$$

and hence, for $X_1, \dots, X_n$ i.i.d. copies of X and $\varepsilon > 0$,

$$\mathbb{P}\big(\bar{X}_n \ge \mu + \varepsilon\big) \ \le\ \exp\Big(-n\mu\, h\Big(\frac{\varepsilon}{\mu}\Big)\Big).$$

Proof. By (2.4) and (2.25), using $m(\lambda(u)) = u$,

$$K_X^*(x) = \int_\mu^x \frac{x-u}{\varsigma^2(\lambda(u))}\,\mathrm{d}u \ \ge\ \int_\mu^x \frac{x-u}{u}\,\mathrm{d}u = x\log\frac{x}{\mu} - (x-\mu),$$

and the last expression equals $\mu h((x-\mu)/\mu)$ by the computation made in the proof of Proposition 2.33. The statement for i.i.d. sums follows from (2.7). □

Condition (2.25) – "the variance never exceeds the mean, under any tilt" – holds with equality for Poisson variables, and holds for every variable with values in $\{0, 1\}$, since a tilted Bernoulli is a Bernoulli.

[V5 – one constant more, and Remark 2.43 becomes a corollary] As stated, the proposition gives only $\mu\, h(\varepsilon/\mu)$ for $\mathrm{Ber}(p)$, and the sharper $\frac{p}{1-p}h(\varepsilon/p)$ has to be redone by hand in Remark 2.43. One constant covers both. Replacement:

Proposition 2.42′. *Let X be a random variable with $0 \in \operatorname{int}\operatorname{dom}(K_X)$ and $\mu = \mathbb{E}X > 0$, and let $\kappa > 0$ be such that*

$$\varsigma^2(\lambda) \ \le\ \kappa\, m(\lambda) \qquad \text{for every } \lambda \ge 0 \text{ in } \operatorname{dom}(K_X).$$

Then, for every $x > \mu$,

$$K_X^*(x) \;\geq\; \frac{1}{\kappa}\Big(x \log \frac{x}{\mu} - (x-\mu)\Big) = \frac{\mu}{\kappa}\, h\Big(\frac{x-\mu}{\mu}\Big),$$

and hence, for $X_1, \dots, X_n$ i.i.d. copies of X and $\varepsilon > 0$,

$$\mathbb{P}\big(\bar{X}_n \geq \mu + \varepsilon\big) \;\leq\; \exp\Big(-\frac{n\mu}{\kappa}\, h\Big(\frac{\varepsilon}{\mu}\Big)\Big).$$

Proof. By (2.4) and the hypothesis, using $m(\lambda(u)) = u$,

$$K_X^*(x) = \int_\mu^x \frac{x-u}{\varsigma^2(\lambda(u))}\,\mathrm{d}u \;\geq\; \frac{1}{\kappa}\int_\mu^x \frac{x-u}{u}\,\mathrm{d}u = \frac{1}{\kappa}\Big(x\log\frac{x}{\mu} - (x-\mu)\Big),$$

and the last expression equals $\frac{\mu}{\kappa} h((x-\mu)/\mu)$ by the computation in the proof of Proposition 2.33; the statement for i.i.d. sums follows from (2.7). □

Two instances. $\kappa = 1$ is the Poisson case, with equality. For $X \sim \mathrm{Ber}(p)$ the tilted law is $\mathrm{Ber}(u)$ at the tilt $\lambda(u)$, so $\varsigma^2(\lambda(u)) = u(1-u) \leq (1-p)u$ for $u \geq p$: the hypothesis holds with $\kappa = 1-p$ and the proposition gives $\mathrm{kl}(p+\varepsilon, p) \geq \frac{p}{1-p} h(\varepsilon/p)$, which is exactly the second line of Remark 2.43 and exactly the exponent that Theorem 2.40 produces for Bernoulli sums. That line of Remark 2.43 then becomes a citation rather than a computation.

Review of 18 September: not yet discussed.

Remark 2.43 (One representation, several inequalities). For $X \sim \mathrm{Ber}(p)$ the tilted law is $\mathrm{Ber}(u)$ at the tilt $\lambda(u)$, so $\varsigma^2(\lambda(u)) = u(1-u)$ and (2.4) is the exact identity

$$\mathrm{kl}(p+\varepsilon,\, p) = \int_p^{p+\varepsilon} \frac{p+\varepsilon-u}{u(1-u)}\,\mathrm{d}u. \tag{2.26}$$

Every inequality of these two chapters is now obtained by bounding the denominator from above:

$$\begin{aligned}
&u(1-u) \leq \tfrac{1}{4} && \rightsquigarrow\quad \mathrm{kl}(p+\varepsilon, p) \;\geq\; 2\varepsilon^2 && \text{(Hoeffding, Pinsker)},\\
&u(1-u) \leq (1-p)u && \rightsquigarrow\quad \mathrm{kl}(p+\varepsilon, p) \;\geq\; \frac{p}{1-p}\, h\Big(\frac{\varepsilon}{p}\Big) && \text{(Bennett)},\\
&\text{then (2.17)} && \rightsquigarrow\quad \mathrm{kl}(p+\varepsilon, p) \;\geq\; \frac{\varepsilon^2}{2\big(p(1-p) + \frac{(1-p)\varepsilon}{3}\big)} && \text{(Bernstein)}.
\end{aligned}$$

The second line uses $u \leq 1$ and $1-u \leq 1-p$ for $u \geq p$, and reproduces exactly the exponent that Theorem 2.40 gives for Bernoulli sums. The hierarchy of the classical inequalities is thus a hierarchy of bounds on a single quantity, the variance of the tilted law.

Corollary 2.44 (Bennett's inequality, inverted)**.** *Under the hypotheses of Theorem 2.40, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$S_n \;<\; \sqrt{2v \log\tfrac{1}{\delta}} + \frac{b}{3}\log\tfrac{1}{\delta}.$$

Proof. By (2.24), $\mathbb{P}(S_n \geq t) \leq \delta$ as soon as $\frac{v}{b^2} h\big(\frac{bt}{v}\big) \geq \log\frac{1}{\delta}$, that is $\frac{bt}{v} \geq h^{-1}\big(\frac{b^2}{v}\log\frac{1}{\delta}\big)$. By (2.18) this holds as soon as $\frac{bt}{v} \geq \sqrt{\frac{2b^2}{v}\log\frac{1}{\delta}} + \frac{b^2}{3v}\log\frac{1}{\delta}$, i.e. $t \geq \sqrt{2v\log\frac{1}{\delta}} + \frac{b}{3}\log\frac{1}{\delta}$. □

Note that Corollary 2.44 is exactly (2.29): the inverted forms of Bennett's and Bernstein's inequalities coincide, although (2.24) is strictly stronger than (2.28). The difference between the two inequalities is invisible at the level of confidence intervals, and this is why Bernstein's form – being explicit – is the one used in practice.

Remark 2.45 (Asymptotics of h, and what Bennett gives). The function h satisfies $h(u) = \frac{u^2}{2} - \frac{u^3}{6} + O(u^4)$ as $u \to 0$ and $h(u) \sim u \log u$ as $u \to \infty$. Consequently, in (2.24):

- for small deviations, $bt/v \to 0$ and the exponent is $\approx \frac{v}{b^2} \cdot \frac{b^2t^2}{2v^2} = \frac{t^2}{2v}$: the Gaussian bound with the *true* variance v – exactly what Hoeffding's inequality failed to give;
- for large deviations, the exponent is $\approx \frac{t}{b} \log \frac{bt}{v}$, a Poissonian tail, strictly heavier than Gaussian but with an extra logarithmic gain over a purely exponential tail.

2.5.3 Bernstein's inequality

Bennett's bound is sharp but awkward: h is not invertible in closed form. The following elementary estimate replaces it by a rational function, at a small cost, and identifies bounded variables as sub-gamma.

Lemma 2.46. *For every $x \in [0, 3)$,*

$$e^x - 1 - x \;\le\; \frac{x^2}{2\big(1 - \frac{x}{3}\big)}. \tag{2.27}$$

Consequently, if Z is centred with $Z \le b$ a.s. and $\mathbb{E}[Z^2] \le \sigma^2$, then $Z \in \Gamma_+\big(\sigma^2, b/3\big)$.

Proof. For $k \ge 2$ one has $k! \ge 2 \cdot 3^{k-2}$, by induction ($k = 2$ and $k = 3$ are equalities, and $(k+1)! = (k+1)k! \ge 3k!$ for $k \ge 2$). Hence, for $0 \le x < 3$,

$$e^x - 1 - x = \sum_{k \ge 2} \frac{x^k}{k!} \le \frac{x^2}{2} \sum_{k \ge 2} \Big(\frac{x}{3}\Big)^{k-2} = \frac{x^2}{2\big(1 - \frac{x}{3}\big)}.$$

Applying this with $x = \lambda b \in [0, 3)$ in Lemma 2.39 gives, for $0 < \lambda < 3/b$,

$$\psi_Z(\lambda) \le \frac{\sigma^2}{b^2} \cdot \frac{\lambda^2 b^2}{2\big(1 - \frac{\lambda b}{3}\big)} = \frac{\lambda^2 \sigma^2}{2\big(1 - \frac{b}{3}\lambda\big)},$$

which is (2.19) with $v = \sigma^2$ and $c = b/3$. □

Theorem 2.47 (Bernstein's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent centred random variables with $Z_i \le b$ almost surely, and let $v = \sum_{i=1}^n \mathbb{E}[Z_i^2]$, $S_n = \sum_i Z_i$. Then for every $t > 0$,*

$$\mathbb{P}\big(S_n \ge t\big) \;\le\; \exp\Big(-\frac{t^2}{2\big(v + \frac{bt}{3}\big)}\Big), \tag{2.28}$$

or equivalently, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$S_n \;<\; \sqrt{2v \log \tfrac{1}{\delta}} + \frac{b}{3} \log \tfrac{1}{\delta}. \tag{2.29}$$

In particular, if $X_1, \dots, X_n$ are i.i.d. with mean μ, variance σ^2 and $X_i - \mu \le b$ almost surely, then with probability at least $1 - \delta$,

$$\bar{X}_n - \mu \;<\; \sigma \sqrt{\frac{2 \log(1/\delta)}{n}} + \frac{b \log(1/\delta)}{3n}. \tag{2.30}$$

Proof. By Lemma 2.46 and (1.23), $S_n \in \Gamma_+(v, b/3)$; apply Theorem 2.37, in the form (2.21) for (2.28) and in the form (2.20) with $y = \log(1/\delta)$ for (2.29). The last statement is the case $Z_i = (X_i - \mu)/n$, for which $v = \sigma^2/n$ and the upper bound is b/n. □

Remark 2.48 (Reading Bernstein's inequality). Formula (2.30) is the single most useful inequality of this chapter, and it should be read as follows. The first term is the Gaussian one, with the *true* standard deviation σ: it is what the central limit theorem predicts, and it is of order $n^{-1/2}$. The second is the price of non-asymptoticity: it involves the range b, not the variance, but it is of order n^{-1}, hence negligible as soon as $n \gtrsim (b/\sigma)^2 \log(1/\delta)$. Compare with Hoeffding's (1.29), which gives $b\sqrt{\log(1/\delta)/(2n)}$ with no σ at all: Bernstein's inequality wins by the factor σ/b, which is exactly the factor that was lost in Remark 1.66. For $\mathrm{Ber}(p)$ with p small, $\sigma/b \approx \sqrt{p}$, whence the factor $\approx 1/p$ in the sample sizes computed in Section 2.1.

Proposition 2.49 (Bernstein's moment condition)**.** *Let Z be centred and suppose that there exist $v > 0$ and $c > 0$ with*

$$\mathbb{E}\big[|Z|^q\big] \;\le\; \frac{q!}{2}\, v\, c^{\,q-2} \qquad \text{for every integer } q \ge 2. \tag{2.31}$$

Then $Z \in \Gamma_+(v, c)$, and Theorem 2.37 applies.

Proof. For $0 < \lambda < 1/c$, using $\mathbb{E}Z = 0$ and (2.31),

$$\mathbb{E}\big[e^{\lambda Z}\big] \le 1 + \sum_{q\ge 2} \frac{\lambda^q \mathbb{E}|Z|^q}{q!} \le 1 + \frac{\lambda^2 v}{2} \sum_{q\ge 2} (\lambda c)^{q-2} = 1 + \frac{\lambda^2 v}{2(1-\lambda c)} \le \exp\!\Big(\frac{\lambda^2 v}{2(1-\lambda c)}\Big). \qquad \square$$

Remark 2.50 (Sub-exponential variables). Condition (2.31) holds, for instance, whenever $\mathbb{E}[e^{|Z|/c}] \le 2$ with $v = 4c^2$ (Exercise 2.4); variables satisfying such a condition are called *sub-exponential*, and the class $\Gamma_+(v, c)$ is the quantitative version of that notion adapted to the Chernoff method. The chi-square variable of Counterexample 2.27 is sub-exponential: if $X \sim \mathcal{N}(0,1)$ then $X^2 - 1 \in \Gamma_+(2, 2)$ (Exercise 2.5), whence $\mathbb{P}(X^2 - 1 \ge 2\sqrt{y} + 2y) \le e^{-y}$, and for a sum of d independent copies, which is sub-gamma with $v = 2d$ and $c = 2$,

$$\mathbb{P}\Big(\|G\|_2^2 - d \;\ge\; 2\sqrt{dy} + 2y\Big) \;\le\; e^{-y}, \qquad G \sim \mathcal{N}(0, I_d),$$

the basic concentration statement for the norm of a Gaussian vector, which is the engine of the Johnson–Lindenstrauss lemma and of random-projection arguments in general.

2.5.4 Four bounds, compared

Let $X_1, \dots, X_n$ be i.i.d. $\mathrm{Ber}(p)$ and let $t \in (0, 1-p)$. The four inequalities of Chapter 1 and of this chapter give

Cramér–Chernoff	$\mathbb{P}(\bar X_n \ge p + t) \le e^{-n\ \mathrm{kl}(p+t,\,p)}$	(2.12)	exact rate
Bennett	$\mathbb{P}(\bar X_n \ge p + t) \le \exp\!\Big(-\dfrac{np}{1-p}\, h\Big(\dfrac{t}{p}\Big)\Big)$	(2.24)	Poisson
Bernstein	$\mathbb{P}(\bar X_n \ge p + t) \le \exp\!\Big(-\dfrac{nt^2}{2\big(p(1-p) + \frac{(1-p)t}{3}\big)}\Big)$	(2.28)	variance-adaptive
Hoeffding	$\mathbb{P}(\bar X_n \ge p + t) \le e^{-2nt^2}$	(1.28)	range only

(for Bennett and Bernstein we used $Z_i = X_i - p \le 1 - p = b$ and $\mathbb{E}Z_i^2 = p(1-p)$). They are in decreasing order of sharpness, and Figure 2.2 shows the four exponents for $p = 0.1$. [*replacement:* "They are in decreasing order of sharpness for small and moderate deviations – kl ≥ Bennett ≥ Bernstein holds for every t, and so does kl ≥ Hoeffding, but Hoeffding's exponent overtakes

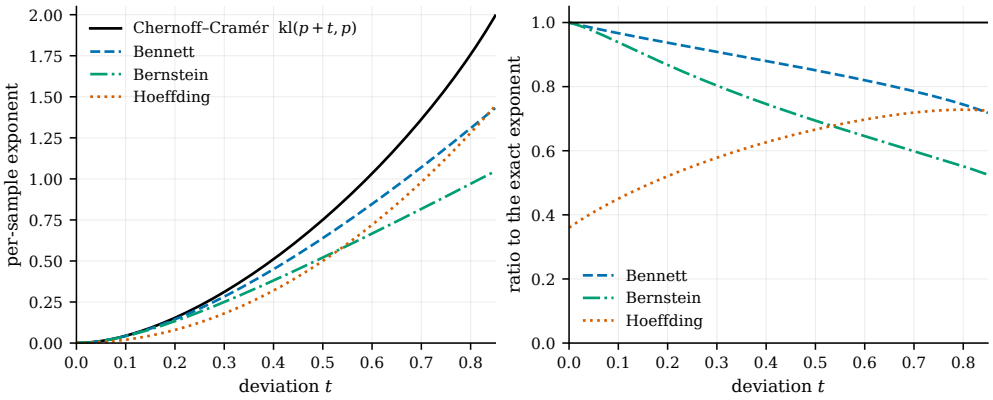


Figure 2.2: The per-sample exponents of the four inequalities for $X_i \sim \mathrm{Ber}(p)$, $p = 0.1$, as a function of the deviation t (left: linear scale; right: the ratio to the exact exponent $\mathrm{kl}(p+t, p)$). Bennett's bound tracks the exact Cramér–Chernoff rate closely over the whole range; Bernstein's is slightly below it; Hoeffding's, which sees only the range of the variables, is off by a factor that grows as p decreases. Note that for very large deviations Hoeffding's exponent *overtakes* Bernstein's: the latter is tuned to the Gaussian regime $t \lesssim \sigma$ and pays for it far in the tail, whereas the former is a uniform quadratic bound. Generated by `figures/fig_bounds_comparison.py`.

Bernstein's far in the tail, already at $t \approx 0.53$ when $p = 0.1$ (see Figure 2.2) – and Figure 2.2 shows the four exponents for $p = 0.1$." **(not yet discussed)**] At $t = p = 0.1$, the per-sample exponents are respectively

$$\mathrm{kl} = 4.44 \cdot 10^{-2}, \quad \text{Bennett} = 4.29 \cdot 10^{-2}, \quad \text{Bernstein} = 4.17 \cdot 10^{-2}, \quad \text{Hoeffding} = 2.00 \cdot 10^{-2}.$$

Bennett and Bernstein are within a few per cent of the truth; Hoeffding loses a factor 2 here, and a factor $\asymp 1/(4p)$ as $p \to 0$.

Remark 2.51 (How tight is the Cramér–Chernoff bound itself?). Very tight, and this can be made non-asymptotic. For $q = k/n$ with $k \in \{0, 1, \dots, n\}$, the combinatorial bound $\binom{n}{k} \geq e^{nH(q)}/(n+1)$, where $H(q) = -q \log q - (1-q)\log(1-q)$, gives

$$\mathbb{P}\big(\mathrm{Bin}(n,p) \geq nq\big) \;\geq\; \mathbb{P}\big(\mathrm{Bin}(n,p) = nq\big) = \binom{n}{k} p^k (1-p)^{n-k} \;\geq\; \frac{e^{-n\ \mathrm{kl}(q,p)}}{n+1},$$

since $\mathrm{kl}(q,p) = -H(q) - q\log p - (1-q)\log(1-p)$. Together with (2.12),

$$\frac{e^{-n\ \mathrm{kl}(q,p)}}{n+1} \;\leq\; \mathbb{P}\big(\mathrm{Bin}(n,p) \geq nq\big) \;\leq\; e^{-n\ \mathrm{kl}(q,p)}\,:$$

the Chernoff bound for binomial tails is exact up to a factor $n+1$, i.e. exact at the exponential scale, in agreement with Theorem 2.19. Everything that can still be gained is polynomial in n (Remark 2.21). This combinatorial argument is specific to the binomial law. The entropic argument of Theorem 3.47 gives a cruder bound, $e^{-n\,\mathrm{kl}(q,p) - O(\sqrt{n \log n})}$ (Corollary 3.48), but it only uses relative entropy and a deviation bound under a tilted law, and so extends to other laws (Remark 3.50); Section 3.4.4 compares the two.

Remark 2.52 (Which inequality to use). A practical summary of Chapter 1 and of this chapter. Let $X_1, \dots, X_n$ be independent with $\bar{X}_n$ their mean.

- *Nothing but a finite variance*: Chebyshev (1.19), with a $1/\delta$ dependence on the confidence – or, if a sub-Gaussian dependence is needed, change estimator (median of means, Exercise 1.14).

- *Bounded variables, no information on the variance*: Hoeffding (1.28). Simple, and optimal in the worst case.
- *Bounded variables with a small variance*: Bernstein (2.30). Always at least as good as Hoeffding up to the additive $b\log(1/\delta)/(3n)$, and much better when $\sigma \ll b$. This is the default choice.
- *Rare events, relative accuracy, Poisson-like counts*: Bennett (2.24), or the exact Chernoff bound (2.12) when the law is known.
- *Unbounded variables with exponential tails*: the sub-gamma class, Theorem 2.37, via Bernstein's moment condition (2.31).
- *A non-linear function of independent variables*: bounded differences (2.33); and if the differences are not uniformly small, one of the variance-sensitive tools of the later chapters.
- *A supremum over a class*: (2.33) for the fluctuation, and an entirely different theory for the expectation (Chapter 4).

Section 3.2 pushes this comparison further – a sharper lower bound on kl by Jensen's inequality, the limit $p \to 0$, and the kl confidence interval – before proving that the exponent kl itself cannot be improved.

2.6 Beyond sums: the bounded-difference inequality

All the inequalities so far concern a *sum* of independent variables, and they all rest on the identity $\psi_{S_n} = \sum_i \psi_{Z_i}$. But the quantities that matter in statistics and in combinatorics are rarely sums: the supremum of an empirical process, the chromatic number of a random graph, the error of a classifier built from the data, the largest eigenvalue of a random matrix are all functions of many independent variables, and none of them is linear. The following theorem, due to McDiarmid [McD89] in this form and going back to Azuma [Azu67] and Hoeffding [Hoe63] through the martingale route, extends Hoeffding's inequality to any function that depends little on each of its arguments.

(From now on $x = (x_1, \dots, x_n)$ denotes a point of a product space, as in Chapter 1; it no longer denotes the level of a Cramér transform.)

Definition 2.53 (Bounded differences)**.** Let $(\mathcal{X}_i, \mathcal{B}_i)_{1\le i\le n}$ be measurable spaces. A measurable function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ has the *bounded-difference property* with constants $c_1, \dots, c_n \ge 0$ if for every i, every $x = (x_1, \dots, x_n)$ and every $x_i' \in \mathcal{X}_i$,

$$\left|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)\right| \;\le\; c_i. \tag{2.32}$$

Changing the coordinates one at a time shows that such an f satisfies $\left|f(x) - f(y)\right| \le \sum_i c_i$ for all x, y; in particular f is bounded, so all the expectations below are well defined.

We first need a conditional version of Hoeffding's lemma. Its proof is the second proof of Lemma 1.62, carried out conditionally.

Lemma 2.54 (Conditional Hoeffding lemma)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-field and let Δ be a bounded random variable with $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$. Suppose there exist $\mathcal{G}$-measurable random variables $A \le B$ with $A \le \Delta \le B$ almost surely and $B - A \le w$ almost surely. Then*

$$\mathbb{E}\left[e^{\lambda\Delta} \mid \mathcal{G}\right] \;\le\; e^{\lambda^2 w^2/8} \qquad \text{almost surely, for every } \lambda \in \mathbb{R}.$$

Proof. On the event $\{A = B\}$ we have $\Delta = A = B$ and, since $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$, $\Delta = 0$ there, so the claim is trivial. On $\{A < B\}$, convexity of $z \mapsto e^{\lambda z}$ on $[A, B]$ gives, pointwise,

$$e^{\lambda\Delta} \;\le\; \frac{B-\Delta}{B-A}e^{\lambda A} + \frac{\Delta - A}{B-A}e^{\lambda B}.$$

All the coefficients are $\mathcal{G}$-measurable, so taking $\mathbb{E}[\cdot \mid \mathcal{G}]$ and using $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$,

$$\mathbb{E}\big[e^{\lambda\Delta} \mid \mathcal{G}\big] \;\le\; \frac{B}{B-A}e^{\lambda A} - \frac{A}{B-A}e^{\lambda B} = (1-p)e^{\lambda A} + p\,e^{\lambda B}, \qquad p := \frac{-A}{B-A} \in [0,1],$$

where $A \le 0 \le B$ because $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$ and $A \le \Delta \le B$. The right-hand side equals $e^{\psi_{\mathrm{Ber}(p)}(u)}$ with $u = \lambda(B-A)$ and $\psi_{\mathrm{Ber}(p)}$ the centred cumulant generating function (1.26) of a $\mathrm{Ber}(p)$ variable, which the second proof of Lemma 1.62 bounded by $u^2/8$, for every $u \in \mathbb{R}$ and every $p \in [0,1]$. Hence $\mathbb{E}[e^{\lambda\Delta} \mid \mathcal{G}] \le e^{\lambda^2(B-A)^2/8} \le e^{\lambda^2 w^2/8}$. □

Theorem 2.55 (Bounded-difference, or McDiarmid's, inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables with X_i taking values in $\mathcal{X}_i$, and let f satisfy (2.32). Write $Z = f(X_1, \dots, X_n)$. Then for every $t > 0$,*

$$\mathbb{P}\big(Z - \mathbb{E}Z \ge t\big) \le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big), \qquad \mathbb{P}\big(|Z - \mathbb{E}Z| \ge t\big) \le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\Big). \tag{2.33}$$

Proof. Let P_i denote the law of X_i, let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$, and define the *Doob martingale* $M_i := \mathbb{E}[Z \mid \mathcal{F}_i]$, so that $M_0 = \mathbb{E}Z$, $M_n = Z$ and

$$Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i, \qquad \Delta_i := M_i - M_{i-1}.$$

Step 1: an explicit form of M_i. By independence and Fubini's theorem, $M_i = g_i(X_1, \dots, X_i)$ almost surely, where

$$g_i(x_1, \dots, x_i) := \int f(x_1, \dots, x_i, y_{i+1}, \dots, y_n)\, \mathrm{d}P_{i+1}(y_{i+1}) \cdots \mathrm{d}P_n(y_n)$$

is measurable and bounded, with $g_n = f$ and $g_0 = \mathbb{E}Z$. Moreover g_i inherits (2.32): for all $x_1, \dots, x_{i-1}$ and all $x, x' \in \mathcal{X}_i$,

$$\big|g_i(x_1, \dots, x_{i-1}, x) - g_i(x_1, \dots, x_{i-1}, x')\big| \;\le\; c_i, \tag{2.34}$$

by integrating (2.32) over the remaining variables.

Step 2: conditional range of Δ_i. Fix i and write $G_i(y) := g_i(X_1, \dots, X_{i-1}, y)$, a random function of $y \in \mathcal{X}_i$ whose randomness is $\mathcal{F}_{i-1}$-measurable. Define

$$B_i := \inf\Big\{r \in \mathbb{Q} :\; \varrho_r = 0\Big\}, \qquad \text{where} \quad \varrho_r := \int \mathbb{1}\big\{G_i(y) > r\big\}\, \mathrm{d}P_i(y),$$

the essential supremum of G_i with respect to P_i, and let A_i be defined symmetrically with sup replaced by inf. For each fixed r the random variable ϱ_r is $\mathcal{F}_{i-1}$-measurable by Fubini's theorem, so $\{B_i < s\} = \bigcup_{r \in \mathbb{Q},\, r<s}\{\varrho_r = 0\}$ is $\mathcal{F}_{i-1}$-measurable: A_i and B_i are $\mathcal{F}_{i-1}$-measurable random variables. Since X_i is independent of $\mathcal{F}_{i-1}$ with law P_i, Fubini's theorem gives $\mathbb{P}\big(g_i(X_1, \dots, X_i) > B_i\big) = 0$ and similarly for A_i, so that

$$A_i \;\le\; g_i(X_1, \dots, X_i) = M_i \;\le\; B_i \qquad \text{a.s.},$$

and, averaging over X_i, also $A_i \le M_{i-1} \le B_i$ a.s. By (2.34), $B_i - A_i \le c_i$ almost surely. Consequently

$$A_i - M_{i-1} \;\le\; \Delta_i \;\le\; B_i - M_{i-1},$$

the two bounds are $\mathcal{F}_{i-1}$-measurable, their difference is at most c_i, and $\mathbb{E}[\Delta_i \mid \mathcal{F}_{i-1}] = 0$ by the tower property.

Step 3: tensorisation. By Lemma 2.54 applied with $\mathcal{G} = \mathcal{F}_{i-1}$, $\mathbb{E}[e^{\lambda\Delta_i} \mid \mathcal{F}_{i-1}] \le e^{\lambda^2 c_i^2/8}$ a.s. Hence, conditioning successively,

$$\mathbb{E}\Big[e^{\lambda \sum_{i\le n}\Delta_i}\Big] = \mathbb{E}\Big[e^{\lambda \sum_{i\le n-1}\Delta_i}\,\mathbb{E}\big[e^{\lambda\Delta_n} \mid \mathcal{F}_{n-1}\big]\Big] \le e^{\lambda^2 c_n^2/8}\,\mathbb{E}\Big[e^{\lambda \sum_{i\le n-1}\Delta_i}\Big] \le \;\dots\; \le\; e^{\lambda^2 \sum_{i\le n} c_i^2/8},$$

using at each step that $\sum_{i\le k-1}\Delta_i$ is $\mathcal{F}_{k-1}$-measurable. Theorem 1.57 and the optimisation already performed in the proof of Theorem 1.64 give (2.33); the lower tail follows by applying the result to $-f$, which has the same bounded differences. □

[V6 – Step 2 can be deleted: freeze the past and use the ordinary Hoeffding lemma] The whole apparatus of Step 2 – essential suprema, their measurability through rational levels, the $\mathcal{F}_{i-1}$-measurable envelopes $A_i \le \Delta_i \le B_i$, and even Lemma 2.54 – is unnecessary for Theorem 2.55. It suffices to freeze the first $i-1$ coordinates at *fixed values* and apply the unconditional Lemma 1.62, whose hypotheses are then about an honest real random variable.

Step 2 (replacement). Fix i and fix $x_1, \dots, x_{i-1}$, and set

$$D(y) := g_i(x_1, \dots, x_{i-1}, y) - g_{i-1}(x_1, \dots, x_{i-1}), \qquad y \in \mathcal{X}_i.$$

By definition of g_{i-1} one has $\int D \, \mathrm{d}P_i = 0$, and by (2.34) the oscillation of D satisfies $\sup D - \inf D \le c_i$ (both are finite, D being bounded). Thus $D(X_i)$ is a centred real random variable taking its values in an interval of length at most c_i, and Lemma 1.62 gives

$$\int_{\mathcal{X}_i} e^{\lambda D(y)}\,\mathrm{d}P_i(y) \;\le\; e^{\lambda^2 c_i^2/8},$$

a bound that does not depend on $x_1, \dots, x_{i-1}$. Since X_i is independent of $\mathcal{F}_{i-1}$, Fubini's theorem identifies $\mathbb{E}\big[e^{\lambda\Delta_i} \mid \mathcal{F}_{i-1}\big]$ with the left-hand side evaluated at $(X_1, \dots, X_{i-1})$, so it is at most $e^{\lambda^2 c_i^2/8}$ almost surely. Step 3 is unchanged. □

This removes about a page, needs no measurable selection, and makes Remark 2.58 immediate: what the proof uses is that $\operatorname{osc}(D) \le c_i$ for $P_1 \otimes \cdots \otimes P_{i-1}$-almost every $(x_1, \dots, x_{i-1})$, which Fubini deduces from the almost-everywhere bounded-difference property. Lemma 2.54 should be *kept*, but stated where it is actually needed – just before Theorem 2.56, whose envelopes A_i, B_i are genuinely random. The sentence of Section 2.8 about "the measurability of the essential suprema in Step 2" should then be deleted. *Review of 18 September: not yet discussed.*

The proof used the martingale structure and nothing else, so it proves more.

Theorem 2.56 (Azuma–Hoeffding). *Let $(\mathcal{F}_i)_{0\le i\le n}$ be a filtration and $(M_i)_{0\le i\le n}$ a martingale with respect to it. Assume that for each i there are $\mathcal{F}_{i-1}$-measurable random variables $A_i \le B_i$ with*

$$A_i \;\le\; M_i - M_{i-1} \;\le\; B_i \quad \text{and} \quad B_i - A_i \le c_i \qquad \text{almost surely.}$$

Then for every $t > 0$, $\mathbb{P}\big(M_n - M_0 \ge t\big) \le \exp\big(-2t^2/\sum_i c_i^2\big)$, and the same bound holds for the lower tail.

Proof. Steps 3 of the proof of Theorem 2.55, verbatim:[*replacement:* "Step 3 of the proof of Theorem 2.55, verbatim:" **(not yet discussed)**] the martingale increments $\Delta_i = M_i - M_{i-1}$ satisfy $\mathbb{E}[\Delta_i \mid \mathcal{F}_{i-1}] = 0$ and lie in an $\mathcal{F}_{i-1}$-measurable interval of length at most c_i, so Lemma 2.54 applies and the conditional moment generating functions multiply. □

Theorem 2.55 is the particular case of the Doob martingale $M_i = \mathbb{E}[f(X_1, \dots, X_n) \mid \mathcal{F}_i]$, and the classical form of Azuma's inequality is the case $|\Delta_i| \le c_i/2$ with deterministic c_i. The martingale formulation is strictly more general: it requires no independence at all, only that the increments be conditionally centred and conditionally bounded, which is what makes it the tool of choice for sequential procedures.

Remark 2.57 (Consistency check). If $f(x_1, \dots, x_n) = \sum_i x_i$ with $x_i \in [a_i, b_i]$, then f has bounded differences with $c_i = b_i - a_i$, and (2.33) is *exactly* Hoeffding's inequality (1.27). Theorem 2.55 is therefore a strict generalisation, obtained at no cost in the constants.

Remark 2.58 (The hypothesis only needs to hold almost everywhere). The proof uses (2.32) only through the inequality $B_i - A_i \le c_i$ of Step 2, which involves essential suprema with respect to P_i. Consequently Theorem 2.55 remains valid under the weaker hypothesis that, for each i,

$$\left|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)\right| \le c_i$$

for $\big(P_1 \otimes \cdots \otimes P_n \otimes P_i\big)$-almost every $(x_1, \dots, x_n, x_i')$. Indeed, by Fubini this implies (2.34) for $P_1 \otimes \cdots \otimes P_{i-1} \otimes P_i \otimes P_i$-almost every $(x_1, \dots, x_{i-1}, x, x')$, which is all that Step 2 requires. This apparently technical refinement is genuinely useful: it allows one to discard configurations that have probability zero – such as samples containing two identical points – and we shall need it in Section 2.7.6.

Remark 2.59 (What bounded differences cost). The constants c_i are *worst-case*: (2.32) must hold for every configuration, including the improbable ones. The resulting variance proxy $\frac{1}{4}\sum_i c_i^2$ can therefore be much larger than the actual variance of Z. For a sum of n i.i.d. Ber(p) variables, for instance, $\sum_i c_i^2 = n$ while $\mathbb{V}\mathrm{ar}(Z) = np(1-p)$: Theorem 2.55 is "Hoeffding-like" and suffers from exactly the defect that Bernstein's inequality repaired in the linear case. Repairing it in the non-linear case requires genuinely different tools – the Efron–Stein inequality [ES81], the entropy method, and Talagrand's convex-distance inequality [Tal95] – which are the subject of later chapters.

[V59 – Pointer to Chapter 5 (Efron–Stein, entropy method)] *Replacement of the previous sentence:* "Repairing it in the non-linear case requires genuinely different tools: the Efron–Stein inequality [ES81], which is the variance-level version of Theorem 2.55 and is exact for sums (Theorem 5.6), the entropy method, whose simplest instance gives Poisson-like concentration for self-bounding functions (Theorem 5.39), and, for sums of dependent terms, negative association (Section 5.4) – all three in Chapter 5, where the missing mass serves as the running example of what Theorem 2.55 cannot do (Remark 5.33) – and Talagrand's convex-distance inequality [Tal95], which is the subject of a later chapter." *Proposed on 15 October 2026 (chapter 5 written): not yet discussed.*

It is also worth noting that Theorem 2.55 is a statement about the concentration of Z *around its mean*, and says nothing whatsoever about the value of that mean: computing or bounding $\mathbb{E}Z$ is a separate problem, and usually the hard one.

2.6.1 Bernstein for martingales: Freedman's inequality

Theorem 2.56 is the martingale counterpart of Hoeffding's inequality, and it inherits its defect: it sees only the range of the increments. The counterpart of Bernstein's inequality exists as well, with the conditional variances playing the role of v. The proof introduces a device – an exponential supermartingale – that is the standard tool for all sequential arguments.

Lemma 2.60 (Conditional Bennett lemma)**.** *Let $\mathcal{G} \subseteq \mathcal{F}$ and let Δ be a bounded random variable with $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$ and $\Delta \le b$ almost surely, $b > 0$. Then, for every $\lambda > 0$,*

$$\mathbb{E}\big[e^{\lambda\Delta} \mid \mathcal{G}\big] \;\le\; \exp\Big(g(\lambda)\,\mathbb{E}\big[\Delta^2 \mid \mathcal{G}\big]\Big), \qquad g(\lambda) := \frac{e^{\lambda b} - 1 - \lambda b}{b^2}.$$

Proof. With φ_2 as in the proof of Lemma 2.39, $e^x = 1 + x + x^2\varphi_2(x)$ and φ_2 is non-decreasing, so $\varphi_2(\lambda\Delta) \le \varphi_2(\lambda b)$ pointwise. Taking conditional expectations and using $\mathbb{E}[\Delta \mid \mathcal{G}] = 0$,

$$\mathbb{E}\big[e^{\lambda\Delta} \mid \mathcal{G}\big] = 1 + \lambda^2\mathbb{E}\big[\Delta^2\varphi_2(\lambda\Delta) \mid \mathcal{G}\big] \le 1 + \lambda^2\varphi_2(\lambda b)\mathbb{E}\big[\Delta^2 \mid \mathcal{G}\big] \le \exp\Big(g(\lambda)\mathbb{E}[\Delta^2 \mid \mathcal{G}]\Big),$$

since $\lambda^2\varphi_2(\lambda b) = g(\lambda)$ and $1 + x \le e^x$. □

Theorem 2.61 (Freedman)**.** *Let $(M_i)_{0\le i\le n}$ be a martingale with respect to $(\mathcal{F}_i)$, with increments $\Delta_i = M_i - M_{i-1}$ satisfying $\Delta_i \le b$ almost surely. Let $V_n := \sum_{i=1}^n \mathbb{E}\big[\Delta_i^2 \mid \mathcal{F}_{i-1}\big]$ be the predictable quadratic variation. If $V_n \le v$ almost surely, then for every $t > 0$,*

$$\mathbb{P}\big(M_n - M_0 \ge t\big) \;\le\; \exp\Big(-\frac{t^2}{2\big(v + \frac{bt}{3}\big)}\Big).$$

Proof. Fix $\lambda \in (0, 3/b)$ and set $W_k := \exp\big(\lambda(M_k - M_0) - g(\lambda)V_k\big)$, with $W_0 = 1$. Since V_k is $\mathcal{F}_{k-1}$-measurable, Lemma 2.60 gives

$$\mathbb{E}\big[W_k \mid \mathcal{F}_{k-1}\big] = W_{k-1}\, e^{-g(\lambda)\mathbb{E}[\Delta_k^2 \mid \mathcal{F}_{k-1}]}\,\mathbb{E}\big[e^{\lambda\Delta_k} \mid \mathcal{F}_{k-1}\big] \;\le\; W_{k-1},$$

so (W_k) is a non-negative supermartingale and $\mathbb{E}[W_n] \le 1$. On the event $\{M_n - M_0 \ge t\}$ we have, using $V_n \le v$, $W_n \ge \exp\big(\lambda t - g(\lambda)v\big)$, whence by Markov's inequality

$$\mathbb{P}\big(M_n - M_0 \ge t\big) \;\le\; \exp\big(-\lambda t + g(\lambda)v\big).$$

By Lemma 2.46, $g(\lambda) \le \frac{\lambda^2}{2(1-\lambda b/3)}$, so the right-hand side is at most $\exp\big(-\lambda t + \frac{\lambda^2 v}{2(1-\lambda b/3)}\big)$; optimising as in Theorem 2.37 gives the announced bound. □

Remark 2.62. Three comments. First, Theorem 2.61 contains Theorem 2.47: for a sum of independent centred variables the predictable quadratic variation is the deterministic quantity $v = \sum_i \mathbb{E}[Z_i^2]$. Second, the hypothesis "$V_n \le v$ almost surely" can be relaxed to a statement on the event $\{V_n \le v\}$, which is what makes the inequality usable when the variance is itself random – as in sequential experiments, where it is the main tool. Third, the exponential supermartingale (W_k) is more than a proof device: by the optional stopping theorem the bound $\mathbb{E}[W_\tau] \le 1$ holds at any stopping time τ, which is exactly what is needed to build the time-uniform confidence sequences alluded to in Remark 1.84.

2.7 Applications

2.7.1 Finite hypothesis classes: the first PAC bounds

We return to the framework of Section 1.3. Let $\mathcal{H}$ be a *finite* set of classifiers – a "hypothesis class" – and let $\widehat{h}_n \in \arg\min_{h\in\mathcal{H}} L_{\mathcal{S}_n}(h)$ be an empirical risk minimiser. Remark 1.71 explained why Corollary 1.70 does not apply to $\widehat{h}_n$; the remedy is to control all the hypotheses at once.

We first give a name to what we are trying to prove.

Definition 2.63 (PAC learnability and sample complexity)**.** A hypothesis class $\mathcal{H}$ is *agnostically PAC learnable – probably approximately correct –* if there exist a learning algorithm $\mathcal{A} = (\mathcal{A}_n)$ and a function $n_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that, for all $\varepsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ and every $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$,

$$\mathbb{P}\Big(L_{\mathcal{D}}(\widehat{h}_n) \;\geq\; \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \Big) \;\leq\; \delta, \qquad \widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n),\ \mathcal{S}_n \sim \mathcal{D}^{\otimes n}.$$

It is *PAC learnable in the realizable case* if the same holds for all $\mathcal{D}$ such that $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$. For a given algorithm, each pair (ε, δ) has a smallest admissible value of $n_{\mathcal{H}}(\varepsilon, \delta)$; the infimum of these values over all learning algorithms, taken separately for each pair, is the *sample complexity* of $\mathcal{H}$.

Remark 2.64. Valiant's original definition [Val84] also requires the algorithm to run in time polynomial in $1/\varepsilon$ and $1/\delta$; we retain only the statistical half of the requirement. Note the order of the quantifiers: the sample size may not depend on $\mathcal{D}$, which is what makes the theory distribution-free, and is also what makes it pessimistic.

Theorem 2.65 (Uniform deviations over a finite class)**.** *For every $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} \big| L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h) \big| \;\leq\; \sqrt{\frac{\log\big(2\,|\mathcal{H}|\,/\delta\big)}{2n}}. \tag{2.35}$$

Consequently, on the same event,

$$L_{\mathcal{D}}(\widehat{h}_n) \;\leq\; \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{2\log\big(2\,|\mathcal{H}|\,/\delta\big)}{n}}, \tag{2.36}$$

and $n \geq 2\varepsilon^{-2}\log(2\,|\mathcal{H}|\,/\delta)$ samples suffice for the excess risk of the empirical risk minimiser over $\mathcal{H}$ to be at most ε with probability $1 - \delta$.

Proof. Fix $h \in \mathcal{H}$. By Corollary 1.70, $\mathbb{P}\big(\big|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\big| > t\big) \leq 2e^{-2nt^2}$. By the union bound over the $|\mathcal{H}|$ hypotheses,

$$\mathbb{P}\Big(\sup_{h \in \mathcal{H}} \big| L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h) \big| > t \Big) \;\leq\; 2\,|\mathcal{H}|\; e^{-2nt^2},$$

which equals δ for $t = \sqrt{\log(2\,|\mathcal{H}|\,/\delta)/(2n)}$. For (2.36), let $\bar{h} \in \arg\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ be a best element of the class – not to be confused with the Bayes classifier h^* of Definition 1.12, which need not belong to $\mathcal{H}$ – and write, on the event of (2.35),

$$L_{\mathcal{D}}(\widehat{h}_n) \leq L_{\mathcal{S}_n}(\widehat{h}_n) + t \leq L_{\mathcal{S}_n}(\bar{h}) + t \leq L_{\mathcal{D}}(\bar{h}) + 2t. \qquad \square$$

Three features of (2.36) deserve emphasis. The dependence on the size of the class is *logarithmic*, so that doubling the number of candidate classifiers costs almost nothing – one may search over millions of hypotheses with a few thousand samples. The dependence on the confidence level is logarithmic as well, thanks to Hoeffding. And the rate is $n^{-1/2}$, which is the rate of the central limit theorem. The next two results show that the last feature is not the end of the story.

Theorem 2.66 (Realizable case: a fast rate)**.** *Assume that some $\bar{h} \in \mathcal{H}$ satisfies $L_{\mathcal{D}}(\bar{h}) = 0$ (for a finite class this is the realizable case $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$ of Definition 2.63), and let $\widehat{h}$ be any element of $\mathcal{H}$ with $L_{\mathcal{S}_n}(\widehat{h}) = 0$ (such an element exists, since $L_{\mathcal{S}_n}(\bar{h}) = 0$ almost surely). Then, with probability at least $1 - \delta$,*

$$L_{\mathcal{D}}(\widehat{h}) \;\leq\; \frac{\log(|\mathcal{H}|\,/\delta)}{n}.$$

Proof. Let $\varepsilon > 0$ and let $\mathcal{H}_\varepsilon = \{h \in \mathcal{H} : L_{\mathcal{D}}(h) > \varepsilon\}$ be the set of "bad" hypotheses. For $h \in \mathcal{H}_\varepsilon$, the probability that h makes no error on the sample is $(1 - L_{\mathcal{D}}(h))^n < (1-\varepsilon)^n \le e^{-n\varepsilon}$. By the union bound,

$$\mathbb{P}\big(\exists h \in \mathcal{H}_\varepsilon : \; L_{\mathcal{S}_n}(h) = 0\big) \;\le\; |\mathcal{H}|\, e^{-n\varepsilon},$$

which is at most δ for $\varepsilon = \log(|\mathcal{H}|/\delta)/n$. On the complementary event, any h with zero empirical risk has $L_{\mathcal{D}}(h) \le \varepsilon$. □

The rate is now n^{-1} instead of $n^{-1/2}$: when the problem is noiseless, learning is quadratically cheaper. Bernstein's inequality explains why, and interpolates between the two regimes.

Theorem 2.67 (Variance-adaptive bound)**.** *Let $y := \log(|\mathcal{H}|/\delta)$. With probability at least $1-\delta$, every $h \in \mathcal{H}$ satisfies*

$$L_{\mathcal{D}}(h) \;\le\; L_{\mathcal{S}_n}(h) + \sqrt{\frac{2L_{\mathcal{S}_n}(h)\, y}{n}} + \frac{4y}{n}. \tag{2.37}$$

Proof. Fix h and write $L = L_{\mathcal{D}}(h)$. The variables $W_i := L - \mathbb{1}\{h(X_i) \neq Y_i\}$, $1 \le i \le n$, are i.i.d., centred, bounded above by $L \le 1$, and $\mathbb{E}[W_i^2] = L(1-L) \le L$. By (2.29) applied to $\sum_i W_i = n\big(L - L_{\mathcal{S}_n}(h)\big)$ with $v = nL(1-L) \le nL$ and $b = 1$, with probability at least $1 - e^{-y}$,

$$L \;<\; L_{\mathcal{S}_n}(h) + \sqrt{\frac{2Ly}{n}} + \frac{y}{3n}. \tag{2.38}$$

A union bound over $\mathcal{H}$ with $e^{-y} = \delta/|\mathcal{H}|$ makes (2.38) hold simultaneously for all $h \in \mathcal{H}$ with probability at least $1-\delta$. It remains to solve the inequality. Put $A = \sqrt{2y/n}$, $B = L_{\mathcal{S}_n}(h) + y/(3n)$ and $u = \sqrt{L}$, so that $u^2 < Au + B$, whence $u < \frac{1}{2}\big(A + \sqrt{A^2 + 4B}\big)$ and, using $\sqrt{A^2+4B} \le A + 2\sqrt{B}$,

$$L = u^2 < \frac{A^2 + 2B}{2} + \frac{A}{2}\sqrt{A^2+4B} \le A^2 + B + A\sqrt{B}.$$

Finally $\sqrt{B} \le \sqrt{L_{\mathcal{S}_n}(h)} + \sqrt{y/(3n)}$ gives

$$L < L_{\mathcal{S}_n}(h) + \sqrt{\frac{2L_{\mathcal{S}_n}(h)y}{n}} + \Big(2 + \frac{1}{3} + \sqrt{\frac{2}{3}}\Big)\frac{y}{n} \;\le\; L_{\mathcal{S}_n}(h) + \sqrt{\frac{2L_{\mathcal{S}_n}(h)y}{n}} + \frac{4y}{n}. \qquad \square$$

Remark 2.68 (Reading (2.37)). The bound is *empirical*: its right-hand side is computable from the data. It interpolates exactly between the two previous theorems. If $L_{\mathcal{S}_n}(h) = 0$ it gives $L_{\mathcal{D}}(h) \le 4\log(|\mathcal{H}|/\delta)/n$, the fast rate of Theorem 2.66 up to the constant 4. If $L_{\mathcal{S}_n}(h)$ is of order one it gives the $n^{-1/2}$ rate of Theorem 2.65, but with the factor $\sqrt{L_{\mathcal{S}_n}(h)}$ in front: a classifier with small empirical risk is certified more precisely than a mediocre one. This is the probabilistic counterpart of Remark 1.14: low noise means small variance, small variance means fast rates, and Bernstein's inequality is the instrument that converts the second statement into the third. Under Massart's or Tsybakov's margin conditions, a refinement of the same argument yields excess-risk bounds of order n^{-1} and $n^{-(1+\alpha)/(2+\alpha)}$ respectively; see [MN06; BBL05].

All the results of this subsection use the finiteness of $\mathcal{H}$ through a union bound, and say nothing about the classes used in practice – half-spaces, intervals, neural networks – which are infinite. Chapter 4 shows that for such classes $\log|\mathcal{H}|$ can be replaced by the Vapnik–Chervonenkis dimension, which characterises learnability and determines the sample complexity up to a logarithmic factor (Theorems 4.59 and 4.65).

2.7.2 Fast rates under Massart's margin condition

Theorem 2.67 shows that a small empirical risk is certified at the rate n^{-1}; Theorem 2.66 shows that a noiseless problem is learned at the rate n^{-1}. We now prove a statement that contains both and explains them: what produces fast rates is not the absence of noise, but the fact that the *variance* of the relevant variables is controlled by their *mean*. This is precisely the situation in which Bernstein's inequality beats Hoeffding's, and the hypothesis that creates it in classification is Massart's margin condition, already met in Remark 1.14.

Assumption 2.69 (Massart's margin condition)**.** There exists $\gamma > 0$ such that $\left|2\eta(X) - 1\right| \geq 2\gamma$ almost surely.

Since $\left|2\eta(X) - 1\right| \leq 1$ always, necessarily $\gamma \leq \frac{1}{2}$. The condition says that the regression function stays away from the critical level $\frac{1}{2}$: at every point, the Bayes rule of Definition 1.12 is right with conditional probability at least $\frac{1}{2} + \gamma$. It is a *low-noise*, not a noiseless, assumption. Indeed $\min\{\eta, 1-\eta\} = \frac{1}{2}(1 - \left|2\eta - 1\right|)$, so (1.7) gives

$$L^*_{\mathcal{D}} = \mathbb{E}\big[\min\big\{\eta(X), 1-\eta(X)\big\}\big] = \frac{1}{2}\Big(1 - \mathbb{E}\big[\left|2\eta(X) - 1\right|\big]\Big) \;\leq\; \frac{1}{2} - \gamma,$$

with equality if and only if $\left|2\eta(X) - 1\right| = 2\gamma$ almost surely: the Bayes risk may be as large as $\frac{1}{2} - \gamma$, and the realizable case $L^*_{\mathcal{D}} = 0$ corresponds to the extreme value $\gamma = \frac{1}{2}$.

Theorem 2.70 (Fast rates for empirical risk minimisation)**.** *Consider binary classification, $\mathcal{Y} = \{0,1\}$, with the notation of Section 1.4: $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$, $h^* = \mathbb{1}\{\eta \geq \frac{1}{2}\}$ the Bayes classifier of Definition 1.12 and $L^*_{\mathcal{D}} = L_{\mathcal{D}}(h^*)$ the Bayes risk. Let $\mathcal{H}$ be finite, assume that $h^* \in \mathcal{H}$ and that Assumption 2.69 holds, and let $\widehat{h}_n \in \arg\min_{h \in \mathcal{H}} L_{\mathcal{S}_n}(h)$. Then, for every $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$L_{\mathcal{D}}(\widehat{h}_n) - L^*_{\mathcal{D}} \;\leq\; \frac{2\log\big(|\mathcal{H}|\,/\delta\big)}{\gamma\, n}. \tag{2.39}$$

In particular the sample complexity of $\mathcal{H}$ satisfies $n_{\mathcal{H}}(\varepsilon, \delta) \leq \dfrac{2\log(|\mathcal{H}|\,/\delta)}{\gamma\,\varepsilon}$.

Proof. For $h \in \mathcal{H}$ and $1 \leq i \leq n$ set

$$Z_i(h) := \mathbb{1}\{h^*(X_i) \neq Y_i\} - \mathbb{1}\{h(X_i) \neq Y_i\},$$

so that the $Z_i(h)$ are i.i.d., $\frac{1}{n}\sum_i Z_i(h) = L_{\mathcal{S}_n}(h^*) - L_{\mathcal{S}_n}(h)$ and $\mathbb{E}[Z_1(h)] = L^*_{\mathcal{D}} - L_{\mathcal{D}}(h)$.

Step 1: the basic inequality. Adding and subtracting,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(h) - \mathbb{E}[Z_i(h)]\Big) = \Big(L_{\mathcal{S}_n}(h^*) - L_{\mathcal{S}_n}(h)\Big) + \Big(L_{\mathcal{D}}(h) - L^*_{\mathcal{D}}\Big).$$

Taking $h = \widehat{h}_n$ and using $L_{\mathcal{S}_n}(\widehat{h}_n) \leq L_{\mathcal{S}_n}(h^*)$, which holds because $h^* \in \mathcal{H}$ and $\widehat{h}_n$ minimises the empirical risk over $\mathcal{H}$,

$$L_{\mathcal{D}}(\widehat{h}_n) - L^*_{\mathcal{D}} \;\leq\; \frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(\widehat{h}_n) - \mathbb{E}[Z_i(\widehat{h}_n)]\Big). \tag{2.40}$$

Step 2: the variance is controlled by the mean. The variable $Z_1(h)$ takes its values in $\{-1, 0, 1\}$ and vanishes unless $h(X_1) \neq h^*(X_1)$, because when the two classifiers agree the two indicators are equal. Hence

$$\mathbb{E}\big[Z_1(h)^2\big] = \mathbb{P}\big(h(X) \neq h^*(X)\big).$$

On the other hand, the excess-risk identity (1.6) together with Assumption 2.69 gives

$$L_{\mathcal{D}}(h) - L^*_{\mathcal{D}} = \mathbb{E}\Big[\big|2\eta(X) - 1\big| \, \mathbb{1}\{h(X) \neq h^*(X)\}\Big] \;\geq\; 2\gamma \, \mathbb{P}\big(h(X) \neq h^*(X)\big),$$

so that

$$\mathbb{E}\big[Z_1(h)^2\big] \;\leq\; \frac{L_{\mathcal{D}}(h) - L^*_{\mathcal{D}}}{2\gamma} \qquad \text{for every } h \in \mathcal{H}. \tag{2.41}$$

Moreover $Z_i(h) - \mathbb{E}[Z_i(h)] \leq 2$ almost surely.

Step 3: Bernstein and a union bound. Fix $h \in \mathcal{H}$ and apply (2.29) to the centred variables $\big(Z_i(h) - \mathbb{E}[Z_i(h)]\big)/n$, whose variance factor is $\mathbb{E}[Z_1(h)^2]/n$ and whose upper bound is $2/n$: with probability at least $1 - \delta/|\mathcal{H}|$,

$$\frac{1}{n}\sum_{i=1}^{n}\Big(Z_i(h) - \mathbb{E}[Z_i(h)]\Big) \;<\; \sqrt{\frac{2\,\mathbb{E}[Z_1(h)^2]\,y}{n}} + \frac{2y}{3n}, \qquad y := \log\frac{|\mathcal{H}|}{\delta}.$$

By the union bound this holds simultaneously for all $h \in \mathcal{H}$ with probability at least $1 - \delta$; on that event, combining with (2.40) and (2.41) and writing $x := L_{\mathcal{D}}(\widehat{h}_n) - L^*_{\mathcal{D}} \geq 0$ and $\beta := y/n$,

$$x \;\leq\; \sqrt{\frac{x\,\beta}{\gamma}} + \frac{2\beta}{3} \;=:\; g(x). \tag{2.42}$$

Step 4: solving the self-bounding inequality. The function $f(x) = g(x) - x$ is concave on $[0, \infty)$ with $f(0) = 2\beta/3 > 0$. Let $x_0 := 2\beta/\gamma$. Then

$$g(x_0) = \sqrt{\frac{2\beta^2}{\gamma^2}} + \frac{2\beta}{3} = \beta\Big(\frac{\sqrt{2}}{\gamma} + \frac{2}{3}\Big) \;\leq\; \frac{2\beta}{\gamma} = x_0,$$

the inequality being equivalent to $\frac{2}{3} \leq \frac{2-\sqrt{2}}{\gamma}$, i.e. to $\gamma \leq \frac{3}{2}(2 - \sqrt{2}) \approx 0.879$, which holds since $\gamma \leq \frac{1}{2}$. Thus $f(x_0) \leq 0$. If we had $x > x_0$, concavity applied to $x_0 = (1-\theta)\cdot 0 + \theta x$ with $\theta = x_0/x \in (0,1)$ would give $f(x_0) \geq (1-\theta)f(0) + \theta f(x)$, hence $\theta f(x) \leq f(x_0) - (1-\theta)f(0) < 0$ and $f(x) < 0$, contradicting (2.42). Therefore $x \leq x_0 = 2y/(\gamma n)$, which is (2.39). □

[V7 – Step 4 in one line, by arithmetic–geometric mean] The concavity argument is not needed. Replacement for Step 4:

Step 4 (replacement): solving the self-bounding inequality. By the arithmetic–geometric mean inequality, $\sqrt{x \cdot \frac{\beta}{\gamma}} \leq \frac{1}{2}\big(x + \frac{\beta}{\gamma}\big)$, *so* (2.42) *gives*

$$x \;\leq\; \frac{x}{2} + \frac{\beta}{2\gamma} + \frac{2\beta}{3}, \qquad \text{hence} \qquad x \;\leq\; \frac{\beta}{\gamma} + \frac{4\beta}{3} \;\leq\; \frac{2\beta}{\gamma} = \frac{2y}{\gamma n},$$

the last step because $\frac{4}{3} \leq \frac{1}{\gamma}$ *for* $\gamma \leq \frac{3}{4}$, *which holds since* $\gamma \leq \frac{1}{2}$. *This is* (2.39). □

This is precisely the general lemma of Exercise 2.8 ($x \leq a\sqrt{x} + b \Rightarrow x \leq a^2 + 2b$) applied with $a = \sqrt{\beta/\gamma}$ and $b = 2\beta/3$; the solution of that exercise currently claims the opposite and must be corrected as well (see the revision there).

Review of 18 September: not yet discussed.

Remark 2.71 (What made the rate fast). Compare (2.39) with (2.36): the rate has improved from $n^{-1/2}$ to n^{-1}, at the price of a factor $1/\gamma$. The mechanism is worth isolating, because it recurs throughout statistical learning:

(i) the quantity to be controlled, $L_{\mathcal{D}}(\widehat{h}_n) - L^*_{\mathcal{D}}$, is the mean of variables whose *variance is bounded by their mean* – inequality (2.41);

(ii) Bernstein's inequality then bounds the deviation by $\sqrt{2\,(\text{mean})\,y/n}$ rather than by $\sqrt{y/n}$, producing an inequality of the form $x \le \sqrt{\kappa x y/n} + \kappa' y/n$;

(iii) such a *self-bounding* inequality forces $x \lesssim y/n$.

Nothing in this scheme is specific to classification. Note also that the hypothesis $h^* \in \mathcal{H}$ is essential: without it, $\widehat{h}_n$ is compared to the best element of $\mathcal{H}$ and an approximation term reappears, which is of order $n^{-1/2}$ in general. Tsybakov's condition $\mathbb{P}\big(\big|2\eta(X)-1\big| \le s\big) \le Cs^{\alpha}$ interpolates between Assumption 2.69 ($\alpha = \infty$) and no assumption at all ($\alpha = 0$), and yields rates $n^{-(1+\alpha)/(2+\alpha)}$; see [MT99; Tsy04; MN06].

2.7.3 Rare events: estimating a small probability

Suppose we want to estimate a probability p known to be small – the error rate of a good classifier, the failure rate of a component, the frequency of a rare word – and that we want a *relative* accuracy: $\big|\hat{p} - p\big| \le \varepsilon p$ rather than $\big|\hat{p} - p\big| \le \varepsilon$. This is where the difference between Hoeffding and Bernstein becomes a difference of nature.

Proposition 2.72. *Let $X_1, \dots, X_n$ be i.i.d.* $\mathrm{Ber}(p)$, $\hat{p} = \bar{X}_n$, *and let $\varepsilon \in (0,1]$. Then*

$$\mathbb{P}\big(\big|\hat{p} - p\big| \ge \varepsilon p\big) \;\le\; 2\exp\Big(-\frac{3n\varepsilon^2 p}{8}\Big),$$

so that $n \ge \frac{8}{3}\,\frac{\log(2/\delta)}{\varepsilon^2 p}$ samples suffice for a relative accuracy ε at confidence $1-\delta$. Hoeffding's inequality, by contrast, requires $n \ge \frac{\log(2/\delta)}{2\varepsilon^2 p^2}$.

Proof. Apply (2.28) to $\sum_i \pm(X_i - p)$ with $v = np(1-p)$, $b = 1$ and $t = n\varepsilon p$:

$$\frac{t^2}{2\big(v + \frac{bt}{3}\big)} = \frac{n^2\varepsilon^2 p^2}{2\big(np(1-p) + \frac{n\varepsilon p}{3}\big)} = \frac{n\varepsilon^2 p}{2(1-p) + \frac{2\varepsilon}{3}} \;\ge\; \frac{n\varepsilon^2 p}{2 + \frac{2}{3}} = \frac{3n\varepsilon^2 p}{8},$$

using $\varepsilon \le 1$. The Hoeffding computation is (1.28) with $t = \varepsilon p$. □

The two sample sizes differ by a factor $\asymp 1/p$. For $p = 10^{-3}$, $\varepsilon = 10\%$ and $\delta = 5\%$, Bernstein's inequality asks for $n \approx 9.8 \cdot 10^5$ observations and Hoeffding's for $n \approx 1.8 \cdot 10^8$. The intuition is the one of Remark 2.38: estimating a rare event is governed by the number np of *occurrences* observed, not by the number n of observations, and only a variance-sensitive inequality can see this. The order $\log(1/\delta)/(\varepsilon^2 p)$ cannot be improved: by Proposition 3.57, proved with the Bretagnolle–Huber inequality, an estimator with this guarantee at every small p needs at least $\frac{3}{32}\log(1/(4\delta))/(\varepsilon^2 p)$ samples, when $p \le \frac{1}{6}$, $\varepsilon \le \frac{1}{2}$ and $\delta < \frac{1}{4}$.

2.7.4 A preview: random projections and the Johnson–Lindenstrauss lemma

The chi-square bound of Remark 2.50 has a spectacular consequence, which we record here because its proof is now two lines, and which is the starting point of the theory of dimension reduction developed in a later chapter.

Theorem 2.73 (Johnson–Lindenstrauss). *Let $x_1, \dots, x_N \in \mathbb{R}^d$, let $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, and let G be an $r \times d$ random matrix with independent $\mathcal{N}(0, 1/r)$ entries. If*

$$r \;\geq\; \frac{8}{\varepsilon^2} \log \frac{N^2}{\delta},$$

then, with probability at least $1 - \delta$,

$$(1-\varepsilon)\left\|x_i - x_j\right\|_2^2 \;\leq\; \left\|Gx_i - Gx_j\right\|_2^2 \;\leq\; (1+\varepsilon)\left\|x_i - x_j\right\|_2^2 \qquad \text{for all } 1 \leq i < j \leq N.$$

Proof. Fix $x \neq 0$ in $\mathbb{R}^d$.[*sentence to add:* "pairs with $x_i = x_j$ satisfy the conclusion trivially, so the union bound below may be restricted to the pairs with $x_i \neq x_j$." **(not yet discussed)**] The rows of G are independent $\mathcal{N}(0, I_d/r)$ vectors, so the coordinates of Gx are independent $\mathcal{N}(0, \|x\|_2^2/r)$ variables and $\|Gx\|_2^2 = \frac{\|x\|_2^2}{r}\,\chi^2$ with χ^2 a chi-square variable with r degrees of freedom. By Exercise 2.5, $\chi^2 - r \in \Gamma_+(2r, 2)$, so Theorem 2.37 with $y = r\varepsilon^2/8$ gives

$$\mathbb{P}\big(\chi^2 - r \geq \varepsilon r\big) \leq e^{-r\varepsilon^2/8}, \qquad \text{because} \qquad 2\sqrt{ry} + 2y = r\varepsilon\left(\tfrac{1}{\sqrt{2}} + \tfrac{\varepsilon}{4}\right) \leq r\varepsilon$$

for $\varepsilon \leq 1$; and the left-tail half of Exercise 2.5 with $y = r\varepsilon^2/4$ gives $\mathbb{P}(\chi^2 - r \leq -\varepsilon r) \leq e^{-r\varepsilon^2/4}$. Hence

$$\mathbb{P}\Big(\big|\|Gx\|_2^2 - \|x\|_2^2\big| > \varepsilon\,\|x\|_2^2\Big) \;\leq\; 2e^{-r\varepsilon^2/8}.$$

Apply this to the $\binom{N}{2}$ vectors $x_i - x_j$ and take a union bound: the failure probability is at most $2\binom{N}{2}e^{-r\varepsilon^2/8} \leq N^2 e^{-r\varepsilon^2/8} \leq \delta$ under the stated condition on r. □

Remark 2.74. The target dimension $r \asymp \varepsilon^{-2}\log N$ does not depend on the ambient dimension d at all: any N points of a Euclidean space of any dimension can be embedded into $O(\varepsilon^{-2}\log N)$ dimensions with all pairwise distances preserved up to a factor $1 \pm \varepsilon$. This is the most useful positive result of high-dimensional geometry, and the counterpoint to the curse of dimensionality of Remark 1.26: dimension is an obstacle when one needs to *fill* space, and a much smaller one when one only needs to *separate* finitely many points. We shall return to random projections, to the optimality of the bound on r, and to what can be said for infinite sets of points.

2.7.5 Bounded differences at work

Suprema of empirical processes. Let $\mathcal{H}$ be a class of classifiers and consider

$$Z := \sup_{h \in \mathcal{H}} \big| L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h) \big|,$$

the quantity that Theorem 2.65 bounded for finite $\mathcal{H}$. As a function of the n independent observations (X_i, Y_i), Z has the bounded-difference property with $c_i = 1/n$: modifying one observation changes each $L_{\mathcal{S}_n}(h)$ by at most $1/n$, hence changes the supremum by at most $1/n$. Theorem 2.55 gives at once

$$\mathbb{P}\big(|Z - \mathbb{E}Z| \geq t\big) \;\leq\; 2e^{-2nt^2} \tag{2.43}$$

– for *any* class $\mathcal{H}$, finite or not, of any complexity whatsoever (assuming Z measurable, which holds e.g. if $\mathcal{H}$ is countable or if the supremum can be restricted to a countable subclass). This is a remarkable statement, and it contains the strategic insight of statistical

learning theory: *the fluctuation of the uniform deviation is always of order $n^{-1/2}$, whatever the class; the whole difficulty is the size of its expectation*. For a finite class, combining (2.43) with Proposition 2.28 – each $L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)$ is sub-Gaussian with proxy $1/(4n)$ – gives back Theorem 2.65 up to constants. For infinite classes, bounding $\mathbb{E}Z$ requires new ideas: symmetrisation, Rademacher complexity and the Vapnik–Chervonenkis dimension, which are the subject of Chapter 4, and chaining, which removes a logarithmic factor (Theorem 4.52, stated there without proof). In Theorem 4.46, the argument of (2.43) is applied to one-sided suprema, and Theorem 4.41 bounds their expectation by a Rademacher complexity.

The chromatic number of a random graph. Let $G \sim G(n,p)$ be an Erdős–Rényi graph with vertex set $\{1,\dots,n\}$ and edge set E, in which each of the $\binom{n}{2}$ edges is present independently with probability p, and let $\chi(G)$ be its chromatic number.

Proposition 2.75 (Shamir–Spencer)**.** *For every $t > 0$, $\mathbb{P}\big(|\chi(G) - \mathbb{E}\chi(G)| \geq t\big) \leq 2e^{-2t^2/(n-1)}$.*

Proof. Use *vertex exposure*: for $2 \leq i \leq n$ let $X_i := \big(\mathbb{1}\{\{i,j\} \in E\}\big)_{1\leq j<i} \in \{0,1\}^{i-1}$. The variables $X_2,\dots,X_n$ are independent and determine G, so $\chi(G) = f(X_2,\dots,X_n)$ for a measurable f. Modify X_i into X_i', obtaining a graph G': from a proper colouring of G' with $\chi(G')$ colours, one obtains a proper colouring of G by giving vertex i a brand-new colour, so $\chi(G) \leq \chi(G') + 1$; exchanging the roles of G and G' gives $|\chi(G) - \chi(G')| \leq 1$. Thus f has bounded differences with $c_i = 1$ for the $n-1$ variables, and Theorem 2.55 applies. □

Remark 2.76 (Two lessons). First, Proposition 2.75 localises $\chi(G)$ in a window of width $O(\sqrt{n})$ around its mean *without any knowledge of that mean* – the determination of $\mathbb{E}\chi(G(n,1/2)) \sim n/(2\log_2 n)$ is a much harder and later result. Concentration and computation of the mean are genuinely independent problems. Second, the choice of the independent coordinates matters: had we used *edge exposure*, taking the $\binom{n}{2}$ edge indicators as the independent variables, each with $c_e = 1$, we would have obtained the far weaker bound $2\exp\big(-2t^2/\binom{n}{2}\big)$, useless for $t \ll n$. The same random object can be written as a function of independent variables in many ways, and Theorem 2.55 rewards the decomposition with the smallest $\sum_i c_i^2$.

2.7.6 Back to nearest neighbours: concentration of the leave-one-out error

Remark 1.30 left us with a question: Theorem 1.21 bounds the *expected* risk $R_n(\widehat{h}^{\mathrm{NN}}) = \mathbb{E}[L_{\mathcal{D}}(\widehat{h}_n^{\mathrm{NN}})]$ of the nearest-neighbour rule, but a practitioner has one sample and wants a guarantee for it. Can Theorem 2.55 be applied to $L_{\mathcal{D}}(\widehat{h}_n^{\mathrm{NN}})$, viewed as a function of the n independent observations?

Counterexample 2.77 (The naive attempt fails)**.** It cannot. Take $d = 1$, $\mathcal{X} = [0,1)$ with $\mathcal{D}_X$ uniform, and $\eta = \mathbb{1}_{[1/2,1)}$, so that $L_{\mathcal{D}}^* = 0$; take $n = 2$ and let $a > 0$ be small. For the sample $(x_1,y_1) = (0,0)$, $(x_2,y_2) = (1-a,1)$, the rule predicts 0 on $[0,\frac{1-a}{2})$ and 1 beyond, so its risk is $a/2$. Replacing the first observation by $(1-2a,1)$ produces a rule that predicts 1 everywhere, of risk $\frac{1}{2}$: one observation has changed the risk by almost $\frac{1}{2}$. The phenomenon persists for every n, since a single sample point can be the nearest neighbour of a region of arbitrarily large probability. The map $\big((x_i,y_i)\big)_{i\leq n} \mapsto L_{\mathcal{D}}(\widehat{h}_n^{\mathrm{NN}})$ is therefore *not* a function with small bounded differences, and Theorem 2.55 gives nothing useful about it.

There is, however, a closely related quantity that does concentrate, and that is moreover what one computes in practice: the *leave-one-out*, or *deleted*, estimate of the risk. Write $\widehat{h}_{n,-j}^{k\text{-NN}}$

for the k-nearest neighbour rule built from the sample with the j-th observation removed (ties broken by index, as in Definition 1.19), and set

$$\widehat{L}_n^{\mathrm{loo}} := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\widehat{h}_{n,-j}^{k\text{-NN}}(X_j) \neq Y_j\}. \tag{2.44}$$

Since (X_j, Y_j) is independent of the $n-1$ observations used to build $\widehat{h}_{n,-j}^{k\text{-NN}}$, each summand has expectation $\mathbb{P}\big(\widehat{h}_{n-1}^{k\text{-NN}}(X) \neq Y\big)$, so that

$$\mathbb{E}\big[\widehat{L}_n^{\mathrm{loo}}\big] = R_{n-1}\big(\widehat{h}^{k\text{-NN}}\big) : \tag{2.45}$$

the leave-one-out estimate is an *unbiased* estimate of the risk of the rule trained on $n-1$ points – no test set required. It remains to show that it concentrates, and this is where geometry enters.

Lemma 2.78 (Cone covering)**.** *For u in the unit sphere S^{d-1} of $\mathbb{R}^d$ let $C(u) := \{z \in \mathbb{R}^d \setminus \{0\} : \langle z, u\rangle \geq \|z\| \cos(\pi/8)\}$ be the cone of half-angle $\pi/8$ around u. Then $\mathbb{R}^d \setminus \{0\}$ can be covered by $\kappa_d \leq 7^d$ such cones.*

Proof. Let $r = 2\sin(\pi/16)$ and let $\{u_1, \dots, u_N\}$ be an r-net of S^{d-1} for the Euclidean distance, of cardinality $N \leq (1 + 2/r)^d$ – the standard volumetric bound, see [Ver18, Cor. 4.2.13]. For $z \neq 0$ pick u_j with $\big\|z/\|z\| - u_j\big\| \leq r$; since the chord subtending an angle θ on the unit sphere has length $2\sin(\theta/2)$, the angle between z and u_j is at most $\pi/8$, i.e. $z \in C(u_j)$. Finally $r = 2\sin(\pi/16) \geq 0.390$, so $N \leq (1 + 2/0.390)^d \leq 6.13^d \leq 7^d$. □

Lemma 2.79 (A point is the neighbour of few points)**.** *Let $x, z_1, \dots, z_N$ be pairwise distinct points of $\mathbb{R}^d$ and suppose that, for each j, x belongs to the set of the k nearest neighbours of z_j among $\{x, z_1, \dots, z_N\} \setminus \{z_j\}$. Then $N \leq k\kappa_d$.*

Proof. By Lemma 2.78 it suffices to show that at most k of the vectors $z_j - x$ lie in a given cone $C(u)$. Suppose on the contrary that $z_1 - x, \dots, z_{k+1} - x \in C(u)$, labelled so that $\|z_1 - x\| \leq \cdots \leq \|z_{k+1} - x\|$. Any two vectors of $C(u)$ make an angle at most $2 \cdot \frac{\pi}{8} = \frac{\pi}{4}$, so for $j \leq k$, writing θ for the angle at x between $z_j - x$ and $z_{k+1} - x$, we have $\cos\theta \geq \cos\frac{\pi}{4} = \frac{\sqrt{2}}{2}$ and

$$\begin{aligned}\big\|z_{k+1} - z_j\big\|^2 &= \|z_{k+1} - x\|^2 + \big\|z_j - x\big\|^2 - 2\,\|z_{k+1} - x\|\,\big\|z_j - x\big\|\cos\theta \\ &\leq \|z_{k+1} - x\|^2 + \big\|z_j - x\big\|\Big(\big\|z_j - x\big\| - \sqrt{2}\,\|z_{k+1} - x\|\Big) \;<\; \|z_{k+1} - x\|^2,\end{aligned}$$

because $0 < \big\|z_j - x\big\| \leq \|z_{k+1} - x\| < \sqrt{2}\,\|z_{k+1} - x\|$. Thus $z_1, \dots, z_k$ are all *strictly* closer to z_{k+1} than x is, so x is not among the k nearest neighbours of z_{k+1}, whatever the tie-breaking rule – a contradiction. □

Theorem 2.80 (Concentration of the leave-one-out estimate)**.** *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. with law $\mathcal{D}$ on $\mathbb{R}^d \times \mathcal{Y}$, and assume that the law $\mathcal{D}_X$ of X is non-atomic. Then for every $t > 0$,*

$$\mathbb{P}\Big(\big|\widehat{L}_n^{\mathrm{loo}} - R_{n-1}(\widehat{h}^{k\text{-NN}})\big| \;\geq\; t\Big) \;\leq\; 2\exp\Big(-\frac{2nt^2}{\big(1 + 2k\kappa_d\big)^2}\Big).$$

Proof. By (2.45) it suffices to prove the concentration of $\widehat{L}_n^{\mathrm{loo}}$ around its mean. Fix j and replace (x_j, y_j) by (x_j', y_j'), all the other observations being unchanged; we bound the resulting variation of (2.44).

- The j-th summand may change, contributing at most $1/n$.
- For $i \neq j$, the summand $\mathbb{1}\{\widehat{h}^{k\text{-NN}}_{n,-i}(x_i) \neq y_i\}$ depends only on y_i and on the labels of the k nearest neighbours of x_i among $\{x_1, \dots, x_n\} \setminus \{x_i\}$. If neither x_j nor x'_j belongs to that set – computed in the old, respectively the new, sample – then the k nearest neighbours of x_i and their labels are unchanged, and the summand is unchanged. By Lemma 2.79, applied once to x_j and once to x'_j (both legitimate as soon as the points involved are pairwise distinct), the number of indices $i \neq j$ for which this can fail is at most $2k\kappa_d$.

Hence the bounded-difference property (2.32) holds with $c_j = (1 + 2k\kappa_d)/n$, for every configuration of pairwise distinct points. The set of $(x_1, \dots, x_n, x'_j)$ with pairwise distinct coordinates has full measure under $\mathcal{D}_X^{\otimes(n+1)}$, because $\mathcal{D}_X$ is non-atomic; so Remark 2.58 applies and yields the announced bound. □

Remark 2.81 (What this does and does not give). Combined with Theorem 1.21, Theorem 2.80 closes the loop opened in Chapter 1: with probability at least $1 - \delta$,

$$R_{n-1}\big(\widehat{h}^{\text{NN}}\big) \;\leq\; \widehat{L}_n^{\text{loo}} + \big(1 + 2\kappa_d\big)\sqrt{\frac{\log(2/\delta)}{2n}} \qquad (k = 1),$$

an entirely computable certificate on the expected risk of the rule, with no test set and no assumption on $\mathcal{D}$ beyond non-atomicity. The catch is the constant: $\kappa_d \leq 7^d$ is exponential in the dimension, and the bound is informative only for small d – the curse of dimensionality of Remark 1.26 in yet another disguise, and here it is a genuine feature of the method rather than an artefact, since a single point of $\mathbb{R}^d$ can be the nearest neighbour of exponentially many others. The classical analyses of the deleted estimate by Rogers and Wagner [RW78] and Devroye and Wagner [DW79] carry the same constant; obtaining dimension-free guarantees for local rules requires additional structure.

2.7.7 Sums of random vectors

McDiarmid's inequality immediately gives a concentration statement for the norm of a sum of independent random vectors – a quantity that is not a sum of real random variables, and to which none of the preceding results applies directly.

Proposition 2.82. *Let $U_1, \dots, U_n$ be independent random vectors of a normed space, with $\|U_i\| \leq b$ almost surely, and let $Z = \big\|\sum_{i=1}^n U_i\big\|$.[replacement: "Let $U_1, \dots, U_n$ be independent random vectors of a* separable *normed space . . . " – without separability $\|\sum_i U_i\|$ need not be measurable. **(not yet discussed)**] Then for every $t > 0$,*

$$\mathbb{P}\big(|Z - \mathbb{E}Z| \geq t\big) \;\leq\; 2\exp\Big(-\frac{t^2}{2nb^2}\Big).$$

If moreover the space is a Hilbert space and the U_i are centred, then $\mathbb{E}[Z] \leq \big(\sum_i \mathbb{E}\,\|U_i\|^2\big)^{1/2}$.

Proof. Replacing w_i by w'_i changes $\big\|\sum_j w_j\big\|$ by at most $\big\|w_i - w'_i\big\| \leq 2b$ by the triangle inequality, so Z has the bounded-difference property with $c_i = 2b$ and (2.33) applies with $\sum_i c_i^2 = 4nb^2$. For the second statement, by independence and centring, $\mathbb{E}\big[\|\sum_i U_i\|^2\big] = \sum_{i,j} \mathbb{E}\,\langle U_i, U_j\rangle = \sum_i \mathbb{E}\,\|U_i\|^2$, and $\mathbb{E}Z \leq (\mathbb{E}Z^2)^{1/2}$ by Jensen. □

Remark 2.83. The two halves of Proposition 2.82 illustrate once more the division of labour of Remark 2.59: concentration is cheap and generic – the fluctuation of Z is $O(b\sqrt{n})$ whatever

the geometry – whereas the expectation depends on the norm, and in a general Banach space its control is a genuinely hard problem, at the origin of the theory of type and cotype. Note also that the bound is dimension-free: nothing in it refers to the dimension of the ambient space.

Remark 2.84 (Empirical Bernstein bounds). Inequality (2.30) involves the unknown variance σ^2, which limits its practical use. It can be replaced by the empirical variance $\widehat{\sigma}_n^2 = \frac{1}{n(n-1)} \sum_{i<j} (X_i - X_j)^2$, at the price of an extra additive term: for $X_i \in [0,1]$ i.i.d., with probability at least $1-\delta$,

$$\bar{X}_n - \mu \;\le\; \widehat{\sigma}_n \sqrt{\frac{2\log(2/\delta)}{n}} + \frac{7\log(2/\delta)}{3(n-1)}.$$

We admit this result, due to Maurer and Pontil [MP09, Thm. 4]; see also [AMS09]. [*sentence to add:* "Maurer and Pontil state the bound for $\mu - \bar{X}_n$; the form printed here follows by applying it to the variables $-X_i$, whose empirical variance is the same." (For a one-sided bound the two directions are not interchangeable in general, so the remark is worth making.) **(not yet discussed)**] Its proof is not a corollary of Theorem 2.47: it requires a concentration inequality for $\widehat{\sigma}_n^2$ itself, obtained from the fact that $\widehat{\sigma}_n$ is a *self-bounding* function of the sample, a notion we shall meet with the entropy method.

[V60 – Pointer to the self-bounding inequality of Chapter 5] *Replacement of the last sentence:* "[. . .] obtained from the fact that $\widehat{\sigma}_n$ is a *self-bounding* function of the sample (Definition 5.10), to which the exponential inequality of Theorem 5.39 applies." *Proposed on 15 October 2026 (chapter 5 written): not yet discussed.*

2.8 Notes and references

The Cramér–Chernoff method and the classes $\mathrm{subG}(v)$ and $\Gamma_+(v,c)$ are treated in [BLM13, Chap. 2], whose organisation and notation we have largely followed; [Ver18, Chap. 2] develops the same material in the Orlicz-norm style of Remark 2.32, and [Wai19, Chap. 2] in an intermediate one. The exponential tilting of Definition 2.2 is the Esscher transform of actuarial mathematics [Ess32] and the Gibbs measure of statistical physics; the variance representations (2.2) and (2.4), which we have used as the organising principle of the chapter, are elementary but seem not to be as widely known as they deserve. Proposition 2.17 is the finite-dimensional shadow of the Gibbs variational principle (Proposition 3.35) and of Sanov's theorem; see [DZ98, Chap. 6].

Cramér's theorem [Cra38] is the origin of large-deviation theory; [DZ98] is the standard modern reference, and [Fel68, Chap. 27] contains a very readable account of the one-dimensional case. The sharp asymptotics of Remark 2.21 are due to [BR60]; see also [Pet75, Chap. 8].

Bennett's inequality is [Ben62]; Bernstein's goes back to the 1920s and is stated in [Ber46]. The characterisations of Theorem 2.30 are [Ver18, Prop. 2.5.2], from which we have taken the structure of the proof while making the constants explicit. The chi-square bound of Remark 2.50 is the Laurent–Massart inequality [LM00]. Empirical versions of Bernstein's inequality, in which the unknown variance is replaced by its empirical counterpart, are due to Audibert, Munos and Szepesvári [AMS09] and to Maurer and Pontil [MP09].

The bounded-difference inequality in the form of Theorem 2.55 is [McD89], building on [Azu67] and [Hoe63, Sec. 2]; the martingale proof given here is the standard one, with the measurability of the essential suprema in Step 2 spelled out. The concentration of

the chromatic number is [SS87]; see [AS16, Chap. 7] and [JŁR00, Chap. 2] for many more applications of the same kind. The variance-sensitive successors of Theorem 2.55 are the Efron–Stein inequality [ES81] and Talagrand's convex distance inequality [Tal95]; both are treated in [BLM13].

The PAC framework is due to Valiant [Val84]; Theorem 2.65 and Theorem 2.66 are in [SB14, Chap. 2–4]. Fast rates under low-noise conditions originate in [MT99; Tsy04] and [MN06]; the elementary argument of Theorem 2.70 – variance controlled by the mean, Bernstein, self-bounding inequality – is the skeleton of all of them, and is presented in [Mas00] and [BBL05]. The concentration of the deleted estimate (Theorem 2.80) goes back to Rogers and Wagner [RW78] and Devroye and Wagner [DW79]; the cone argument of Lemmas 2.78 and 2.79 is classical in the nearest-neighbour literature, see [DGL96, Chap. 11] and [BD15].

2.9 Exercises

Exercise 2.1 (Tilting the classical laws)**.** For each of the laws of Example 2.8, compute the tilted law P_λ, the functions $m(\lambda)$ and $\varsigma^2(\lambda)$, and check Proposition 2.6 and Proposition 2.11 by direct computation. For which of them is $\operatorname{dom}(K) \neq \mathbb{R}$, and how is this visible on the tails?

Exercise 2.2 (Domain of the cumulant generating function)**.** Let Z be a real random variable. Show that $\lambda_+ := \sup \operatorname{dom}(K_Z) > 0$ if and only if there exist $C, C' > 0$ with $\mathbb{P}(Z \geq t) \leq Ce^{-C't}$ for all $t > 0$, and that in that case $\lambda_+ = \sup\{C' > 0 : \limsup_t e^{C't}\mathbb{P}(Z \geq t) < \infty\}$. Deduce that no variable with a polynomially decaying tail can satisfy a Chernoff bound.

Exercise 2.3 (★ The sub-Gaussian norm is a norm)**.** Show that $\|\cdot\|_{\mathrm{sG}}$ of Definition 2.31 is a norm on the vector space of sub-Gaussian random variables (modulo almost-sure equality), and that the infimum defining it is attained. *Hint: for the triangle inequality, use the convexity of $x \mapsto e^{x^2}$ and the fact that $(a+b)^2/(s_1+s_2)^2 \leq \frac{s_1}{s_1+s_2}\frac{a^2}{s_1^2} + \frac{s_2}{s_1+s_2}\frac{b^2}{s_2^2}$.*

Exercise 2.4 (★ Sub-exponential implies Bernstein's condition)**.** Let Z be centred with $\mathbb{E}\big[e^{|Z|/c}\big] \leq 2$ for some $c > 0$. Show that $\mathbb{E}|Z|^q \leq 2\,q!\,c^q$ for every integer $q \geq 2$, hence that (2.31) holds with $v = 4c^2$ and scale c, and conclude that $Z \in \Gamma_+(4c^2, c)$.

Exercise 2.5 (★ Chi-square concentration)**.** Let $X \sim \mathcal{N}(0,1)$.

(a) Show that $\psi_{X^2-1}(\lambda) = -\frac{1}{2}\log(1-2\lambda) - \lambda$ for $\lambda < \frac{1}{2}$, and that $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}$ for $u \in [0,1)$. Deduce that $X^2 - 1 \in \Gamma_+(2,2)$.

(b) Let $G \sim \mathcal{N}(0, I_d)$. Show that $\|G\|_2^2 - d \in \Gamma_+(2d, 2)$ and deduce the Laurent–Massart bounds

$$\mathbb{P}\big(\|G\|_2^2 \geq d + 2\sqrt{dy} + 2y\big) \leq e^{-y}, \qquad \mathbb{P}\big(\|G\|_2^2 \leq d - 2\sqrt{dy}\big) \leq e^{-y}.$$

(c) Deduce that $\|G\|_2 = \sqrt{d}\,(1 + o(1))$ with overwhelming probability, and comment on the geometric meaning: in high dimension, a Gaussian vector lives on a thin spherical shell.

Exercise 2.6 (Maximum of sub-gamma variables)**.** Let $Z_1, \dots, Z_N$ be centred, not necessarily independent, with $Z_i \in \Gamma_+(v, c)$ for every i. Show that

$$\mathbb{E}\Big[\max_{1 \leq i \leq N} Z_i\Big] \;\leq\; \sqrt{2v \log N} + c \log N.$$

Hint: adapt the proof of Proposition 2.28, optimising λ in $(0, 1/c)$. Compare with Proposition 2.28 and comment on the regime in which the second term dominates.

Exercise 2.7 (The exponential distribution against the sub-gamma bound)**.** Let $Z = Y - 1$ with $Y \sim \mathrm{Exp}(1)$.

(a) Compute ψ_Z and show that $Z \in \Gamma_+(1, 1)$. Is the inequality (2.19) an equality?

(b) Compute $\psi_Z^*(t)$ and compare $e^{-\psi_Z^*(t)}$ with the exact tail $\mathbb{P}(Z \geq t) = e^{-(1+t)}$. What is the ratio, and how does it illustrate Remark 2.21?

(c) Compare both with the sub-gamma bound (2.21). Show that for large t the latter loses a factor 2 in the exponent, and explain why: which of the two regimes of Remark 2.38 is at work?

Exercise 2.8 (A self-bounding lemma)**.** Let $a, b \geq 0$ and let $x \geq 0$ satisfy $x \leq a\sqrt{x} + b$. Show that $x \leq a^2 + 2b$, and that the constant 2 cannot be replaced by 1. Where was this used in Section 2.7.2, and why was a more careful argument needed there?

Exercise 2.9 (Two-sided Bernstein)**.** Let $Z_1, \dots, Z_n$ be independent with $|Z_i| \leq b$ almost surely, and let $v = \sum_i \mathbb{V}\mathrm{ar}(Z_i)$. Show that for every $t > 0$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n (Z_i - \mathbb{E}Z_i)\Big| \geq t\Big) \;\leq\; 2\exp\Big(-\frac{t^2}{2\big(v + \frac{2bt}{3}\big)}\Big),$$

and explain where the factor 2 in front of $bt/3$ comes from.

Exercise 2.10 (Balls in bins)**.** Throw n balls independently and uniformly into n bins, and let M_n be the maximum number of balls in a bin.

(a) Show that $\mathbb{P}\big(\text{bin 1 receives} \geq k \text{ balls}\big) \leq \binom{n}{k} n^{-k} \leq \frac{1}{k!} \leq \big(\frac{e}{k}\big)^k$.

(b) Deduce that $\mathbb{P}\big(M_n \geq 3\log n / \log\log n\big) \to 0$.

(c) Why does Hoeffding's inequality give a much weaker result here, and which of the inequalities of this chapter captures the right behaviour?

Exercise 2.11 (McDiarmid is sharp)**.** (a) Show that the constant 2 in the exponent of (2.33) cannot be improved, by considering $f(x_1, \dots, x_n) = \sum_i x_i$ with $x_i \in \{-1, 1\}$ and comparing with the exact behaviour of a Rademacher sum.

(b) Show that the bound can nevertheless be very far from the truth: for $f(x_1, \dots, x_n) = \max_i x_i$ with X_i i.i.d. uniform on $[0, 1]$, compare the fluctuation of Z with what (2.33) predicts, and relate your finding to Remark 2.59.

Exercise 2.12 (Concentration of a histogram estimator)**.** Let $X_1, \dots, X_n$ be i.i.d. with density f on $\mathbb{R}^d$, and let $\hat{f}_n$ be the histogram estimator associated with a fixed partition of $\mathbb{R}^d$. Show that $Z = \int |\hat{f}_n - f|$ satisfies the bounded-difference property with $c_i = 2/n$, and deduce that $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2e^{-nt^2/2}$. Note that no assumption whatsoever is made on f or on the partition.

Exercise 2.13 (★ The lower bound in Cramér's theorem, by hand)**.** Let $X_1, \dots, X_n$ be i.i.d. $\mathrm{Ber}(p)$ and let $q \in (p, 1)$ with nq an integer. Using Stirling's formula in the form $\sqrt{2\pi j}\,(j/e)^j \leq j! \leq e\sqrt{j}\,(j/e)^j$, show that

$$\mathbb{P}\big(\mathrm{Bin}(n, p) = nq\big) \;\geq\; \frac{C}{\sqrt{n}}\, e^{-n\ \mathrm{kl}(q,p)}$$

for an explicit constant $C = C(q) > 0$. Compare with Remark 2.51 and with Remark 2.21.

Exercise 2.14 (Concentration of the empirical variance)**.** Let $X_1, \dots, X_n$ be i.i.d. with values in $[0,1]$ and let $\widetilde{\sigma}_n^2 = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2$.

(a) Show that $\widetilde{\sigma}_n^2 = \frac{1}{2n^2}\sum_{i,j}(X_i - X_j)^2$ and deduce that, as a function of $(X_1, \dots, X_n)$, it has bounded differences with $c_i \le 1/n$.

(b) Deduce that $\mathbb{P}\big(\big|\widetilde{\sigma}_n^2 - \mathbb{E}\widetilde{\sigma}_n^2\big| \ge t\big) \le 2e^{-2nt^2}$.

(c) Compute $\mathbb{E}\widetilde{\sigma}_n^2$ and comment on the bias.

Exercise 2.15 (Bounded differences with a bad constant)**.** Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be independent uniform letters from a finite alphabet, and let Z be the length of the longest common subsequence of the two words.

(a) Show that Z has bounded differences with $c_i = 1$ in each of the $2n$ variables, and deduce that $\mathbb{P}(|Z - \mathbb{E}Z| \ge t) \le 2e^{-t^2/n}$.

(b) It is known that $\mathbb{E}Z \sim \kappa n$ for a constant κ depending on the alphabet, whose value is unknown even for a binary alphabet. Comment on what Theorem 2.55 does and does not tell us here.

Exercise 2.16 (A sharper union bound)**.** Let $Z_1, \dots, Z_N$ be centred with $Z_i \in \mathrm{subG}(v_i)$, the v_i being possibly different. Show that $\mathbb{E}[\max_i Z_i] \le \sqrt{2 \max_i v_i \, \log N}$, and that this cannot be improved to $\sqrt{2\,\overline{v}\log N}$ with $\overline{v}$ the average of the v_i. What does this say about union bounds over heterogeneous families?

Exercise 2.17 (Numerical comparison)**.** For $p \in \{0.5, 0.1, 0.01\}$ and $n \in \{10^2, 10^4\}$, tabulate the four bounds of Section 2.5.4 and the exact binomial tail $\mathbb{P}(\mathrm{Bin}(n,p) \ge n(p+t))$ for a range of t. For which values of (p,t) is Hoeffding's inequality within a factor 2 of the truth in the exponent? Draw the analogous picture for the *left* tail and comment.

Chapter 3

Lower Bounds: Relative Entropy, Large Deviations and No Free Lunch

Summary. The first two chapters produced upper bounds: on the probability of a deviation, on the risk of a learning rule. An upper bound is only half of a statement; the other half says that it cannot be improved. This chapter develops the tool that produces the second half – relative entropy and the change of measure – and applies it at three levels: the deviations of a sum (the Chernoff exponent is exact, and we prove it, closing the part of Cramér's theorem left open in Chapter 2 for every variable with exponential moments of all positive orders, and more generally for steep cumulant generating functions), statistical estimation (Le Cam, Fano and Assouad lower bounds on the minimax risk), and learning (no algorithm learns without prior knowledge: the no-free-lunch theorem, and the curse of dimensionality for Lipschitz problems).

3.1 Why lower bounds

Every inequality proved so far is an upper bound. Hoeffding's inequality says that $\mathbb{P}(\bar{X}_n - \mu \geq t)$ is *at most* e^{-2nt^2}; the nearest-neighbour theorem says that the expected risk of $\widehat{h}^{\mathrm{NN}}$ is *at most* $2L^*_{\mathcal{D}} + 3c\sqrt{d}\, n^{-1/(d+1)}$; the bounds of Section 2.7.1 say that the excess risk of empirical risk minimisation over a finite class is *at most* $\sqrt{2\log(2|\mathcal{H}|/\delta)/n}$. None of these statements, taken alone, tells us whether the bound is good. A bound can be loose because the proof was clumsy, because the inequality used was the wrong one, or because the problem itself is hard; only a *lower* bound can tell these three situations apart. One asks for optimality at three different levels, and this chapter treats them in turn.

(i) Is the exponent right? For a fixed law and a fixed deviation, is the exponential rate given by the Cramér–Chernoff inequality the true rate of decay of $\mathbb{P}(\bar{X}_n \geq x)$? Theorem 2.19 answered yes, but only the upper half of it was proved. This is a question about *one* probability distribution, and a lower bound here is a statement of the form $\mathbb{P}(\bar{X}_n \geq x) \geq e^{-n(K^*(x)+o(1))}$, with K^* the Cramér transform of the law – for Bernoulli variables, $e^{-n(\mathrm{kl}(x,\mu)+o(1))}$.

(ii) Can any estimator do better? When one estimates an unknown parameter – a mean, a probability, a regression function – from n observations, the risk of a particular estimator is an upper bound on the *minimax risk*, the best worst-case risk achievable. A lower bound is now a statement about *all* estimators simultaneously: whatever one does with the data, there is a parameter for which the error is at least so much. Such a statement cannot be proved by analysing one procedure; it must exploit the fact that two different parameters may produce data that are hard to tell apart.

(iii) Can any learning algorithm do better? In learning, the same question takes an extreme

form. The theorems of Chapters 1 and 2 bound the risk of specific rules under specific assumptions – a Lipschitz regression function, a finite hypothesis class. Is there a single algorithm that learns every problem? The no-free-lunch theorem answers no: for every algorithm and every sample size there is a problem, perfectly learnable in principle, on which the algorithm fails with constant probability. Assumptions are not a technical convenience; without them there is nothing to prove.

The three questions have the same answer, and the answer has the same structure. To show that an event is not too unlikely under a law Q, one *changes the measure* to a law P under which the event is typical, and pays for the change with the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$, whose typical value under P is $e^{-\mathrm{KL}(P\|Q)}$: its logarithm has mean $-\mathrm{KL}(P\|Q)$, minus the relative entropy. To show that no procedure can distinguish two parameters, one shows that the laws of the data under the two parameters have a small relative entropy, hence are close in total variation, hence cannot be told apart by any test. In both cases the decisive quantity is the relative entropy, and the decisive properties are the ones that make it computable for samples of size n – *tensorisation*, $\mathrm{KL}(P^{\otimes n}\|Q^{\otimes n}) = n\,\mathrm{KL}(P\|Q)$ – and the one that makes it applicable to arbitrary procedures – *data processing*: no function of the data, however clever, can increase it.

Plan of the chapter. Section 3.2 revisits Bernoulli deviations: it collects the upper bounds of Chapters 1 and 2 in a single picture, organised around an integral representation of kl, compares them on worked examples, and turns the Chernoff bound into a confidence interval. Section 3.3 develops relative entropy for general laws: definition, examples, convexity, tensorisation, data processing, the inequalities of Pinsker and of Bretagnolle and Huber, and its interpretation as the optimal error exponent in hypothesis testing (the Neyman–Pearson lemma and Stein's lemma). Section 3.4 proves the large-deviation lower bound by change of measure, first asymptotically, then non-asymptotically with entropic arguments. Section 3.5 builds the minimax toolkit – the reduction from estimation to testing, Le Cam's two-point lemma, Fano's inequality, Assouad's lemma – and applies it to Bernoulli and Gaussian means. Section 3.6 proves the no-free-lunch theorem and its quantitative companion for Lipschitz problems, the curse of dimensionality, which we compare honestly with the nearest-neighbour upper bounds of Chapter 1. The tools of Sections 3.5 and 3.6 are reused in Chapter 4 to prove that the Vapnik–Chervonenkis dimension governs the sample complexity of learning, from below as well as from above.

3.2 Bernoulli deviations revisited

Bernoulli variables are the simplest non-trivial case and the one that matters most in learning, since the empirical risk of a fixed classifier is a mean of Bernoulli variables. Before proving that the Chernoff bound for them cannot be improved at the exponential scale, we collect what is known from above, in a form that makes every classical inequality a corollary of one identity.

3.2.1 The Chernoff–Hoeffding bound

Throughout this section $X_1, \dots, X_n$ are i.i.d. $\mathrm{Ber}(\mu)$ with $\mu \in (0,1)$, unless stated otherwise, and kl is the Bernoulli relative entropy of Notation 2.22: $\mathrm{kl}(x,\mu) = x\log\frac{x}{\mu} + (1-x)\log\frac{1-x}{1-\mu}$. In this chapter the parameter of a Bernoulli law is written μ, since it is the mean $\mathbb{E}X_1$, and the level x, as the argument of the Cramér transform $K^*(x)$ of Chapter 2: $\mathrm{kl}(x,\mu)$ is the function $\mathrm{kl}(q,p)$ of Notation 2.22 with $q = x$ and $p = \mu$.

Theorem 3.1 (Chernoff–Hoeffding deviation bound). *(i) For every $x \in (\mu, 1]$,*

$$\mathbb{P}\big(\bar{X}_n \geq x\big) \;\leq\; e^{-n\,\mathrm{kl}(x,\mu)}, \tag{3.1}$$

and for every $x \in [0, \mu)$, $\mathbb{P}(\bar{X}_n \leq x) \leq e^{-n\,\mathrm{kl}(x,\mu)}$.

(ii) For all $x, \mu \in (0,1)$, with $\tilde{x} := (2\mu + x)/3$,

$$\mathrm{kl}(x,\mu) \;\geq\; \frac{(x-\mu)^2}{2\,\tilde{x}(1-\tilde{x})} \;\geq\; \frac{(x-\mu)^2}{2\max_{u\in[\mu\wedge x,\,\mu\vee x]} u(1-u)} \;\geq\; 2(x-\mu)^2. \tag{3.2}$$

(iii) (Hoeffding's bound for Bernoulli variables) For every $x \in (\mu, 1]$, $\mathbb{P}(\bar{X}_n \geq x) \leq e^{-2n(x-\mu)^2}$.

(iv) Inequality (3.1), hence also (iii), remains true if $X_1, \dots, X_n$ are only assumed independent, with values in $[0,1]$ and expectations $\mu_1, \dots, \mu_n$, provided μ is replaced by $\bar{\mu} := (\mu_1 + \cdots + \mu_n)/n \in (0,1)$.

Proof. (i) is (2.12), which was derived from Proposition 2.23(b) and (2.7); the case $x = 1$ is included, since $\mathrm{kl}(1,\mu) = \log(1/\mu)$ is the value at 1 of the Cramér transform of $\mathrm{Ber}(\mu)$, and indeed $\mathbb{P}(\bar{X}_n \geq 1) = \mu^n$. The left deviations follow by applying the right ones to the variables $1 - X_i \sim \mathrm{Ber}(1-\mu)$, since $\mathrm{kl}(1-x, 1-\mu) = \mathrm{kl}(x,\mu)$.

(ii) is proved in Corollary 3.5 below, and (iii), which is the case $a = 0$, $b = 1$ of (1.28), also follows from (i) and the inequality $\mathrm{kl}(x,\mu) \geq 2(x-\mu)^2$, given by (ii) – or directly by Pinsker's inequality (2.5) – for $x \in (\mu, 1)$, and at $x = 1$ by continuity of $\mathrm{kl}(\cdot,\mu)$ on $[0,1]$.

(iv) For $u \in [0,1]$ and $\lambda \in \mathbb{R}$, the convexity of $u \mapsto e^{\lambda u}$ gives $e^{\lambda u} \leq (1-u)e^0 + ue^\lambda$. Taking expectations, $\mathbb{E}e^{\lambda X_i} \leq 1 - \mu_i + \mu_i e^\lambda$, and by independence and the concavity of the logarithm (the arithmetic–geometric mean inequality),

$$\log \mathbb{E}e^{\lambda n \bar{X}_n} \;\leq\; \sum_{i=1}^{n} \log\big(1 - \mu_i + \mu_i e^\lambda\big) \;\leq\; n \log\big(1 - \bar{\mu} + \bar{\mu} e^\lambda\big) = n\, K_{\mathrm{Ber}(\bar{\mu})}(\lambda).$$

The Chernoff bound (Theorem 1.57) thus gives, for $x > \bar{\mu}$, $\mathbb{P}(\bar{X}_n \geq x) \leq \exp\big(-n \sup_{\lambda>0}\{\lambda x - K_{\mathrm{Ber}(\bar{\mu})}(\lambda)\}\big)$, and the supremum is $\mathrm{kl}(x,\bar{\mu})$ by Proposition 2.10(iii) and Proposition 2.23(b). □

Remark 3.2 (Bernoulli variables are the worst case). Part (iv) says that among all independent $[0,1]$-valued variables with a given average mean, the Chernoff bound is the largest for Bernoulli variables: the inequality $e^{\lambda u} \leq 1 - u + ue^\lambda$ is the statement that, for a convex function on $[0,1]$, the extremal law with given mean is carried by the endpoints. (This is the first step of the second proof of Lemma 1.62.) This observation is Hoeffding's [Hoe63, Thm. 1], and it has a converse: since Bernoulli variables belong to the class, *no* inequality valid for all independent $[0,1]$-valued variables can beat the exponent $\mathrm{kl}(x,\bar{\mu})$, once we have shown in Section 3.4 that this exponent is the exact one for Bernoulli variables. In that sense, (3.1) is the optimal exponential inequality for bounded variables with known means and no further information.

3.2.2 One identity, and Pollard's lemma

The function $x \mapsto \mathrm{kl}(x,\mu)$ is smooth on $(0,1)$, with

$$\mathrm{kl}(\mu,\mu) = \partial_x\,\mathrm{kl}(x,\mu)\big|_{x=\mu} = 0, \qquad \partial_x\,\mathrm{kl}(x,\mu) = \log\frac{x(1-\mu)}{\mu(1-x)}, \qquad \partial_x^2\,\mathrm{kl}(x,\mu) = \frac{1}{x(1-x)}.$$

Taylor's formula with integral remainder, after the substitution $u = \mu + s(x-\mu)$, therefore gives the identity

$$\mathrm{kl}(x,\mu) = \frac{(x-\mu)^2}{2} \int_0^1 \frac{2(1-s)\,\mathrm{d}s}{\big(\mu + s(x-\mu)\big)\big(1-\mu-s(x-\mu)\big)}, \tag{3.3}$$

which is (2.26) in another parametrisation, and a special case of the representation (2.4) of the Cramér transform as an average of inverse tilted variances: the law $\mathrm{Ber}(\mu + s(x-\mu))$ is the tilt of $\mathrm{Ber}(\mu)$ with mean $\mu + s(x-\mu)$. The weight $2(1-s)\,\mathrm{d}s$ is a probability measure on $[0,1]$, which suggests applying Jensen's inequality. The principle is general, and we isolate it.

Lemma 3.3 (Pollard's lemma)**.** *Let g be twice continuously differentiable on an open interval $I \ni 0$, and define, for $x \in I$,*

$$G(x) := \frac{g(x) - g(0) - xg'(0)}{x^2/2} \quad (x \neq 0), \qquad G(0) := g''(0).$$

Then $G(x) = \int_0^1 g''(sx)\,2(1-s)\,\mathrm{d}s$ for every $x \in I$, an average with respect to the probability measure $2(1-s)\,\mathrm{d}s$ on $[0,1]$, whose mean is $\int_0^1 s \cdot 2(1-s)\,\mathrm{d}s = \frac{1}{3}$; consequently:

(a) if $g'' \geq 0$ then $G \geq 0$;

(b) if g'' is non-decreasing on I, then G is non-decreasing on I;

(c) if g'' is convex on I, then G is convex on I and $G(x) \geq g''(x/3)$ for every $x \in I$.

Proof. For $x \neq 0$, Taylor's formula with integral remainder gives $g(x) - g(0) - xg'(0) = \int_0^x g''(u)(x-u)\,\mathrm{d}u = x^2 \int_0^1 g''(sx)(1-s)\,\mathrm{d}s$ (substitute $u = sx$), which is the representation; at $x = 0$ it is trivial. For each $s \in [0,1]$, the function $x \mapsto g''(sx)$ is non-negative, non-decreasing or convex on I when g'' is (for (b), note that $sx \leq sy$ when $x \leq y$ and $s \geq 0$), and these properties are preserved by averaging with respect to the probability measure $2(1-s)\,\mathrm{d}s$ on $[0,1]$. Finally, under (c), Jensen's inequality for the convex function $s \mapsto g''(sx)$ and this probability measure gives $G(x) \geq g''\big(x \int_0^1 s \cdot 2(1-s)\,\mathrm{d}s\big) = g''(x/3)$; and $\int_0^1 s \cdot 2(1-s)\,\mathrm{d}s = 1 - \frac{2}{3} = \frac{1}{3}$. □

Remark 3.4 (Three instances). (a) $g = \exp$: $g'' = \exp$ is non-decreasing and convex, so $u \mapsto (e^u - 1 - u)/u^2$ is non-decreasing and convex on $\mathbb{R}$. The first property is exactly what the proof of Bennett's lemma (Lemma 2.39) extracts from (2.23), which is the representation of Lemma 3.3 for this g.

(b) $g = h$, Bennett's function (2.11), on $I = (-1,\infty)$: $h''(u) = 1/(1+u)$ is convex, so (c) gives, for every $u > -1$,

$$h(u) \;\geq\; \frac{u^2}{2} \cdot \frac{1}{1 + u/3},$$

which is the inequality (2.17) that turns Bennett's inequality into Bernstein's. This is not a new proof: the representation of Lemma 3.3 for this g is (2.16), and (c) is the Jensen step by which (2.17) was derived from it. The lemma makes plain, however, that the bound holds on the whole domain; for $u < 0$ it improves on the bound $h(u) \geq u^2/2$ stated in (2.17).

(c) $g(s) = \mathrm{kl}(\mu+s,\mu)$ on $I = (-\mu, 1-\mu)$: $g''(s) = 1/\big((\mu+s)(1-\mu-s)\big)$ is convex, being the sum $\frac{1}{\mu+s} + \frac{1}{1-\mu-s}$ of two convex functions; this is the source of the first inequality of (3.2).

The lemma is taken from Pollard's notes on empirical processes [Pol01, Ex. 14]; it is the natural tool whenever an inequality of the form "$g(x) \geq$ quadratic" must be proved with a good constant. In this book it serves only for Corollary 3.5 below, instances (a) and (b) recasting computations of Chapter 2.

Corollary 3.5 (Lower bounds on kl)**.** *Inequality* (3.2) *holds for all* $x, \mu \in (0,1)$.

Proof. Apply Lemma 3.3(c) to $g(s) = \mathrm{kl}(\mu+s,\mu)$ at the point $s = x-\mu$, using $g(0) = g'(0) = 0$ and Remark 3.4(c): $\mathrm{kl}(x,\mu) = \frac{(x-\mu)^2}{2} G(x-\mu) \geq \frac{(x-\mu)^2}{2} g''\big(\frac{x-\mu}{3}\big)$, and $g''\big(\frac{x-\mu}{3}\big) = 1/(\tilde{x}(1-\tilde{x}))$ since $\mu + \frac{x-\mu}{3} = \tilde{x}$. The second inequality holds because $\tilde{x}$ lies between μ and x, and the third because $u(1-u) \leq \frac{1}{4}$. □

Remark 3.6 (The Jensen bound beats Bernstein's). For $x = \mu + t > \mu$, Bernstein's inequality (2.28) applied to the centred variables $X_i - \mu \leq 1-\mu$, with $\sum_i \mathbb{E}[(X_i-\mu)^2] = n\mu(1-\mu)$, gives the per-sample exponent $t^2/\big(2(\mu(1-\mu) + (1-\mu)t/3)\big) = t^2/\big(2(1-\mu)\tilde{x}\big)$ – which is the Jensen bound applied, not to the tilted variance $u(1-u)$ itself, but to its Bennett majorant $(1-\mu)u$ of Remark 2.43. Since $1-\tilde{x} \leq 1-\mu$,

$$\tilde{x}(1-\tilde{x}) \;\leq\; (1-\mu)\,\tilde{x} = \mu(1-\mu) + \frac{(1-\mu)t}{3},$$

strictly for $t > 0$: the first lower bound of (3.2) always beats Bernstein's exponent. The same holds for left deviations, $x = \mu - t$, where Bernstein is applied to $\mu - X_i \leq \mu$ and $\tilde{x}(1-\tilde{x}) \leq \mu(1-\tilde{x}) = \mu(1-\mu) + \frac{\mu t}{3}$. Both reduce to the Gaussian exponent $t^2/(2\mu(1-\mu))$ as $t \to 0$. The Jensen bound is even exact to third order, $\mathrm{kl}(\mu+t,\mu) - \frac{t^2}{2\tilde{x}(1-\tilde{x})} = O(t^4)$, because Jensen's inequality is an equality for affine functions, whereas Bernstein's exponent is off at order t^3; the second bound of (3.2), which uses the worst variance on the segment, is sometimes called the *improved Hoeffding* bound.

3.2.3 Worked comparisons

The value of an inequality is best judged on examples where the truth is known. Three classical ones follow; in each, the reader should keep in mind that kl itself is the exact exponent (Theorem 3.42 below), so that a lower bound on kl measures exactly what an inequality loses.

Example 3.7 (Majority vote)**.** Let $\mu < \frac{1}{2}$. Then

$$\mathbb{P}\Big(\bar{X}_n \geq \frac{1}{2}\Big) \;\leq\; \big(4\mu(1-\mu)\big)^{n/2} \;\leq\; \exp\Big(-\frac{n}{2}(1-2\mu)^2\Big). \tag{3.4}$$

The first bound is (3.1), since $\mathrm{kl}(\frac{1}{2},\mu) = \frac{1}{2}\log\frac{1}{4\mu(1-\mu)}$; it also has a two-line direct proof, $\mathbb{P}(\bar{X}_n \geq \frac{1}{2}) = \sum_{k \geq n/2} \binom{n}{k} \mu^k (1-\mu)^{n-k} \leq (\mu(1-\mu))^{n/2} \sum_k \binom{n}{k} = (4\mu(1-\mu))^{n/2}$, using $\mu^k(1-\mu)^{n-k} \leq (\mu(1-\mu))^{n/2}$ for $k \geq n/2$. The second follows from $4\mu(1-\mu) = 1-(1-2\mu)^2$ and $1-u \leq e^{-u}$; it is also exactly Hoeffding's bound $e^{-2n(1/2-\mu)^2}$. The interpretation: a committee of n independent experts, each wrong with probability $\mu < \frac{1}{2}$, is wrong by majority with probability exponentially small in $n(1-2\mu)^2$. Near $\mu = \frac{1}{2}$ the two bounds of (3.4) agree to leading order; for $\mu \to 0$ the first decays like $(4\mu)^{n/2}$, much faster than the second.

Example 3.8 (Halving a small probability from below)**.** Let us bound $\mathbb{P}(\bar{X}_n \le \mu/2)$, a left deviation of relative size $\frac{1}{2}$, when μ is small. By (i) of Theorem 3.1, $\mathbb{P}(\bar{X}_n \le \mu/2) \le e^{-n\,\mathrm{kl}(\mu/2,\mu)}$, and

$$\mathrm{kl}\Big(\frac{\mu}{2},\mu\Big) = -\frac{\mu}{2}\log 2 + \Big(1-\frac{\mu}{2}\Big)\log\frac{1-\mu/2}{1-\mu} \;\ge\; -\frac{\mu}{2}\log 2 + \frac{\mu}{2} = \frac{1-\log 2}{2}\,\mu \approx 0.153\,\mu,$$

using $\log w \ge (w-1)/w$ with $w = \frac{1-\mu/2}{1-\mu}$, for which $(w-1)/w = \frac{\mu/2}{1-\mu/2}$. Hence $\mathbb{P}(\bar{X}_n \le \mu/2) \le e^{-0.153\,n\mu} \le e^{-n\mu/7}$. The Jensen bound of (3.2), with $\tilde{x} = \frac{5}{6}\mu$, gives $\mathrm{kl}(\frac{\mu}{2},\mu) \ge \frac{\mu^2/4}{2\cdot\frac{5}{6}\mu} = \frac{3}{20}\mu = 0.15\,\mu$, hence $\mathbb{P}(\bar{X}_n \le \mu/2) \le e^{-3n\mu/20} \le e^{-n\mu/8}$, while Hoeffding's inequality gives only $e^{-n\mu^2/2}$: to make the bound smaller than δ, Hoeffding requires $n \ge 2\log(1/\delta)/\mu^2$ and the other two $n \ge \log(1/\delta)/\mathrm{kl}(\mu/2,\mu) \approx 6.5\log(1/\delta)/\mu$. For $\mu = 10^{-3}$ the ratio of the two sample sizes is about 300. This is the regime of Section 2.7.3 and of the realizable case of Section 2.7.1, where one needs to certify that a small empirical frequency reflects a small probability.

Example 3.9 (Doubling a small probability: six bounds)**.** Consider now $\mathbb{P}(\bar{X}_n \ge 2\mu)$, for $\mu < \frac{1}{2}$. We write each bound as $e^{-n\gamma(\mu)}$ and compare the per-sample exponents $\gamma(\mu)$.

- *Chernoff, with* $\log w \ge (w-1)/w$. With $w = \frac{1-2\mu}{1-\mu}$, $\log w \ge \frac{w-1}{w} = -\frac{\mu}{1-2\mu}$, so

$$\mathrm{kl}(2\mu,\mu) = 2\mu\log 2 + (1-2\mu)\log\frac{1-2\mu}{1-\mu} \;\ge\; (2\log 2 - 1)\,\mu \approx 0.386\,\mu.$$

- *Chernoff, with the Jensen bound* (3.2). Here $\tilde{x} = \frac{4}{3}\mu$ and $\mathrm{kl}(2\mu,\mu) \ge \frac{\mu^2/2}{\frac{4}{3}\mu(1-\frac{4}{3}\mu)} \ge \frac{3}{8}\mu$.

- *Improved Hoeffding* (second bound of (3.2)): $\max_{u\in[\mu,2\mu]} u(1-u) \le 2\mu$, so $\mathrm{kl}(2\mu,\mu) \ge \frac{\mu^2/2}{2\mu} = \frac{\mu}{4}$.

- *Bennett* (Theorem 2.40) with $Z_i = X_i - \mu \le b = 1-\mu$, variance factor $n\mu(1-\mu)$ and deviation $n\mu$, so that $b\cdot n\mu/(n\mu(1-\mu)) = 1$: the exponent is $\frac{\mu}{1-\mu}h(1) \ge (2\log 2-1)\mu \approx 0.386\,\mu$.

- *Bernstein* (Theorem 2.47) with the same parameters: $\frac{(n\mu)^2}{2(n\mu(1-\mu)+(1-\mu)n\mu/3)} = \frac{3n\mu}{8(1-\mu)}$, so $\gamma(\mu) \ge \frac{3}{8}\mu$.

- *Hoeffding*: $\gamma(\mu) = 2\mu^2$.

The exact exponent satisfies $\mathrm{kl}(2\mu,\mu) = (2\log 2-1)\mu + O(\mu^2)$ as $\mu \to 0$, so the six bounds compare as follows.

bound	$\gamma(\mu) \ge$	$\lim_{\mu\to 0}\frac{\gamma(\mu)}{\mathrm{kl}(2\mu,\mu)}$	$\frac{\gamma(\mu)}{\mathrm{kl}(2\mu,\mu)}$ at $\mu = 0.1$
Chernoff + $\log w \ge (w-1)/w$	$(2\log 2-1)\mu$	1	0.870
Chernoff + Jensen	$\frac{3}{8}\mu$	0.971	0.974
improved Hoeffding	$\frac{1}{4}\mu$	0.647	0.704
Bennett	$(2\log 2-1)\mu$	1	0.967
Bernstein	$\frac{3}{8}\mu$	0.971	0.938
Hoeffding	$2\mu^2$	0	0.450

(The last column uses the exact value of each bound rather than its simplified lower bound – e.g. $\frac{\mu}{1-\mu}h(1)$ for Bennett – the two coinciding for the first and last rows.) For $\mu = 0.1$ this is the event $\{\bar{X}_n \geq 2\mu\}$, written $\{\bar{X}_n \geq p + t\}$ with $t = p = 0.1$ in Section 2.5.4, and the entries of the last column for Bennett, Bernstein and Hoeffding are the ratios to kl of the exponents computed there (Figure 2.2); the table adds the two Chernoff-based rows, the improved Hoeffding bound and the limit $\mu \to 0$. Three lessons. First, as $\mu \to 0$ the Bennett exponent is *exact*: the binomial tail is then a Poisson tail, which is what Bennett's inequality is built on (Remark 2.41). Second, Bernstein and the Jensen bound lose only 3% in the exponent as $\mu \to 0$; the improved Hoeffding bound, which uses the largest variance on the segment $[\mu, 2\mu]$, loses a third. Third, Hoeffding's bound is not merely worse by a constant: its exponent is of the wrong *order*, μ^2 instead of μ, because it uses the worst possible variance $\frac{1}{4}$ instead of $\mu(1-\mu)$.

3.2.4 A divergence on the set of means

The Chernoff bound measures the distance from the empirical mean to the true one with kl, not with $|\cdot|$. Read this way, it becomes a confidence statement. We write $\mathbb{P}_\mu$ for the law under which $X_1, X_2, \ldots$ are i.i.d. $\mathrm{Ber}(\mu)$; the same notation is used with other parameters in Sections 3.4 and 3.5.

Proposition 3.10 (The kl confidence region)**.** *For every $\delta \in (0,1)$ and $n \geq 1$,*

$$\mathbb{P}_\mu\Big(n\ \mathrm{kl}\big(\bar{X}_n, \mu\big) \geq \log\frac{1}{\delta}\Big)\ \leq\ 2\delta. \tag{3.5}$$

Consequently the random set

$$I_n(\delta) := \Big\{\mu' \in [0,1] :\ \mathrm{kl}\big(\bar{X}_n, \mu'\big) \leq \frac{\log(1/\delta)}{n}\Big\}$$

is an interval that contains μ with probability at least $1-2\delta$, and it is contained in the Hoeffding interval $\big[\bar{X}_n - \sqrt{\log(1/\delta)/(2n)},\ \bar{X}_n + \sqrt{\log(1/\delta)/(2n)}\big]$.

Proof. Let $y = \log(1/\delta)/n > 0$. The function $\mathrm{kl}(\cdot, \mu)$ is continuous on $[0,1]$, strictly decreasing on $[0,\mu]$ and strictly increasing on $[\mu,1]$, with $\mathrm{kl}(\mu,\mu) = 0$. If $y \leq \mathrm{kl}(1,\mu)$, let $x_\delta \in (\mu,1]$ be the unique point with $\mathrm{kl}(x_\delta,\mu) = y$; then $\{\bar{X}_n \geq \mu,\ \mathrm{kl}(\bar{X}_n,\mu) \geq y\} = \{\bar{X}_n \geq x_\delta\}$, whose probability is at most $e^{-n\,\mathrm{kl}(x_\delta,\mu)} = \delta$ by (3.1). If $y > \mathrm{kl}(1,\mu)$ that event is empty. The same argument on $[0,\mu]$ with the left deviation bound handles $\{\bar{X}_n \leq \mu,\ \mathrm{kl}(\bar{X}_n,\mu) \geq y\}$, and the union bound gives (3.5). The set $I_n(\delta)$ is a sublevel set of the convex function $\mu' \mapsto \mathrm{kl}(\bar{X}_n,\mu')$, hence an interval, and $\mu \notin I_n(\delta)$ implies the event in (3.5), so that $\mathbb{P}_\mu(\mu \notin I_n(\delta)) \leq 2\delta$. Finally, if $\mu' \in I_n(\delta)$ then, by Pinsker's inequality (2.5) (extended by continuity to $\bar{X}_n \in \{0,1\}$), $2(\bar{X}_n - \mu')^2 \leq \mathrm{kl}(\bar{X}_n,\mu') \leq y$. □

Remark 3.11 (Reading the kl interval). Figure 3.1 draws $\mu' \mapsto \mathrm{kl}(\bar{X}_n,\mu')$ and its sublevel set. Two features make $I_n(\delta)$ the right interval in practice. It is *asymmetric*: near the boundary of $[0,1]$ the function $\mathrm{kl}(\bar{X}_n,\cdot)$ is steep on the side of the boundary and flat on the other, so that when $\bar{X}_n$ is small the interval extends further to the right than to the left, as it should, for a binomial with small mean has a long right tail. And it is *adaptive*: by (3.2), its half-width is of order $\sqrt{2\bar{X}_n(1-\bar{X}_n)\log(1/\delta)/n}$ when this is larger than $\log(1/\delta)/n$, and of order $\log(1/\delta)/n$ otherwise – in particular, when $\bar{X}_n = 0$, $I_n(\delta) = [0, 1-\delta^{1/n}] \subset [0, \log(1/\delta)/n]$, the confidence bound of the realizable case (Theorem 2.66). The Hoeffding interval, of half-width $\sqrt{\log(1/\delta)/(2n)}$ regardless of $\bar{X}_n$, is the worst case of both features. The factor 2 in

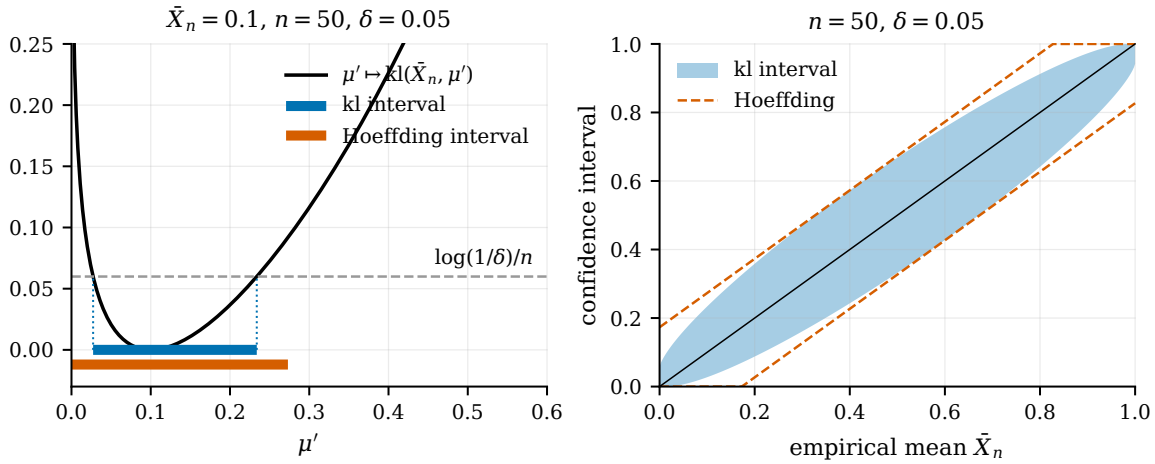


Figure 3.1: The kl confidence interval of Proposition 3.10. *Left:* the function $\mu' \mapsto \mathrm{kl}(\bar{X}_n, \mu')$ for $\bar{X}_n = 0.1$, its sublevel set at height $\log(1/\delta)/n$ (the kl interval) and the Hoeffding interval $\bar{X}_n \pm \sqrt{\log(1/\delta)/(2n)}$, which contains it. *Right:* both intervals as functions of the empirical mean, for $n = 50$ and $\delta = 0.05$. The kl interval is asymmetric and shrinks near the boundary of $[0, 1]$, where the variance $\mu'(1 - \mu')$ is small; the Hoeffding interval has constant width and sticks out of $[0, 1]$. Generated by `figures/fig_kl_confidence.py`.

(3.5) comes from the union of two tails; for a one-sided bound, it disappears. The kl interval is the basis of the KL-UCB algorithm for bandit problems [GC11]. Its exact counterpart is the Clopper–Pearson interval [CP34], which inverts the exact binomial tails instead of their Chernoff bounds: since $\mathbb{P}(\mathrm{Bin}(n, \mu') \le k) \le e^{-n\,\mathrm{kl}(k/n, \mu')}$ for $\mu' \ge k/n$ by Theorem 3.1(i), and symmetrically for the right tail, the Clopper–Pearson interval for $k = n\bar{X}_n$ successes, with δ in each tail, is contained in $I_n(\delta)$, hence slightly narrower; the two coincide when $\bar{X}_n \in \{0, 1\}$.

3.3 Relative entropy

The function kl has appeared in two roles: as the Cramér transform of a Bernoulli law (Proposition 2.23) and, through Proposition 2.17, as the relative entropy of the tilted law with respect to the original one. We now develop relative entropy for arbitrary laws, with complete proofs of the properties used in the rest of the chapter and in Chapter 4.

3.3.1 Definition and examples

Throughout, $(\Omega, \mathcal{F})$ is a measurable space and P, Q are probability measures on it. We write $\Phi(u) := u \log u$ for $u > 0$, $\Phi(0) := 0$; Φ is continuous and strictly convex on $[0, \infty)$, and $\Phi(u) \ge -1/e$, with equality at $u = 1/e$ only.

Definition 3.12 (Relative entropy)**.** The *relative entropy*, or *Kullback–Leibler divergence*, of P with respect to Q is

$$\mathrm{KL}(P\|Q) := \begin{cases} \displaystyle\int_\Omega \Phi\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\,\mathrm{d}Q = \int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q & \text{if } P \ll Q, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon–Nikodym derivative of P with respect to Q.

The integral is well defined in $(-\infty, +\infty]$: the negative part of $\Phi(\mathrm{d}P/\mathrm{d}Q)$ is at most $1/e$, hence Q-integrable. It does not depend on the version of $\mathrm{d}P/\mathrm{d}Q$. It can also be written

$$\mathrm{KL}(P\|Q) = \int_\Omega \log \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}P \qquad (P \ll Q), \tag{3.6}$$

where the integrand is defined P-almost surely, since $P(\mathrm{d}P/\mathrm{d}Q = 0) = \int_{\{\mathrm{d}P/\mathrm{d}Q=0\}} \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q = 0$; indeed, $\int g\ \mathrm{d}P = \int g \frac{\mathrm{d}P}{\mathrm{d}Q}\, \mathrm{d}Q$ for every non-negative measurable g, and one applies this to the positive and negative parts of $\log \frac{\mathrm{d}P}{\mathrm{d}Q}$. In words: $\mathrm{KL}(P\|Q)$ is the expectation, under P, of the log-likelihood ratio of P against Q. It will be shown in Proposition 3.16 to be non-negative.

In practice one computes with densities.

Lemma 3.13 (Relative entropy through densities)**.** *Let ν be a σ-finite measure on $(\Omega, \mathcal{F})$ such that $P \ll \nu$ and $Q \ll \nu$ (for instance $\nu = P + Q$), with densities $p = \mathrm{d}P/\mathrm{d}\nu$ and $q = \mathrm{d}Q/\mathrm{d}\nu$. Then $P \ll Q$ if and only if $P(q = 0) = 0$, in which case $\mathrm{d}P/\mathrm{d}Q = p/q$ on $\{q > 0\}$, and in all cases*

$$\mathrm{KL}(P\|Q) = \int_\Omega p \log \frac{p}{q}\, \mathrm{d}\nu, \tag{3.7}$$

with the conventions $0 \log \frac{0}{b} = 0$ for $b \geq 0$ and $a \log \frac{a}{0} = +\infty$ for $a > 0$. The integral is well defined in $(-\infty, +\infty]$.

Proof. If $P(q = 0) = 0$ then, for every $A \in \mathcal{F}$, $P(A) = \int_{A\cap\{q>0\}} p\ \mathrm{d}\nu = \int_A \frac{p}{q} \mathbb{1}\{q > 0\}\, q\ \mathrm{d}\nu = \int_A \frac{p}{q} \mathbb{1}\{q > 0\}\, \mathrm{d}Q$, so $P \ll Q$ with the stated density. Conversely if $P \ll Q$ then $P(q = 0) = 0$ because $Q(q = 0) = 0$. In that case $p = 0$ ν-almost everywhere on $\{q = 0\}$, and

$$\mathrm{KL}(P\|Q) = \int_{\{q>0\}} \frac{p}{q} \log \frac{p}{q}\, q\ \mathrm{d}\nu = \int_{\{q>0\}} p \log \frac{p}{q}\ \mathrm{d}\nu = \int_\Omega p \log \frac{p}{q}\ \mathrm{d}\nu.$$

If $P \not\ll Q$, then $\nu(\{p > 0, q = 0\}) > 0$ and the integrand is $+\infty$ on that set, so both sides of (3.7) are $+\infty$. In every case the integrand equals $q\Phi(p/q) \geq -q/e$ on $\{q > 0\}$ and is non-negative elsewhere, so its negative part is integrable. □

Proposition 3.14 (Examples)**.** *(a)* Exponential families. *Let Z be a real random variable with law P and cumulant generating function $K = K_Z$, and for $\lambda \in \mathrm{dom}(K)$ let P_λ be the law of Z under the tilted measure $\mathbb{P}_\lambda$ of Definition 2.2. For $\lambda, \lambda' \in \mathrm{int}\,\mathrm{dom}(K)$,*

$$\mathrm{KL}(P_\lambda\|P_{\lambda'}) = K(\lambda') - K(\lambda) - (\lambda' - \lambda)K'(\lambda), \tag{3.8}$$

the Bregman divergence *of the convex function K. For $\lambda' = 0$ this is $\lambda m(\lambda) - K(\lambda) = K^*(m(\lambda))$, the identity (2.8).*

(b) Bernoulli: $\mathrm{KL}\big(\mathrm{Ber}(x)\|\mathrm{Ber}(\mu)\big) = \mathrm{kl}(x, \mu)$ *for $x, \mu \in [0, 1]$ (with $\mathrm{kl}(x, \mu) = +\infty$ if $\mu \in \{0, 1\}$ and $x \neq \mu$).*

(c) Gaussian, same variance: *in $\mathbb{R}$ and in $\mathbb{R}^d$,*

$$\mathrm{KL}\big(\mathcal{N}(\mu_1, \sigma^2)\|\mathcal{N}(\mu_2, \sigma^2)\big) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2},$$
$$\mathrm{KL}\big(\mathcal{N}(\theta_1, \sigma^2 I_d)\|\mathcal{N}(\theta_2, \sigma^2 I_d)\big) = \frac{\|\theta_1 - \theta_2\|^2}{2\sigma^2}.$$

(d) Poisson: $\mathrm{KL}\big(\mathrm{Poi}(a)\|\mathrm{Poi}(b)\big) = a \log \frac{a}{b} - a + b = b\, h\big(\frac{a}{b} - 1\big)$ *for $a, b > 0$, with h as in (2.11).*

(e) Exponential: $\mathrm{KL}\big(\mathrm{Exp}(a)\|\mathrm{Exp}(b)\big) = \log\frac{a}{b} + \frac{b}{a} - 1$ *for rates* $a, b > 0$.

Proof. (a) By (2.1), P_λ and $P_{\lambda'}$ are both equivalent to P and $\log\frac{\mathrm{d}P_\lambda}{\mathrm{d}P_{\lambda'}}(z) = (\lambda-\lambda')z - K(\lambda) + K(\lambda')$. Integrating with respect to P_λ, under which Z has mean $m(\lambda) = K'(\lambda)$ by Theorem 2.3, gives (3.8). (b) For $\mu \in (0,1)$ this is Lemma 3.13 with ν the counting measure on $\{0,1\}$; if $\mu \in \{0,1\}$ and $x \neq \mu$, then $\mathrm{Ber}(x) \not\ll \mathrm{Ber}(\mu)$. (c) The log-likelihood ratio is $\frac{(z-\mu_2)^2-(z-\mu_1)^2}{2\sigma^2} = \frac{(\mu_1-\mu_2)(2z-\mu_1-\mu_2)}{2\sigma^2}$, whose mean under $\mathcal{N}(\mu_1,\sigma^2)$ is $\frac{(\mu_1-\mu_2)^2}{2\sigma^2}$. The d-dimensional statement follows from the one-dimensional one by tensorisation (Proposition 3.20 below), the coordinates being independent. (d) $\log\frac{\mathrm{dPoi}(a)}{\mathrm{dPoi}(b)}(k) = k\log\frac{a}{b} - a + b$, with mean $a\log\frac{a}{b} - a + b$ under $\mathrm{Poi}(a)$; the second form is the computation of Proposition 2.23(c). (e) $\log\frac{ae^{-ax}}{be^{-bx}} = \log\frac{a}{b} - (a-b)x$, with mean $\log\frac{a}{b} - \frac{a-b}{a}$ under $\mathrm{Exp}(a)$. □

Remark 3.15 (Reading the examples). Items (b)–(e) are all instances of (a): each family is the exponential family generated by one of its members, and the relative entropy between two members is the gap between the cumulant generating function and its tangent. Taylor's formula then gives the local behaviour $\mathrm{KL}(P_\lambda\|P_{\lambda'}) = \frac{1}{2}(\lambda'-\lambda)^2K''(\lambda) + O(|\lambda'-\lambda|^3)$: *locally, relative entropy is half a squared distance, measured with the Fisher information* $K''(\lambda)$ *of the family.* In the mean parametrisation $x = m(\lambda)$ the Fisher information becomes $1/\varsigma^2(\lambda)$ and one recovers, for the Bernoulli family, $\mathrm{kl}(x,\mu) \approx \frac{(x-\mu)^2}{2\mu(1-\mu)}$, in agreement with (3.3). Relative entropy is not symmetric – $\mathrm{kl}(x,\mu) \neq \mathrm{kl}(\mu,x)$ in general, and $\mathrm{KL}(P\|Q)$ may be finite while $\mathrm{KL}(Q\|P)$ is infinite (Exercise 3.2) – and it does not satisfy the triangle inequality; it is a *divergence*, not a distance. The order of the arguments always has the same meaning in this chapter: the first argument is the law under which the expectation is taken, i.e. the law that is *supposed true* for the purpose of the argument.

3.3.2 Positivity, convexity, tensorisation

Proposition 3.16 (Gibbs' inequality)**.** $\mathrm{KL}(P\|Q) \geq 0$, *with equality if and only if* $P = Q$.

Proof. If $P \not\ll Q$ there is nothing to prove. Otherwise let $F = \mathrm{d}P/\mathrm{d}Q$; since Φ is convex and $\int F\,\mathrm{d}Q = 1$, Jensen's inequality gives $\mathrm{KL}(P\|Q) = \int \Phi(F)\,\mathrm{d}Q \geq \Phi\big(\int F\,\mathrm{d}Q\big) = \Phi(1) = 0$. Since Φ is strictly convex, equality forces F to be Q-almost surely constant, hence equal to 1, i.e. $P = Q$. □

Lemma 3.17 (Log-sum inequality)**.** *For* $a_1,\dots,a_N \geq 0$ *and* $b_1,\dots,b_N \geq 0$, *with* $A = \sum_j a_j$ *and* $B = \sum_j b_j$, *and the conventions of Lemma 3.13,*

$$A\log\frac{A}{B} \;\leq\; \sum_{j=1}^{N} a_j \log\frac{a_j}{b_j}.$$

Proof. If $a_j > 0 = b_j$ for some j, the right-hand side is $+\infty$. Otherwise, the indices with $a_j = 0$ contribute 0 to the right-hand side, and removing them can only decrease B, hence increase the left-hand side (as $B \mapsto A\log\frac{A}{B}$ is non-increasing); so we may assume $a_j > 0$ and $b_j > 0$ for every j. Jensen's inequality for Φ and the weights b_j/B then gives $\sum_j \frac{b_j}{B}\Phi\big(\frac{a_j}{b_j}\big) \geq \Phi\big(\frac{A}{B}\big)$, and multiplying by B yields the claim. □

Proposition 3.18 (Joint convexity). *Let $P_1, \dots, P_N$ and $Q_1, \dots, Q_N$ be probability measures on $(\Omega, \mathcal{F})$ and let $w_1, \dots, w_N \geq 0$ with $\sum_j w_j = 1$. Then*

$$\mathrm{KL}\Big(\sum_j w_j P_j \,\Big\|\, \sum_j w_j Q_j\Big) \;\leq\; \sum_j w_j \,\mathrm{KL}(P_j \| Q_j).$$

In particular $\mathrm{KL}(P\|Q)$ *is convex in each argument separately; for instance,* $\mathrm{KL}\big(P\| \sum_j w_j Q_j\big) \leq \sum_j w_j \,\mathrm{KL}(P\|Q_j)$.

Proof. Let $\nu = \sum_j (P_j + Q_j)$ and let p_j, q_j be the densities. By Lemma 3.13 the left-hand side is $\int \big(\sum_j w_j p_j\big) \log \frac{\sum_j w_j p_j}{\sum_j w_j q_j} \,\mathrm{d}\nu$, and Lemma 3.17 applied pointwise with $a_j = w_j p_j$ and $b_j = w_j q_j$ bounds the integrand by $\sum_j w_j p_j \log \frac{p_j}{q_j}$ (the factors w_j cancel inside the logarithms, and the terms with $w_j = 0$ vanish). Integrating, which is legitimate since all the negative parts are integrable, gives the claim. □

Proposition 3.19 (Chain rule). *Let $(\mathcal{X}, \mathcal{B}_\mathcal{X}, \nu)$ and $(\mathcal{Y}, \mathcal{B}_\mathcal{Y}, \nu')$ be σ-finite measure spaces, let p, q be probability densities on $\mathcal{X}$ with respect to ν and let $k, k' : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ be measurable functions such that $k(x, \cdot)$ and $k'(x, \cdot)$ are probability densities with respect to ν' for every x. Let P and Q be the laws on $\mathcal{X} \times \mathcal{Y}$ with densities $p(x)k(x, y)$ and $q(x)k'(x, y)$ with respect to $\nu \otimes \nu'$, and $P_\mathcal{X}$, $Q_\mathcal{X}$ their first marginals (with densities p, q). Writing $P(\cdot \mid x)$ and $Q(\cdot \mid x)$ for the laws with densities $k(x, \cdot)$ and $k'(x, \cdot)$,*

$$\mathrm{KL}(P\|Q) = \mathrm{KL}(P_\mathcal{X}\|Q_\mathcal{X}) + \int_\mathcal{X} \mathrm{KL}\big(P(\cdot \mid x)\|Q(\cdot \mid x)\big) \,\mathrm{d}P_\mathcal{X}(x), \tag{3.9}$$

all terms being in $[0, +\infty]$ (and $x \mapsto \mathrm{KL}\big(P(\cdot \mid x)\|Q(\cdot \mid x)\big)$ being measurable).

Proof. For $a, b, a', b' \geq 0$ one checks the identity $aa' \log \frac{aa'}{bb'} = a' \cdot a \log \frac{a}{b} + a \cdot a' \log \frac{a'}{b'}$, with the conventions of Lemma 3.13 and $0 \cdot \infty = 0$: if $a = 0$ or $a' = 0$ both sides vanish, and if $a, a' > 0$ both sides are infinite exactly when $bb' = 0$, and equal otherwise. Apply it pointwise with $a = p(x)$, $b = q(x)$, $a' = k(x, y)$, $b' = k'(x, y)$. The two terms on the right have integrable negative parts with respect to $\nu \otimes \nu'$, since $a' \, a \log \frac{a}{b} \geq -a'\, b/e$ and $a \, a' \log \frac{a'}{b'} \geq -a \, b'/e$, and $\iint k(x, y) q(x) \,\mathrm{d}\nu \,\mathrm{d}\nu' = \iint p(x) k'(x, y) \,\mathrm{d}\nu \,\mathrm{d}\nu' = 1$. We may therefore integrate term by term and use Fubini–Tonelli on positive and negative parts separately. Since $\int k(x, y) \, \nu'(\mathrm{d}y) = 1$, the first term integrates to $\int p \log \frac{p}{q} \,\mathrm{d}\nu = \mathrm{KL}(P_\mathcal{X}\|Q_\mathcal{X})$; the second integrates to $\int p(x) \big(\int k \log \frac{k}{k'} \,\mathrm{d}\nu'\big) \nu(\mathrm{d}x) = \int \mathrm{KL}\big(P(\cdot \mid x)\|Q(\cdot \mid x)\big) \,\mathrm{d}P_\mathcal{X}(x)$ by Lemma 3.13 applied for each x (which also gives the measurability). The left-hand side is $\mathrm{KL}(P\|Q)$ by Lemma 3.13. □

Proposition 3.20 (Tensorisation). *For probability measures P_i, Q_i on measurable spaces $(\Omega_i, \mathcal{F}_i)$, $i = 1, \dots, n$,*

$$\mathrm{KL}\big(P_1 \otimes \cdots \otimes P_n \,\|\, Q_1 \otimes \cdots \otimes Q_n\big) = \sum_{i=1}^n \mathrm{KL}(P_i\|Q_i).$$

In particular $\mathrm{KL}(P^{\otimes n}\|Q^{\otimes n}) = n \,\mathrm{KL}(P\|Q)$.

Proof. By induction it suffices to treat $n = 2$. Let $\nu_i = P_i + Q_i$, with densities p_i, q_i. The product laws have densities $p_1(x)p_2(y)$ and $q_1(x)q_2(y)$ with respect to $\nu_1 \otimes \nu_2$, which is the situation of Proposition 3.19 with $k(x, y) = p_2(y)$ and $k'(x, y) = q_2(y)$ (normalise ν_1, ν_2, which are finite measures, if one wishes). The conditional laws of Proposition 3.19 for the joint laws $P_1 \otimes P_2$ and $Q_1 \otimes Q_2$ are P_2 and Q_2 for every x, and (3.9) gives $\mathrm{KL}(P_1\|Q_1) + \mathrm{KL}(P_2\|Q_2)$. □

Remark 3.21 (Why tensorisation is the decisive property). The relative entropy between the laws of two n-samples grows *linearly* in n. Other measures of discrepancy do not behave so simply: the total variation distance between $P^{\otimes n}$ and $Q^{\otimes n}$ (Definition 3.26 below) has no closed form in terms of that between P and Q, and only satisfies $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le n\,\mathrm{TV}(P, Q)$, which is vacuous for $n \ge 1/\mathrm{TV}(P, Q)$. All the lower bounds of this chapter follow the same route: compute a relative entropy between product laws by tensorisation, then convert it into a statement about probabilities of events by the inequalities of the next subsection. The chain rule (3.9) is the form of tensorisation that survives when the observations are *not* independent – for instance when they are collected sequentially by an adaptive procedure, as in bandit problems [LS20, Chap. 15]; the density hypothesis of Proposition 3.19 can be removed when $\mathcal{Y}$ is a standard Borel space, using regular conditional distributions [PW25], but we shall not need this.

3.3.3 Data processing

Proposition 3.22 (Data-processing inequality, or contraction of relative entropy)**.** *Let $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ be measurable, and let P^X, Q^X denote the laws of X under P and Q (the push-forward measures). Then*

$$\mathrm{KL}\big(P^X \| Q^X\big) \;\le\; \mathrm{KL}(P\|Q).$$

Proof. If $\mathrm{KL}(P\|Q) = +\infty$ there is nothing to prove. Otherwise $P \ll Q$; let $F = \mathrm{d}P/\mathrm{d}Q$, which is Q-integrable with $\int F\,\mathrm{d}Q = 1$. By the Doob–Dynkin lemma there is a measurable $\gamma : \mathcal{X} \to [0, \infty)$ with $\mathbb{E}_Q[F \mid X] = \gamma(X)$ Q-almost surely.

Step 1: $P^X \ll Q^X$ and $\mathrm{d}P^X/\mathrm{d}Q^X = \gamma$. For every $B \in \mathcal{B}_{\mathcal{X}}$,

$$P^X(B) = \mathbb{E}_Q\big[F\,\mathbb{1}\{X \in B\}\big] = \mathbb{E}_Q\big[\mathbb{E}_Q[F \mid X]\,\mathbb{1}\{X \in B\}\big] = \mathbb{E}_Q\big[\gamma(X)\mathbb{1}\{X \in B\}\big] = \int_B \gamma\;\mathrm{d}Q^X.$$

Step 2: Jensen. The function Φ is convex and bounded below, and $\Phi(F)$ is Q-integrable because $\mathrm{KL}(P\|Q) < \infty$; the conditional Jensen inequality gives $\Phi(\mathbb{E}_Q[F \mid X]) \le \mathbb{E}_Q[\Phi(F) \mid X]$ Q-almost surely. Hence

$$\begin{aligned}\mathrm{KL}\big(P^X\|Q^X\big) &= \int_{\mathcal{X}} \Phi(\gamma)\;\mathrm{d}Q^X = \mathbb{E}_Q\big[\Phi\big(\mathbb{E}_Q[F \mid X]\big)\big] \\ &\le \mathbb{E}_Q\big[\mathbb{E}_Q[\Phi(F) \mid X]\big] = \mathbb{E}_Q[\Phi(F)] = \mathrm{KL}(P\|Q).\end{aligned}$$

□

Data processing extends at once to *randomised* transformations. A *Markov kernel* from $(\Omega, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ is a map $\kappa : \Omega \times \mathcal{B}_{\mathcal{Y}} \to [0, 1]$ such that $\kappa(\omega, \cdot)$ is a probability measure for every ω and $\kappa(\cdot, B)$ is measurable for every B; it describes an output Y drawn from $\kappa(\omega, \cdot)$ when the input is ω. We write $P\kappa(B) := \int \kappa(\omega, B)\,P(\mathrm{d}\omega)$ for the law of the output when the input has law P, and $P \otimes \kappa$ for the joint law of (input, output) on $\Omega \times \mathcal{Y}$.

Corollary 3.23 (Data processing through a kernel)**.** *For every Markov kernel κ, $\mathrm{KL}(P\kappa\|Q\kappa) \le \mathrm{KL}(P \otimes \kappa\|Q \otimes \kappa) = \mathrm{KL}(P\|Q)$.*

Proof. If $P \ll Q$ with $F = \mathrm{d}P/\mathrm{d}Q$, then for every measurable $E \subset \Omega \times \mathcal{Y}$, $(P \otimes \kappa)(E) = \iint \mathbb{1}_E(\omega, y)\,\kappa(\omega, \mathrm{d}y)F(\omega)\,Q(\mathrm{d}\omega)$, so $P \otimes \kappa \ll Q \otimes \kappa$ with density $(\omega, y) \mapsto F(\omega)$, and $\mathrm{KL}(P \otimes \kappa\|Q \otimes \kappa) = \int \Phi(F(\omega))\;\mathrm{d}(Q \otimes \kappa) = \mathrm{KL}(P\|Q)$. If $P \not\ll Q$, a set A with $Q(A) = 0 < P(A)$ gives the set $A \times \mathcal{Y}$, null for $Q \otimes \kappa$ and not for $P \otimes \kappa$, and both sides are infinite. The inequality is Proposition 3.22 applied to the projection $(\omega, y) \mapsto y$. □

Corollary 3.24 (Data processing for expectations)**.** *For every measurable* $Z : \Omega \to [0,1]$*,*

$$\mathrm{kl}\big(\mathbb{E}_P[Z], \mathbb{E}_Q[Z]\big) \;\le\; \mathrm{KL}(P\|Q).$$

In particular $\mathrm{kl}\big(P(A), Q(A)\big) \le \mathrm{KL}(P\|Q)$ *for every event* $A \in \mathcal{F}$*.*

Proof. Apply Corollary 3.23 to the kernel $\kappa(\omega, \cdot) = \mathrm{Ber}(Z(\omega))$ on $\{0,1\}$: then $P\kappa = \mathrm{Ber}(\mathbb{E}_P Z)$ and $Q\kappa = \mathrm{Ber}(\mathbb{E}_Q Z)$, and Proposition 3.14(b) concludes. For $Z = \mathbb{1}_A$ the kernel is deterministic. □

Remark 3.25 (What data processing says, and what it is used for). Proposition 3.22 is the mathematical form of the principle that *no processing of the data can create information*: if two laws are hard to distinguish from an observation ω, they are at least as hard to distinguish from any statistic $X(\omega)$, even a randomised one. Equality holds when X is *sufficient* for the pair $\{P, Q\}$, i.e. when $\mathrm{d}P/\mathrm{d}Q$ is a function of X – this is the origin of the subject, since Kullback and Leibler introduced their divergence precisely as a measure of "information for discrimination" that is preserved by sufficient statistics and only by them [KL51]. Corollary 3.24, the form we shall use most, reduces a statement about the laws of a whole sample to a statement about *two numbers*: the probabilities of one event under the two hypotheses. Any learning algorithm, any estimator, any test is a measurable – possibly randomised – function of the data; that is why the lower bounds derived from Corollary 3.24 hold for all of them at once. This formulation, with a general $[0,1]$-valued Z, is emphasised in [GMS19], where it is used to derive lower bounds for sequential decision problems.

3.3.4 From relative entropy to probabilities of events

Relative entropy is computable; what one wants to bound are probabilities. Three conversions are used throughout: Pinsker's inequality, which compares KL with the total variation distance and is good when KL is small; the inequality $\mathrm{kl}(x, \mu) \ge x \log \frac{1}{\mu} - \log 2$, good when one of the two probabilities is close to 1; and the Bretagnolle–Huber inequality, good when KL is large.

Definition 3.26 (Total variation distance)**.** $\mathrm{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$.

Lemma 3.27 (Expressions of the total variation)**.** *Let* ν*,* p*,* q *be as in Lemma 3.13. Then*

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int \big|p - q\big|\,\mathrm{d}\nu = 1 - \int \min(p,q)\,\mathrm{d}\nu = P(A^*) - Q(A^*), \qquad A^* := \{p > q\}.$$

Moreover, for every measurable $\phi : \Omega \to [0,1]$*,* $\int (1-\phi)\,\mathrm{d}P + \int \phi\,\mathrm{d}Q \;\ge\; \int \min(p,q)\,\mathrm{d}\nu = 1 - \mathrm{TV}(P,Q)$*, with equality for* $\phi = \mathbb{1}_{A^*}$*.*

Proof. For every A, $P(A) - Q(A) = \int_A (p-q)\,\mathrm{d}\nu \le \int (p-q)_+\,\mathrm{d}\nu = P(A^*) - Q(A^*)$, and symmetrically $Q(A) - P(A) \le \int (q-p)_+\,\mathrm{d}\nu$. Since $\int (p-q)\,\mathrm{d}\nu = 0$, the two positive parts have the same integral, equal to $\frac{1}{2}\int \big|p-q\big|$; this proves the first and last expressions. The second follows from $\min(p,q) = p - (p-q)_+$. For the last statement, $(1-\phi)p + \phi q \ge \min(p,q)$ pointwise, with equality when $\phi = 1$ on $\{q < p\}$ and $\phi = 0$ on $\{q > p\}$. □

Proposition 3.28 (Pinsker's inequality)**.** $\mathrm{TV}(P,Q) \le \sqrt{\mathrm{KL}(P\|Q)/2}$.

Proof. Let $A \in \mathcal{F}$. By Corollary 3.24 and Pinsker's inequality for Bernoulli laws (2.5), $\mathrm{KL}(P\|Q) \ge \mathrm{kl}\big(P(A), Q(A)\big) \ge 2\big(P(A) - Q(A)\big)^2$. Inequality (2.5) was stated for parameters in $(0,1)$; it extends to $x \in [0,1]$, $\mu \in (0,1)$ by continuity of $\mathrm{kl}(\cdot, \mu)$, and is trivial when $\mu \in \{0,1\}$, since $\mathrm{kl}(x,\mu)$ is then $+\infty$ unless $x = \mu$. Take the supremum over A. □

Remark 3.29 (On Pinsker's inequality). The proof is a perfect illustration of the strategy described in Remark 3.25: data processing reduces the general inequality to the case of two-point spaces, where it is a calculus exercise (Corollary 2.12 derived it from Popoviciu's bound $\varsigma^2 \le \frac{1}{4}$ on the tilted variance). Pinsker [Pin64] proved the inequality with a worse constant; the optimal constant $\frac{1}{2}$ is due independently to Csiszár [Csi67], Kullback [Kul67] and Kemperman [Kem69], and it is sharp: for $P = \mathrm{Ber}(\frac{1}{2}+\varepsilon)$ and $Q = \mathrm{Ber}(\frac{1}{2})$, $\mathrm{TV} = \varepsilon$ and $\mathrm{KL} = 2\varepsilon^2 + O(\varepsilon^4)$. Pinsker's inequality is *useless* when $\mathrm{KL}(P\|Q) \ge 2$, since its right-hand side then exceeds 1; but KL between product laws grows like n, and this is exactly the regime of large deviations. The next two results do not have this defect.

Lemma 3.30 (A lower bound on kl for unlikely events)**.** *For $x \in [0,1]$ and $\mu \in (0,1]$,*

$$\mathrm{kl}(x,\mu) \;\ge\; x\log\frac{1}{\mu} - H(x) \;\ge\; x\log\frac{1}{\mu} - \log 2,$$

where $H(x) = -x\log x - (1-x)\log(1-x) \in [0,\log 2]$ is the binary entropy.

Proof. $\mathrm{kl}(x,\mu) = -H(x) + x\log\frac{1}{\mu} + (1-x)\log\frac{1}{1-\mu}$, and the last term is non-negative (it is $+\infty$ if $\mu = 1 > x$, and 0 if $x = 1$). The maximum of H is $H(\frac{1}{2}) = \log 2$. □

Corollary 3.31 (Change of measure for events)**.** *For every event A with $P(A) > 0$,*

$$Q(A) \;\ge\; \exp\!\Big(-\frac{\mathrm{KL}(P\|Q) + \log 2}{P(A)}\Big). \tag{3.10}$$

Proof. If $Q(A) = 0$ then $P \not\ll Q$ and the right-hand side is 0. Otherwise Corollary 3.24 and Lemma 3.30 give $\mathrm{KL}(P\|Q) \ge P(A)\log\frac{1}{Q(A)} - \log 2$; rearrange. □

Read (3.10) as follows: *if an event is likely under P, it cannot be much less likely than $e^{-\mathrm{KL}(P\|Q)}$ under Q*. This is the non-asymptotic form of the change-of-measure argument, and the engine of Theorem 3.47 and of Fano's inequality.

Proposition 3.32 (Bretagnolle–Huber inequality)**.** *For all probability measures P, Q on $(\Omega,\mathcal{F})$,*

$$\int \min(p,q)\,\mathrm{d}\nu \;\ge\; 1 - \sqrt{1 - e^{-\mathrm{KL}(P\|Q)}} \;\ge\; \frac{1}{2}\,e^{-\mathrm{KL}(P\|Q)}. \tag{3.11}$$

Equivalently, $\mathrm{TV}(P,Q) \le \sqrt{1-e^{-\mathrm{KL}(P\|Q)}}$, and for every event A,

$$P(A^c) + Q(A) \;\ge\; \frac{1}{2}\,e^{-\mathrm{KL}(P\|Q)}.$$

Proof. We may assume $\mathrm{KL}(P\|Q) < \infty$, hence $P \ll Q$. Let $a = \int \min(p,q)\,\mathrm{d}\nu$ and $b = \int \max(p,q)\,\mathrm{d}\nu$; since $\min + \max = p + q$, $a + b = 2$. By the Cauchy–Schwarz inequality, as $\sqrt{pq} = \sqrt{\min(p,q)\max(p,q)}$,

$$\Big(\int \sqrt{pq}\,\mathrm{d}\nu\Big)^2 \;\le\; a\,b = a(2-a).$$

On the other hand, $\int \sqrt{pq}\,\mathrm{d}\nu \ge \int_{\{p>0\}} \sqrt{q/p}\;\mathrm{d}P$, and Jensen's inequality for the concave logarithm gives

$$2\log\int_{\{p>0\}} \sqrt{\frac{q}{p}}\;\mathrm{d}P \;\ge\; \int_{\{p>0\}} \log\frac{q}{p}\;\mathrm{d}P = -\mathrm{KL}(P\|Q),$$

using Lemma 3.13. Hence $a(2-a) \geq e^{-\mathrm{KL}(P\|Q)}$, that is $(1-a)^2 \leq 1 - e^{-\mathrm{KL}(P\|Q)}$, which is the first inequality of (3.11) since $a \leq 1$. The second follows from $1 - \sqrt{1-u} \geq u/2$ for $u \in [0,1]$, a consequence of $\sqrt{1-u} \leq 1 - u/2$. The total variation form is Lemma 3.27, and the last statement follows from $P(A^c) + Q(A) = \int_{A^c} p \, \mathrm{d}\nu + \int_A q \, \mathrm{d}\nu \geq \int \min(p,q) \, \mathrm{d}\nu$. □

Remark 3.33 (Pinsker or Bretagnolle–Huber?). Both inequalities bound TV by a function of $D = \mathrm{KL}(P\|Q)$. Pinsker's bound $\sqrt{D/2}$ is better for small D (it is sharp as $D \to 0$, while $\sqrt{1-e^{-D}} \approx \sqrt{D}$ loses a factor $\sqrt{2}$), and the Bretagnolle–Huber bound $\sqrt{1-e^{-D}} < 1$ is better for $D > 1.594$ – the root of $D/2 = 1 - e^{-D}$ – and never vacuous. In terms of the sum of two errors, which is what lower bounds use, Pinsker gives $1 - \sqrt{D/2}$, positive only for $D < 2$, while Bretagnolle–Huber gives $\frac{1}{2}e^{-D}$, positive for every D: it is the right tool when the divergence is of order n and one wants an exponentially small but non-zero lower bound. The inequality is due to Bretagnolle and Huber [BH79], in a paper on minimax density estimation; it was popularised by Tsybakov's book [Tsy09, Lemma 2.6] and by the bandit literature [LS20, Thm. 14.2].

For upper bounds on KL one often uses the χ^2-divergence, $\chi^2(P\|Q) := \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q = \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 \mathrm{d}Q - 1$ if $P \ll Q$, and $+\infty$ otherwise.

Lemma 3.34 (Relative entropy and χ^2)**.** $\mathrm{KL}(P\|Q) \leq \log\bigl(1 + \chi^2(P\|Q)\bigr) \leq \chi^2(P\|Q)$. *In particular, for $x \in [0,1]$ and $\mu \in (0,1)$,*

$$\mathrm{kl}(x,\mu) \;\leq\; \frac{(x-\mu)^2}{\mu(1-\mu)}. \tag{3.12}$$

Proof. If $P \ll Q$ with $F = \mathrm{d}P/\mathrm{d}Q$, (3.6) and Jensen's inequality for the concave logarithm give $\mathrm{KL}(P\|Q) = \int \log F \, \mathrm{d}P \leq \log \int F \, \mathrm{d}P = \log \int F^2 \, \mathrm{d}Q$. For Bernoulli laws, $\chi^2 = \frac{(x-\mu)^2}{\mu} + \frac{(x-\mu)^2}{1-\mu} = \frac{(x-\mu)^2}{\mu(1-\mu)}$. □

Comparing (3.12) with (3.2), $\mathrm{kl}(x,\mu)$ is squeezed between $(x-\mu)^2/(2\max u(1-u))$ and $(x-\mu)^2/(\mu(1-\mu))$: up to a factor 2 and the choice of the point where the variance is evaluated, *the relative entropy between two Bernoulli laws is the squared difference of the means divided by the variance.*

3.3.5 The variational formula

The last general property is the duality between relative entropy and cumulant generating functions, already met in Chapter 2 as the Donsker–Varadhan inequality (2.10).

Proposition 3.35 (Gibbs variational principle, Donsker–Varadhan formula)**.** *For every bounded measurable $g : \Omega \to \mathbb{R}$,*

$$\log \mathbb{E}_Q\left[e^g\right] = \max_P \Bigl\{\mathbb{E}_P[g] - \mathrm{KL}(P\|Q)\Bigr\}, \tag{3.13}$$

the maximum over all probability measures P on $(\Omega, \mathcal{F})$ being attained exactly at the Gibbs measure Q_g defined by $\mathrm{d}Q_g/\mathrm{d}Q = e^g/\mathbb{E}_Q[e^g]$. Dually,

$$\mathrm{KL}(P\|Q) = \sup_g \Bigl\{\mathbb{E}_P[g] - \log \mathbb{E}_Q\left[e^g\right]\Bigr\}, \tag{3.14}$$

the supremum being over bounded measurable functions.

Proof. Let P with $\mathrm{KL}(P\|Q) < \infty$. Since Q_g and Q are equivalent with $\log \frac{\mathrm{d}Q_g}{\mathrm{d}Q} = g - \log \mathbb{E}_Q e^g$ bounded, $P \ll Q_g$ and $\log \frac{\mathrm{d}P}{\mathrm{d}Q} = \log \frac{\mathrm{d}P}{\mathrm{d}Q_g} + g - \log \mathbb{E}_Q e^g$ P-almost surely; integrating with respect to P,

$$\mathrm{KL}(P\|Q) = \mathrm{KL}(P\|Q_g) + \mathbb{E}_P[g] - \log \mathbb{E}_Q\left[e^g\right]. \tag{3.15}$$

By Proposition 3.16, $\mathbb{E}_P[g] - \mathrm{KL}(P\|Q) \le \log \mathbb{E}_Q e^g$, with equality if and only if $P = Q_g$; measures with $\mathrm{KL}(P\|Q) = \infty$ give $-\infty$. This proves (3.13), and also the inequality "$\ge$" in (3.14).

For the reverse inequality in (3.14), suppose first $P \ll Q$, with $F = \mathrm{d}P/\mathrm{d}Q$, and let $g_M := \max\bigl(\min(\log F, M), -M\bigr)$ for $M > 0$ (with $\log 0 = -\infty$). As $M \to \infty$, $\mathbb{E}_P[g_M] \to \int \log F \, \mathrm{d}P = \mathrm{KL}(P\|Q)$, by monotone convergence for the positive parts and dominated convergence for the negative parts (dominated by $(\log F)^-$, which is P-integrable by (3.6)); and $e^{g_M} = \min\bigl(\max(F, e^{-M}), e^M\bigr) \to F$ pointwise, with $e^{g_M} \le F + 1$, so $\mathbb{E}_Q e^{g_M} \to \mathbb{E}_Q F = 1$. Hence the supremum is at least $\mathrm{KL}(P\|Q)$, whether finite or not. If $P \not\ll Q$, choose A with $Q(A) = 0 < P(A)$ and $g = M\mathbb{1}_A$: then $\mathbb{E}_P g - \log \mathbb{E}_Q e^g = MP(A) \to \infty$. □

Remark 3.36. With $g = \lambda Z$, (3.15) is the identity used in the proof of Proposition 2.17, and (3.13) says that the cumulant generating function and the relative entropy are convex conjugates of each other – the infinite-dimensional version of the duality between K_Z and K_Z^*. The formula is due, in this generality, to Donsker and Varadhan [DV75]; in statistical mechanics (3.13) is the statement that the Gibbs measure minimises the free energy. It is the starting point of the *entropy method* – in the form of the duality formula of Proposition 5.35, it is what makes the entropy tensorise (Theorem 5.36) – and of PAC-Bayesian bounds, which appear in a later chapter.

3.3.6 The testing interpretation

Why should relative entropy be *the* measure of how far apart two laws are? The most convincing answer is operational: it is the exponential rate at which one can tell them apart from a sample.

Let P_0, P_1 be two probability measures on $(\Omega, \mathcal{F})$, with densities p_0, p_1 with respect to some σ-finite ν. A (randomised) *test* is a measurable $\phi : \Omega \to [0,1]$, interpreted as the probability of rejecting the null hypothesis H_0 : "the law is P_0" in favour of H_1 : "the law is P_1". Its *type I* and *type II* errors are

$$\alpha(\phi) := \int \phi \, \mathrm{d}P_0, \qquad \beta(\phi) := \int (1-\phi) \, \mathrm{d}P_1.$$

A deterministic test is the indicator $\phi = \mathbb{1}_R$ of a rejection region R. One cannot make both errors small at once unless P_0 and P_1 are far apart: by Lemma 3.27, $\alpha(\phi)+\beta(\phi) \ge 1-\mathrm{TV}(P_0, P_1)$ for every test. The classical question is how small β can be made when α is held below a prescribed level.

Lemma 3.37 (Neyman–Pearson). *(i) Let $t \ge 0$ and let ϕ^* be a test with $\phi^* = 1$ on $\{p_1 > t p_0\}$ and $\phi^* = 0$ on $\{p_1 < t p_0\}$ (a* likelihood-ratio test*). Every test ϕ with $\alpha(\phi) \le \alpha(\phi^*)$ satisfies $\beta(\phi) \ge \beta(\phi^*)$.*

(ii) For every $\alpha \in (0,1)$ there is a likelihood-ratio test of the form $\phi^* = \mathbb{1}\{p_1 > t p_0\} + \gamma\, \mathbb{1}\{p_1 = t p_0\}$, with $t \ge 0$ and $\gamma \in [0,1]$, such that $\alpha(\phi^*) = \alpha$.

(iii) $\inf_\phi\{\alpha(\phi) + \beta(\phi)\} = 1 - \mathrm{TV}(P_0, P_1)$, attained by $\phi = \mathbb{1}\{p_1 > p_0\}$.

Proof. (i) By construction $(\phi^* - \phi)(p_1 - tp_0) \geq 0$ pointwise: where $p_1 > tp_0$, $\phi^* = 1 \geq \phi$; where $p_1 < tp_0$, $\phi^* = 0 \leq \phi$. Integrating with respect to ν,

$$\big(\beta(\phi) - \beta(\phi^*)\big) = \int (\phi^* - \phi)p_1 \, \mathrm{d}\nu \;\geq\; t \int (\phi^* - \phi)p_0 \, \mathrm{d}\nu = t\big(\alpha(\phi^*) - \alpha(\phi)\big) \;\geq\; 0.$$

(ii) Let $F = p_1/p_0$ on $\{p_0 > 0\}$, a random variable under P_0, and $G(s) = P_0(F > s)$, a non-increasing right-continuous function with $G(s) \to 0$ as $s \to \infty$. Let $t = \inf\{s \geq 0 : G(s) \leq \alpha\}$; by right continuity $G(t) \leq \alpha$, and $P_0(F \geq t) \geq \alpha$: for $t > 0$ because $P_0(F \geq t) = \lim_{s \uparrow t} G(s)$ and $G(s) > \alpha$ for $s < t$, and for $t = 0$ because $P_0(F \geq 0) = 1$. If $P_0(F = t) = 0$ take $\gamma = 0$, so that $\alpha(\phi^*) = G(t) = P_0(F \geq t) = \alpha$; otherwise $\gamma = (\alpha - G(t))/P_0(F = t) \in [0, 1]$ gives $\alpha(\phi^*) = G(t) + \gamma P_0(F = t) = \alpha$. (The set $\{p_0 = 0\}$ is P_0-null, so it does not affect $\alpha(\phi^*)$.) (iii) This is the last statement of Lemma 3.27 with $P = P_1$, $Q = P_0$. □

Remark 3.38. The lemma is from 1933 [NP33], and its proof is one line once one thinks of writing $(\phi^* - \phi)(p_1 - tp_0) \geq 0$. Its message is that *the likelihood ratio is the only statistic that matters* for testing a simple hypothesis against a simple alternative: every other test is dominated by one that thresholds p_1/p_0. For n i.i.d. observations the likelihood ratio is the product $\prod_i p_1(X_i)/p_0(X_i)$, and its logarithm is a sum of i.i.d. variables whose mean under P_0 is $-\mathrm{KL}(P_0\|P_1)$ and under P_1 is $\mathrm{KL}(P_1\|P_0)$. The law of large numbers and large deviations for that sum govern the performance of the optimal test, and this is how relative entropy enters.

For $n \geq 1$ and $\alpha \in (0, 1)$, consider n i.i.d. observations and define the smallest type II error achievable at level α,

$$\beta_n^*(\alpha) := \inf\Big\{ \int (1 - \phi) \, \mathrm{d}P_1^{\otimes n} : \; \phi : \Omega^n \to [0, 1] \text{ measurable}, \; \int \phi \, \mathrm{d}P_0^{\otimes n} \leq \alpha \Big\}.$$

Theorem 3.39 (Stein's lemma)**.** *Assume* $D := \mathrm{KL}(P_0\|P_1) < \infty$. *Then for every* $\alpha \in (0, 1)$,

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n^*(\alpha) = -\mathrm{KL}(P_0\|P_1).$$

Moreover, for every $n \geq 1$ *and every* $\alpha \in (0, 1)$,

$$\beta_n^*(\alpha) \;\geq\; \exp\Big(-\frac{nD + \log 2}{1 - \alpha}\Big). \tag{3.16}$$

Proof. Write $\mathbb{P}_j = P_j^{\otimes n}$, $\mathbb{E}_j$ for the corresponding expectations, and $X_1, \ldots, X_n$ for the coordinates. Since $D < \infty$, $P_0 \ll P_1$; let

$$\Lambda_n := \sum_{i=1}^{n} \log \frac{p_0(X_i)}{p_1(X_i)} \in [-\infty, +\infty],$$

with the conventions $\log \frac{0}{b} = -\infty$ for $b \geq 0$, $\log \frac{a}{0} = +\infty$ for $a > 0$, and $\Lambda_n := -\infty$ as soon as $p_0(X_i) = 0$ for some i (so that $\Lambda_n > -\infty$ implies $\prod_i p_0(X_i) > 0$). Under $\mathbb{P}_0$ the summands are almost surely finite, i.i.d., with integrable negative part (Lemma 3.13) and mean $D < \infty$, hence integrable: by the law of large numbers (Theorem 1.37), $\Lambda_n/n \to D$ in $\mathbb{P}_0$-probability. The change of measure between $\mathbb{P}_1$ and $\mathbb{P}_0$ reads: for every measurable $g : \Omega^n \to [0, \infty)$ vanishing on $\{\Lambda_n = -\infty\}$,

$$\mathbb{E}_1[g] = \mathbb{E}_0\big[g \, e^{-\Lambda_n}\big], \tag{3.17}$$

because on $\{\Lambda_n > -\infty\}$ one has $\prod_i p_0(X_i) > 0$ and $\prod_i p_1(X_i) = e^{-\Lambda_n} \prod_i p_0(X_i)$.

Achievability. Let $\varepsilon > 0$ and let $\phi_n := \mathbb{1}\{\Lambda_n < n(D-\varepsilon)\}$: reject H_0 when the likelihood ratio is not as large as it typically is under P_0. Its type I error $\mathbb{P}_0(\Lambda_n < n(D-\varepsilon))$ tends to 0, hence is at most α for n large enough. Its type II error is, by (3.17) applied to $g = 1-\phi_n = \mathbb{1}\{\Lambda_n \geq n(D-\varepsilon)\}$,

$$\mathbb{E}_1[1-\phi_n] = \mathbb{E}_0\big[e^{-\Lambda_n}\mathbb{1}\{\Lambda_n \geq n(D-\varepsilon)\}\big] \leq e^{-n(D-\varepsilon)}.$$

Hence $\limsup_n \frac{1}{n}\log\beta_n^*(\alpha) \leq -D+\varepsilon$ for every $\varepsilon > 0$.

Converse. Let ϕ be any test with $\mathbb{E}_0[\phi] \leq \alpha$, and $\varepsilon > 0$. Applying (3.17) to $g = (1-\phi)\mathbb{1}\{-\infty < \Lambda_n \leq n(D+\varepsilon)\}$,

$$\mathbb{E}_1[1-\phi] \;\geq\; \mathbb{E}_0\big[(1-\phi)e^{-\Lambda_n}\mathbb{1}\{\Lambda_n \leq n(D+\varepsilon)\}\big] \;\geq\; e^{-n(D+\varepsilon)}\Big(\mathbb{E}_0[1-\phi] - \mathbb{P}_0\big(\Lambda_n > n(D+\varepsilon)\big)\Big),$$

using $\mathbb{P}_0(\Lambda_n = -\infty) = 0$. The bracket is at least $1-\alpha-\mathbb{P}_0(\Lambda_n > n(D+\varepsilon)) \to 1-\alpha > 0$, so $\liminf_n \frac{1}{n}\log\beta_n^*(\alpha) \geq -D-\varepsilon$.

The non-asymptotic bound. For a test ϕ with $\mathbb{E}_0\phi \leq \alpha$, apply Corollary 3.24 to $Z = 1-\phi$, then Lemma 3.30 and Proposition 3.20:

$$nD = \mathrm{KL}(\mathbb{P}_0\|\mathbb{P}_1) \;\geq\; \mathrm{kl}\big(\mathbb{E}_0[1-\phi], \mathbb{E}_1[1-\phi]\big) \;\geq\; (1-\alpha)\log\frac{1}{\mathbb{E}_1[1-\phi]} - \log 2,$$

where we used that $x \mapsto x\log\frac{1}{\mu}$ is non-decreasing and $\mathbb{E}_0[1-\phi] \geq 1-\alpha$. Rearranging gives (3.16). □

Remark 3.40 (Stein's lemma is Cramér's theorem in disguise). The two halves of the proof are exactly the two halves of the large-deviation theory of Chapter 2. Under P_1, the variable $Z = \log\frac{p_0}{p_1}(X)$ has cumulant generating function $K(\lambda) = \log\int p_0^\lambda p_1^{1-\lambda}\,\mathrm{d}\nu$, finite on $[0,1]$, with $K(0^+) = \log P_1(p_0 > 0) \leq 0$, $K(1) = \log P_0(p_1 > 0) = 0$ and, for the tilt $\lambda = 1$, tilted law P_0 and tilted mean $K'(1) = \mathbb{E}_{P_0}[Z] = D$ *(formally; the derivative at the endpoint of the domain needs a justification we skip)*. The type II error of the optimal test is the probability, under P_1, that Λ_n/n reaches the level D typical of P_0; its exponent is $K^*(D) = 1\cdot D - K(1) = D$. Achievability is the Chernoff upper bound, and the converse is the change-of-measure lower bound – the same argument that proves Cramér's lower bound in Section 3.4. Note also the role of the two halves of the converse: the entropic bound (3.16) is non-asymptotic but gives the exponent only up to the factor $1/(1-\alpha)$ (a *weak converse*), while the law of large numbers gives the exact exponent for every fixed α (the *strong converse*). This dichotomy will reappear in Theorems 3.42 and 3.47.

Remark 3.41 (History, and the symmetric problem). The lemma is attributed to Charles Stein, who never published it; the first printed account is in Chernoff's 1956 survey [Che56]. It gives an operational meaning to $\mathrm{KL}(P_0\|P_1)$ *including its asymmetry*: if one insists on a fixed type I error, the type II error decays like $e^{-n\,\mathrm{KL}(P_0\|P_1)}$, and exchanging the roles of the hypotheses exchanges the arguments. If instead both errors are weighted equally, the optimal exponent of $\inf_\phi\{\alpha(\phi)+\beta(\phi)\}$ is the *Chernoff information* $-\log\min_{\lambda\in[0,1]}\int p_0^\lambda p_1^{1-\lambda}\,\mathrm{d}\nu$, which lies between 0 and $\min\{\mathrm{KL}(P_0\|P_1), \mathrm{KL}(P_1\|P_0)\}$ [Che52], [CT06, Sec. 11.9]; we do not prove this. The Bretagnolle–Huber inequality gives the cruder but non-asymptotic statement $\inf_\phi\{\alpha+\beta\} \geq \frac{1}{2}e^{-n\,\mathrm{KL}(P_0\|P_1)}$. Relative entropy itself was introduced by Kullback and Leibler [KL51] as the mean information per observation for discrimination in favour of P_0 against P_1, and developed in Kullback's book [Kul59]; its use in information theory goes back to Shannon's entropy, and Sanov's theorem [San57] identifies it as the rate function for large deviations of empirical measures. For a modern treatment of all these topics, see [CT06; PW25].

3.4 Lower bounds for deviations

We now prove that the Chernoff bound (3.1) is exact at the exponential scale, first asymptotically by the change-of-measure argument already sketched in the proof of Theorem 2.19, then non-asymptotically with the entropy inequalities of Section 3.3. Throughout, $\mathbb{P}_a$ denotes the law under which $X_1, X_2, \dots$ are i.i.d. $\mathrm{Ber}(a)$, and $\mathbb{E}_a$ the corresponding expectation.

3.4.1 Change of measure: the asymptotic lower bound

Theorem 3.42 (Large-deviation lower bound for Bernoulli variables)**.** *Let $\mu \in (0,1)$ and $x \in [\mu, 1)$. Then*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu\big(\bar{X}_n \geq x\big) \;\geq\; -\,\mathrm{kl}(x,\mu).$$

Consequently, for every $x \in (\mu, 1)$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu\big(\bar{X}_n \geq x\big) = -\,\mathrm{kl}(x,\mu), \tag{3.18}$$

and for $x = 1$ one has exactly $\frac{1}{n}\log \mathbb{P}_\mu(\bar{X}_n \geq 1) = \log \mu = -\,\mathrm{kl}(1,\mu)$.

Proof. Fix $a \in (x, 1)$ and $\varepsilon > 0$. For $n \geq 1$, the laws $\mathbb{P}_\mu$ and $\mathbb{P}_a$ of $(X_1, \dots, X_n)$ are both carried by $\{0,1\}^n$ and charge every point, and

$$\Lambda_n := \log \frac{\mathrm{d}\mathbb{P}_a}{\mathrm{d}\mathbb{P}_\mu}(X_1, \dots, X_n) = \sum_{i=1}^{n} \Big(X_i \log \frac{a}{\mu} + (1 - X_i) \log \frac{1-a}{1-\mu} \Big)$$

is a sum of i.i.d. bounded variables whose mean under $\mathbb{P}_a$ is $\mathrm{kl}(a,\mu)$. Changing the measure, and then restricting the expectation to the event where Λ_n is not much larger than its typical value,

$$\begin{aligned} \mathbb{P}_\mu\big(\bar{X}_n \geq x\big) &= \mathbb{E}_a\Big[e^{-\Lambda_n} \mathbb{1}\{\bar{X}_n \geq x\} \Big] \\ &\geq \mathbb{E}_a\Big[e^{-\Lambda_n} \mathbb{1}\{\bar{X}_n \geq x\}\, \mathbb{1}\big\{\Lambda_n \leq n\big(\mathrm{kl}(a,\mu) + \varepsilon\big)\big\} \Big] \\ &\geq e^{-n(\mathrm{kl}(a,\mu)+\varepsilon)} \Big(1 - \mathbb{P}_a\big(\bar{X}_n < x\big) - \mathbb{P}_a\Big(\frac{\Lambda_n}{n} > \mathrm{kl}(a,\mu) + \varepsilon \Big) \Big). \end{aligned}$$

By the law of large numbers under $\mathbb{P}_a$ (Theorem 1.37), $\bar{X}_n \to a > x$ and $\Lambda_n / n \to \mathrm{kl}(a,\mu)$ in probability, so both probabilities in the bracket tend to 0. Taking logarithms and dividing by n,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu\big(\bar{X}_n \geq x\big) \;\geq\; -\,\mathrm{kl}(a,\mu) - \varepsilon.$$

Let $\varepsilon \downarrow 0$ and then $a \downarrow x$, using the continuity of $\mathrm{kl}(\cdot,\mu)$ on $[0,1)$. The limit (3.18) follows by combining this with (3.1), and the case $x = 1$ is $\mathbb{P}_\mu(X_1 = \cdots = X_n = 1) = \mu^n$. □

Remark 3.43 (The proof in one picture). The proof follows Remark 2.16 to the letter. Under $\mathbb{P}_\mu$ the event $\{\bar{X}_n \geq x\}$ is rare; under $\mathbb{P}_a$, a slightly above x, it is typical. The price of passing from one measure to the other is the likelihood ratio $e^{-\Lambda_n}$, and on the typical event its logarithm is $-n\,\mathrm{kl}(a,\mu) + o(n)$. The tilted law $\mathrm{Ber}(a)$ is the *cheapest* way to make the event typical: by Proposition 2.17, $\mathrm{kl}(x,\mu)$ is the smallest relative entropy $\mathrm{KL}(Q\|\mathrm{Ber}(\mu))$ over the laws Q with mean at least x. Here the likelihood ratio is even an explicit function of $\bar{X}_n$, $\Lambda_n/n = \mathrm{kl}(a,\mu) + \lambda_a(\bar{X}_n - a)$ with $\lambda_a = \log \frac{a(1-\mu)}{\mu(1-a)} > 0$, so that a single application of the law of large numbers would suffice; we wrote the proof in the form that generalises.

Corollary 3.44 (Cramér's theorem for Bernoulli variables)**.** *Theorem 2.19 holds, upper and lower bounds included, when* $X \sim \mathrm{Ber}(\mu)$ *with* $\mu \in (0,1)$.

Proof. The upper bound was proved in Chapter 2. By Proposition 2.23(b), $K_X^*(x) = \mathrm{kl}(x,\mu)$ on $[0,1]$ and $+\infty$ outside, so the interior of $\{K_X^* < \infty\}$ is $(0,1)$, and we must show, with $S_n = X_1 + \cdots + X_n$ for i.i.d. copies X_i of X, that $\lim \frac{1}{n} \log \mathbb{P}(S_n \geq nx) = -\mathrm{kl}(x,\mu)$ for $x \in [\mu, 1)$. For $x > \mu$ this is (3.18); for $x = \mu$, the lower bound of Theorem 3.42 gives lim inf ≥ 0, and the upper bound lim sup ≤ 0 is trivial. □

The argument uses nothing specific to Bernoulli variables, except the existence of a tilted law with the right mean. We record the general version, which proves Theorem 2.19 in a much larger generality.

Proposition 3.45 (Cramér's lower bound by tilting)**.** *Let* $X, X_1, X_2, \ldots$ *be i.i.d. real random variables, not almost surely constant, with finite mean* $\mu = \mathbb{E}X$*, and let* $K = K_X$*. Assume that* $\lambda_+ := \sup \mathrm{dom}(K) > 0$*, so that* $(0, \lambda_+) \subset \mathrm{int}\,\mathrm{dom}(K)$*; let* $m = K'$ *on* $(0, \lambda_+)$ *and* $x_+ := \lim_{\lambda \uparrow \lambda_+} m(\lambda) \in (\mu, +\infty]$*. Then, for every* $x \in [\mu, x_+)$*,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big(S_n \geq nx\big) = -K^*(x).$$

If moreover $\lambda_+ = +\infty$*, then* $x_+ = \operatorname{ess\,sup} X$*, and Theorem 2.19 holds for* X *in full.*

Proof. The upper bound is (2.7). By Theorem 2.3, m is continuous and increasing on $(0, \lambda_+)$, strictly since X is not almost surely constant, and $m(\lambda) \to \mu$ as $\lambda \downarrow 0$: for $0 < \lambda \leq \lambda_0 < \lambda_1 < \lambda_+$, $e^{\lambda X} \leq 1 + e^{\lambda_0 X}$ and $|X|\, e^{\lambda X} \leq |X| + X^+ e^{\lambda_0 X} \leq |X| + \frac{e^{\lambda_1 X}}{e(\lambda_1 - \lambda_0)}$, so that $\mathbb{E}e^{\lambda X} \to 1$ and $\mathbb{E}[Xe^{\lambda X}] \to \mu$ by dominated convergence. Hence m is a bijection from $(0, \lambda_+)$ onto (μ, x_+); write $\lambda(\cdot)$ for its inverse. By Proposition 2.10(iv), $K^*(a) = \lambda(a)\, a - K(\lambda(a)) < \infty$ for $a \in (\mu, x_+)$, so the convex function K^* is finite on $[\mu, x_+)$, hence continuous on (μ, x_+), and also at μ from the right, since $0 \leq K^*(a) \leq \frac{a-\mu}{a'-\mu} K^*(a')$ for $\mu < a < a' < x_+$.

For the lower bound, let $a \in (x, x_+)$ and $\lambda = \lambda(a) \in (0, \lambda_+)$. Let $\mathbb{P}_\lambda$ be the law under which the X_i are i.i.d. with the tilted law of Definition 2.2; by (2.1), $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}_\lambda}(X_1, \ldots, X_n) = e^{-\lambda S_n + nK(\lambda)}$. For $\varepsilon > 0$, on the event $\{nx \leq S_n \leq n(a+\varepsilon)\}$ this density is at least $e^{-n((a+\varepsilon)\lambda - K(\lambda))} = e^{-n(K^*(a)+\lambda\varepsilon)}$. Hence

$$\mathbb{P}\big(S_n \geq nx\big) \;\geq\; e^{-n(K^*(a)+\lambda\varepsilon)}\, \mathbb{P}_\lambda\big(x \leq \bar{X}_n \leq a + \varepsilon\big).$$

Under $\mathbb{P}_\lambda$, X is integrable with mean a (Theorem 2.3, λ being interior), so the last probability tends to 1 by Theorem 1.37. Hence lim inf $\frac{1}{n} \log \mathbb{P}(S_n \geq nx) \geq -K^*(a) - \lambda\varepsilon$; let $\varepsilon \to 0$, then $a \downarrow x$, using the continuity of K^* on $[\mu, x_+)$.

Suppose now $\lambda_+ = +\infty$, and let $b < \operatorname{ess\,sup} X$; choose $b' \in (b, \operatorname{ess\,sup} X)$. For $\lambda > 0$, since $(b - X)^+ e^{\lambda X} \leq (b - X)^+ e^{\lambda b}$ and $\mathbb{E}e^{\lambda X} \geq e^{\lambda b'} \mathbb{P}(X \geq b')$,

$$b - m(\lambda) = \mathbb{E}_\lambda[b - X] \;\leq\; \mathbb{E}_\lambda\big[(b-X)^+\big] \;\leq\; \frac{\mathbb{E}[(b-X)^+]}{\mathbb{P}(X \geq b')}\, e^{-\lambda(b'-b)} \xrightarrow[\lambda\to\infty]{} 0,$$

where $\mathbb{E}[(b-X)^+]$ is finite since X is integrable; so $x_+ \geq b$; hence $x_+ \geq \operatorname{ess\,sup} X$, and the reverse inequality is clear. Finally $K^*(s) = +\infty$ for $s > \operatorname{ess\,sup} X$ (see the proof of Corollary 2.12), so the interior of $\{K^* < \infty\}$ lies below x_+, and the limit statement of Theorem 2.19 is the first part of the proposition. □

Remark 3.46 (What remains of Theorem 2.19). Proposition 3.45 proves Theorem 2.19 completely for every variable whose cumulant generating function is finite on $[0, \infty)$: bounded

variables, finitely-valued variables (in particular Bernoulli ones), Gaussian, Poisson and binomial laws, and every sub-Gaussian variable. It also proves it whenever K is *steep*, i.e. $m(\lambda) \to \sup\{K^* < \infty\}$ as $\lambda \uparrow \lambda_+$ – for instance for the exponential law, where $m(\lambda) = 1/(1-\lambda) \to \infty$. What is left are two degenerate situations: the case $\lambda_+ = 0$, i.e. $K = +\infty$ on $(0,\infty)$ (heavy right tail, where $K^* = 0$ on $[\mu,\infty)$ and the claim is that $\mathbb{P}(S_n \geq nx)$ decays subexponentially), and the non-steep case, where $\lambda_+ < \infty$, $x_+ < \infty$ and the deviations $x \in [x_+,\infty)$ cannot be made typical by any interior tilt – think of a density proportional to $e^{-y}(1+y)^{-3}$ on $[0,\infty)$. There one needs a truncation argument: tilt the law conditioned on $\{X \leq M\}$ and let $M \to \infty$. We do not reproduce it, and refer to [DZ98, Thm. 2.2.3], where the lower bound is proved for every real law; this residual case is the only part of Theorem 2.19 that this book admits.

3.4.2 The entropic way: a non-asymptotic lower bound

The change-of-measure proof is asymptotic: it uses the law of large numbers under $\mathbb{P}_a$ and gives no information for a fixed n. The inequalities of Section 3.3 give a version with explicit constants, in three lines, and without ever writing a likelihood ratio.

Theorem 3.47 (Non-asymptotic entropic lower bound)**.** *Let $\mu \in (0,1)$, $x \in [\mu,1)$ and $n \geq 1$. For every $\varepsilon \in (0, 1-x)$,*

$$\mathbb{P}_\mu\big(\bar{X}_n \geq x\big) \;\geq\; \exp\Big(-\frac{n\ \mathrm{kl}(x+\varepsilon,\mu)+\log 2}{1-e^{-n\,\mathrm{kl}(x,\,x+\varepsilon)}}\Big) \;\geq\; \exp\Big(-\frac{n\ \mathrm{kl}(x+\varepsilon,\mu)+\log 2}{1-e^{-2n\varepsilon^2}}\Big). \tag{3.19}$$

Proof. Let $a = x+\varepsilon \in (x,1)$ and $A = \{\bar{X}_n \geq x\}$, an event of $\{0,1\}^n$. By tensorisation (Proposition 3.20) and Proposition 3.14(b), $\mathrm{KL}(\mathbb{P}_a\|\mathbb{P}_\mu) = n\ \mathrm{kl}(a,\mu)$ for the laws of $(X_1,\dots,X_n)$, and the change-of-measure inequality (3.10) gives

$$\mathbb{P}_\mu(A) \;\geq\; \exp\Big(-\frac{n\ \mathrm{kl}(a,\mu)+\log 2}{\mathbb{P}_a(A)}\Big).$$

Under $\mathbb{P}_a$ the event A is typical: by the left-deviation bound of Theorem 3.1(i), applied at the level $x < a$, $\mathbb{P}_a(A^c) \leq \mathbb{P}_a(\bar{X}_n \leq x) \leq e^{-n\,\mathrm{kl}(x,a)}$, and $\mathrm{kl}(x,a) \geq 2\varepsilon^2$ by (2.5). Since the right-hand side above is non-decreasing in $\mathbb{P}_a(A)$, both inequalities of (3.19) follow. □

Unpacked, the proof is the chain of slide-rule steps

$$n\ \mathrm{kl}(x+\varepsilon,\mu) = \mathrm{KL}\big(\mathbb{P}_{x+\varepsilon}\|\mathbb{P}_\mu\big) \;\geq\; \mathrm{kl}\big(\mathbb{P}_{x+\varepsilon}(A),\mathbb{P}_\mu(A)\big) \;\geq\; \mathbb{P}_{x+\varepsilon}(A)\log\frac{1}{\mathbb{P}_\mu(A)} - \log 2\,:$$

tensorisation, data processing, and Lemma 3.30. Nothing in it refers to the binomial law except the computation of KL and the upper bound on $\mathbb{P}_{x+\varepsilon}(A^c)$; this is what makes the method portable.

Corollary 3.48 (Explicit second-order term)**.** *Let $\mu \in (0,1)$, $x \in [\mu,1)$, and let $n \geq 2$ be such that $\varepsilon_n := \sqrt{\log n/n} < 1-x$. Then, with $\lambda_y := \log\frac{y(1-\mu)}{\mu(1-y)}$,*

$$\log\frac{1}{\mathbb{P}_\mu(\bar{X}_n \geq x)} \;\leq\; \Big(1+\frac{2}{n^2}\Big)\Big(n\ \mathrm{kl}(x,\mu) + \lambda_{x+\varepsilon_n}\sqrt{n\log n} + \log 2\Big).$$

In particular $\frac{1}{n}\log\mathbb{P}_\mu(\bar{X}_n \geq x) \geq -\mathrm{kl}(x,\mu) - O\big(\sqrt{\log n/n}\big)$, the constant in the O depending only on x and μ.

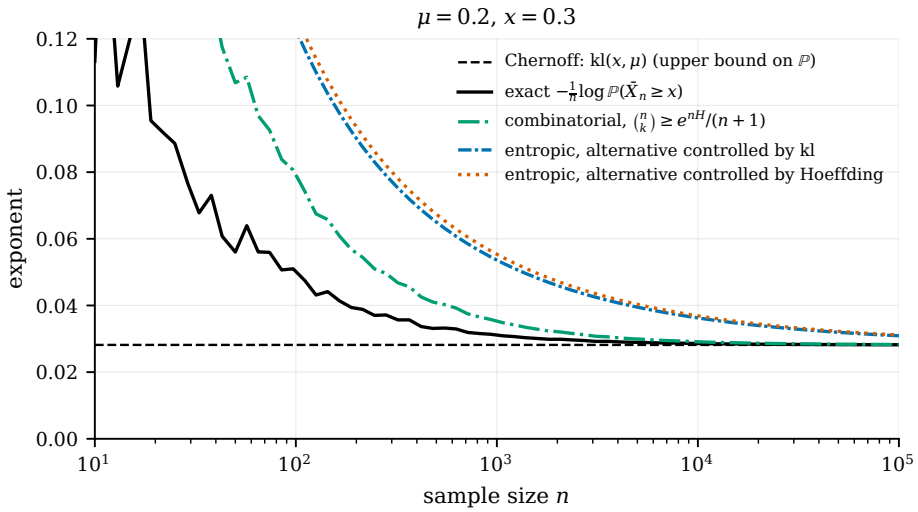


Figure 3.2: Per-sample exponents $-\frac{1}{n}\log(\cdot)$ for $\mathbb{P}_\mu(\bar{X}_n \geq x)$, $\mu = 0.2$, $x = 0.3$: the exact value (computed from the binomial law), the Chernoff upper bound on the probability (a lower bound on the exponent, the horizontal line $\mathrm{kl}(x, \mu)$), and three lower bounds on the probability (upper bounds on the exponent): the combinatorial bound of Remark 2.51 and the entropic bound (3.19), optimised over ε, with the probability of the alternative controlled by the exact Chernoff bound or by Hoeffding's. All converge to $\mathrm{kl}(x, \mu)$, the entropic ones at the slow rate $\sqrt{\log n/n}$ of Corollary 3.48. Generated by `figures/fig_ld_lower.py`.

Proof. Take $\varepsilon = \varepsilon_n$ in the second bound of (3.19), so that $e^{-2n\varepsilon_n^2} = n^{-2} \leq \frac{1}{4}$ and $\frac{1}{1-n^{-2}} \leq 1 + \frac{2}{n^2}$ (as $\frac{1}{1-u} \leq 1 + 2u$ for $u \leq \frac{1}{2}$). By convexity of $\mathrm{kl}(\cdot, \mu)$, whose derivative at y is λ_y, $\mathrm{kl}(x + \varepsilon, \mu) \leq \mathrm{kl}(x, \mu) + \varepsilon\lambda_{x+\varepsilon}$. □

Remark 3.49 (Upper and lower bounds side by side). Together with (3.1), Corollary 3.48 gives, for $x \in (\mu, 1)$,

$$e^{-n\,\mathrm{kl}(x,\mu) - O(\sqrt{n \log n})} \;\leq\; \mathbb{P}_\mu\big(\bar{X}_n \geq x\big) \;\leq\; e^{-n\,\mathrm{kl}(x,\mu)}\,;$$

the exponent is matched exactly, and the discrepancy is confined to a lower-order term. The choice $\varepsilon \asymp \sqrt{\log n/n}$ balances the two losses in (3.19): a larger ε makes $\mathbb{P}_{x+\varepsilon}(A)$ closer to 1 but costs $n\varepsilon\lambda_{x+\varepsilon}$ in the numerator, a smaller one makes the denominator small. Optimising numerically over ε does not change the order $\sqrt{n \log n}$ of the correction, which is the price of the method; Figure 3.2 shows how slowly the entropic exponent converges. For binomial tails the truth is much closer to the upper bound: the combinatorial argument of Remark 2.51, applied with $k = \lceil nx \rceil$ and $x' = k/n \in [x, x + \frac{1}{n}]$, gives $\mathbb{P}_\mu(\bar{X}_n \geq x) \geq e^{-n\,\mathrm{kl}(x',\mu)}/(n+1) \geq e^{-n\,\mathrm{kl}(x,\mu) - \lambda_{x+1/n}}/(n+1)$, a correction of order $\log n$ only, and the lattice version of the Bahadur–Rao asymptotics [BR60] (Remark 2.21 states the non-lattice one) shows that the truth is $e^{-n\,\mathrm{kl}(x,\mu)}$ times a factor of order $n^{-1/2}$, whose constant oscillates with the fractional part of nx. The entropic bound is therefore not the sharpest one for binomial tails. Its merit lies elsewhere: it uses neither the explicit form of the law nor any combinatorics, only a computation of KL and an *upper* bound on a deviation under the alternative. That is exactly the situation of the minimax lower bounds of Section 3.5 and of the learning lower bounds of Chapter 4, where the event A is defined by an arbitrary estimator or algorithm and no explicit computation is possible.

Remark 3.50 (Beyond Bernoulli variables). The same three steps give, for any real law P with

a tilted law P_λ of mean $a > x$ as in Proposition 3.45, the bound

$$\mathbb{P}\big(\bar{X}_n \geq x\big) \;\geq\; \exp\Big(-\frac{nK^*(a)+\log 2}{1-\mathbb{P}_\lambda(\bar{X}_n < x)}\Big),$$

since $\mathrm{KL}(P_\lambda^{\otimes n}\|P^{\otimes n}) = nK^*(a)$ by (2.8) and tensorisation. Any upper bound on the left deviation $\mathbb{P}_\lambda(\bar{X}_n < x)$ of the tilted sample then yields an explicit non-asymptotic lower bound, and letting $n \to \infty$ and then $a \downarrow x$ re-proves Proposition 3.45. The quality of the second-order term depends on that upper bound. With $a = x + \varepsilon$, Chebyshev's inequality $\mathbb{P}_\lambda(\bar{X}_n < x) \leq \varsigma^2(\lambda)/(n\varepsilon^2)$ costs about $n\lambda\varepsilon + K^*(x)\varsigma^2(\lambda)/\varepsilon^2$ in the exponent, and the best choice $\varepsilon \asymp n^{-1/3}$ only gives $e^{-nK^*(x)-O(n^{2/3})}$. But an exponential bound is always available: λ is interior to $\mathrm{dom}(K)$, so the cumulant generating function $s \mapsto K(\lambda+s)-K(\lambda)$ of the tilted law is finite near 0, and the Chernoff bound under P_λ, with $\varepsilon \asymp \sqrt{\log n/n}$ as in Corollary 3.48, gives $e^{-nK^*(x)-O(\sqrt{n\log n})}$. More generally still, (3.10) holds for *every* event and every pair of laws, which is what we use next.

3.4.3 An alternative route through Bretagnolle–Huber

The Bretagnolle–Huber inequality also produces a lower bound on $\mathbb{P}_\mu(\bar{X}_n \geq x)$. With $a = x+\varepsilon$ and $A = \{\bar{X}_n \geq x\}$, Proposition 3.32 and tensorisation give $\mathbb{P}_\mu(A) + \mathbb{P}_a(A^c) \geq \frac{1}{2}e^{-n\,\mathrm{kl}(a,\mu)}$, hence

$$\mathbb{P}_\mu\big(\bar{X}_n \geq x\big) \;\geq\; \frac{1}{2}e^{-n\,\mathrm{kl}(x+\varepsilon,\mu)} - \mathbb{P}_{x+\varepsilon}\big(\bar{X}_n < x\big). \tag{3.20}$$

This looks attractive – there is no division by $\mathbb{P}_a(A)$ – but it is weak. If the second term is controlled by Hoeffding's inequality, the bound is *always vacuous*: $\mathrm{kl}(x+\varepsilon,\mu) \geq 2(x+\varepsilon-\mu)^2 \geq 2\varepsilon^2$ by Pinsker, so that $\frac{1}{2}e^{-n\,\mathrm{kl}(x+\varepsilon,\mu)} < e^{-2n\varepsilon^2}$. If it is controlled by the exact Chernoff bound $e^{-n\,\mathrm{kl}(x,x+\varepsilon)}$, the right-hand side of (3.20) is positive for large n only if $\mathrm{kl}(x,x+\varepsilon) > \mathrm{kl}(x+\varepsilon,\mu)$; as $\varepsilon \to 0$ the left-hand side tends to 0 and the right-hand side to $\mathrm{kl}(x,\mu) > 0$, so this forces $\varepsilon \geq \varepsilon_0$ for some $\varepsilon_0 > 0$ independent of n, and the exponent obtained, $\mathrm{kl}(x+\varepsilon_0,\mu)$, is strictly worse than $\mathrm{kl}(x,\mu)$. The lesson is general. (3.20) lower-bounds the *sum* of the two error probabilities, and it is the right tool when both are small and comparable – in testing and minimax problems, where one wants $\max\{\mathbb{P}_\mu(A),\mathbb{P}_a(A^c)\} \geq \frac{1}{4}e^{-n\,\mathrm{kl}(a,\mu)}$. For a single tail, where one of the two probabilities is close to 1, Lemma 3.30 is the better conversion.

3.4.4 What each level of precision costs

It is instructive to line up the lower bounds on deviation probabilities met so far, from the crudest to the sharpest, together with the tool each one requires. For i.i.d. variables with mean μ and variance σ^2:

statement	tool	reference
$\mathbb{P}(\lvert\bar{X}_n-\mu\rvert>\gamma\sigma/\sqrt{n})\geq\frac{(1-\gamma^2)^2}{3+M_4/(n\sigma^4)}$ ($\gamma\in[0,1]$, $M_4=\mathbb{E}[(X-\mu)^4]$)	second and fourth moments (Paley–Zygmund)	Remark 3.51
$\mathbb{P}(\bar{X}_n-\mu>\gamma\sigma/\sqrt{n})\to\mathbb{P}(G>\gamma)$ ($G\sim\mathcal{N}(0,1)$)	central limit theorem	Theorem 1.38
$\frac{1}{n}\log\mathbb{P}(\bar{X}_n\geq x)\to -K^*(x)$	change of measure + LLN	Theorem 3.42 and Proposition 3.45
$\mathbb{P}(\bar{X}_n\geq x)\geq e^{-nK^*(x)-O(\sqrt{n\log n})}$	data processing + an upper bound under the alternative	Theorem 3.47 and Remark 3.50
$\mathbb{P}(\bar{X}_n\geq x)\geq e^{-n\,\mathrm{kl}(x,\mu)}/(n+1)$ (binomial, $nx\in\mathbb{N}$)	counting (method of types)	Remark 2.51
$\mathbb{P}(\bar{X}_n\geq x)\geq\frac{\text{const}}{\sqrt{n}}e^{-n\,\mathrm{kl}(x,\mu)}$ (binomial, constant depending on x,μ)	Stirling's formula	Exercise 2.13
$\mathbb{P}(\bar{X}_n\geq x)\sim\frac{e^{-nK^*(x)}}{\lambda_x\varsigma(\lambda_x)\sqrt{2\pi n}}$ (non-lattice laws)	tilting + local CLT (Bahadur–Rao)	Remark 2.21

The first two lines are *anti-concentration* at the scale of the standard deviation, the next ones are lower bounds at the scale of large deviations. The two scales require different tools: anti-concentration is a matter of moments, as in Section 1.7.2; large deviations are a matter of changing the measure, i.e. of relative entropy. Moving down the table, each line buys a better subexponential factor with more information about the law. (When $nx\notin\mathbb{N}$, the counting bound holds with $x'=\lceil nx\rceil/n\in[x,x+\frac{1}{n}]$ in place of x, hence with the extra factor $e^{-\lambda_{x+1/n}}$ of Remark 3.49; the same factor is absorbed into the constant of the next line, since Exercise 2.13 assumes $nx\in\mathbb{N}$.)

Remark 3.51 (Anti-concentration at the $\sqrt{n}$ scale). Let $Z=(S_n-n\mu)^2$ with $\mathbb{E}[(X-\mu)^4]=M_4<\infty$. Then $\mathbb{E}Z=n\sigma^2$ and, expanding the fourth power and using independence and centring, $\mathbb{E}Z^2=nM_4+3n(n-1)\sigma^4\leq nM_4+3n^2\sigma^4$. The Paley–Zygmund inequality (Proposition 1.53) applied to Z at level γ^2 gives, for $\gamma\in[0,1]$,

$$\mathbb{P}\Big(\lvert\bar{X}_n-\mu\rvert>\frac{\gamma\sigma}{\sqrt{n}}\Big)\;\geq\;\frac{(1-\gamma^2)^2}{3+M_4/(n\sigma^4)},$$

a non-asymptotic, distribution-dependent version of what the central limit theorem says: fluctuations of order $\sigma/\sqrt{n}$ occur with probability bounded below. For Bernoulli variables, $M_4/\sigma^4\leq 1/(\mu(1-\mu))$. This is the kind of lower bound that shows that Hoeffding's and Bernstein's inequalities have the right scale $\sqrt{\log(1/\delta)/n}$ *up to the logarithm*, and statements of this type are what agnostic learning lower bounds require at the scale $n^{-1/2}$; at the exponential scale, the change of measure takes over.

3.5 The minimax toolkit

We now turn from one law to a family of laws. The question is no longer how rare an event is under a given distribution, but how well *any* procedure can perform *uniformly* over a class of distributions. The answer is organised around one idea: *estimation is at least as hard as testing*, and testing is hard when the hypotheses have close laws.

3.5.1 Minimax risk and the reduction to testing

A *statistical model* is a family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ of probability measures on a measurable space $(\Omega,\mathcal{F})$, the space of observations; typically $\Omega = \mathcal{X}^n$ and $\mathbb{P}_\theta = P_\theta^{\otimes n}$ is the law of an n-sample. An *estimator* is a measurable map $\hat\theta : \Omega \to \Theta$; a *randomised estimator* is a Markov kernel from Ω to Θ, or equivalently a measurable function $\hat\theta(\omega, U)$ of the data and of an independent uniform variable U. Given a measurable *loss* $\ell : \Theta\times\Theta \to [0,\infty)$, the *minimax risk* is

$$\mathcal{R}^* := \inf_{\hat\theta}\sup_{\theta\in\Theta} \mathbb{E}_\theta\big[\ell(\hat\theta,\theta)\big], \tag{3.21}$$

the infimum being over all (possibly randomised) estimators. Any particular estimator gives an upper bound on $\mathcal{R}^*$; a lower bound must hold for all of them, and it is in this sense that it certifies optimality.

The typical loss is $\ell(t,\theta) = w(\rho(t,\theta))$, where ρ is a *semi-distance* on Θ (symmetric, satisfying the triangle inequality, possibly vanishing off the diagonal) and $w:[0,\infty)\to[0,\infty)$ is non-decreasing – for instance $w(u)=u^2$, the quadratic loss.

As in Section 3.3.6, the letter ϕ denotes a *test*; from now on a test may decide between finitely many hypotheses, and a randomised test is a measurable function $\phi(\omega,U)$ of the observation and of an independent uniform variable U, exactly as a randomised estimator.

Lemma 3.52 (From estimation to testing)**.** *Let $\theta_1,\dots,\theta_N\in\Theta$ satisfy $\rho(\theta_j,\theta_k)\ge 2s$ for $j\neq k$, for some $s>0$. For every estimator $\hat\theta$, define the test $\phi := \arg\min_j \rho(\hat\theta,\theta_j)$ (ties broken by taking the smallest index). Then $\{\phi\neq j\}\subset\{\rho(\hat\theta,\theta_j)\ge s\}$ for every j, and consequently*

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\big[w\big(\rho(\hat\theta,\theta)\big)\big] \;\ge\; w(s)\max_j \mathbb{P}_{\theta_j}\big(\rho(\hat\theta,\theta_j)\ge s\big) \;\ge\; w(s)\inf_{\phi'}\max_j \mathbb{P}_{\theta_j}(\phi'\neq j),$$

the last infimum being over all (possibly randomised) tests $\phi':\Omega\to\{1,\dots,N\}$.

Proof. If $\rho(\hat\theta,\theta_j)<s$ then, for $k\neq j$, the triangle inequality gives $\rho(\hat\theta,\theta_k)\ge\rho(\theta_j,\theta_k)-\rho(\hat\theta,\theta_j)>2s-s=s>\rho(\hat\theta,\theta_j)$, so $\phi=j$. The first inequality is Markov's inequality in the form $w(\rho)\ge w(s)\mathbb{1}\{\rho\ge s\}$, and the second holds because ϕ is a test (randomised if $\hat\theta$ is). The measurability of ϕ follows from that of the finitely many maps $\omega\mapsto\rho(\hat\theta(\omega),\theta_j)$, which we assume – it holds as soon as ρ is measurable on $\Theta\times\Theta$. □

The problem is thus reduced to a lower bound on the error of tests between finitely many hypotheses that are $2s$*-separated* in the parameter space. The art lies in choosing them: far enough apart that the separation s is large, close enough that their laws cannot be distinguished. The three methods below differ in the number of hypotheses and in the way their errors are aggregated.

3.5.2 Le Cam's two-point method

Lemma 3.53 (Le Cam)**.** *Let P_0,P_1 be probability measures on $(\Omega,\mathcal{F})$. For every (possibly randomised) test $\phi:\Omega\to\{0,1\}$,*

$$\begin{aligned}\max_{j\in\{0,1\}} P_j(\phi\neq j) \;&\ge\; \frac12\big(P_0(\phi\neq 0)+P_1(\phi\neq 1)\big) \;\ge\; \frac{1-\mathrm{TV}(P_0,P_1)}{2}\\ &\ge\; \max\Big\{\frac12-\sqrt{\frac{\mathrm{KL}(P_0\|P_1)}{8}},\ \frac14 e^{-\mathrm{KL}(P_0\|P_1)}\Big\}.\end{aligned} \tag{3.22}$$

Consequently, if $\theta_0, \theta_1 \in \Theta$ satisfy $\rho(\theta_0, \theta_1) \geq 2s$, then for every estimator,

$$\max_{j \in \{0,1\}} \mathbb{E}_{\theta_j}\big[w\big(\rho(\hat{\theta}, \theta_j)\big)\big] \;\geq\; w(s)\,\frac{1 - \mathrm{TV}(\mathbb{P}_{\theta_0}, \mathbb{P}_{\theta_1})}{2}.$$

Proof. A randomised test is described by $g(\omega) = \mathbb{P}(\phi = 1 \mid \omega) \in [0,1]$, and $P_0(\phi \neq 0) + P_1(\phi \neq 1) = \int g \,\mathrm{d}P_0 + \int (1-g)\,\mathrm{d}P_1 \geq 1 - \mathrm{TV}(P_0, P_1)$ by Lemma 3.27. The last inequality of (3.22) combines Pinsker's inequality (Proposition 3.28) and the Bretagnolle–Huber inequality (Proposition 3.32). The estimation statement is Lemma 3.52 with $N = 2$. □

Remark 3.54 (The two-point method in practice). For an n-sample, tensorisation gives $\mathrm{KL}(P_{\theta_0}^{\otimes n} \| P_{\theta_1}^{\otimes n}) = n\,\mathrm{KL}(P_{\theta_0} \| P_{\theta_1})$, and the recipe is: *choose θ_1 at distance $2s_n$ from θ_0, with s_n as large as possible subject to $n\,\mathrm{KL}(P_{\theta_0} \| P_{\theta_1}) \leq 1$.* The lower bound is then of order $w(s_n)$. In regular parametric models KL is locally quadratic (Remark 3.15), so $s_n \asymp n^{-1/2}$ and the quadratic risk is $\gtrsim 1/n$: the parametric rate. The method is due to Le Cam [Le 73]; the formulation with the total variation of the product laws and its conversion through KL are classical [Tsy09, Chap. 2]. Its limitation is visible in its structure: two hypotheses can only reveal the difficulty of a *one-dimensional* sub-problem. In dimension d, the rate d/n requires many hypotheses, and this is what Fano's and Assouad's methods provide.

Proposition 3.55 (Estimating a Bernoulli mean)**.** *Let $X_1, \dots, X_n$ be i.i.d.* $\mathrm{Ber}(\mu)$, $\mu \in [0,1]$, *and consider the quadratic loss. Then*

$$\frac{1}{220\,n} \;\leq\; \inf_{\hat{\mu}} \sup_{\mu \in [0,1]} \mathbb{E}_{\mu}\big[(\hat{\mu} - \mu)^2\big] \;\leq\; \frac{1}{4n}.$$

Proof. Upper bound. $\mathbb{E}_{\mu}[(\bar{X}_n - \mu)^2] = \mu(1-\mu)/n \leq 1/(4n)$.

Lower bound. Let $r = \frac{1}{4\sqrt{n}}$, $\mu_0 = \frac{1}{2}$ and $\mu_1 = \frac{1}{2} + r \leq \frac{3}{4}$, so that $\big|\mu_0 - \mu_1\big| = 2s$ with $s = r/2$. By tensorisation and (3.12),

$$\mathrm{KL}\big(\mathrm{Ber}(\mu_0)^{\otimes n} \| \mathrm{Ber}(\mu_1)^{\otimes n}\big) = n\;\mathrm{kl}\Big(\frac{1}{2}, \frac{1}{2} + r\Big) \leq \frac{nr^2}{\frac{1}{4} - r^2} = \frac{1/16}{\frac{1}{4} - \frac{1}{16n}} \leq \frac{1}{3}.$$

By Pinsker's inequality the total variation between the two product laws is at most $\sqrt{1/6}$, and Lemma 3.53 with $w(u) = u^2$ gives, for every estimator,

$$\max_{j} \mathbb{E}_{\mu_j}\big[(\hat{\mu} - \mu_j)^2\big] \;\geq\; \frac{r^2}{4} \cdot \frac{1 - \sqrt{1/6}}{2} = \frac{1 - \sqrt{1/6}}{128\,n} \;\geq\; \frac{1}{220\,n}.$$ □

Remark 3.56 (Constants, and the local picture). The constant $\frac{1}{220}$ is far from the truth: the exact minimax risk is $\frac{1}{4(1+\sqrt{n})^2}$, attained by the shrunken estimator $(S_n + \sqrt{n}/2)/(n + \sqrt{n})$, whose risk does not depend on μ – a classical result of Hodges and Lehmann, see [LC98, Chap. 5]. Two-point arguments are designed to capture the *rate* and the dependence on the parameters, not the constant. They also capture the local behaviour: the same computation with $\mu_0 = \mu$ and $\mu_1 = \mu + \frac{1}{4}\sqrt{\mu(1-\mu)/n}$ shows that no estimator can achieve a risk $o(\mu(1-\mu)/n)$ uniformly on a neighbourhood of μ (Exercise 3.10), so that the variance $\mu(1-\mu)/n$ of the empirical mean is the correct benchmark at every point. This is the non-asymptotic shadow of the Hájek–Le Cam local asymptotic minimax theorem [Vaa98, Chap. 8].

Minimax bounds in expectation are not the only ones of interest. In the regime of high confidence, the Bretagnolle–Huber form of Lemma 3.53 shows that the $\log(1/\delta)$ of the upper bounds of Chapters 1 and 2 is necessary. We illustrate this on the rare-event problem of Section 2.7.3.

Proposition 3.57 (Estimating a small probability: a lower bound). *Let $\mu \in (0, \frac{1}{6}]$, $\varepsilon \in (0, \frac{1}{2}]$ and $\delta \in (0, \frac{1}{4})$, let $\mu_0 = \mu$ and $\mu_1 = \mu\frac{1+\varepsilon}{1-\varepsilon}$, and let $\hat{\mu}$ be any estimator based on $X_1, \dots, X_n$ i.i.d.* $\mathrm{Ber}(\mu_j)$ *such that*

$$\mathbb{P}_{\mu_j}\big(|\hat{\mu} - \mu_j| \geq \varepsilon\mu_j\big) \leq \delta \qquad \text{for both } j = 0 \text{ and } j = 1.$$

Then $n \geq \dfrac{3}{32} \cdot \dfrac{\log(1/(4\delta))}{\varepsilon^2\mu}$.

Proof. Here $\mu_1 \leq 3\mu \leq \frac{1}{2}$. The intervals $(\mu_0(1-\varepsilon), \mu_0(1+\varepsilon))$ and $(\mu_1(1-\varepsilon), \mu_1(1+\varepsilon))$ are disjoint, since $\mu_1(1-\varepsilon) = \mu_0(1+\varepsilon)$. Hence the event $A = \{|\hat{\mu} - \mu_0| < \varepsilon\mu_0\}$ satisfies $\mathbb{P}_{\mu_0}(A^c) \leq \delta$ and $\mathbb{P}_{\mu_1}(A) \leq \mathbb{P}_{\mu_1}(|\hat{\mu} - \mu_1| \geq \varepsilon\mu_1) \leq \delta$. By Proposition 3.32 and tensorisation, $2\delta \geq \mathbb{P}_{\mu_0}(A^c) + \mathbb{P}_{\mu_1}(A) \geq \frac{1}{2}e^{-n\,\mathrm{kl}(\mu_0,\mu_1)}$, that is $n\,\mathrm{kl}(\mu_0, \mu_1) \geq \log\frac{1}{4\delta}$. Finally, by (3.12), using $\mu_1(1-\mu_1) \geq \mu_1/2 = \mu(1+\varepsilon)/(2(1-\varepsilon))$,

$$\mathrm{kl}(\mu_0, \mu_1) \leq \frac{(\mu_1 - \mu_0)^2}{\mu_1(1-\mu_1)} \leq \mu^2\Big(\frac{2\varepsilon}{1-\varepsilon}\Big)^2 \cdot \frac{2(1-\varepsilon)}{\mu(1+\varepsilon)} = \frac{8\mu\varepsilon^2}{1-\varepsilon^2} \leq \frac{32}{3}\,\mu\varepsilon^2. \qquad \square$$

Remark 3.58. Compared with Proposition 2.72, which shows that $n \geq \frac{8}{3}\log(2/\delta)/(\varepsilon^2\mu)$ samples suffice, the sample complexity of estimating a small probability with relative accuracy ε and confidence $1-\delta$ is therefore $\asymp \log(1/\delta)/(\varepsilon^2\mu)$, up to a numerical factor: Bernstein's inequality is rate-optimal in all three parameters, and Hoeffding's bound $\log(2/\delta)/(2\varepsilon^2\mu^2)$ is off by the factor $1/\mu$. The dependence on δ could not have been obtained with Pinsker's inequality, which only yields constant-probability statements; it is the exponential form $\frac{1}{2}e^{-\mathrm{KL}}$ of the Bretagnolle–Huber inequality that converts the linear growth of KL in n into the logarithmic dependence on δ.

3.5.3 Adaptive sampling: comparing two treatments

Section 1.10.3 showed that $n \geq 2\log(2/\delta)/\varepsilon^2$ subjects per arm suffice to identify the better of two treatments whose means differ by at least ε, and claimed that the order ε^{-2} is a feature of the problem. We can now prove it, and even for experimenters who allocate the subjects *adaptively*, choosing each treatment in view of the previous outcomes. The observations are then no longer independent, and the chain rule replaces tensorisation.

Consider Bernoulli outcomes with means $\theta = (\mu_A, \mu_B) \in \Theta = (0,1)^2$. An *adaptive design* with budget T draws U uniform on $[0,1]$, then, for $t = 1, \dots, T$, chooses a treatment $I_t = \pi_t(U, Y_1, \dots, Y_{t-1}) \in \{A, B\}$ and observes Y_t, which given the past has law $\mathrm{Ber}(\mu_{I_t})$; finally it outputs $\phi = \phi(U, Y_1, \dots, Y_T) \in \{A, B\}$. The maps π_t and ϕ are measurable and arbitrary. We write $\mathbb{P}_\theta$ for the law of $(U, Y_1, \dots, Y_T)$ and $N_a = \sum_{t \leq T} \mathbb{1}\{I_t = a\}$ for the number of subjects allocated to a.

Lemma 3.59 (Divergence decomposition). *For two instances $\theta = (\mu_A, \mu_B)$ and $\theta' = (\mu'_A, \mu'_B)$ in* $(0,1)^2$,

$$\mathrm{KL}\big(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}\big) = \mathbb{E}_\theta[N_A]\ \mathrm{kl}(\mu_A, \mu'_A) + \mathbb{E}_\theta[N_B]\ \mathrm{kl}(\mu_B, \mu'_B).$$

Proof. With respect to the product of Lebesgue measure on $[0,1]$ and counting measure on $\{0,1\}^T$, $\mathbb{P}_\theta$ has density $\prod_{t \leq T} \mu_{I_t}^{y_t}(1-\mu_{I_t})^{1-y_t}$, where I_t is the function of $(u, y_1, \dots, y_{t-1})$ given by π_t; similarly for θ'. Both densities are positive, so by Lemma 3.13 $\mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) = \mathbb{E}_\theta\big[\sum_t Z_t\big]$ with $Z_t = Y_t \log\frac{\mu_{I_t}}{\mu'_{I_t}} + (1-Y_t)\log\frac{1-\mu_{I_t}}{1-\mu'_{I_t}}$, a bounded variable. Conditionally on $\mathcal{F}_{t-1} = \sigma(U, Y_1, \dots, Y_{t-1})$, I_t is fixed and Y_t has law $\mathrm{Ber}(\mu_{I_t})$, so $\mathbb{E}_\theta[Z_t \mid \mathcal{F}_{t-1}] = \mathrm{kl}(\mu_{I_t}, \mu'_{I_t})$. Taking expectations and summing over t, $\mathbb{E}_\theta\big[\sum_t \mathrm{kl}(\mu_{I_t}, \mu'_{I_t})\big] = \sum_a \mathbb{E}_\theta[N_a]\,\mathrm{kl}(\mu_a, \mu'_a)$. $\square$

Proposition 3.60 (Comparing two treatments: a lower bound). *Let $\varepsilon \in (0, \frac{1}{4}]$ and $\delta \in (0, \frac{1}{4})$, and let $\theta = (\frac{1+\varepsilon}{2}, \frac{1-\varepsilon}{2})$ and $\theta' = (\frac{1-\varepsilon}{2}, \frac{1+\varepsilon}{2})$, two instances in which the means differ by ε and the better treatment is A, respectively B. If an adaptive design with budget T satisfies $\mathbb{P}_\theta(\phi \neq A) \le \delta$ and $\mathbb{P}_{\theta'}(\phi \neq B) \le \delta$, then*

$$T \;\ge\; \frac{1-\varepsilon^2}{2\varepsilon^2} \log \frac{1}{4\delta} \;\ge\; \frac{15}{32} \cdot \frac{\log(1/(4\delta))}{\varepsilon^2}.$$

Proof. By Lemma 3.59, and since the two arms play symmetric roles, $\mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) = (\mathbb{E}_\theta N_A + \mathbb{E}_\theta N_B)\ \mathrm{kl}(\frac{1+\varepsilon}{2}, \frac{1-\varepsilon}{2}) = T\varepsilon \log \frac{1+\varepsilon}{1-\varepsilon}$. Since $\log \frac{1+\varepsilon}{1-\varepsilon} = 2\sum_{k\ge 0} \frac{\varepsilon^{2k+1}}{2k+1} \le \frac{2\varepsilon}{1-\varepsilon^2}$, this is at most $2T\varepsilon^2/(1-\varepsilon^2)$. The Bretagnolle–Huber inequality (Proposition 3.32) applied to the event $\{\phi \neq A\}$ gives $2\delta \ge \mathbb{P}_\theta(\phi \neq A) + \mathbb{P}_{\theta'}(\phi = A) \ge \frac{1}{2} e^{-\mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}_{\theta'})}$, whence $\mathrm{KL}(\mathbb{P}_\theta \| \mathbb{P}_{\theta'}) \ge \log \frac{1}{4\delta}$; and $1 - \varepsilon^2 \ge \frac{15}{16}$. □

Remark 3.61. The non-adaptive design of Proposition 1.83 uses $T = 2n = 4\log(2/\delta)/\varepsilon^2$ subjects in total; Proposition 3.60 shows that no design, adaptive or not, can do better than $\frac{15}{32}\log(1/(4\delta))/\varepsilon^2$: the sample size $\asymp \log(1/\delta)/\varepsilon^2$ is a property of the problem, as claimed in Section 1.10.3, and adaptivity can only improve the constant. The decomposition of Lemma 3.59 extends to random budgets (stopping times), with the same proof and Wald-type bookkeeping, and to any number of treatments; it is the basic tool for lower bounds in bandit problems and sequential testing [LS20, Chap. 15], [GMS19].

3.5.4 Fano's inequality

Theorem 3.62 (Fano's inequality). *Let $P_1, \dots, P_N$ be probability measures on $(\Omega, \mathcal{F})$, with $N \ge 2$, and let Q be any probability measure on $(\Omega, \mathcal{F})$. Let*

$$\bar{D} := \frac{1}{N} \sum_{j=1}^{N} \mathrm{KL}(P_j \| Q).$$

For every (possibly randomised) test $\phi : \Omega \to \{1, \dots, N\}$, the average error $\bar{p}_e := \frac{1}{N}\sum_j P_j(\phi \neq j)$ satisfies

$$\max_j P_j(\phi \neq j) \;\ge\; \bar{p}_e \;\ge\; 1 - \frac{\bar{D} + \log 2}{\log N}. \tag{3.23}$$

More precisely,

$$H(\bar{p}_e) + \bar{p}_e \log(N-1) \;\ge\; \log N - \bar{D}, \tag{3.24}$$

where H is the binary entropy of Lemma 3.30.

Proof. Write $g_j(\omega) := \mathbb{P}(\phi = j \mid \omega)$, so that $g_j \ge 0$ and $\sum_j g_j = 1$ (for a deterministic test, $g_j = \mathbb{1}\{\phi = j\}$), and let $p_j = \int g_j \, \mathrm{d}P_j = P_j(\phi = j)$ and $q_j = \int g_j \, \mathrm{d}Q$, so that $\sum_j q_j = 1$. By Corollary 3.24, $\mathrm{KL}(P_j \| Q) \ge \mathrm{kl}(p_j, q_j)$ for every j. Averaging, and using the joint convexity of kl – Proposition 3.18 applied to the Bernoulli laws $\mathrm{Ber}(p_j)$, $\mathrm{Ber}(q_j)$ with weights $\frac{1}{N}$ –

$$\bar{D} \;\ge\; \frac{1}{N}\sum_{j=1}^{N} \mathrm{kl}(p_j, q_j) \;\ge\; \mathrm{kl}(\bar{p}, \bar{q}), \qquad \bar{p} := \frac{1}{N}\sum_j p_j = 1 - \bar{p}_e, \quad \bar{q} := \frac{1}{N}\sum_j q_j = \frac{1}{N}.$$

Lemma 3.30 gives $\mathrm{kl}(\bar{p}, \frac{1}{N}) \ge \bar{p}\log N - \log 2$, that is $1 - \bar{p}_e \le (\bar{D} + \log 2)/\log N$, which is (3.23). For (3.24), compute directly

$$\mathrm{kl}\Big(1 - \bar{p}_e, \frac{1}{N}\Big) = (1-\bar{p}_e)\log\big(N(1-\bar{p}_e)\big) + \bar{p}_e \log \frac{N\bar{p}_e}{N-1} = \log N - H(\bar{p}_e) - \bar{p}_e \log(N-1),$$

and note that $\bar{q}$ equals $\frac{1}{N}$ exactly, because the g_j sum to one, so that no monotonicity argument is needed. □

Remark 3.63 (On the statement). (a) *The choice of Q.* Any Q is allowed. With $Q = P_1$ one gets the average of $\mathrm{KL}(P_j \| P_1)$; with the mixture $Q = \bar{P} := \frac{1}{N}\sum_k P_k$, $\bar{D}$ is the *mutual information* $I(J;\omega)$ between a uniformly random index J and the observation $\omega \sim P_J$, and by convexity (Proposition 3.18) it is at most $\frac{1}{N^2}\sum_{j,k}\mathrm{KL}(P_j\|P_k) \le \max_{j,k}\mathrm{KL}(P_j\|P_k)$. The mixture is the optimal choice, but $Q = P_1$ is often simpler.

(b) *Reading the inequality.* The right-hand side of (3.23) is positive – in fact at least $\frac{1}{2}$ – as soon as $\bar{D} \le \frac{1}{2}\log N - \log 2$. To identify one of N hypotheses one must acquire about $\log N$ nats of information, and each observation brings at most KL of it: this is the information-theoretic content of the inequality, and $\log N$ is the number of nats needed to write down the index j.

(c) *History.* Inequality (3.24), in the equivalent form $H(J \mid \omega) \le H(\bar{p}_e) + \bar{p}_e \log(N-1)$ involving the conditional entropy, is due to Fano [Fan61], in the context of channel coding, where it proves the converse of Shannon's coding theorem. Its use for statistical lower bounds goes back to Has'minskii and Ibragimov [IH81] and Birgé [Bir83]; see [Yu97] for a comparison with Le Cam's and Assouad's methods. The proof given here, through the data-processing inequality for the single number $\bar{p}$ and joint convexity, avoids conditional entropies altogether and is close in spirit to [GMS19].

Fano's inequality requires a large family of hypotheses that are pairwise far apart in the parameter space but close in KL. The standard supply comes from the Hamming cube $\{0,1\}^d$, with the Hamming distance $\rho_H(b,b') := \sum_{i=1}^d \mathbb{1}\{b_i \ne b'_i\}$.

Lemma 3.64 (Gilbert–Varshamov)**.** *For every $d \ge 1$ there exist $N \ge \frac{1}{2}e^{d/16}$ points of $\{0,1\}^d$ whose pairwise Hamming distances are all at least $d/4$.*

Proof. Let $N = \lfloor e^{d/16} \rfloor \ge \frac{1}{2}e^{d/16}$ (as $\lfloor y \rfloor \ge y/2$ for $y \ge 1$), and draw $B_1, \dots, B_N$ independently and uniformly in $\{0,1\}^d$. For $j \ne k$, $\rho_H(B_j, B_k)$ has the binomial law $\mathrm{Bin}(d, \frac{1}{2})$, so by Hoeffding's inequality (1.28) $\mathbb{P}(\rho_H(B_j,B_k) < d/4) \le e^{-2d(1/4)^2} = e^{-d/8}$. By the union bound, the probability that some pair is at distance less than $d/4$ is at most $\binom{N}{2}e^{-d/8} < \frac{N^2}{2}e^{-d/8} \le \frac{1}{2}$. Hence some realisation has all pairwise distances at least $d/4$. □

Remark 3.65. The lemma is a first instance of the *probabilistic method*: an object with the desired property exists because a random one has it with positive probability, and the probability is controlled by a concentration inequality. The bound $N \ge \frac{1}{2}e^{d/16}$ is far from the best possible – the greedy (sphere-covering) argument of Gilbert [Gil52] and Varshamov [Var57], which takes a maximal set of points at mutual distance at least $d/4$, so that the Hamming balls of radius less than $d/4$ around them cover the cube, gives $N \ge 2^d / \sum_{k<d/4}\binom{d}{k} \ge e^{d\,\mathrm{kl}(1/4,1/2)}$ – but only the exponential dependence on d matters: it gives $\log N \asymp d$ hypotheses at mutual distance $\asymp d$, and Fano's inequality then produces lower bounds proportional to d. Exercise 3.11 carries this out for the Gaussian mean.

3.5.5 Assouad's lemma

When the parameter set contains a hypercube and the loss decomposes along its coordinates, there is a more economical method, which only requires comparing *neighbouring* vertices.

Lemma 3.66 (Assouad). *Let $(P_b)_{b\in\{0,1\}^d}$ be probability measures on $(\Omega,\mathcal{F})$, and for $b\in\{0,1\}^d$ and $i\le d$ let $b^{(i)}$ be b with its i-th coordinate flipped. For every (possibly randomised) estimator $\hat b:\Omega\to\{0,1\}^d$,*

$$\max_b \mathbb{E}_b\big[\rho_H(\hat b,b)\big] \;\ge\; \frac{1}{2^d}\sum_b \mathbb{E}_b\big[\rho_H(\hat b,b)\big] \;\ge\; \frac{d}{2}\min_{b,\,i}\Big(1-\mathrm{TV}\big(P_b,P_{b^{(i)}}\big)\Big). \tag{3.25}$$

In particular, if $\mathrm{KL}(P_b\|P_{b^{(i)}})\le D$ for all b and i, the right-hand side is at least $\frac{d}{2}\max\{1-\sqrt{D/2},\ \frac12 e^{-D}\}$.

Proof. Write $\rho_H(\hat b,b)=\sum_i \mathbb{1}\{\hat b_i\neq b_i\}$ and exchange the sums:

$$\frac{1}{2^d}\sum_b \mathbb{E}_b\big[\rho_H(\hat b,b)\big]=\sum_{i=1}^d \frac{1}{2^d}\sum_{b:\,b_i=0}\Big(P_b(\hat b_i\neq 0)+P_{b^{(i)}}(\hat b_i\neq 1)\Big),$$

since $b\mapsto b^{(i)}$ is a bijection from $\{b_i=0\}$ onto $\{b_i=1\}$. For fixed i and b, $\hat b_i$ is a (randomised) test between P_b and $P_{b^{(i)}}$, so by Lemma 3.27 the bracket is at least $1-\mathrm{TV}(P_b,P_{b^{(i)}})$. There are 2^{d-1} vertices with $b_i=0$, whence (3.25). The last statement follows from Pinsker's and the Bretagnolle–Huber inequalities. □

Corollary 3.67 (Assouad's lemma for estimation). *Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a model, ℓ a loss, and suppose that Θ contains points $(\theta_b)_{b\in\{0,1\}^d}$ for which there are $\beta>0$ and a measurable map $\Pi:\Theta\to\{0,1\}^d$ such that*

$$\ell(t,\theta_b)\;\ge\;\beta\,\rho_H\big(\Pi(t),b\big)\qquad\text{for all } t\in\Theta,\ b\in\{0,1\}^d. \tag{3.26}$$

Then, writing $P_b=\mathbb{P}_{\theta_b}$,

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{E}_\theta\big[\ell(\hat\theta,\theta)\big]\;\ge\;\frac{\beta d}{2}\min_{b,\,i}\Big(1-\mathrm{TV}\big(P_b,P_{b^{(i)}}\big)\Big).$$

Proof. For any estimator $\hat\theta$, $\sup_\theta \mathbb{E}_\theta\ell(\hat\theta,\theta)\ge\max_b\mathbb{E}_b\ell(\hat\theta,\theta_b)\ge\beta\max_b\mathbb{E}_b\rho_H(\Pi(\hat\theta),b)$, and $\Pi\circ\hat\theta$ is an estimator of b, to which Lemma 3.66 applies. □

Proposition 3.68 (Estimating a Gaussian mean in $\mathbb{R}^d$). *Let $X_1,\dots,X_n$ be i.i.d. $\mathcal{N}(\theta,\sigma^2 I_d)$ with $\theta\in\mathbb{R}^d$ unknown and $\sigma>0$ known. Then*

$$\frac{2}{27}\cdot\frac{d\sigma^2}{n}\;\le\;\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}_\theta\big[\|\hat\theta-\theta\|^2\big]\;\le\;\frac{d\sigma^2}{n},$$

the upper bound being attained by $\bar X_n$.

Proof. The upper bound is $\mathbb{E}_\theta\|\bar X_n-\theta\|^2=\sum_{i\le d}\mathbb{V}\mathrm{ar}(\bar X_{n,i})=d\sigma^2/n$. For the lower bound, let $r>0$ and $\theta_b=rb$ for $b\in\{0,1\}^d$, and define $\Pi(t)_i:=\mathbb{1}\{t_i\ge r/2\}$ for $t\in\mathbb{R}^d$. If $\Pi(t)_i\neq b_i$ then $|t_i-rb_i|\ge r/2$, so $\|t-\theta_b\|^2\ge\frac{r^2}{4}\rho_H(\Pi(t),b)$: (3.26) holds with $\beta=r^2/4$. Neighbouring vertices differ in one coordinate, and by tensorisation and Proposition 3.14(c), $\mathrm{KL}(\mathbb{P}_{\theta_b}\|\mathbb{P}_{\theta_{b^{(i)}}})=nr^2/(2\sigma^2)$; Pinsker's inequality gives $1-\mathrm{TV}\ge 1-\sqrt{nr^2/(4\sigma^2)}$. With $r=\gamma\sigma/\sqrt{n}$, Corollary 3.67 yields the lower bound $\frac{d\sigma^2}{n}\cdot\frac{\gamma^2}{8}\big(1-\frac{\gamma}{2}\big)$, and the choice $\gamma=\frac43$ gives $\frac{\gamma^2}{8}(1-\frac{\gamma}{2})=\frac{16}{9}\cdot\frac18\cdot\frac13=\frac{2}{27}$. □

Remark 3.69 (What Assouad's lemma captures, and the exact answer). The point of Proposition 3.68 is the factor d: the difficulty of estimating a d-dimensional mean is the sum of the difficulties of d one-dimensional problems, and Assouad's lemma turns this heuristic into a proof by averaging over the cube, so that each coordinate contributes its own two-point lower bound. Le Cam's method with two points would give only σ^2/n, whatever the dimension. The constant $\frac{2}{27}$ is again not sharp. In fact $\bar{X}_n$ is exactly minimax: for the prior $\theta \sim \mathcal{N}(0, \sigma_0^2 I_d)$, the joint law of $(\theta, \bar{X}_n)$ is Gaussian, the posterior mean $\mathbb{E}[\theta \mid X_1, \dots, X_n]$ minimises $\mathbb{E}\left\|\hat{\theta} - \theta\right\|^2$ among all (even randomised) estimators, being an orthogonal projection in L^2, and its risk is $d\left(\sigma_0^{-2} + n\sigma^{-2}\right)^{-1}$; since a maximum dominates an average, the minimax risk is at least this Bayes risk for every σ_0, and letting $\sigma_0 \to \infty$ gives exactly $d\sigma^2/n$. Assouad's lemma is from [Ass83]; see [Tsy09, Chap. 2] and [Yu97]. It is the tool of choice for the lower bounds of Chapter 4: the agnostic lower bound of order $\mathrm{VCdim}(\mathcal{H})/\varepsilon^2$ (Theorem 4.61) is proved there by applying Lemma 3.66 to distributions indexed by the vertices of a cube, built on a shattered set of $\mathrm{VCdim}(\mathcal{H})$ points, while its factor $\log(1/\delta)/\varepsilon^2$ comes from Lemma 3.53; the realizable lower bound of order $\mathrm{VCdim}(\mathcal{H})/\varepsilon$ (Theorem 4.63) comes instead from a direct averaging argument in the spirit of Theorem 3.72.

Remark 3.70 (Comparing the three methods). Le Cam's method uses two hypotheses and bounds the *sum* of two errors; it is optimal for one-dimensional problems and for testing. Fano's inequality uses N hypotheses, separated in the parameter space and close in KL, and bounds the *average* error over all of them; it is the most flexible method, since it needs no product structure – it handles sparse vectors, low-rank matrices and nonparametric classes, where the packing comes from metric entropy – but its constants are poor. Assouad's lemma needs a hypercube and a loss that decomposes along the coordinates, and in return it requires only *local* comparisons (neighbouring vertices) and yields good constants. [Yu97] shows how the three are related; all three rest on the same two facts, tensorisation and data processing.

3.6 No free lunch

We return to classification. Chapters 1 and 2 produced algorithms that learn under assumptions: the nearest-neighbour rule when the regression function is Lipschitz, empirical risk minimisation when the best classifier lies in a known finite class. Can the assumptions be dispensed with – is there an algorithm that learns *every* distribution, with a sample size that does not depend on the distribution? The answer is no, and the proof is a counting argument in the spirit of Assouad's lemma.

We work in the framework of Section 1.3 with $\mathcal{Y} = \{0, 1\}$, assuming that the singletons of $\mathcal{X}$ are measurable. A *randomised* learning algorithm is a sequence of maps $\mathcal{A}_n : (\mathcal{X} \times \mathcal{Y})^n \times [0, 1] \to \mathcal{M}(\mathcal{X}, \mathcal{Y})$, jointly measurable as in Definition 1.5, and it outputs $\widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n, U)$ where U is uniform on $[0, 1]$ and independent of $\mathcal{S}_n$; deterministic algorithms are the special case where $\mathcal{A}_n$ does not depend on its last argument. Probabilities below are with respect to $(\mathcal{S}_n, U)$. In this section, as in Chapter 1, x denotes a point of $\mathcal{X}$ or a finite sequence of points, no longer a level.

Lemma 3.71 (From expectation to probability)**.** *Let Z be a random variable with values in $[0, 1]$, and $\tau \in [0, 1)$. Then $\mathbb{P}(Z \geq \tau) \geq (\mathbb{E}Z - \tau)/(1 - \tau)$. In particular, if $\mathbb{E}Z \geq \frac{1}{4}$ then $\mathbb{P}(Z \geq \frac{1}{8}) \geq \frac{1}{7}$.*

Proof. $\mathbb{E}Z \leq \tau\,\mathbb{P}(Z < \tau) + 1 \cdot \mathbb{P}(Z \geq \tau) = \tau + (1 - \tau)\mathbb{P}(Z \geq \tau)$. With $\tau = \frac{1}{8}$ and $\mathbb{E}Z \geq \frac{1}{4}$, this gives $\mathbb{P}(Z \geq \frac{1}{8}) \geq \frac{1/8}{7/8} = \frac{1}{7}$. □

The lemma is a reverse Markov inequality, valid because Z is bounded; its constant cannot be improved (Exercise 3.3).

Theorem 3.72 (No free lunch)**.** *Let $\mathcal{A}$ be any (possibly randomised) learning algorithm for binary classification over $\mathcal{X}$, let $n \geq 1$, and let $C \subset \mathcal{X}$ be a set of $2n$ points (which exists if $|\mathcal{X}| \geq 2n$). For $f : C \to \{0,1\}$, let $\mathcal{D}_f$ be the law of $(X, f(X))$ with X uniform on C. Then*

$$\frac{1}{2^{2n}} \sum_{f \in \{0,1\}^C} \mathbb{E}\big[L_{\mathcal{D}_f}(\widehat{h}_n)\big] \;\geq\; \frac{1}{4}, \qquad \widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n, U),\ \mathcal{S}_n \sim \mathcal{D}_f^{\otimes n}. \tag{3.27}$$

Consequently there is a distribution $\mathcal{D} = \mathcal{D}_f$ on $\mathcal{X} \times \{0,1\}$ such that

(i) there exists a classifier $\bar{f} : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(\bar{f}) = 0$ (for instance the extension of f by 0 outside C, which is measurable since C is a finite union of measurable singletons), so that $L_{\mathcal{D}}^* = 0$;

(ii) $\mathbb{E}[L_{\mathcal{D}}(\widehat{h}_n)] \geq \frac{1}{4}$ and, with probability at least $\frac{1}{7}$ over the choice of $\mathcal{S}_n \sim \mathcal{D}^{\otimes n}$ and of U,

$$L_{\mathcal{D}}(\widehat{h}_n) \;\geq\; \frac{1}{8}.$$

Proof. For $f \in \{0,1\}^C$ and a classifier h, $L_{\mathcal{D}_f}(h) = \frac{1}{2n}\sum_{z \in C} \mathbb{1}\{h(z) \neq f(z)\}$. Under $\mathcal{D}_f^{\otimes n}$ the sample is $s_f(X_1, \dots, X_n)$, where $s_f(x) := \big((x_1, f(x_1)), \dots, (x_n, f(x_n))\big)$ for $x = (x_1, \dots, x_n) \in C^n$, and $X_1, \dots, X_n$ are i.i.d. uniform on C, independent of U. Hence, by Fubini's theorem,

$$\frac{1}{2^{2n}} \sum_{f} \mathbb{E}\big[L_{\mathcal{D}_f}(\widehat{h}_n)\big] = \int_0^1 \frac{1}{(2n)^n} \sum_{x \in C^n} \left[\frac{1}{2^{2n}} \sum_{f} L_{\mathcal{D}_f}\big(\mathcal{A}_n(s_f(x), u)\big) \right] \mathrm{d}u,$$

and it suffices to show that the bracket is at least $\frac{1}{4}$ for every fixed $x \in C^n$ and $u \in [0,1]$. Fix them, write $h_f := \mathcal{A}_n(s_f(x), u)$, and let $W := C \setminus \{x_1, \dots, x_n\}$, a set of $|W| \geq 2n - n = n$ points that the sample has not visited. Keeping only these points in the risk,

$$\frac{1}{2^{2n}} \sum_{f} L_{\mathcal{D}_f}(h_f) \;\geq\; \frac{1}{2n} \sum_{v \in W} \frac{1}{2^{2n}} \sum_{f} \mathbb{1}\big\{h_f(v) \neq f(v)\big\}.$$

Fix $v \in W$ and pair each f with the labelling $f^{(v)}$ that agrees with f everywhere on C except at v. Since v does not appear in x, the two samples coincide, $s_f(x) = s_{f^{(v)}}(x)$, hence $h_f = h_{f^{(v)}}$; and since $f(v) \neq f^{(v)}(v)$, exactly one of $\mathbb{1}\{h_f(v) \neq f(v)\}$ and $\mathbb{1}\{h_{f^{(v)}}(v) \neq f^{(v)}(v)\}$ equals 1. The map $f \mapsto f^{(v)}$ being an involution without fixed points, the inner average equals $\frac{1}{2}$, and the bracket is at least $\frac{|W|}{2n} \cdot \frac{1}{2} \geq \frac{1}{4}$. This proves (3.27).

An average is at most a maximum, so some f satisfies $\mathbb{E}[L_{\mathcal{D}_f}(\widehat{h}_n)] \geq \frac{1}{4}$; for that f, (i) is clear and (ii) follows from Lemma 3.71 applied to $Z = L_{\mathcal{D}_f}(\widehat{h}_n) \in [0,1]$. □

Remark 3.73 (What the theorem says, and what it does not). (a) *The order of the quantifiers.* The bad distribution depends on the algorithm and on n. For every fixed distribution there are excellent algorithms – the constant algorithm that outputs the Bayes classifier, for instance – but no algorithm is good for all distributions at once. The problem is noiseless ($L_{\mathcal{D}}^* = 0$) and the target is a deterministic function: the difficulty is not noise but *ignorance*. On the n or more points of C that it has not seen, the sample carries no information on the labels, and the algorithm can only guess.

(b) *Prior knowledge restores learnability.* Suppose the algorithm knows that the target belongs to a finite class $\mathcal{H} \ni \bar{f}$, and runs empirical risk minimisation over $\mathcal{H}$. A hypothesis h with $L_{\mathcal{D}}(h) \geq \frac{1}{8}$ has zero empirical risk with probability $(1 - L_{\mathcal{D}}(h))^n \leq (\frac{7}{8})^n$, and there are at most $|\mathcal{H}| - 1$ of them ($\bar{f}$ is not one), so the proof of Theorem 2.66 gives $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \geq \frac{1}{8}) \leq (|\mathcal{H}|-1)(\frac{7}{8})^n < \frac{1}{7}$ as soon as $n > \log(7(|\mathcal{H}|-1))/\log\frac{8}{7}$, and in particular as soon as $n \geq 8\log(7(|\mathcal{H}| - 1))$ (since $\log\frac{8}{7} > \frac{1}{8}$). For $\mathcal{H} = \{\bar{f}\}$ the risk is 0 for every n. The theorem is therefore not a statement about the hardness of the distribution $\mathcal{D}_f$, but about the size of the class over which the algorithm is required to succeed: *the class of all functions on a set of* $2n$ *points is too large to be learned from* n *examples.*

(c) *Randomisation does not help*: the proof bounds the average risk for each fixed value u of the internal randomness.

(d) *Averaging over targets.* Inequality (3.27) is stronger than the existence of one bad distribution: if the labelling is drawn uniformly at random, every algorithm has expected risk at least $\frac{1}{4}$. On the points outside the sample, its *off-training-set* error is exactly $\frac{1}{2}$, the error of a coin toss. This is the form of the principle popularised by Wolpert [Wol96], whose no-free-lunch theorems state that, averaged over all targets, all algorithms have the same off-training-set performance; the optimisation version is [WM97].

Remark 3.74 (No free lunch as an instance of Assouad's lemma). Index the distributions $\mathcal{D}_f$ by the vertex $f \in \{0,1\}^C$ of a cube of dimension $2n$, and let $P_f := \mathcal{D}_f^{\otimes n}$ be the law of the sample. Since $L_{\mathcal{D}_f}(h) = \frac{1}{2n}\rho_H(h|_C, f)$, Lemma 3.66 gives

$$\frac{1}{2^{2n}}\sum_f \mathbb{E}\big[L_{\mathcal{D}_f}(\widehat{h}_n)\big] \;\geq\; \frac{1}{2}\min_{f,v}\Big(1 - \mathrm{TV}\big(P_f, P_{f^{(v)}}\big)\Big),$$

where $f^{(v)}$ flips the label of $v \in C$. The laws P_f and $P_{f^{(v)}}$ coincide on samples that do not contain v and are mutually singular on those that do, so their total variation is the probability that v is sampled, $1 - (1 - \frac{1}{2n})^n$. Hence the average risk is at least $\frac{1}{2}(1 - \frac{1}{2n})^n$, which is at least $\frac{1}{4}$ by Bernoulli's inequality $(1 - \frac{1}{2n})^n \geq 1 - \frac{n}{2n}$, and tends to $\frac{1}{2}e^{-1/2} \approx 0.303$ as $n \to \infty$. The pairing argument in the proof of Theorem 3.72 is exactly the proof of Assouad's lemma, and this point of view is the one that extends to noisy problems and yields the agnostic lower bounds of Chapter 4. The same computation with X uniform on any finite set C gives an average risk at least $\frac{1}{2}(1 - \frac{1}{|C|})^n$, which is at least $\frac{1}{4}$ as soon as $n \leq (|C| - 1)\log 2$; applied to a largest $\frac{1}{c}$-separated subset of $[0,1]^d$, which has $N_d(c)$ points (Lemma 3.77 below), it would raise the threshold of Theorem 3.78 from $N_d(c)/2$ to $(N_d(c) - 1)\log 2$, a gain of about 38% for large $N_d(c)$.

Corollary 3.75 (There is no universal learner)**.** *If $\mathcal{X}$ is infinite, the class of all classifiers $\{0,1\}^{\mathcal{X}}$ is not PAC learnable, even in the realizable case (Definition 2.63): for every algorithm and every n there is a distribution, realizable by some classifier, for which $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \geq \frac{1}{8}) \geq \frac{1}{7}$.*

Proof. For every n there is a set C of $2n$ points; apply Theorem 3.72. A function $n_{\mathcal{H}}(\varepsilon, \delta)$ as in Definition 2.63 cannot exist for $\varepsilon \leq \frac{1}{8}$ and $\delta < \frac{1}{7}$. □

The same argument applies to any class $\mathcal{H}$ that contains all the labellings of arbitrarily large finite sets – a class that *shatters* sets of every size. That this is the only obstruction, and that the size of the largest shattered set quantifies the sample complexity, is the content of the Vapnik–Chervonenkis theory of Chapter 4, where the consequence "infinite VC dimension implies non-learnability" is drawn from Theorem 3.72.

Remark 3.76 (Slow rates). Theorem 3.72 uses distributions that change with n. A deeper statement, due to Devroye, holds for a *single* distribution per algorithm, provided $\mathcal{X}$ is infinite (for instance $\mathcal{X} = \mathbb{R}^d$): for every sequence of classification rules and every sequence of positive numbers $\frac{1}{16} \geq \varepsilon_1 \geq \varepsilon_2 \geq \dots$ converging to 0, there is a distribution $\mathcal{D}$, with $L^*_{\mathcal{D}} = 0$, such that $\mathbb{E}[L_{\mathcal{D}}(\widehat{h}_n)] \geq \varepsilon_n$ for all n [DGL96, Thm. 7.2]. Universally consistent rules exist (Stone's theorem [Sto77], recalled in Section 1.5.1), but their convergence can be arbitrarily slow: *consistency is free, rates are not*. We do not prove this result, whose proof is a more delicate version of the construction above.

3.6.1 The curse of dimensionality

The no-free-lunch theorem uses arbitrary labellings of a finite set, which look nothing like the smooth problems of Section 1.5. But a labelling of a well-separated set of points can be realised by a Lipschitz regression function, and the theorem then says that Lipschitz problems in dimension d require a sample size exponential in d. We use the Euclidean distance on $[0,1]^d$, as in Assumption 1.20, and call a set C *r-separated* if $\|z - z'\| \geq r$ for all distinct $z, z' \in C$.

Lemma 3.77 (Separated sets in the cube). *Let $c > 0$, $d \geq 1$, and let $N_d(c)$ be the largest cardinality of a $\frac{1}{c}$-separated subset of $[0,1]^d$. Then*

$$N_d(c) \;\geq\; \max\Big\{(\lfloor c \rfloor + 1)^d,\; \frac{c^d}{V_d}\Big\} \qquad \text{and} \qquad \frac{c^d}{V_d} \;\geq\; \Big(c\sqrt{\frac{d}{2\pi e}}\Big)^d,$$

where V_d is the volume of the Euclidean unit ball of $\mathbb{R}^d$.

Proof. If $c \geq 1$, the grid $\{0, \frac{1}{\lfloor c \rfloor}, \frac{2}{\lfloor c \rfloor}, \dots, 1\}^d$ has $(\lfloor c \rfloor + 1)^d$ points at mutual distance at least $\frac{1}{\lfloor c \rfloor} \geq \frac{1}{c}$; if $c < 1$ the first bound is 1 and trivial. For the second, the open balls of radius $\frac{1}{2c}$ centred at the points of a $\frac{1}{c}$-separated set are disjoint and contained in $[-\frac{1}{2c}, 1 + \frac{1}{2c}]^d$, so such a set has at most $(1 + \frac{1}{c})^d (2c)^d / V_d = (2c+2)^d / V_d$ points and a maximal one exists. If C is maximal, every $x \in [0,1]^d$ lies within distance $\frac{1}{c}$ of C (otherwise $C \cup \{x\}$ would still be separated), so the balls of radius $\frac{1}{c}$ centred at C cover the cube and $|C|\; V_d c^{-d} \geq 1$. Finally $V_d^{-1} = \Gamma(\frac{d}{2} + 1)\pi^{-d/2} \geq (d/(2\pi e))^{d/2}$, as in the proof of Proposition 1.27. □

Theorem 3.78 (No free lunch for Lipschitz problems). *Let $d \geq 1$, $c > 0$, $\mathcal{X} = [0,1]^d$, and let $n \geq 1$ satisfy $2n \leq N_d(c)$ – in particular, any*

$$n \;\leq\; \frac{1}{2} \max\Big\{(\lfloor c \rfloor + 1)^d,\; \frac{c^d}{V_d}\Big\}$$

will do. For every (possibly randomised) learning algorithm there is a distribution $\mathcal{D}$ on $[0,1]^d \times \{0,1\}$ such that

(i) its regression function $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ admits a version that is c-Lipschitz on the whole cube $[0,1]^d$;

(ii) its Bayes risk is $L^*_{\mathcal{D}} = 0$;

(iii) $\mathbb{E}[L_{\mathcal{D}}(\widehat{h}_n)] \geq \frac{1}{4}$, and $L_{\mathcal{D}}(\widehat{h}_n) \geq \frac{1}{8}$ with probability at least $\frac{1}{7}$.

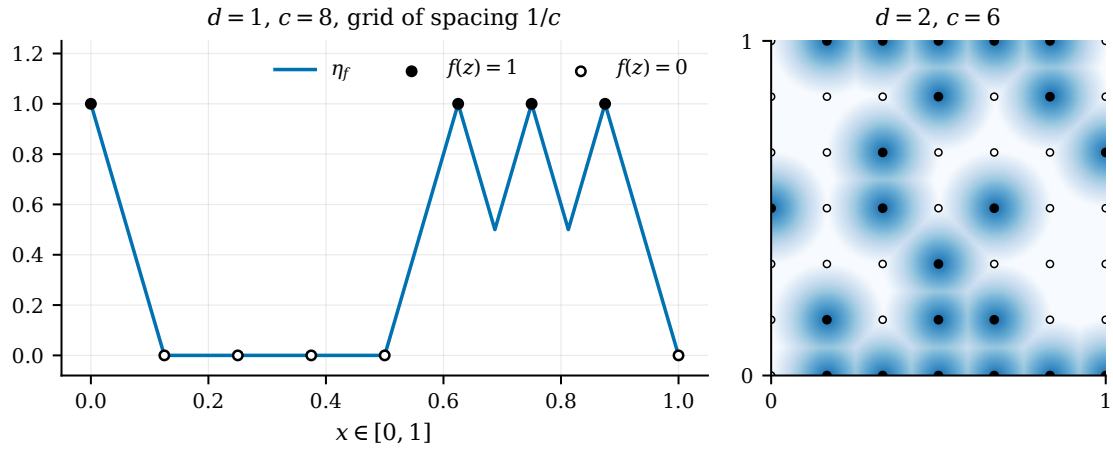


Figure 3.3: The regression functions $\eta_f = \max\{0, 1 - c\,\mathrm{dist}(\cdot, f^{-1}(1))\}$ of the proof of Theorem 3.78, for a random labelling f of a $\frac{1}{c}$-separated grid (black: label 1, white: label 0). Each η_f is c-Lipschitz on the whole cube and takes the values $f(z) \in \{0, 1\}$ on the grid, so the problem is noiseless; yet a sample that misses a grid point carries no information on its label. In dimension d the grid has $(\lfloor c \rfloor + 1)^d$ points. Generated by `figures/fig_nfl_lipschitz.py`.

Proof. Let C be a $\frac{1}{c}$-separated subset of $[0,1]^d$ with $2n$ points, and let $\mathcal{D} = \mathcal{D}_f$ be the distribution given by Theorem 3.72 for this C, so that (iii) holds. Let $C_1 := f^{-1}(1)$, for this f, and

$$\eta_f(x) := \max\Big\{0,\ 1 - c\,\mathrm{dist}(x, C_1)\Big\}, \qquad x \in [0,1]^d,$$

with $\eta_f \equiv 0$ if $C_1 = \emptyset$. The function $x \mapsto \mathrm{dist}(x, C_1)$ is 1-Lipschitz and $r \mapsto \max\{0, 1 - cr\}$ is c-Lipschitz, so η_f is c-Lipschitz with values in $[0,1]$. If $z \in C_1$ then $\eta_f(z) = 1 = f(z)$; if $z \in C \setminus C_1$ then $\mathrm{dist}(z, C_1) \geq \frac{1}{c}$ by separation, so $\eta_f(z) = 0 = f(z)$. Under $\mathcal{D}_f$, X is uniform on C and $Y = f(X)$, so $\mathbb{P}(Y = 1 \mid X) = f(X) = \eta_f(X)$ almost surely: η_f is a version of the regression function, which proves (i). Since $\eta_f \in \{0,1\}$ on C, $L^*_{\mathcal{D}} = \mathbb{E}[\min(\eta_f(X), 1 - \eta_f(X))] = 0$ by (1.7). □

Remark 3.79 (Lower and upper bounds, compared honestly). Every distribution of Theorem 3.78 has $L^*_{\mathcal{D}} = 0$ and satisfies Assumption 1.20, with the closed cube $[0,1]^d$ in place of $[0,1)^d$; to stay in $[0,1)^d$, use the grid $\{k/\lfloor c \rfloor : 0 \leq k < \lfloor c \rfloor\}^d$, which has $\lfloor c \rfloor^d$ points, or the volume bound, which is unaffected. For such distributions, Theorem 1.21 and Remark 1.29 give

$$\mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n)\big] \ \leq\ 3c\sqrt{d}\, n^{-1/(d+1)} \qquad \text{and} \qquad \mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}^{\mathrm{NN}}_n)\big] \ \leq\ 4c\sqrt{d}\, n^{-1/d} \quad (d \geq 2).$$

These upper bounds guarantee an expected risk below $\frac{1}{4}$ for the nearest-neighbour rule, uniformly over the class, as soon as $n > (16c\sqrt{d})^d$ (for $d \geq 2$), while Theorem 3.78 and Lemma 3.77 show that *every* algorithm needs

$$n \ >\ \frac{c^d}{2V_d} \ \geq\ \frac{1}{2}\Big(\frac{c\sqrt{d}}{\sqrt{2\pi e}}\Big)^d.$$

Both thresholds are of the form $(\zeta\, c\sqrt{d})^d$, with ζ between $1/\sqrt{2\pi e} \approx 0.24$ and 16: the sample complexity of learning a c-Lipschitz problem to constant accuracy is *exponential in* d, the base of the exponential grows like $c\sqrt{d}$, and the nearest-neighbour rule achieves it up to the

value of ζ. This is the precise sense in which the curse of dimensionality of Remark 1.26 is not a weakness of the nearest-neighbour rule but a property of the class. The theorem says nothing about small accuracies $\varepsilon \to 0$: the lower bound concerns the constant accuracy $\frac{1}{8}$ and is obtained in the noiseless case; showing that the *rate* $n^{-1/(d+2)}$ of Remark 1.78 is optimal for noisy Lipschitz problems requires Assouad's lemma with hypotheses whose regression functions are close to $\frac{1}{2}$ [AT07]. The way out of the curse is the same as the way out of Theorem 3.72: restrict the class, using structure – low intrinsic dimension (Section 1.5.2), margin conditions (Section 2.7.2), or a hypothesis class of finite capacity (Chapter 4).

3.7 Notes and references

Bernoulli deviations. The bound (3.1) for bounded variables, together with the observation that Bernoulli variables are the extremal case, is in Hoeffding's paper [Hoe63, Thm. 1], which also contains the comparison with Bennett's and Bernstein's bounds that we carried out in Example 3.9. Lemma 3.3 and its use for the inequalities of Bennett and Bernstein are taken from Pollard's notes [Pol01]. The kl confidence interval of Proposition 3.10 is implicit in the Chernoff bound and was put to systematic use in the analysis of bandit algorithms [GC11]; the exact binomial interval is due to Clopper and Pearson [CP34].

Relative entropy. The divergence was introduced by Kullback and Leibler [KL51] and developed in [Kul59]; its information-theoretic role is the subject of [CT06], and [PW25] is a comprehensive modern treatment, including the general chain rule and f-divergences, of which KL, total variation, χ^2 and the squared Hellinger distance are instances. Pinsker's inequality is [Pin64], with the sharp constant in [Csi67; Kul67; Kem69]; the Bretagnolle–Huber inequality is [BH79]. The proof of the data-processing inequality by conditional Jensen is classical; the reduction of everything to Corollary 3.24, that is, to the contraction of KL onto the means of a single $[0,1]$-valued statistic, is emphasised in [GMS19]. The Gibbs variational principle in the form of Proposition 3.35 is [DV75]. The Neyman–Pearson lemma is [NP33]; Stein's lemma first appeared in [Che56], and Hoeffding [Hoe65] developed the corresponding theory for multinomial models, where the likelihood-ratio test is shown to be optimal in the sense of error exponents.

Large deviations. The change-of-measure proof of the lower bound goes back to Cramér [Cra38]; [DZ98, Sec. 2.2] gives the general statement for real variables and [DZ98, Chap. 6] Sanov's theorem [San57], which identifies $\mathrm{KL}(Q\|P)$ as the cost of observing an empirical distribution close to Q when sampling from P. The entropic lower bound of Theorem 3.47 is in the spirit of the "fundamental inequality" of [GMS19], where the same three steps (tensorisation, data processing, Lemma 3.30) are used to prove lower bounds for bandit problems.

Minimax lower bounds. The reduction to testing and the three methods of Section 3.5 are the core of Chapter 2 of Tsybakov's book [Tsy09], which is the best introduction to the subject; Yu's survey [Yu97] compares the three methods; the original sources are Le Cam [Le 73; Le 86], Fano [Fan61] (in coding theory), Has'minskii and Ibragimov [IH81] and Birgé [Bir83] (Fano in statistics), and Assouad [Ass83]. For the modern information-theoretic point of view, with many examples in high-dimensional statistics, see [Wai19, Chap. 15].

No free lunch. Theorem 3.72 and its proof follow [SB14, Chap. 5], with the constants $\frac{1}{8}$ and $\frac{1}{7}$ of that book. The off-training-set formulation is Wolpert's [Wol96; WM97]; the slow-rates theorem of Remark 3.76 and many related results are in [DGL96, Chap. 7]. Minimax rates for classification with smooth regression functions, including the role of margin conditions, are established in [AT07].

3.8 Exercises

Exercise 3.1 (Gaussian relative entropies)**.** (a) Show that for $\sigma_1, \sigma_2 > 0$,

$$\mathrm{KL}\big(\mathcal{N}(\mu_1, \sigma_1^2) \| \mathcal{N}(\mu_2, \sigma_2^2)\big) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

(b) Show that for a positive definite matrix Σ, $\mathrm{KL}(\mathcal{N}(\theta_1, \Sigma) \| \mathcal{N}(\theta_2, \Sigma)) = \frac{1}{2}(\theta_1 - \theta_2)^\top \Sigma^{-1}(\theta_1 - \theta_2)$.

Exercise 3.2 (⋆ Relative entropy is not a distance)**.** (a) Compute $\mathrm{kl}(\frac{1}{2}, \frac{1}{4})$ and $\mathrm{kl}(\frac{1}{4}, \frac{1}{2})$. (b) Find two laws P, Q with $\mathrm{KL}(P\|Q) < \infty = \mathrm{KL}(Q\|P)$. (c) Find $\mu_1, \mu_2, \mu_3 \in [0, 1)$ with $\mathrm{kl}(\mu_1, \mu_3) > \mathrm{kl}(\mu_1, \mu_2) + \mathrm{kl}(\mu_2, \mu_3)$.

Exercise 3.3 (⋆ The reverse Markov inequality is sharp)**.** (a) Show that for every $\varepsilon \in (0, \frac{1}{8})$ there is a random variable Z with values in $[0, 1]$, $\mathbb{E}Z = \frac{1}{4}$ and $\mathbb{P}(Z \geq \frac{1}{8}) = \frac{1+8\varepsilon}{7+8\varepsilon}$. Conclude that the constant $\frac{1}{7}$ in Lemma 3.71 cannot be improved. (b) Show that $\mathbb{E}Z \geq \frac{1}{4}$ gives no positive lower bound on $\mathbb{P}(Z \geq \frac{1}{4})$.

Exercise 3.4 (Hellinger distance)**.** With the notation of Lemma 3.13, let $A(P, Q) = \int \sqrt{pq}\, \mathrm{d}\nu$ and $\mathrm{Hel}^2(P, Q) = \int (\sqrt{p} - \sqrt{q})^2\, \mathrm{d}\nu = 2(1 - A(P, Q))$ (we write Hel rather than the usual H, which denotes the binary entropy in this book). (a) Show that $\mathrm{Hel}^2(P, Q) \leq \mathrm{KL}(P\|Q)$. (b) Show that $\mathrm{TV}(P, Q) \leq \mathrm{Hel}(P, Q)\sqrt{1 - \mathrm{Hel}^2(P, Q)/4}$. (c) Show that $A(P_1 \otimes P_2, Q_1 \otimes Q_2) = A(P_1, Q_1)A(P_2, Q_2)$, and deduce a formula for $\mathrm{Hel}^2(P^{\otimes n}, Q^{\otimes n})$. (d) Let $P = \mathrm{Unif}[0, 1]$ and $Q_n = \mathrm{Unif}[0, 1 + \frac{1}{n}]$. Compute $\mathrm{KL}(P\|Q_n)$, $\mathrm{KL}(Q_n\|P)$ and $\mathrm{TV}(P^{\otimes n}, Q_n^{\otimes n})$. Deduce that for the model $\{\mathrm{Unif}[0, \theta]^{\otimes n} : \theta \in [1, 2]\}$ the minimax quadratic risk is at least of order n^{-2}, faster than the parametric rate n^{-1}. Which of the two divergences can be used in a KL-based two-point argument, and why does the other one fail?

Exercise 3.5 (Pinsker from Hoeffding's lemma)**.** Let $g : \Omega \to [0, 1]$ be measurable. Using Proposition 3.35 and Hoeffding's lemma (Lemma 1.62), show that $\mathbb{E}_P[g] - \mathbb{E}_Q[g] \leq \sqrt{\mathrm{KL}(P\|Q)/2}$, and deduce Proposition 3.28. Explain in what sense Pinsker's inequality and Hoeffding's lemma are dual to each other.

Exercise 3.6 (Sufficiency and equality in data processing)**.** (a) In the setting of Proposition 3.22, show that if $\mathrm{d}P/\mathrm{d}Q = g(X)$ Q-almost surely for some measurable g, then $\mathrm{KL}(P^X\|Q^X) = \mathrm{KL}(P\|Q)$. (b) Deduce that the law $\mathrm{Bin}(n, a)$ of $n\bar{X}_n$ under $\mathbb{P}_a$ satisfies $\mathrm{KL}(\mathrm{Bin}(n, a)\|\mathrm{Bin}(n, \mu)) = n\,\mathrm{kl}(a, \mu)$, and check it by direct computation. (c) Give an example where the inequality of Proposition 3.22 is strict.

Exercise 3.7 (Poisson large deviations, from below)**.** Let $X_1, X_2, \ldots$ be i.i.d. $\mathrm{Poi}(\sigma^2)$ and $x > \sigma^2$. (a) Adapt the proof of Theorem 3.42 to show that $\frac{1}{n}\log \mathbb{P}(\bar{X}_n \geq x) \to -(x \log \frac{x}{\sigma^2} - x + \sigma^2)$. (b) Show that for every $\varepsilon > 0$ and every n with $n\varepsilon^2 > x + \varepsilon$,

$$\mathbb{P}\big(\bar{X}_n \geq x\big) \;\geq\; \exp\Big(-\frac{n\ \mathrm{KL}(\mathrm{Poi}(x + \varepsilon)\|\mathrm{Poi}(\sigma^2)) + \log 2}{1 - (x + \varepsilon)/(n\varepsilon^2)}\Big).$$

Exercise 3.8 (Stein's lemma for Gaussian shifts)**.** Let $P_0 = \mathcal{N}(0, 1)$ and $P_1 = \mathcal{N}(\mu, 1)$ with $\mu > 0$, and $\alpha \in (0, 1)$. (a) Using Lemma 3.37, show that the optimal level-α test based on $X_1, \ldots, X_n$ rejects H_0 when $\sqrt{n}\bar{X}_n > z_\alpha$, where $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim \mathcal{N}(0, 1)$, and that $\beta_n^*(\alpha) = \mathbb{P}(Z \leq z_\alpha - \mu\sqrt{n})$. (b) Deduce $\frac{1}{n}\log \beta_n^*(\alpha) = -\frac{\mu^2}{2} + \frac{\mu z_\alpha}{\sqrt{n}} + O\big(\frac{\log n}{n}\big)$, and compare with Theorem 3.39 and with (3.16).

Exercise 3.9 (Bayesian testing: the Bhattacharyya bound)**.** Let $A = \int \sqrt{p_0 p_1}\, \mathrm{d}\nu$ be the affinity of Exercise 3.4. Show that for every n,

$$\frac{1}{2}\, e^{-n \min\{\mathrm{KL}(P_0\|P_1), \mathrm{KL}(P_1\|P_0)\}} \;\le\; \inf_{\phi}\Big\{ \int \phi \, \mathrm{d}P_0^{\otimes n} + \int (1-\phi)\, \mathrm{d}P_1^{\otimes n} \Big\} \;\le\; A^n,$$

and that $-\log A \le \frac{1}{2}\min\{\mathrm{KL}(P_0\|P_1), \mathrm{KL}(P_1\|P_0)\}$. Compute both sides for two Gaussian laws with the same variance, and compare with the Chernoff information of Remark 3.41.

Exercise 3.10 (⋆ Local minimax lower bound for a Bernoulli mean)**.** Let $\mu \in (0,1)$, $n \ge \frac{\mu}{4(1-\mu)}$ and $r = \frac{1}{4}\sqrt{\mu(1-\mu)/n}$. Show that every estimator $\hat{\mu}$ satisfies

$$\max_{\mu' \in \{\mu, \mu+r\}} \mathbb{E}_{\mu'}\big[(\hat{\mu} - \mu')^2\big] \;\ge\; \frac{3}{512} \cdot \frac{\mu(1-\mu)}{n}.$$

Exercise 3.11 (⋆ The Gaussian mean by Fano's method)**.** In the setting of Proposition 3.68, assume $d \ge 89$. Using Lemma 3.64, Theorem 3.62 with $Q = \mathcal{N}(0, \sigma^2 I_d)^{\otimes n}$ and Lemma 3.52, show that

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta\big[\|\hat{\theta} - \theta\|^2\big] \;\ge\; \frac{d\sigma^2}{2048\, n}.$$

Compare the constant with that of Proposition 3.68.

Exercise 3.12 (No free lunch with a large domain)**.** (a) In the proof of Theorem 3.72, take for C a set of $|C| \ge n$ points and show that the average in (3.27) is at least $\frac{1}{2}\big(1 - \frac{n}{|C|}\big)$. (b) Deduce that if $\mathcal{X}$ is infinite, then for every algorithm, every n and every $\varepsilon > 0$ there is a distribution with $L^*_{\mathcal{D}} = 0$ and $\mathbb{E}[L_{\mathcal{D}}(\widehat{h}_n)] \ge \frac{1}{2} - \varepsilon$, and that $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \ge \frac{1}{4}) \ge \frac{1}{3} - \frac{4\varepsilon}{3}$: on an infinite domain, no algorithm does better than a coin toss, uniformly over noiseless problems. This is a convenient form for classes of infinite VC dimension (Chapter 4).

Chapter 4

PAC Learning and Vapnik–Chervonenkis Theory

Summary. Chapter 2 showed that a finite class can be learned with a number of examples logarithmic in its size. We now answer the question for arbitrary classes of binary classifiers. The answer is combinatorial: a class is learnable if and only if its Vapnik–Chervonenkis dimension is finite, and the VC dimension then determines the sample complexity, up to constants and a factor $\log(1/\varepsilon)$ – from above through symmetrisation, Rademacher complexities, Massart's lemma and the Sauer–Shelah lemma, from below through the no-free-lunch theorem and the minimax tools of Chapter 3. This is the fundamental theorem of statistical learning, which we prove in qualitative and in quantitative form.

4.1 Introduction

Section 2.7.1 answered the learning question for *finite* classes. If $\mathcal{H}$ is a finite set of classifiers, the empirical risk minimiser $\widehat{h}_n$ over $\mathcal{H}$ satisfies, with probability at least $1-\delta$,

$$L_{\mathcal{D}}(\widehat{h}_n) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \sqrt{\frac{2\log(2\,|\mathcal{H}|\,/\delta)}{n}}$$

(Theorem 2.65), and in the realizable case the excess term improves to $\log(|\mathcal{H}|\,/\delta)/n$ (Theorem 2.66). In the language of Definition 2.63, finite classes are PAC learnable, and their sample complexity grows like the *logarithm* of their cardinality.

This is a satisfactory answer to a question that is not quite the right one. The classes used in practice are infinite: thresholds on the real line, half-spaces of $\mathbb{R}^k$, axis-parallel boxes, neural networks with real weights. Discretising the parameters and invoking Theorem 2.65 gives something, but the resulting bound depends on an arbitrary discretisation step, and it cannot explain why, say, the class of all half-planes is learnable while the class of all convex polygons is not. Both are infinite; both are described by real parameters; and Theorem 2.65 sees no difference between them. What is needed is a measure of the size of a class that is finite for the first and infinite for the second, and that controls the sample complexity in the way $\log|\mathcal{H}|$ did for finite classes.

The measure is the *Vapnik–Chervonenkis dimension*, and the purpose of this chapter is to prove that it characterises learnability completely, both qualitatively and quantitatively. The main result, the *fundamental theorem of statistical learning* (Theorems 4.59 and 4.65), states that for a class $\mathcal{H}$ of binary classifiers the following are equivalent: $\mathcal{H}$ is PAC learnable; $\mathcal{H}$ is agnostically PAC learnable; the empirical risk minimiser learns $\mathcal{H}$; the empirical risks

converge to the risks uniformly over $\mathcal{H}$; and $\mathrm{VCdim}(\mathcal{H}) < \infty$. Moreover, if $d = \mathrm{VCdim}(\mathcal{H})$, the sample complexity is of order

$$\frac{d + \log(1/\delta)}{\varepsilon^2} \quad \text{(agnostic case)}, \qquad \frac{d + \log(1/\delta)}{\varepsilon} \quad \text{(realizable case)},$$

up to constants and a factor $\log(1/\varepsilon)$ in the upper bounds: *the VC dimension plays exactly the role that* $\log|\mathcal{H}|$ *played for finite classes*. (As is customary in VC theory, d denotes the VC dimension in this chapter and the next; the dimension of a Euclidean domain, written d in Chapters 1 to 3, is written k, as in $\mathbb{R}^k$ above.)

The proof has a pleasant architecture, which is also the plan of the chapter.

- Section 4.2 recalls the definitions of PAC learnability and introduces the *uniform convergence property*, which is the analytic form of learnability: if empirical risks are uniformly close to true risks, empirical risk minimisation works (Corollary 4.10).
- Section 4.3 defines shattering, the growth function and the VC dimension, computes the VC dimension of the standard classes – half-lines, intervals, boxes, half-spaces (through Radon's theorem), the sine class – and proves the easy half of the fundamental theorem: a class of infinite VC dimension is not learnable. The proof is the no-free-lunch theorem of Chapter 3 (Theorem 3.72), applied to a shattered set.
- Section 4.4 proves the Sauer–Shelah lemma: if $\mathrm{VCdim}(\mathcal{H}) = d$, the number of ways $\mathcal{H}$ can label n points is not 2^n but at most $(en/d)^d$. A combinatorial dichotomy – exponential or polynomial, nothing in between – turns an infinite class into an effectively finite one on every sample.
- Section 4.5 converts this combinatorial fact into a uniform law of large numbers. The expected uniform deviation is bounded by the *Rademacher complexity* through *symmetrisation*; the Rademacher complexity of a class with finitely many patterns is bounded by *Massart's lemma*, which is the maximal inequality of Proposition 2.28; and the deviation of the uniform deviation from its mean is controlled by McDiarmid's inequality (Theorem 2.55), exactly as announced in Section 2.7.5. This gives the upper bounds, in both the agnostic and the realizable cases.
- Section 4.6 states and proves the fundamental theorem, and proves the matching *lower* bounds with the minimax tools of Chapter 3: Assouad's lemma (Lemma 3.66) for the factor d/ε^2, Le Cam's two-point method (Lemma 3.53) for the factor $\log(1/\delta)/\varepsilon^2$, and a direct Bayesian argument in the spirit of the no-free-lunch theorem for the realizable factor d/ε.

Throughout, the guiding principle is the one formulated after (2.43): the fluctuations of the uniform deviation $\sup_{h\in\mathcal{H}} \left|L_{S_n}(h) - L_{\mathcal{D}}(h)\right|$ are *always* of order $n^{-1/2}$, and the whole difficulty is to bound its expectation. The VC dimension is what makes this expectation small.

4.2 PAC learning and uniform convergence

4.2.1 Setting and standing assumptions

We work in the framework of Section 1.3 with binary labels, $\mathcal{Y} = \{0, 1\}$, and the 0–1 loss, so that $L_{\mathcal{D}}(h) = \mathbb{P}(h(X) \neq Y)$ and $L_{S_n}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}$. A *hypothesis class* is a

non-empty set $\mathcal{H}$ of measurable maps $\mathcal{X} \to \{0,1\}$. We identify a classifier h with the set $\{h = 1\} \subseteq \mathcal{X}$ whenever it is convenient, and speak indifferently of a class of functions or of a class of sets. A distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$ is *realizable by* $\mathcal{H}$ if $\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$.

Suprema over infinite classes raise the measurability issues described in Remark 1.6. We settle them once and for all with the first of the two remedies discussed there.

Assumption 4.1 (Countable approximation)**.** There is a countable subset $\mathcal{H}_0 \subseteq \mathcal{H}$ such that every $h \in \mathcal{H}$ is the pointwise limit of a sequence of elements of $\mathcal{H}_0$.

For $\{0,1\}$-valued functions pointwise convergence means that $h_j(x) = h(x)$ for all j large enough, for every x. Two consequences will be used repeatedly. First, on a finite set of points every $h \in \mathcal{H}$ coincides with some element of $\mathcal{H}_0$; hence every supremum over $\mathcal{H}$ of a quantity that depends on h only through finitely many values $h(x_i)$ and continuously (by dominated convergence) through $L_{\mathcal{D}}(h)$ – such as $\sup_h(L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h))$ – equals the corresponding supremum over $\mathcal{H}_0$, and is therefore a random variable. Second, an empirical risk minimiser always exists, because $L_{\mathcal{S}_n}(h)$ takes only the $n+1$ values $0, \frac{1}{n}, \dots, 1$, and it can be chosen *measurably*: enumerate $\mathcal{H}_0 = \{h_1, h_2, \dots\}$ and let $\widehat{h}_n$ be the first h_j attaining $\min_{h\in\mathcal{H}} L_{\mathcal{S}_n}(h)$, which exists by the first remark. The classes of finite VC dimension of Section 4.3.2 satisfy Assumption 4.1 (take parameters in $\mathbb{Q}$, approximating from the appropriate side); for the sine class, see Remark 4.25. We assume Assumption 4.1 throughout Sections 4.2, 4.5 and 4.6, and we say so again in the statements where it is used. The combinatorial results of Sections 4.3 and 4.4 and all the lower bounds need no measurability at all.

4.2.2 PAC learnability, revisited

Definition 2.63 introduced PAC learnability and the sample complexity for finite classes. Since the whole chapter is about sample complexities, and in particular about *lower* bounds on them, we need the definition in a more explicit form, in which the dependence on the algorithm is visible.

Definition 4.2 (Sample complexity of an algorithm and of a class)**.** Let $\mathcal{H}$ be a hypothesis class and $\mathcal{A} = (\mathcal{A}_n)_{n\geq 1}$ a learning algorithm (Definition 1.5); write $\widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n)$. For $\varepsilon, \delta \in (0,1)$ let $n_{\mathcal{A}}(\varepsilon,\delta) \in \mathbb{N} \cup \{+\infty\}$ be the smallest integer n_0 such that for every $n \geq n_0$ and every distribution $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$,

$$\mathbb{P}_{\mathcal{S}_n\sim\mathcal{D}^{\otimes n}}\left(L_{\mathcal{D}}(\widehat{h}_n) \;\geq\; \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\right) \;\leq\; \delta, \tag{4.1}$$

and let $n^{\mathrm{re}}_{\mathcal{A}}(\varepsilon,\delta)$ be defined in the same way with (4.1) required only for the distributions realizable by $\mathcal{H}$, for which it reads $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \geq \varepsilon) \leq \delta$.

(a) $\mathcal{A}$ *agnostically PAC learns* $\mathcal{H}$ if $n_{\mathcal{A}}(\varepsilon,\delta) < \infty$ for all $\varepsilon,\delta \in (0,1)$, and *PAC learns* $\mathcal{H}$ (in the realizable case) if $n^{\mathrm{re}}_{\mathcal{A}}(\varepsilon,\delta) < \infty$ for all ε,δ.

(b) The *agnostic* and *realizable sample complexities* of $\mathcal{H}$ are

$$n_{\mathcal{H}}(\varepsilon,\delta) := \inf_{\mathcal{A}} n_{\mathcal{A}}(\varepsilon,\delta), \qquad n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) := \inf_{\mathcal{A}} n^{\mathrm{re}}_{\mathcal{A}}(\varepsilon,\delta),$$

where the infimum is over all learning algorithms. $\mathcal{H}$ is *agnostically PAC learnable*, respectively *PAC learnable*, if some algorithm agnostically PAC learns $\mathcal{H}$, respectively PAC learns $\mathcal{H}$, in the sense of (a); then $n_{\mathcal{H}}$, respectively $n^{\mathrm{re}}_{\mathcal{H}}$, is finite for all $\varepsilon,\delta \in (0,1)$, and conversely (Remark 4.3).

Several comments are in order; the first one explains the relation with Definition 2.63.

Remark 4.3 (Relation with Definition 2.63). Definition 4.2 is Definition 2.63 with the quantifiers written out: an algorithm $\mathcal{A}$ satisfies the requirement of Definition 2.63 with a function $n_{\mathcal{H}}$ if and only if $n_{\mathcal{A}} \le n_{\mathcal{H}}$ pointwise, so the two notions of learnability coincide. The only new point is the definition of "the smallest such function": we take the pointwise infimum over algorithms, so that the algorithm achieving $n_{\mathcal{H}}(\varepsilon,\delta)$ may depend on (ε,δ). This is the convention under which *lower* bounds are meaningful and strongest: to prove $n_{\mathcal{H}}(\varepsilon,\delta) \ge N$ we shall show that for *every* algorithm $\mathcal{A}$ and *every* $n < N$ there is a distribution $\mathcal{D}$ for which (4.1) fails at sample size n; this forces $n_{\mathcal{A}}(\varepsilon,\delta) > n$ for all $n < N$, i.e. $n_{\mathcal{A}}(\varepsilon,\delta) \ge N$. Upper bounds, on the other hand, will always be achieved by one algorithm that does not depend on (ε,δ): empirical risk minimisation.

Allowing the algorithm to depend on (ε,δ) does not enlarge the set of learnable classes: if $n_{\mathcal{H}}$ is finite everywhere, a single algorithm agnostically PAC learns $\mathcal{H}$. Indeed, for $j \ge 1$ choose an algorithm $\mathcal{A}^{(j)}$ with $N_j := n_{\mathcal{A}^{(j)}}\big(\frac{1}{j+1},\frac{1}{j+1}\big) < \infty$. The integers $M_j := \max(N_1,\dots,N_j,j)$ are non-decreasing and tend to infinity; for $n \ge M_1$ let $j(n)$ be the largest j with $M_j \le n$, and let $\mathcal{A}_n := \mathcal{A}_n^{(j(n))}$ (and $\mathcal{A}_n := \mathcal{A}_n^{(1)}$ for $n < M_1$). Given $\varepsilon,\delta \in (0,1)$, choose j with $\frac{1}{j+1} \le \min(\varepsilon,\delta)$. Every $n \ge M_j$ satisfies $j(n) \ge j$ and $n \ge M_{j(n)} \ge N_{j(n)}$, so that, for every $\mathcal{D}$, since $\frac{1}{j(n)+1} \le \frac{1}{j+1} \le \min(\varepsilon,\delta)$,

$$\mathbb{P}\Big(L_{\mathcal{D}}(\widehat{h}_n) \ge \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\Big) \;\le\; \mathbb{P}\Big(L_{\mathcal{D}}(\widehat{h}_n) \ge \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \tfrac{1}{j(n)+1}\Big) \;\le\; \tfrac{1}{j(n)+1} \;\le\; \delta;$$

that is, $n_{\mathcal{A}}(\varepsilon,\delta) \le M_j < \infty$. The same construction applies verbatim to $n_{\mathcal{H}}^{\mathrm{re}}$. (Under Assumption 4.1 there is a second proof: if $n_{\mathcal{H}}^{\mathrm{re}}$ is finite everywhere, then $\mathrm{VCdim}(\mathcal{H}) < \infty$ by Theorem 4.28, and every measurable empirical risk minimiser agnostically PAC learns $\mathcal{H}$ by Theorem 4.59.)

Remark 4.4 (Distribution-free, and the order of the quantifiers). As already noted in Remark 2.64 – and, for the no-free-lunch theorem, in Remark 3.73(a) – the sample size n_0 of Definition 4.2 is chosen *before* the distribution: the same n_0 must work for every $\mathcal{D}$. This is what makes the theory distribution-free – a guarantee holds whatever the unknown law of the data – and also what makes it pessimistic, since a single bad distribution suffices to make the sample complexity large. Exchanging the quantifiers gives a much weaker notion: if for every $\mathcal{D}$ there is an $n_0(\varepsilon,\delta,\mathcal{D})$, one speaks of *universal consistency* when $\mathcal{H}$ is the set of all classifiers. The k-nearest-neighbour rule of Section 1.10.1 is universally consistent [Sto77], yet the class of all classifiers on an infinite domain is not PAC learnable (Theorem 4.28 below): the two notions are genuinely different, and the no-free-lunch theorem of Chapter 3 is precisely the statement that no rate can be uniform over all distributions. Between the two extremes lie *non-uniform* learnability, where n_0 may depend on the target hypothesis but not on $\mathcal{D}$ otherwise [SB14, Chap. 7], and the recent theory of *universal learning rates* [Bou+21].

Remark 4.5 (Computational requirements). As recalled in Remark 2.64, Valiant's definition [Val84] also asks for an algorithm running in time polynomial in $1/\varepsilon$, $1/\delta$ and the size of the examples. This chapter is purely statistical: we ask how many examples are necessary and sufficient, not how much computation. The two questions are very different. Empirical risk minimisation over half-spaces is NP-hard in the agnostic case [HSV95], while it is a linear program in the realizable case; and, under cryptographic assumptions, there are classes of finite VC dimension – hence learnable with few examples, by the results of this chapter – that cannot be learned in polynomial time by any algorithm [KV94].

Remark 4.6 (Proper and improper learners). Definition 4.2 does not require $\widehat{h}_n \in \mathcal{H}$: the algorithm may output any measurable classifier, and $\mathcal{H}$ only serves as the benchmark.

Learners with values in $\mathcal{H}$ are called *proper*; empirical risk minimisation is proper. Our upper bounds are achieved by proper learners and our lower bounds hold for all learners, improper ones included, so the distinction plays no role in the fundamental theorem. It does matter for the finer question of the *optimal* realizable sample complexity (Remark 4.57).

4.2.3 Uniform convergence

The proof of Theorem 2.65 had two steps: a uniform bound on $\left|L_{S_n}(h) - L_{\mathcal{D}}(h)\right|$, then a three-term comparison showing that a uniform bound makes empirical risk minimisation work. We isolate the second step, which has nothing to do with finiteness.

Definition 4.7 (Representative sample)**.** Let $\varepsilon > 0$. A sample $s = ((x_1, y_1), \dots, (x_n, y_n))$ is *ε-representative* for $\mathcal{H}$ and $\mathcal{D}$ if

$$\sup_{h \in \mathcal{H}} \left| L_s(h) - L_{\mathcal{D}}(h) \right| \;\le\; \varepsilon .$$

Lemma 4.8 (Representative samples make ERM work)**.** *If s is $\frac{\varepsilon}{2}$-representative, then every $\widehat{h} \in \mathcal{H}$ with $L_s(\widehat{h}) = \min_{h \in \mathcal{H}} L_s(h)$ satisfies $L_{\mathcal{D}}(\widehat{h}) \le \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$.*

Proof. For every $h \in \mathcal{H}$, $L_{\mathcal{D}}(\widehat{h}) \le L_s(\widehat{h}) + \frac{\varepsilon}{2} \le L_s(h) + \frac{\varepsilon}{2} \le L_{\mathcal{D}}(h) + \varepsilon$, where the first and last inequalities use the representativeness and the middle one the minimality of $\widehat{h}$. Take the infimum over h. □

Note that only half of the representativeness has been used for $\widehat{h}$ – the one-sided bound $L_{\mathcal{D}}(\widehat{h}) - L_s(\widehat{h}) \le \frac{\varepsilon}{2}$ – and the other half only for a single, fixed comparison hypothesis. We shall exploit this asymmetry in Theorem 4.50.

Definition 4.9 (Uniform convergence property)**.** A hypothesis class $\mathcal{H}$ has the *uniform convergence property* if for all $\varepsilon, \delta \in (0,1)$ the quantity

$$n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) := \min\Big\{ n_0 \in \mathbb{N} : \; \forall n \ge n_0, \; \forall \mathcal{D}, \; \mathbb{P}_{S_n \sim \mathcal{D}^{\otimes n}}\big(S_n \text{ is } \varepsilon\text{-representative}\big) \ge 1 - \delta \Big\}$$

is finite. The function $n_{\mathcal{H}}^{\mathrm{UC}}$ is the *uniform convergence sample complexity* of $\mathcal{H}$.

In words: the empirical risks converge to the risks uniformly over $\mathcal{H}$, in probability, at a rate that does not depend on $\mathcal{D}$. The classical name for this property is a *uniform Glivenko–Cantelli* property of the class of sets $\{(x,y) : h(x) \neq y\}$, $h \in \mathcal{H}$: for $\mathcal{H}$ the class of half-lines $\mathbb{1}_{(-\infty, a]}$ on $\mathbb{R}$ and $Y \equiv 0$, the empirical risks are the values of the empirical distribution function, and uniform convergence is the Glivenko–Cantelli theorem, with a rate uniform over the distribution of X.

Corollary 4.10 (Uniform convergence implies agnostic learnability)**.** *Under Assumption 4.1, if $\mathcal{H}$ has the uniform convergence property, then $\mathcal{H}$ is agnostically PAC learnable by any empirical risk minimiser, and for all $\varepsilon, \delta \in (0,1)$ and every $\varepsilon' \in (0, \varepsilon)$,*

$$n_{\mathcal{H}}(\varepsilon, \delta) \;\le\; n_{ERM}(\varepsilon, \delta) \;\le\; n_{\mathcal{H}}^{\mathrm{UC}}\big(\tfrac{\varepsilon'}{2}, \delta\big),$$

where n_{ERM} denotes the sample complexity $n_{\mathcal{A}}$ of any (measurable) empirical risk minimiser $\mathcal{A}$.

Proof. Let $n \ge n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon'/2, \delta)$ and let $\mathcal{D}$ be arbitrary. With probability at least $1 - \delta$ the sample is $\frac{\varepsilon'}{2}$-representative, and on that event $L_{\mathcal{D}}(\widehat{h}_n) \le \inf_h L_{\mathcal{D}}(h) + \varepsilon' < \inf_h L_{\mathcal{D}}(h) + \varepsilon$ by Lemma 4.8. Hence $\mathbb{P}\big(L_{\mathcal{D}}(\widehat{h}_n) \ge \inf_h L_{\mathcal{D}}(h) + \varepsilon\big) \le \delta$. □

The auxiliary $\varepsilon' < \varepsilon$ only reconciles the non-strict inequality in (4.1) with the non-strict inequality in Definition 4.7; since all the explicit bounds on $n_{\mathcal{H}}^{\mathrm{UC}}$ below are continuous in ε, it plays no role, and we shall write $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2, \delta)$ loosely.

Remark 4.11 (Uniform convergence is sufficient; is it necessary?). For binary classification with the 0–1 loss, the fundamental theorem will show that uniform convergence is also *necessary* for learnability: the two properties are equivalent. This is a special feature of the setting. For general learning problems – e.g. stochastic convex optimisation – there are learnable problems without the uniform convergence property, and even problems that are learnable but not by any empirical risk minimiser [Sha+10]. In such problems the right notion is *stability* of the learning algorithm rather than uniform convergence over the class.

4.2.4 Finite classes

The results of Section 2.7.1 read as follows in the language of this chapter.

Proposition 4.12 (Sample complexities of a finite class)**.** *Let $\mathcal{H}$ be finite. Then*

$$n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) \le \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil, \qquad n_{\mathcal{H}}(\varepsilon, \delta) \le \left\lfloor \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rfloor + 1,$$
$$n_{\mathcal{H}}^{\mathrm{re}}(\varepsilon, \delta) \le \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil,$$

the last two being achieved by empirical risk minimisation.

Proof. The first bound follows from (2.35), whose right-hand side is at most ε as soon as $n \ge \log(2|\mathcal{H}|/\delta)/(2\varepsilon^2)$. For the second, if $n > 2\log(2|\mathcal{H}|/\delta)/\varepsilon^2$, there is an $\varepsilon' < \varepsilon$ such that $n \ge 2\log(2|\mathcal{H}|/\delta)/\varepsilon'^2$, and we may apply Corollary 4.10 with this ε'. For the third, let $\mathcal{D}$ be realizable; since $\mathcal{H}$ is finite the infimum is attained, by some $\bar{h}$ with $L_{\mathcal{D}}(\bar{h}) = 0$, and $L_{S_n}(\bar{h}) = 0$ almost surely, so that an empirical risk minimiser has zero empirical risk. If $\mathcal{H}_\varepsilon = \{h \in \mathcal{H} : L_{\mathcal{D}}(h) \ge \varepsilon\}$, then for each $h \in \mathcal{H}_\varepsilon$

$$\mathbb{P}\big(L_{S_n}(h) = 0\big) = \big(1 - L_{\mathcal{D}}(h)\big)^n \le (1-\varepsilon)^n \le e^{-n\varepsilon},$$

and the union bound gives $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \ge \varepsilon) \le \mathbb{P}(\exists h \in \mathcal{H}_\varepsilon : L_{S_n}(h) = 0) \le |\mathcal{H}|\, e^{-n\varepsilon}$, which is at most δ for $n \ge \log(|\mathcal{H}|/\delta)/\varepsilon$. This is the proof of Theorem 2.66. □

The realizable bound is better by a factor $1/\varepsilon$: in a noiseless problem a bad hypothesis is eliminated as soon as it makes one mistake, which happens after about $1/L_{\mathcal{D}}(h)$ examples, whereas in the agnostic case two hypotheses whose risks differ by ε can only be told apart after about $1/\varepsilon^2$ examples. The same dichotomy will reappear, with $\log|\mathcal{H}|$ replaced by the VC dimension, in Theorem 4.65.

4.3 Shattering and the Vapnik–Chervonenkis dimension

4.3.1 Definitions

An infinite class may behave, on any given sample, like a finite one: what matters for the empirical risks is not the class itself but the set of labellings it induces on the sample points.

Definition 4.13 (Traces, shattering)**.** Let $\mathcal{H}$ be a hypothesis class. For a finite sequence $x = (x_1, \dots, x_n) \in \mathcal{X}^n$, the *trace* of $\mathcal{H}$ on x is

$$\mathcal{H}(x) := \big\{(h(x_1), \dots, h(x_n)) : \ h \in \mathcal{H}\big\} \subseteq \{0,1\}^n.$$

For a finite set $C \subseteq \mathcal{X}$ we write $\mathcal{H}_C := \{h|_C : h \in \mathcal{H}\} \subseteq \{0,1\}^C$, which is the trace on any enumeration of C. The class $\mathcal{H}$ *shatters* C if $\mathcal{H}_C = \{0,1\}^C$, that is, if every labelling of the points of C is realised by some $h \in \mathcal{H}$.

Definition 4.14 (Growth function)**.** The *growth function* of $\mathcal{H}$ is

$$\tau_{\mathcal{H}}(n) := \max_{x \in \mathcal{X}^n} |\mathcal{H}(x)|, \qquad n \geq 1.$$

The maximum exists since $|\mathcal{H}(x)| \leq 2^n$. Repeated points do not help: if x has only $m < n$ distinct coordinates, $|\mathcal{H}(x)|$ equals the size of the trace on these m points. Hence $\tau_{\mathcal{H}}(n)$ is also the maximum of $|\mathcal{H}_C|$ over sets C of size *at most* n, and it is the maximum over sets of size exactly n when $|\mathcal{X}| \geq n$. Obviously $\tau_{\mathcal{H}}(n) \leq \min(2^n, |\mathcal{H}|)$; $\tau_{\mathcal{H}}$ is non-decreasing; and $\tau_{\mathcal{H}}(n+n') \leq \tau_{\mathcal{H}}(n)\tau_{\mathcal{H}}(n')$, since a pattern on $n+n'$ points is determined by its restrictions to the first n and to the last n'.

Definition 4.15 (Vapnik–Chervonenkis dimension)**.** The *VC dimension* of a hypothesis class $\mathcal{H}$ is

$$\operatorname{VCdim}(\mathcal{H}) := \sup\bigl\{|C| : C \subseteq \mathcal{X} \text{ finite, shattered by } \mathcal{H}\bigr\} \in \mathbb{N} \cup \{+\infty\},$$

with the convention that the empty set is shattered by every non-empty class (its only labelling is the empty function).

Since a subset of a shattered set is shattered, $\operatorname{VCdim}(\mathcal{H}) \geq d$ if and only if some set of size d is shattered, and $\operatorname{VCdim}(\mathcal{H}) = d < \infty$ if and only if

$$\tau_{\mathcal{H}}(n) = 2^n \text{ for } n \leq d \qquad \text{and} \qquad \tau_{\mathcal{H}}(n) < 2^n \text{ for } n > d.$$

To prove $\operatorname{VCdim}(\mathcal{H}) = d$ one must therefore do two different things: exhibit *one* set of d points that is shattered, and show that *no* set of $d+1$ points is. The asymmetry of the quantifiers – "there exists" for the lower bound, "for all" for the upper bound – is the source of most mistakes in computing VC dimensions. Note also that $\operatorname{VCdim}(\mathcal{H}) = 0$ means that no point is shattered, i.e. all $h \in \mathcal{H}$ take the same value at every point: $\mathcal{H}$ is reduced to a single function.

Remark 4.16 (What the VC dimension is not). It is tempting to read $\operatorname{VCdim}(\mathcal{H})$ as "the number of parameters" of $\mathcal{H}$. The examples below show that this is often numerically true (half-spaces of $\mathbb{R}^k$ have $k+1$ parameters and VC dimension $k+1$) and sometimes spectacularly false: the class $\{x \mapsto \mathbb{1}\{\sin(\omega x) > 0\}\}$ has one parameter and infinite VC dimension (Proposition 4.24), while the class $\{\mathbb{1}_{(-\infty, a_1+\cdots+a_k]} : a \in \mathbb{R}^k\}$ has k parameters and VC dimension 1 (it is the class of half-lines of Example 4.17), and a finite class may have VC dimension much smaller than $\log_2 |\mathcal{H}|$ (Example 4.23). The VC dimension is a combinatorial notion of size, blind to parametrisation.

4.3.2 Examples

Example 4.17 (Half-lines)**.** Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\}$. Then $\operatorname{VCdim}(\mathcal{H}) = 1$ and $\tau_{\mathcal{H}}(n) = n+1$.

Indeed, a single point x is shattered ($a < x$ and $a \geq x$ give the two labels). If $x_1 < x_2$, the labelling $(0,1)$ is impossible, since $h(x_2) = 1$ means $a \geq x_2 > x_1$, whence $h(x_1) = 1$. On n distinct points $x_{(1)} < \cdots < x_{(n)}$ the traces are exactly the patterns $(1,\dots,1,0,\dots,0)$ with $j \in \{0,\dots,n\}$ ones, so $\tau_{\mathcal{H}}(n) = n+1$. Assumption 4.1 holds with $a \in \mathbb{Q}$, approximating a from above.

Example 4.18 (Intervals)**.** Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{\mathbb{1}_{[a,b]} : a \le b\}$. Then $\mathrm{VCdim}(\mathcal{H}) = 2$ and $\tau_{\mathcal{H}}(n) = 1 + n + \binom{n}{2}$.

Two points $x_1 < x_2$ are shattered: take $[a, b]$ containing none, one or both of them. Three points $x_1 < x_2 < x_3$ are not, since an interval containing x_1 and x_3 contains x_2: the labelling $(1, 0, 1)$ is impossible. On n distinct ordered points the traces are the empty pattern and the patterns whose ones form a block $\{i, i+1, \dots, j\}$, $1 \le i \le j \le n$; there are $1 + \binom{n+1}{2} = 1 + n + \binom{n}{2}$ of them.

In both examples $\tau_{\mathcal{H}}(n) = \sum_{i=0}^{d} \binom{n}{i}$ exactly. This is no coincidence: we shall see in Section 4.4 that this is the largest growth function compatible with VC dimension d.

Proposition 4.19 (Axis-parallel boxes)**.** *Let $\mathcal{X} = \mathbb{R}^k$ and let $\mathcal{H}$ be the class of indicators of boxes $\prod_{j=1}^{k} [a_j, b_j]$, $a_j \le b_j$. Then* $\mathrm{VCdim}(\mathcal{H}) = 2k$.

Proof. Lower bound. Let $C = \{\pm e_1, \dots, \pm e_k\}$, where (e_j) is the canonical basis. Let $A \subseteq C$. If $A = \emptyset$, a box far away from C realises the empty labelling. Otherwise let B_A be the smallest box containing A: its j-th side is the interval between the smallest and the largest j-th coordinate of the points of A. These coordinates belong to $\{-1, 0, 1\}$, and the value 1 (resp. -1) is reached only if $e_j \in A$ (resp. $-e_j \in A$). So if $e_j \notin A$, the j-th side of B_A lies in $(-\infty, 0]$ and $e_j \notin B_A$; similarly for $-e_j$. Hence $B_A \cap C = A$, and C is shattered.

Upper bound. Let C be a set of $2k + 1$ distinct points. For each coordinate j, choose a point of C with maximal j-th coordinate and one with minimal j-th coordinate. The set A of chosen points has at most $2k$ elements, so some $c \in C \setminus A$ remains, and c lies in the smallest box containing A (each of its coordinates is between the chosen minimum and maximum). Any box containing A contains this smallest box, hence c: the labelling $\mathbb{1}_A$ is not realised, and C is not shattered. □

The case of half-spaces requires a classical result of convex geometry.

Lemma 4.20 (Radon's theorem)**.** *Any $k + 2$ points $x_1, \dots, x_{k+2}$ of $\mathbb{R}^k$ (not necessarily distinct) can be partitioned into two sets $I \sqcup J = \{1, \dots, k + 2\}$ whose convex hulls intersect:* $\mathrm{conv}\{x_i : i \in I\} \cap \mathrm{conv}\{x_j : j \in J\} \neq \emptyset$.

Proof. The $k + 2$ vectors $(x_i, 1) \in \mathbb{R}^{k+1}$ are linearly dependent: there exist $\alpha_1, \dots, \alpha_{k+2}$, not all zero, with $\sum_i \alpha_i x_i = 0$ and $\sum_i \alpha_i = 0$. Let $I = \{i : \alpha_i > 0\}$ and $J = \{j : \alpha_j \le 0\}$. Since the α_i sum to zero and are not all zero, both I and $\{j : \alpha_j < 0\} \subseteq J$ are non-empty, and $s := \sum_{i \in I} \alpha_i = \sum_{j \in J} (-\alpha_j) > 0$. The point

$$p := \sum_{i \in I} \frac{\alpha_i}{s} x_i = \sum_{j \in J} \frac{-\alpha_j}{s} x_j$$

is a convex combination of the x_i, $i \in I$, and also of the x_j, $j \in J$. □

Proposition 4.21 (Half-spaces)**.** *Let $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{H} = \{x \mapsto \mathbb{1}\{\langle w, x\rangle + b \ge 0\} : w \in \mathbb{R}^k, b \in \mathbb{R}\}$ be the class of affine half-spaces. Then* $\mathrm{VCdim}(\mathcal{H}) = k + 1$.

Proof. Lower bound. Let $C = \{0, e_1, \dots, e_k\}$ and let $(y_0, y_1, \dots, y_k) \in \{0, 1\}^{k+1}$ be a labelling. Put $b = y_0 - \frac{1}{2}$ and $w_j = 2y_j - 1 - b$. Then $\langle w, 0\rangle + b = b$ is ≥ 0 if and only if $y_0 = 1$, and $\langle w, e_j\rangle + b = 2y_j - 1$ is ≥ 0 if and only if $y_j = 1$. So C is shattered.

Upper bound. Let $x_1, \dots, x_{k+2}$ be distinct points and let $I \sqcup J$ be a Radon partition (Lemma 4.20). Suppose some $h = \mathbb{1}\{\langle w, \cdot\rangle + b \ge 0\}$ labels the x_i, $i \in I$, by 1 and the x_j, $j \in J$, by 0. The sets $\{\langle w, \cdot\rangle + b \ge 0\}$ and $\{\langle w, \cdot\rangle + b < 0\}$ are convex and disjoint; the first contains $\mathrm{conv}\{x_i : i \in I\}$ and the second $\mathrm{conv}\{x_j : j \in J\}$, which contradicts Lemma 4.20. So no set of $k + 2$ points is shattered. □

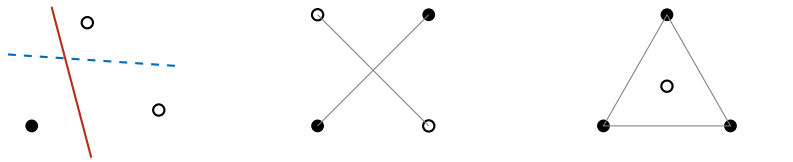


Figure 4.1: Half-planes in $\mathbb{R}^2$ ($k = 2$). Left: three points in general position are shattered; the solid line realises the labelling shown (filled point labelled 1), the dashed line the labelling in which only the top point is labelled 1. Middle and right: the two possible configurations of four points (up to degenerate cases) and, for each, a labelling that no half-plane realises. In both cases the filled and the empty points form a Radon partition: their convex hulls meet.

Assumption 4.1 holds for half-spaces: approximating (w, b) by rational (w_m, b_m) with $\|w_m - w\| \le 1/m^2$ and $b_m - b \in [1/m, 2/m]$, the inequality $\langle w_m, x\rangle + b_m \ge 0$ holds eventually at every x with $\langle w, x\rangle + b \ge 0$ (because $\langle w_m - w, x\rangle + b_m - b \ge 1/m - \|x\|/m^2 \ge 0$ for $m \ge \|x\|$) and fails eventually at every x with $\langle w, x\rangle + b < 0$. The same argument applies to boxes.

Remark 4.22 (Counting dichotomies: Cover's theorem). For homogeneous half-spaces $\{x \mapsto \mathbb{1}\{\langle w, x\rangle \ge 0\}\}$ in $\mathbb{R}^k$ the VC dimension is k (Exercise 4.5), and the number of labellings of n points in general position (any k of them linearly independent) by *strict* homogeneous half-spaces is exactly $2\sum_{i=0}^{k-1}\binom{n-1}{i}$, independently of the configuration [Cov65]; we admit this formula. For $n = 2k$ this is 2^{n-1}: a random labelling of $2k$ points in general position is linearly separable with probability exactly $\frac{1}{2}$. This "capacity" $2k$ of the perceptron is one of the oldest results in the field, and a precursor of the Sauer–Shelah lemma.

Example 4.23 (Finite classes)**.** If $\mathcal{H}$ is finite, $\mathrm{VCdim}(\mathcal{H}) \le \lfloor \log_2 |\mathcal{H}| \rfloor$, since shattering a set of size m requires 2^m distinct functions. The inequality can be very far from an equality: the class of singletons $\{\mathbb{1}_{\{z\}} : z \in \{1, \dots, N\}\}$ on $\mathcal{X} = \{1, \dots, N\}$, $N \ge 3$, has VC dimension 1 (a point is shattered, but no pair, since the labelling $(1, 1)$ is impossible), while $\log_2 |\mathcal{H}| = \log_2 N$ is arbitrarily large. In this sense the VC dimension is a much finer measure of size than the cardinality.

Proposition 4.24 (A one-parameter class of infinite VC dimension)**.** *Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{x \mapsto \mathbb{1}\{\sin(\omega x) > 0\} : \omega \in \mathbb{R}\}$. Then $\mathrm{VCdim}(\mathcal{H}) = \infty$. More precisely, for every $m \ge 1$ the set $\{2^{-1}, 2^{-2}, \dots, 2^{-m}\}$ is shattered, by frequencies ω that are integer multiples of π.*

Proof. Let $y \in \{0, 1\}^m$ be a labelling, put $c_i = 1 - y_i$ and

$$\omega := \pi\Big(1 + \sum_{i=1}^{m} c_i 2^i\Big).$$

Fix $j \in \{1, \dots, m\}$. Then

$$\omega 2^{-j} = \pi \sum_{i>j} c_i 2^{i-j} + \pi c_j + \pi\Big(2^{-j} + \sum_{i<j} c_i 2^{i-j}\Big).$$

The first term is an even multiple of π, since 2^{i-j} is an even integer for $i > j$. The last term, $\pi\theta_j$, satisfies $0 < \theta_j \le 2^{-j} + \sum_{i=1}^{j-1} 2^{i-j} = 2^{-j} + 1 - 2^{1-j} = 1 - 2^{-j} < 1$. Hence $\sin(\omega 2^{-j}) = \sin(\pi c_j + \pi\theta_j) = (-1)^{c_j}\sin(\pi\theta_j)$, with $\sin(\pi\theta_j) > 0$: it is positive if and only if $c_j = 0$, i.e. $y_j = 1$. Every labelling is realised. □

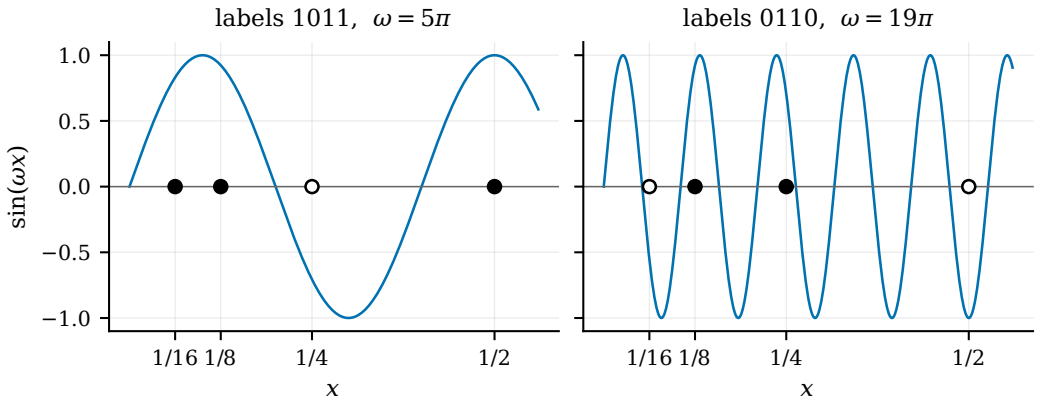


Figure 4.2: Two of the sixteen labellings of $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$ realised by $x \mapsto \mathbb{1}\{\sin(\omega x) > 0\}$, with the frequencies of Proposition 4.24 (filled points are labelled 1). For the labelling $y = (y_1, \dots, y_4)$ of $(2^{-1}, \dots, 2^{-4})$, $\omega/\pi = 1 + \sum_i (1 - y_i)2^i$.

The proof is a binary-expansion trick: the integer $\sum_i c_i 2^i$ encodes the labelling, and the point 2^{-j} reads its j-th binary digit through the parity of $\lfloor \omega 2^{-j}/\pi \rfloor$ (Figure 4.2). A single real parameter can store infinitely many bits, which is why counting parameters does not measure complexity (Remark 4.16).

Remark 4.25 (Measurability of the sine class). The proof uses only the countable subclass $\mathcal{H}' = \{x \mapsto \mathbb{1}\{\sin(\pi j x) > 0\} : j \in \mathbb{N}\}$, which therefore also has infinite VC dimension and satisfies Assumption 4.1 trivially. For the full class, pointwise approximation by rational frequencies fails at the zeros of $\sin(\omega \cdot)$; but since the only use we make of the sine class is the negative result of Theorem 4.28, whose proof requires no measurability, this is immaterial.

Example 4.26 (Convex sets)**.** Let $\mathcal{X} = \mathbb{R}^2$ and let $\mathcal{H}$ be the class of indicators of convex polygons (with any number of vertices), or of all closed convex sets. Then $\mathrm{VCdim}(\mathcal{H}) = \infty$: any finite set C of points on the unit circle is shattered, since for $A \subseteq C$ the polygon $\mathrm{conv}(A)$ contains no point of $C \setminus A$ (a point of the circle is an extreme point of the disc, hence is not a convex combination of other points of the disc). By contrast, convex polygons with at most m vertices form a class of finite VC dimension: Proposition 4.37 below bounds it by $6m \log_2(3m)$, with no geometry at all.

Table 4.1 summarises the examples, together with some of the exercises.

Remark 4.27 (Neural networks). For feed-forward networks with W real weights and the threshold activation $t \mapsto \mathbb{1}\{t \geq 0\}$, the VC dimension is $O(W \log W)$ [BH89], by a growth-function argument of the same nature as Proposition 4.37 applied layer by layer; for networks with piecewise linear activations such as $t \mapsto \max(t, 0)$ and q layers, it is of order $Wq \log W$, up to the replacement of $\log W$ by $\log(W/q)$ in the lower bound [Bar+19] (both admitted). Modern networks are trained in the regime $W \gg n$, where these bounds – and therefore all the upper bounds of this chapter – are vacuous, although such networks often generalise well. This does not contradict the lower bounds of Section 4.6.2, which are worst-case over distributions; it shows that the VC dimension of the architecture is not the quantity that explains their behaviour, and motivates the norm- and margin-based complexity measures of Section 6.1 and of later chapters.

class	domain	VCdim	reference
half-lines $\mathbb{1}_{(-\infty,a]}$	$\mathbb{R}$	1	Example 4.17
two-sided thresholds	$\mathbb{R}$	2	Exercise 4.2
intervals $\mathbb{1}_{[a,b]}$	$\mathbb{R}$	2	Example 4.18
unions of N intervals	$\mathbb{R}$	$2N$	Exercise 4.3
axis-parallel boxes	$\mathbb{R}^k$	$2k$	Proposition 4.19
homogeneous half-spaces	$\mathbb{R}^k$	k	Exercise 4.5
affine half-spaces	$\mathbb{R}^k$	$k+1$	Proposition 4.21
Euclidean balls	$\mathbb{R}^k$	$k+1$ (*)	Exercise 4.6
$\{g \geq 0\}$, $g \in V$, $\dim V = m$	any	$\leq m$	Exercise 4.7
singletons	$\mathbb{R}$	1	Exercise 4.1
J-fold combinations of VC-d classes	any	$< 2dJ \log_2(3J)$	Proposition 4.37
finite class	any	$\leq \log_2 \lvert\mathcal{H}\rvert$	Example 4.23
$\mathbb{1}\{\sin(\omega x) > 0\}$	$\mathbb{R}$	∞	Proposition 4.24
convex polygons, convex sets	$\mathbb{R}^2$	∞	Example 4.26
finite subsets	$\mathbb{N}$	∞	Exercise 4.4

Table 4.1: VC dimensions of the classes met in this chapter. (*): only the bounds $k+1 \leq \mathrm{VCdim} \leq k+2$ are proved here; the value $k+1$ is due to Dudley [Dud79].

4.3.3 Infinite VC dimension forbids learning

The no-free-lunch theorem of Chapter 3 (Theorem 3.72) states that for every learning algorithm $\mathcal{A}$ over a domain $\mathcal{X}'$ whose singletons are measurable and every sample size $n \leq |\mathcal{X}'|/2$, there exist a distribution $\mathcal{D}'$ on $\mathcal{X}' \times \{0,1\}$ and a function $f : \mathcal{X}' \to \{0,1\}$ with $L_{\mathcal{D}'}(f) = 0$ such that $\mathbb{P}_{S_n \sim \mathcal{D}'^{\otimes n}}\left(L_{\mathcal{D}'}(\mathcal{A}_n(S_n)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$. Applied to a shattered set, it gives the "only if" half of the fundamental theorem.

Theorem 4.28 (Infinite VC dimension implies non-learnability)**.** *Let $\mathcal{H}$ be a hypothesis class and $n \geq 1$. If $\mathcal{H}$ shatters a set of size $2n$, then for every learning algorithm $\mathcal{A}$ there is a distribution $\mathcal{D}$ realizable by $\mathcal{H}$ – indeed with $L_{\mathcal{D}}(h) = 0$ for some $h \in \mathcal{H}$ – such that*

$$\mathbb{P}_{S_n \sim \mathcal{D}^{\otimes n}}\left(L_{\mathcal{D}}(\mathcal{A}_n(S_n)) \geq \tfrac{1}{8}\right) \;\geq\; \tfrac{1}{7}.$$

Consequently, if $\mathrm{VCdim}(\mathcal{H}) = \infty$ then $n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) = \infty$ for all $\varepsilon \leq \frac{1}{8}$ and $\delta < \frac{1}{7}$: $\mathcal{H}$ is not PAC learnable, and a fortiori not agnostically PAC learnable. If $\mathrm{VCdim}(\mathcal{H}) = d < \infty$, then $n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) > \lfloor d/2 \rfloor$ for the same (ε,δ).

Proof. Let C be shattered with $|C| = 2n$, and let $\mathcal{A}^C$ be the algorithm on the domain C defined by $\mathcal{A}^C_n(s) := \mathcal{A}_n(s)|_C$ for $s \in (C \times \{0,1\})^n$. We equip C with its discrete σ-field: its singletons are then measurable, as Theorem 3.72 requires, $\mathcal{A}^C$ is a learning algorithm on C, and the inclusion $C \hookrightarrow \mathcal{X}$ is measurable whatever the σ-field of $\mathcal{X}$. Theorem 3.72 applied to $\mathcal{X}' = C$ and $\mathcal{A}^C$ provides $\mathcal{D}'$ and $f : C \to \{0,1\}$ as above. Let $\mathcal{D}$ be the image of $\mathcal{D}'$ under the inclusion $C \times \{0,1\} \hookrightarrow \mathcal{X} \times \{0,1\}$, so that a sample from $\mathcal{D}$ is a sample from $\mathcal{D}'$ viewed in $\mathcal{X} \times \{0,1\}$, and for every classifier g on $\mathcal{X}$, $L_{\mathcal{D}}(g) = L_{\mathcal{D}'}(g|_C)$; in particular $L_{\mathcal{D}}(\mathcal{A}_n(S_n)) = L_{\mathcal{D}'}(\mathcal{A}^C_n(S_n))$, which is at least $\frac{1}{8}$ with probability at least $\frac{1}{7}$. Since C is shattered there is $h \in \mathcal{H}$ with $h|_C = f$, so $L_{\mathcal{D}}(h) = L_{\mathcal{D}'}(f) = 0$ and $\mathcal{D}$ is realizable.

If $\mathrm{VCdim}(\mathcal{H}) = \infty$, this holds for every n, so for $\varepsilon \leq \frac{1}{8}$ and $\delta < \frac{1}{7}$ condition (4.1) fails at every sample size, for every algorithm. If $\mathrm{VCdim}(\mathcal{H}) = d$, it holds for every $n \leq d/2$, so $n^{\mathrm{re}}_{\mathcal{A}}(\varepsilon,\delta) > \lfloor d/2 \rfloor$ for every $\mathcal{A}$. □

Remark 4.29 (What the argument does and does not say). The bad distribution depends on n, as it must: for each *fixed* $\mathcal{D}$, many algorithms (for instance a nearest-neighbour rule

on a countable domain) eventually learn. Theorem 4.28 says that no single sample size works for all distributions simultaneously – a statement about the order of the quantifiers of Remark 4.4. The second part, $n^{\mathrm{re}}_{\mathcal{H}} > \lfloor d/2 \rfloor$, is a first lower bound for classes of finite VC dimension, but a weak one: it does not involve ε. The correct lower bounds, d/ε in the realizable case and d/ε^2 in the agnostic case, are proved in Section 4.6.2 with the same idea – a shattered set on which the labels are chosen at random – refined by the minimax tools of Chapter 3.

4.4 The Sauer–Shelah lemma

By definition, $\tau_{\mathcal{H}}(n) = 2^n$ for $n \le \mathrm{VCdim}(\mathcal{H})$ and $\tau_{\mathcal{H}}(n) < 2^n$ beyond. The remarkable fact is that the growth function does not merely fall short of 2^n after $d = \mathrm{VCdim}(\mathcal{H})$: it collapses to a polynomial of degree d. Throughout, we write

$$\Phi_d(n) := \sum_{i=0}^{d} \binom{n}{i},$$

with the convention $\binom{n}{i} = 0$ for $i > n$, so that $\Phi_d(n) = 2^n$ for $n \le d$.

4.4.1 The lemma

We prove a statement about arbitrary families of binary vectors, which is stronger than what is needed and has a simpler induction.

Theorem 4.30 (Sauer–Shelah lemma, Pajor's form)**.** *Let C be a finite set and $\mathcal{V} \subseteq \{0,1\}^C$. Say that $\mathcal{V}$* shatters *$B \subseteq C$ if $\{f|_B : f \in \mathcal{V}\} = \{0,1\}^B$. Then*

$$|\mathcal{V}| \;\le\; \bigl|\{B \subseteq C : \mathcal{V} \text{ shatters } B\}\bigr|.$$

Proof. Write $\mathrm{sh}(\mathcal{V})$ for the set of subsets of C shattered by $\mathcal{V}$. We argue by induction on $|C|$. If $C = \emptyset$, then $\{0,1\}^C$ has a single element; either $\mathcal{V} = \emptyset$ and the claim is $0 \le |\mathrm{sh}(\mathcal{V})|$, or $|\mathcal{V}| = 1$, $\mathcal{V}$ shatters $\emptyset$, and the claim is $1 \le 1$.

Let $|C| \ge 1$, fix $c \in C$ and put $C' = C \setminus \{c\}$. We write elements of $\{0,1\}^C$ as pairs (g,t) with $g \in \{0,1\}^{C'}$ the restriction to C' and $t \in \{0,1\}$ the value at c. Define two families on C':

$$\mathcal{V}_0 := \bigl\{g : (g,0) \in \mathcal{V} \text{ or } (g,1) \in \mathcal{V}\bigr\}, \qquad \mathcal{V}_1 := \bigl\{g : (g,0) \in \mathcal{V} \text{ and } (g,1) \in \mathcal{V}\bigr\}.$$

Each $g \in \mathcal{V}_0$ has one or two extensions in $\mathcal{V}$, and two exactly when $g \in \mathcal{V}_1$; hence

$$|\mathcal{V}| = |\mathcal{V}_0| + |\mathcal{V}_1|\,. \tag{4.2}$$

The family $\mathcal{V}_0$. It is the restriction of $\mathcal{V}$ to C', so for $B \subseteq C'$, $\mathcal{V}_0$ shatters B if and only if $\mathcal{V}$ does. By the induction hypothesis,

$$|\mathcal{V}_0| \le |\mathrm{sh}(\mathcal{V}_0)| = \bigl|\{B \in \mathrm{sh}(\mathcal{V}) : c \notin B\}\bigr|.$$

The family $\mathcal{V}_1$. If $\mathcal{V}_1$ shatters $B \subseteq C'$, then $\mathcal{V}$ shatters $B \cup \{c\}$: given a labelling (u,t) of $B \cup \{c\}$, pick $g \in \mathcal{V}_1$ with $g|_B = u$; then $(g,t) \in \mathcal{V}$ by definition of $\mathcal{V}_1$, and it realises (u,t). The map $B \mapsto B \cup \{c\}$ being injective, the induction hypothesis gives

$$|\mathcal{V}_1| \le |\mathrm{sh}(\mathcal{V}_1)| \le \bigl|\{B \in \mathrm{sh}(\mathcal{V}) : c \in B\}\bigr|.$$

Adding the two inequalities and using (4.2) completes the induction. □

Corollary 4.31 (Sauer–Shelah lemma)**.** *If* $\mathrm{VCdim}(\mathcal{H}) = d < \infty$*, then for every* $n \geq 1$*,*

$$\tau_{\mathcal{H}}(n) \;\leq\; \Phi_d(n) = \sum_{i=0}^{d} \binom{n}{i},$$

and consequently $\tau_{\mathcal{H}}(n) \leq (en/d)^d$ *for* $n \geq d \geq 1$*, and* $\tau_{\mathcal{H}}(n) \leq (n+1)^d$ *for all* n*.*

Proof. Let C be a set of at most n points; recall that $\tau_{\mathcal{H}}(n)$ is the maximum of $|\mathcal{H}_C|$ over such sets. Apply Theorem 4.30 to $\mathcal{V} = \mathcal{H}_C$: the sets it shatters are the subsets of C shattered by $\mathcal{H}$, which have at most d elements, and there are at most $\Phi_d(|C|) \leq \Phi_d(n)$ of them. The bound $(en/d)^d$ is Lemma 4.34 below. For the last bound, $B = \{i_1 < \cdots < i_q\} \mapsto (i_1, \ldots, i_q, 0, \ldots, 0) \in \{0, \ldots, n\}^d$ injects the subsets of $\{1, \ldots, n\}$ of size at most d into a set of size $(n+1)^d$. □

Remark 4.32 (The induction in the language of hypotheses). Written for $\mathcal{V} = \mathcal{H}_C$, the family $\mathcal{V}_1$ is the trace on C' of

$$\mathcal{H}' := \big\{h \in \mathcal{H} : \; \exists h' \in \mathcal{H},\; h'(c) = 1 - h(c) \text{ and } h' = h \text{ on } C'\big\},$$

the set of hypotheses whose value at c can be flipped without changing anything else on C. This is the formulation used in [SB14, Lemma 6.10]. The decomposition (4.2) says that removing a point loses exactly the patterns that come in "flippable pairs", and the key observation is that the flippable part, being shattered together with c, is charged to the sets containing c.

Proposition 4.33 (The Sauer–Shelah bound is attained)**.** *Let* $\mathcal{X}$ *be infinite and* $d \in \mathbb{N}$*. The class* $\mathcal{H}_{\leq d} := \{\mathbb{1}_B : \; B \subseteq \mathcal{X},\; |B| \leq d\}$ *has VC dimension* d *and* $\tau_{\mathcal{H}_{\leq d}}(n) = \Phi_d(n)$ *for every* n*.*

Proof. On a set C of n points the traces of $\mathcal{H}_{\leq d}$ are exactly the indicators of the subsets of C of size at most d; there are $\Phi_d(n)$ of them. A set of size d is shattered, a set of size $d+1$ is not (the labelling $\mathbb{1}_C$ is missing). □

The half-lines and intervals of Examples 4.17 and 4.18 are two further cases of equality, with a totally different structure. Classes for which equality holds in Theorem 4.30 for every finite C are called *shattering-extremal*, or *maximum* classes when moreover they attain $\Phi_d(n)$; their structure is a subject of active combinatorial research.

4.4.2 Two proofs of the polynomial bound

Lemma 4.34 (Binomial sums)**.** *For integers* $1 \leq d \leq n$*,*

$$\Phi_d(n) \;\leq\; \left(\frac{en}{d}\right)^d.$$

Moreover, if $d \leq n/2$*, then* $\Phi_d(n) \leq e^{nH(d/n)}$*, where* $H(p) = -p\log p - (1-p)\log(1-p)$ *is the binary entropy (in nats) of Lemma 3.30, and* $e^{nH(d/n)} \leq (en/d)^d$*.*

First proof: a change of measure. Since $d/n \leq 1$, we have $(d/n)^d \leq (d/n)^i$ for $i \leq d$, so

$$\left(\frac{d}{n}\right)^d \Phi_d(n) \leq \sum_{i=0}^{d} \binom{n}{i}\left(\frac{d}{n}\right)^i \leq \sum_{i=0}^{n} \binom{n}{i}\left(\frac{d}{n}\right)^i = \left(1 + \frac{d}{n}\right)^n \leq e^d.$$

□

Second proof: Chernoff's bound for the binomial. Let $N \sim \mathrm{Bin}(n, \frac{1}{2})$, so that $\Phi_d(n) = 2^n \mathbb{P}(N \le d)$.

Case $d < n/2$. The variable $n - N$ is also $\mathrm{Bin}(n, \frac{1}{2})$, and $\{N \le d\} = \{(n-N)/n \ge 1 - d/n\}$ with $1 - d/n \in (\frac{1}{2}, 1)$. By the Chernoff bound (2.12) and the symmetry $\mathrm{kl}(1-q, 1-p) = \mathrm{kl}(q, p)$,

$$\mathbb{P}(N \le d) \le e^{-n\,\mathrm{kl}(1-d/n,\,1/2)} = e^{-n\,\mathrm{kl}(d/n,\,1/2)}.$$

Now $\mathrm{kl}(p, \frac{1}{2}) = \log 2 - H(p)$, hence $\Phi_d(n) \le 2^n e^{-n\log 2 + nH(d/n)} = e^{nH(d/n)}$. Finally,

$$n\,H\Big(\frac{d}{n}\Big) = d\log\frac{n}{d} + (n-d)\log\Big(1 + \frac{d}{n-d}\Big) \le d\log\frac{n}{d} + d,$$

by $\log(1+u) \le u$. This is $\log\big((en/d)^d\big)$.

Case $d = n/2$. Then $H(d/n) = \log 2$ and $\Phi_d(n) \le 2^n = e^{nH(1/2)}$; the comparison with $(en/d)^d$ is the next case.

Case $n/2 \le d \le n$. It suffices to show $(en/d)^d \ge 2^n$, i.e. $f(d/n) \ge \log 2$ where $f(x) := x\log(e/x) = x(1 - \log x)$. Since $f'(x) = -\log x \ge 0$ on $(0, 1]$, f is non-decreasing there, and for $x \in [\frac{1}{2}, 1]$, $f(x) \ge f(\frac{1}{2}) = \frac{1}{2}(1 + \log 2) \approx 0.85 > \log 2 \approx 0.69$. □

The second proof is longer but more informative. It shows where the bound comes from – $\Phi_d(n)/2^n$ is a binomial lower tail, and its logarithm is a relative entropy – and it gives the sharper entropy bound, which is asymptotically exact: $\frac{1}{n}\log \Phi_{\lfloor pn \rfloor}(n) \to H(p)$ for $p < \frac{1}{2}$ (by Stirling's formula, or by the lower bound of Theorem 3.42). The loss in the step $\log(1+u) \le u$ is small when $d \ll n$, which is the relevant regime.

Remark 4.35 (On the second proof). The Chernoff argument is often written for $n \ge 2d$ only, the remaining range $d \le n \le 2d$ being handled by the monotonicity of f, as above. Two small points deserve care. The first is that the Chernoff bound for the *lower* tail is needed, which is why we pass to $n - N$. The second is the inequality used to simplify $(n-d)\log\frac{n}{n-d}$: either $\log u \ge 1 - 1/u$ applied to $u = (n-d)/n$, or $\log(1+u) \le u$ applied to $u = d/(n-d)$ – the two are the same inequality.

Corollary 4.36 (The VC dichotomy)**.** *For every hypothesis class $\mathcal{H}$, exactly one of the following holds:*

(i) $\tau_{\mathcal{H}}(n) = 2^n$ *for every n (and* $\mathrm{VCdim}(\mathcal{H}) = \infty$*);*

(ii) $\mathrm{VCdim}(\mathcal{H}) = d < \infty$ *and* $\tau_{\mathcal{H}}(n) \le (en/d)^d$ *for all $n \ge d$ (with $\tau_{\mathcal{H}} \equiv 1$ if $d = 0$).*

In particular, $\frac{1}{n}\log \tau_{\mathcal{H}}(n)$ tends either to $\log 2$ or to 0, and the growth function is either exponential or polynomial: nothing in between is possible.

4.4.3 First consequences: combining classes

The growth function behaves well under combinations, and the Sauer–Shelah lemma translates this into a bound on the VC dimension.

Proposition 4.37 (Boolean combinations)**.** *Let $\mathcal{H}_1, \dots, \mathcal{H}_J$ be hypothesis classes on $\mathcal{X}$ with* $\mathrm{VCdim}(\mathcal{H}_j) \le d$ *for all j, where $d \ge 1$ and $J \ge 1$, and let $\beta : \{0,1\}^J \to \{0,1\}$ be any Boolean function. The class*

$$\mathcal{H}_\beta := \big\{ x \mapsto \beta\big(h_1(x), \dots, h_J(x)\big) : \; h_j \in \mathcal{H}_j \big\}$$

satisfies $\tau_{\mathcal{H}_\beta}(n) \le \prod_{j=1}^{J} \tau_{\mathcal{H}_j}(n)$ for all n, and

$$\mathrm{VCdim}(\mathcal{H}_\beta) \;<\; 2dJ\log_2(3J).$$

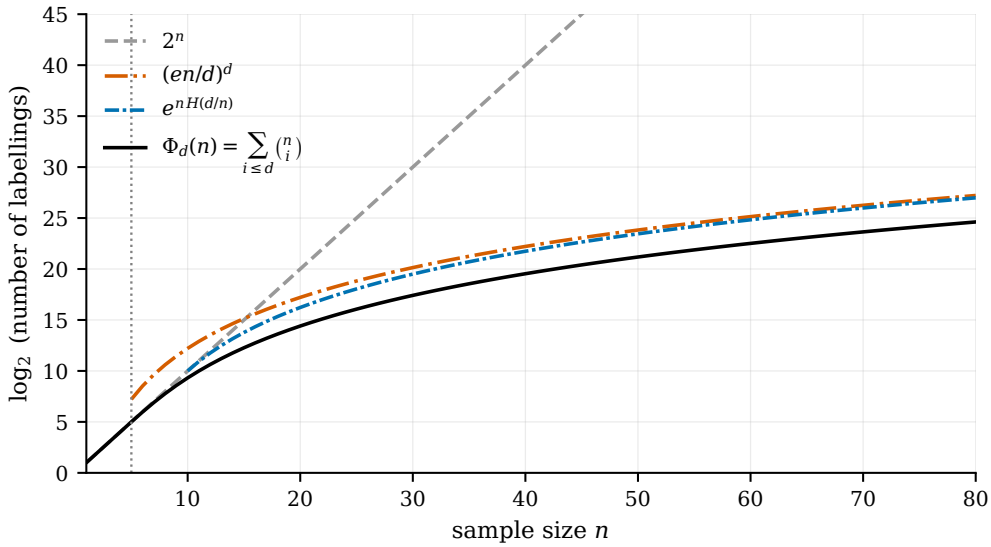


Figure 4.3: The Sauer–Shelah dichotomy for $d = 5$: $\log_2$ of 2^n, of the largest possible growth function $\Phi_5(n)$ (attained by Proposition 4.33), of the entropy bound $e^{nH(d/n)}$ (for $n \geq 2d$) and of the polynomial bound $(en/d)^d$ (for $n \geq d$). The curves of 2^n and $\Phi_5(n)$ coincide up to $n = d$; after that, Φ_5 bends from linear to logarithmic growth in the $\log_2$ scale. The dotted line marks $n = d$.

In particular this applies to unions ($\beta = \max$) and intersections ($\beta = \min$) of J sets taken from classes of VC dimension at most d.

Proof. For $x \in \mathcal{X}^n$, the trace $\mathcal{H}_\beta(x)$ is the image of $\mathcal{H}_1(x) \times \cdots \times \mathcal{H}_J(x)$ under the coordinatewise application of β, which gives the first claim. Let now $m \geq 1$ be the size of a set shattered by $\mathcal{H}_\beta$. If $m < d$, then $m < 2dJ \log_2(3J)$ since $2J \log_2(3J) \geq 2 \log_2 3 > 1$. If $m \geq d$, Corollary 4.31 gives $2^m = \tau_{\mathcal{H}_\beta}(m) \leq (em/d)^{dJ}$. Writing $u = m/(dJ)$, this reads $g(u) \leq 0$, where

$$g(u) := u \log 2 - \log(eJu).$$

The function g is convex, and $g'(u) = \log 2 - 1/u > 0$ for $u > 1/\log 2 \approx 1.44$. Let $u_0 = 2 \log_2(3J) \geq 2 \log_2 3 \approx 3.17$. Then

$$g(u_0) = \log(9J^2) - \log\bigl(2eJ \log_2(3J)\bigr) = \log \frac{9J}{2e \log_2(3J)} > 0,$$

because $\frac{9}{2e} J - \log_2(3J)$ is positive at $J = 1$ ($1.655 > 1.585$) and non-decreasing in $J \geq 1$ (its derivative is $\frac{9}{2e} - \frac{1}{J \log 2} \geq 1.655 - 1.443 > 0$). Hence $g > 0$ on $[u_0, \infty)$, and $g(u) \leq 0$ forces $u < u_0$, i.e. $m < 2dJ \log_2(3J)$. □

For instance, intersections of J half-planes – which include all convex polygons with at most J vertices – have VC dimension less than $6J \log_2(3J)$, by Proposition 4.21 with $d = 3$. The logarithmic factor is not an artefact: Eisenstat and Angluin [EA07] exhibit classes of VC dimension d whose J-fold unions have VC dimension of order $dJ \log J$, and the same order is attained by unions of J half-spaces in spaces of fixed but large enough dimension [CMK19].

Remark 4.38 (History). The lemma was proved independently, and for different reasons, three times around 1970. Vapnik and Chervonenkis [VC71] (announced in 1968) proved that the growth function of a class of finite VC dimension is bounded by a polynomial in n, as the combinatorial core of their uniform law of large numbers; [SB14, Sec. 6.7] credits them

with the lemma itself. Sauer [Sau72] proved the sharp bound $\Phi_d(n)$, answering a question of Erdős on the density of families of sets, and Shelah [She72] proved it in model theory, crediting M. Perles – whence the name Sauer–Shelah–Perles lemma used by combinatorialists. The strengthening of Theorem 4.30, in which the right-hand side counts the shattered sets themselves, is due to Pajor [Paj85], and was found independently by several authors; the elegant alternative proof by *shifting* (compressing the family coordinate by coordinate, which does not increase the set of shattered sets and makes the family hereditary) is due to Frankl [Fra83] and Alon [Alo83].

4.5 Uniform convergence through Rademacher complexity

We now prove that finite VC dimension implies uniform convergence, with explicit constants. The argument is organised in four steps, each of which is of independent interest and holds far beyond binary classification:

(1) *concentration*: the uniform deviation is within $O(n^{-1/2})$ of its mean, by McDiarmid's inequality;

(2) *symmetrisation*: the mean is at most twice the *Rademacher complexity* of the class, an average over random signs;

(3) *conditioning*: given the sample, the Rademacher complexity is the expected maximum of finitely many sub-Gaussian variables, which *Massart's lemma* bounds by $\sqrt{2\log(\text{number of patterns})}$;

(4) *counting*: the number of patterns is at most $\tau_{\mathcal{H}}(n)$, which the Sauer–Shelah lemma bounds by $(en/d)^d$.

4.5.1 Rademacher complexity

It costs nothing, and clarifies the structure, to work with a general class of bounded functions. Let $\mathcal{Z}$ be a measurable space – for us $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ – let $Z, Z_1, \dots, Z_n$ be i.i.d. with law P on $\mathcal{Z}$, and let $\mathcal{F}$ be a class of measurable functions $\mathcal{Z} \to [0,1]$. We write

$$Pf := \mathbb{E} f(Z), \qquad P_n f := \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

The case of interest is the *loss class*

$$\ell \circ \mathcal{H} := \big\{ f_h : (x,y) \mapsto \mathbb{1}\{h(x) \neq y\} \,:\, h \in \mathcal{H} \big\},$$

for which $P = \mathcal{D}$, $P f_h = L_{\mathcal{D}}(h)$ and $P_n f_h = L_{\mathcal{S}_n}(h)$. We assume that $\mathcal{F}$ satisfies the analogue of Assumption 4.1: there is a countable $\mathcal{F}_0 \subseteq \mathcal{F}$ such that every $f \in \mathcal{F}$ is a pointwise limit of elements of $\mathcal{F}_0$. By dominated convergence, every supremum over $\mathcal{F}$ appearing below is then equal to the supremum over $\mathcal{F}_0$, hence measurable. If $\mathcal{H}$ satisfies Assumption 4.1, so does $\ell \circ \mathcal{H}$, since $h_j \to h$ pointwise implies $f_{h_j} \to f_h$ pointwise.

Throughout, $\xi = (\xi_1, \dots, \xi_n)$ denotes a vector of i.i.d. Rademacher signs, $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \tfrac{1}{2}$, independent of everything else, as in Exercise 1.6 and Example 2.26(d). (Learning-theory texts often write them with the letter σ; in this book σ^2 is a variance and ε an accuracy.)

Definition 4.39 (Rademacher complexity)**.** For a bounded set $A \subseteq \mathbb{R}^n$, the *Rademacher average* of A is

$$\mathrm{Rad}(A) := \mathbb{E}_\xi\Big[\sup_{a\in A} \frac{1}{n}\sum_{i=1}^{n} \xi_i a_i\Big].$$

For $z = (z_1, \dots, z_n) \in \mathcal{Z}^n$, let $\mathcal{F}(z) := \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\} \subseteq [0,1]^n$. The *empirical Rademacher complexity* of $\mathcal{F}$ on z and the *Rademacher complexity* of $\mathcal{F}$ at sample size n are

$$\widehat{\mathfrak{R}}_z(\mathcal{F}) := \mathrm{Rad}\big(\mathcal{F}(z)\big) = \mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \xi_i f(z_i)\Big], \qquad \mathfrak{R}_n(\mathcal{F}) := \mathbb{E}\big[\widehat{\mathfrak{R}}_{(Z_1,\dots,Z_n)}(\mathcal{F})\big].$$

For a hypothesis class $\mathcal{H}$ and $x \in \mathcal{X}^n$ we write likewise $\widehat{\mathfrak{R}}_x(\mathcal{H}) = \mathrm{Rad}(\mathcal{H}(x))$.

The Rademacher complexity depends on the law P, although the notation does not show it. It measures how well the class can *correlate with pure noise*: $\frac{1}{n}\sum_i \xi_i f(z_i)$ is the empirical correlation between f and a random labelling of the sample, and $\widehat{\mathfrak{R}}_z(\mathcal{F})$ is the best correlation achievable within the class, on average over the labelling. A class that can fit any labelling – one that shatters $\{z_1, \dots, z_n\}$, for binary functions – has $\widehat{\mathfrak{R}}_z = \mathbb{E}\frac{1}{n}\sum_i (\xi_i)_+ = \frac{1}{2}$, the largest possible value for $[0,1]$-valued functions; a single function has $\widehat{\mathfrak{R}}_z = 0$.

Lemma 4.40 (Elementary properties)**.** *Let $A, B \subseteq \mathbb{R}^n$ be bounded.*

(i) $\mathrm{Rad}(A) \geq 0$*, with equality if A is a singleton; and $A \subseteq B$ implies* $\mathrm{Rad}(A) \leq \mathrm{Rad}(B)$.

(ii) For $c \in \mathbb{R}$ and $b \in \mathbb{R}^n$, $\mathrm{Rad}(cA + b) = |c|\,\mathrm{Rad}(A)$.

(iii) $\mathrm{Rad}(\mathrm{conv}\, A) = \mathrm{Rad}(A)$.

(iv) $\mathrm{Rad}(A) \leq \frac{1}{\sqrt{n}} \sup_{a\in A} \|a\|_2$.

(v) For the 0–1 loss and any sample $s = ((x_1, y_1), \dots, (x_n, y_n))$, $\widehat{\mathfrak{R}}_s(\ell \circ \mathcal{H}) = \widehat{\mathfrak{R}}_x(\mathcal{H})$, where $x = (x_1, \dots, x_n)$. In particular $\mathfrak{R}_n(\ell \circ \mathcal{H}) = \mathbb{E}\,\widehat{\mathfrak{R}}_{(X_1,\dots,X_n)}(\mathcal{H})$ does not depend on the conditional law of Y given X.

Proof. (i) For any fixed $a \in A$, $\sup_{a'\in A} \sum_i \xi_i a'_i \geq \sum_i \xi_i a_i$, whose expectation is 0. Monotonicity is obvious. (ii) $\sum_i \xi_i(ca_i + b_i) = c\sum_i \xi_i a_i + \sum_i \xi_i b_i$; the second term has mean zero and does not depend on a, and if $c < 0$ we use that $-\xi$ has the same law as ξ. (iii) A linear functional has the same supremum on A and on its convex hull. (iv) By Cauchy–Schwarz, $\sum_i \xi_i a_i \leq \|\xi\|_2 \|a\|_2 = \sqrt{n}\,\|a\|_2$ for every $a \in A$ and every $\xi \in \{-1,1\}^n$; take the supremum over a, then the expectation, and divide by n. (v) For $y \in \{0,1\}$ and $t \in \{0,1\}$, $\mathbb{1}\{t \neq y\} = y + (1-2y)t$. Hence $\ell \circ \mathcal{H}(s) = y + D\,\mathcal{H}(x)$, where D is the diagonal matrix with entries $1 - 2y_i \in \{\pm 1\}$. The vector $(\xi_i(1-2y_i))_i$ has the same law as ξ, so $\mathrm{Rad}(D\,\mathcal{H}(x)) = \mathrm{Rad}(\mathcal{H}(x))$, and the translation by y is harmless by (ii). □

4.5.2 Symmetrisation

Theorem 4.41 (Symmetrisation)**.** *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions satisfying the countable approximation property. Then*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} (Pf - P_n f)\Big] \;\leq\; 2\,\mathfrak{R}_n(\mathcal{F}) \qquad \textit{and} \qquad \mathbb{E}\Big[\sup_{f\in\mathcal{F}} (P_n f - Pf)\Big] \;\leq\; 2\,\mathfrak{R}_n(\mathcal{F}).$$

Proof. We may replace $\mathcal{F}$ by the countable $\mathcal{F}_0$. Let $Z'_1, \dots, Z'_n$ be an independent copy of the sample – a *ghost sample* – and $P'_n f = \frac{1}{n}\sum_i f(Z'_i)$. Write $\mathcal{G} = \sigma(Z_1, \dots, Z_n)$.

Step 1: introducing the ghost sample. For each f, $Pf = \mathbb{E}[P'_n f \mid \mathcal{G}]$ almost surely, hence $Pf - P_n f = \mathbb{E}[P'_n f - P_n f \mid \mathcal{G}] \le \mathbb{E}[\sup_g (P'_n g - P_n g) \mid \mathcal{G}]$ almost surely. Since $\mathcal{F}_0$ is countable, the inequality holds simultaneously for all f outside a null set, and taking the supremum over f and then expectations,

$$\mathbb{E}\sup_f \big(Pf - P_n f\big) \;\le\; \mathbb{E}\sup_f \big(P'_n f - P_n f\big) = \mathbb{E}\sup_f \frac{1}{n}\sum_{i=1}^n \big(f(Z'_i) - f(Z_i)\big).$$

Step 2: introducing the signs. Fix $r \in \{\pm 1\}^n$. Exchanging Z_i and Z'_i for the indices with $r_i = -1$ does not change the joint law of $(Z_1, \dots, Z_n, Z'_1, \dots, Z'_n)$, because the $2n$ variables are i.i.d. Therefore the family $\big(\sum_i r_i(f(Z'_i) - f(Z_i))\big)_{f\in\mathcal{F}_0}$ has the same law as $\big(\sum_i (f(Z'_i) - f(Z_i))\big)_{f\in\mathcal{F}_0}$, and so do their suprema. Averaging over r – that is, taking $r = \xi$ random and independent of the samples, and using Fubini –

$$\mathbb{E}\sup_f \frac{1}{n}\sum_{i=1}^n \big(f(Z'_i) - f(Z_i)\big) = \mathbb{E}\sup_f \frac{1}{n}\sum_{i=1}^n \xi_i\big(f(Z'_i) - f(Z_i)\big).$$

Step 3: splitting. The supremum of a sum is at most the sum of the suprema, and $-\xi$ has the same law as ξ:

$$\mathbb{E}\sup_f \frac{1}{n}\sum_i \xi_i\big(f(Z'_i) - f(Z_i)\big) \le \mathbb{E}\sup_f \frac{1}{n}\sum_i \xi_i f(Z'_i) + \mathbb{E}\sup_f \frac{1}{n}\sum_i (-\xi_i) f(Z_i) = 2\mathfrak{R}_n(\mathcal{F}).$$

The second inequality of the theorem is proved identically, or by applying the first to $\{1 - f : f \in \mathcal{F}\}$, whose Rademacher complexity is that of $\mathcal{F}$ by Lemma 4.40(ii). □

Remark 4.42 (What symmetrisation achieves). The quantity $\sup_f (Pf - P_n f)$ involves the unknown P and a supremum of non-centred, dependent variables. After symmetrisation, everything is expressed through $\frac{1}{n}\sum_i \xi_i f(Z_i)$, which, *conditionally on the sample*, is a Rademacher sum with fixed coefficients: a sub-Gaussian variable with variance proxy $\frac{1}{n^2}\sum_i f(Z_i)^2 \le \frac{1}{n}$ (Example 2.26(d)). The only property of the class that survives is the geometry of the finite-dimensional set $\mathcal{F}(Z_1, \dots, Z_n) \subseteq [0,1]^n$. This is the conceptual point: *the population is eliminated*, and an infinite-dimensional problem becomes a family of n-dimensional ones. The argument in this form goes back to Giné and Zinn [GZ84]; the double-sample idea of Step 1 is already in Vapnik and Chervonenkis [VC71], where it is used with random permutations of the $2n$ points instead of random signs.

The factor 2 is the price of the ghost sample, and symmetrisation loses nothing more: the Rademacher complexity is *also a lower bound* for the uniform deviation, up to a term $n^{-1/2}$ that any non-degenerate class pays anyway.

Proposition 4.43 (Desymmetrisation)**.** *Under the assumptions of Theorem 4.41,*

$$\mathfrak{R}_n(\mathcal{F}) \;\le\; 2\,\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \big|P_n f - Pf\big|\Big] + \frac{1}{\sqrt{n}}.$$

Proof. Writing $f(Z_i) = (f(Z_i) - Pf) + Pf$ and using $0 \le Pf \le 1$,

$$\sup_f \frac{1}{n}\sum_i \xi_i f(Z_i) \le \sup_f \frac{1}{n}\sum_i \xi_i\big(f(Z_i) - Pf\big) + \Big|\frac{1}{n}\sum_i \xi_i\Big|.$$

The expectation of the last term is at most $\big(\mathbb{E}(\frac{1}{n}\sum_i \xi_i)^2\big)^{1/2} = n^{-1/2}$. For the first term, introduce a ghost sample as in Step 1 of the previous proof, now conditionally on $(\xi, Z_1, \dots, Z_n)$: since $\mathbb{E}[f(Z_i') \mid \xi, Z_1, \dots, Z_n] = Pf$,

$$\mathbb{E}\sup_f \frac{1}{n}\sum_i \xi_i\big(f(Z_i) - Pf\big) \le \mathbb{E}\sup_f \frac{1}{n}\sum_i \xi_i\big(f(Z_i) - f(Z_i')\big) = \mathbb{E}\sup_f \frac{1}{n}\sum_i \big(f(Z_i) - f(Z_i')\big),$$

the equality by the exchangeability argument of Step 2. Finally $\frac{1}{n}\sum_i(f(Z_i) - f(Z_i')) = (P_n f - Pf) + (Pf - P_n' f)$, and the supremum of this sum has expectation at most $2\mathbb{E}\sup_f \big|P_n f - Pf\big|$. □

Together, Theorem 4.41 and Proposition 4.43 show that $\mathbb{E}\sup_f \big|P_n f - Pf\big| \to 0$ if and only if $\mathfrak{R}_n(\mathcal{F}) \to 0$: the Rademacher complexity is not merely a convenient upper bound but the right measure of the uniform deviation. (For the upper bound with an absolute value, fix $f_0 \in \mathcal{F}$. Since $\sup_f(Pf - P_n f) \ge Pf_0 - P_n f_0$ and $\sup_f(P_n f - Pf) \ge P_n f_0 - Pf_0$,

$$\sup_f \big|P_n f - Pf\big| \le \sup_f (Pf - P_n f) + \sup_f (P_n f - Pf) + \big|P_n f_0 - Pf_0\big|,$$

and Theorem 4.41 gives $\mathbb{E}\sup_f \big|P_n f - Pf\big| \le 4\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2\sqrt{n}}$; see also Exercise 4.10.)

4.5.3 Massart's lemma

Lemma 4.44 (Massart's finite class lemma)**.** *Let $A \subseteq \mathbb{R}^n$ be a finite set with $N \ge 1$ elements, and $r := \max_{a\in A}\|a\|_2$. Then*

$$\mathbb{E}_\xi\Big[\max_{a\in A}\sum_{i=1}^n \xi_i a_i\Big] \;\le\; r\sqrt{2\log N}, \qquad \textit{i.e.} \qquad \mathrm{Rad}(A) \le \frac{r\sqrt{2\log N}}{n}.$$

Proof. For each $a \in A$, the variable $W_a := \sum_i \xi_i a_i$ is centred and belongs to $\mathrm{subG}(\|a\|_2^2) \subseteq \mathrm{subG}(r^2)$ by Example 2.26(d). The W_a are not independent, but Proposition 2.28 does not require independence, and gives $\mathbb{E}\max_a W_a \le \sqrt{2r^2\log N}$. □

The proof is two lines because the work was done in Chapter 2: the moment generating function of a Rademacher sum is dominated by a Gaussian one, and the maximum of N sub-Gaussian variables costs $\sqrt{2\log N}$ times their scale. Note that by Lemma 4.40(ii) we may first translate A, so that r may be replaced by $\min_{b\in\mathbb{R}^n}\max_{a\in A}\|a - b\|_2$, the radius of the smallest Euclidean ball containing A.

Corollary 4.45 (Rademacher complexity of a class of finite VC dimension)**.** *Let $\mathcal{H}$ be a hypothesis class. For every $x \in \mathcal{X}^n$,*

$$\widehat{\mathfrak{R}}_x(\mathcal{H}) \;\le\; \sqrt{\frac{\log|\mathcal{H}(x)|}{2n}} \;\le\; \sqrt{\frac{\log \tau_{\mathcal{H}}(n)}{2n}},$$

and if $\mathrm{VCdim}(\mathcal{H}) = d \ge 1$ and $n \ge d$,

$$\widehat{\mathfrak{R}}_x(\mathcal{H}) \;\le\; \sqrt{\frac{d\log(en/d)}{2n}}.$$

The same bounds hold for $\widehat{\mathfrak{R}}_s(\ell\circ\mathcal{H})$, for every sample s, and for $\mathfrak{R}_n(\ell\circ\mathcal{H})$, whatever the distribution $\mathcal{D}$.

Proof. The set $\mathcal{H}(x) \subseteq \{0,1\}^n$ is finite. Translating it by $-\frac{1}{2}(1,\dots,1)$ (Lemma 4.40(ii)) puts it inside $\{-\frac{1}{2},\frac{1}{2}\}^n$, whose elements have Euclidean norm $\sqrt{n}/2$. Lemma 4.44 gives $\widehat{\mathfrak{R}}_x(\mathcal{H}) \le \frac{\sqrt{n}}{2}\cdot\frac{\sqrt{2\log|\mathcal{H}(x)|}}{n} = \sqrt{\log|\mathcal{H}(x)|/(2n)}$. Then use $|\mathcal{H}(x)| \le \tau_{\mathcal{H}}(n)$ and Corollary 4.31, and Lemma 4.40(v) for the loss class. □

For a finite class, Corollary 4.45 gives $\mathfrak{R}_n \le \sqrt{\log|\mathcal{H}|/(2n)}$, the expression that appears in Theorem 2.65: the Rademacher approach contains the finite-class theory, with the cardinality of the class replaced by the number of its patterns on the sample. This is the precise sense in which a class of finite VC dimension is "effectively finite", of effective cardinality $\tau_{\mathcal{H}}(n) \le (en/d)^d$, and effective log-cardinality $d\log(en/d)$.

4.5.4 Uniform convergence with high probability

Theorem 4.46 (Rademacher bounds)**.** *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions with the countable approximation property, and $\delta \in (0,1)$. Each of the following holds with probability at least $1-\delta$:*

$$\sup_{f\in\mathcal{F}}\big(Pf - P_nf\big) \le 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{4.3}$$

$$\sup_{f\in\mathcal{F}}\big|Pf - P_nf\big| \le 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}, \tag{4.4}$$

$$\sup_{f\in\mathcal{F}}\big(Pf - P_nf\big) \le 2\widehat{\mathfrak{R}}_{(Z_1,\dots,Z_n)}(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4.5}$$

Proof. Let $U^+(z_1,\dots,z_n) := \sup_f\big(Pf - \frac{1}{n}\sum_i f(z_i)\big)$. Replacing one z_i by z_i' changes each $\frac{1}{n}\sum_i f(z_i)$ by at most $\frac{1}{n}$, hence changes U^+ by at most $\frac{1}{n}$: U^+ has the bounded-difference property with $c_i = \frac{1}{n}$, $\sum_i c_i^2 = \frac{1}{n}$, and Theorem 2.55 gives $\mathbb{P}\big(U^+ \ge \mathbb{E}U^+ + t\big) \le e^{-2nt^2}$. With $t = \sqrt{\log(1/\delta)/(2n)}$ and $\mathbb{E}U^+ \le 2\mathfrak{R}_n(\mathcal{F})$ (Theorem 4.41), this is (4.3). The same argument applies to $U^- = \sup_f(P_nf - Pf)$; a union bound with $\delta/2$ for each of U^+ and U^- gives (4.4), since $\sup_f\big|Pf - P_nf\big| = \max(U^+,U^-)$. Finally, the map $z \mapsto \widehat{\mathfrak{R}}_z(\mathcal{F})$ also has bounded differences $\frac{1}{n}$ (for each fixed ξ the supremum moves by at most $\frac{1}{n}$, and so does its average over ξ), so with probability at least $1-\delta/2$, $\mathfrak{R}_n(\mathcal{F}) \le \widehat{\mathfrak{R}}_{(Z_1,\dots,Z_n)}(\mathcal{F}) + \sqrt{\log(2/\delta)/(2n)}$. Combining with (4.3) at level $\delta/2$ gives (4.5). □

Remark 4.47 (Data-dependent bounds). The right-hand side of (4.5) is computable from the data: the empirical Rademacher complexity is an expectation over the signs only, which can be computed exactly for small classes, and estimated by Monte Carlo otherwise – with a controlled error, since it is itself a bounded-difference function of ξ. Such a bound adapts to the distribution: if $\mathcal{D}_{\mathcal{X}}$ is concentrated on a few points, or on a region where the class's patterns are few, then $\widehat{\mathfrak{R}}$ is small even if $\mathrm{VCdim}(\mathcal{H})$ is large. This is the main reason for the success of Rademacher complexities in learning theory [Kol01; BM02]; another is that they are defined for real-valued classes (Chapter 6), where the combinatorial VC dimension has no direct meaning.

We now specialise to classes of finite VC dimension. First, a technical lemma for inverting bounds of the form $\sqrt{d\log n/n}$.

Lemma 4.48 (Inverting $x \ge a\log x + b$)**.** *(i) Let $a > 0$. If $a \le e$, then $x \ge a\log x$ for every $x > 0$. If $a > e$, then $x \ge a\log x$ for every $x \ge 2a\log a$.*

(ii) Let $a \geq 1$ and $b > 0$. If $x \geq 4a\log(2a) + 2b$, then $x > a\log x + b$.

Proof. (i) The function $g(x) = x - a\log x$ is minimal on $(0,\infty)$ at $x = a$, with value $a(1-\log a)$, which is non-negative when $a \leq e$. If $a > e$, then $2a\log a > 2a > a$, so g is increasing on $[2a\log a, \infty)$ and it suffices to check $g(2a\log a) \geq 0$, i.e. $a\log a \geq a\log(2\log a)$, i.e. $a \geq 2\log a$; this holds for every $a > 0$, since $a - 2\log a$ is minimal at $a = 2$, where it equals $2 - 2\log 2 > 0$. (ii) Since $x \geq 4a\log(2a) = 2(2a)\log(2a)$, part (i) applied to $2a$ gives $x \geq 2a\log x$. Since $4a\log(2a) > 0$, also $x > 2b$. Adding $\frac{x}{2} \geq a\log x$ and $\frac{x}{2} > b$ gives the claim. □

Lemma 4.48(i) is sharp: for $a > e$ the inequality $x \geq a\log x$ fails on an interval around $x = a$.

Theorem 4.49 (Uniform convergence for VC classes)**.** *Let $\mathcal{H}$ satisfy Assumption 4.1, with $\mathrm{VCdim}(\mathcal{H}) = d \geq 1$. For every distribution $\mathcal{D}$, every $n \geq d$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{h\in\mathcal{H}} \left| L_{S_n}(h) - L_{\mathcal{D}}(h) \right| \;\leq\; \sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4.6}$$

Consequently $\mathcal{H}$ has the uniform convergence property, with

$$n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon,\delta) \;\leq\; \left\lceil \frac{16\,d\log(8e/\varepsilon^2) + 2\log(2/\delta)}{\varepsilon^2} \right\rceil.$$

Proof. Apply (4.4) to $\mathcal{F} = \ell\circ\mathcal{H}$; by Corollary 4.45, $2\mathfrak{R}_n(\ell\circ\mathcal{H}) \leq 2\sqrt{d\log(en/d)/(2n)} = \sqrt{2d\log(en/d)/n}$. For the sample complexity, write the right-hand side of (4.6) as $\alpha + \beta$. Since $(\alpha+\beta)^2 \leq 2(\alpha^2+\beta^2)$, it is less than ε as soon as

$$n\varepsilon^2 \;>\; 4d\log\frac{en}{d} + \log\frac{2}{\delta}.$$

In terms of $u = en/d$ this reads $u > a\log u + b$ with $a = 4e/\varepsilon^2 \geq 1$ and $b = e\log(2/\delta)/(d\varepsilon^2) > 0$, which by Lemma 4.48(ii) holds as soon as $u \geq 4a\log(2a) + 2b$, i.e. as soon as $n \geq \big(16d\log(8e/\varepsilon^2) + 2\log(2/\delta)\big)/\varepsilon^2$; this also guarantees $n \geq d$. □

For learning, Lemma 4.8 and Theorem 4.49 bound the excess risk of an empirical risk minimiser by twice the right-hand side of (4.6). A bound better by a factor 2 on the complexity term – a factor 4 on its contribution to the sample complexity – is obtained by using the one-sided inequality (4.3) for the empirical risk minimiser and Hoeffding's inequality for a fixed comparison hypothesis, as noticed after Lemma 4.8.

Theorem 4.50 (Empirical risk minimisation over a VC class)**.** *Under the assumptions of Theorem 4.49, let $\widehat{h}_n$ be a (measurable) empirical risk minimiser over $\mathcal{H}$. Then*

$$\mathbb{E}\big[L_{\mathcal{D}}(\widehat{h}_n)\big] - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \;\leq\; \sqrt{\frac{2d\log(en/d)}{n}},$$

and with probability at least $1-\delta$,

$$L_{\mathcal{D}}(\widehat{h}_n) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \;\leq\; \sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}. \tag{4.7}$$

Consequently $\mathcal{H}$ is agnostically PAC learnable by empirical risk minimisation, with

$$n_{\mathcal{H}}(\varepsilon,\delta) \;\leq\; n_{\mathrm{ERM}}(\varepsilon,\delta) \;\leq\; \left\lceil \frac{16\,d\log(8e/\varepsilon^2) + 8\log(2/\delta)}{\varepsilon^2} \right\rceil.$$

Proof. Let $\bar{h} \in \mathcal{H}$ be fixed. Since $L_{\mathcal{S}_n}(\widehat{h}_n) \le L_{\mathcal{S}_n}(\bar{h})$,

$$L_{\mathcal{D}}(\widehat{h}_n) - L_{\mathcal{D}}(\bar{h}) \le \big(L_{\mathcal{D}}(\widehat{h}_n) - L_{\mathcal{S}_n}(\widehat{h}_n)\big) + \big(L_{\mathcal{S}_n}(\bar{h}) - L_{\mathcal{D}}(\bar{h})\big) \le U^+ + \big(L_{\mathcal{S}_n}(\bar{h}) - L_{\mathcal{D}}(\bar{h})\big), \quad (4.8)$$

with $U^+ = \sup_h (L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h))$. The second term has mean zero, and $\mathbb{E}U^+ \le 2\mathfrak{R}_n(\ell \circ \mathcal{H}) \le \sqrt{2d\log(en/d)/n}$ by Theorem 4.41 and Corollary 4.45; taking expectations and then the infimum over $\bar{h}$ gives the first claim. For the second, (4.3) at level $\delta/2$ and the one-sided Hoeffding inequality (first half of (1.28)), applied to the variables $\mathbb{1}\{\bar{h}(X_i) \neq Y_i\}$ at level $\delta/2$, bound the two terms of (4.8) by $\sqrt{2d\log(en/d)/n} + \sqrt{\log(2/\delta)/(2n)}$ and $\sqrt{\log(2/\delta)/(2n)}$ respectively, outside an event of probability at most δ. If the infimum is attained we take $\bar{h}$ a minimiser; otherwise we apply the bound to $\bar{h}$ with $L_{\mathcal{D}}(\bar{h}) \le \inf_h L_{\mathcal{D}}(h) + \varepsilon'$ and let $\varepsilon' \downarrow 0$, the events $\{L_{\mathcal{D}}(\widehat{h}_n) - \inf_h L_{\mathcal{D}}(h) > B + \varepsilon'\}$, where B is the right-hand side of (4.7), increasing to $\{L_{\mathcal{D}}(\widehat{h}_n) - \inf_h L_{\mathcal{D}}(h) > B\}$. The sample complexity follows as in the proof of Theorem 4.49: the right-hand side of (4.7) is less than ε as soon as $n\varepsilon^2 > 4d\log(en/d) + 4\log(2/\delta)$, and Lemma 4.48(ii), with $a = 4e/\varepsilon^2$ and $b = 4e\log(2/\delta)/(d\varepsilon^2)$, shows that this holds for $n \ge \big(16d\log(8e/\varepsilon^2) + 8\log(2/\delta)\big)/\varepsilon^2$. □

Remark 4.51 (Why not Markov's inequality). A shorter route bounds $\mathbb{E}\sup_h \big|L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h)\big|$ by symmetrisation and then concludes with Markov's inequality, which yields, with probability $1-\delta$, the bound

$$\sup_{h\in\mathcal{H}} \big|L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h)\big| \le \frac{1 + \sqrt{\log(2\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{n/2}}$$

and a sample complexity of order $d\log(d/(\delta\varepsilon))/(\delta\varepsilon)^2$. The dependence on the confidence level is $1/\delta^2$ instead of $\log(1/\delta)$: for $\delta = 10^{-3}$, a factor of order 10^6 instead of 7. Markov's inequality is wasteful here because it uses only the first moment of a variable whose fluctuations around its mean are sub-Gaussian with variance proxy $1/(4n)$ – as McDiarmid's inequality shows for *every* class, with no assumption at all (Section 2.7.5). This is the strategic insight recalled in the introduction: once the expectation is controlled, the deviations are free. Note also that the route through $\tau_{\mathcal{H}}(2n)$ – symmetrising with the ghost sample and taking a union bound over the patterns of $\mathcal{H}$ on the $2n$ points – is replaced above by a bound involving only $\tau_{\mathcal{H}}(n)$, thanks to the splitting of Step 3 in the proof of Theorem 4.41.

4.5.5 How sharp is this? The logarithmic factor and the rate

The bound $\mathfrak{R}_n(\ell \circ \mathcal{H}) \le \sqrt{d\log(en/d)/(2n)}$ of Corollary 4.45 is off by a factor $\sqrt{\log(n/d)}$. Massart's lemma treats the $\tau_{\mathcal{H}}(n)$ patterns as if they were spread out in all directions, like $\tau_{\mathcal{H}}(n)$ orthogonal vectors; but the patterns of a VC class are clustered, most of them being close to many others in Hamming distance. Taking this geometry into account requires *chaining*: one applies Massart's lemma not once but at every scale, to successive approximations of the class by finite nets, and sums the contributions (Dudley's entropy integral). Combined with Haussler's bound on the packing numbers of VC classes, this removes the logarithm.

Theorem 4.52 (Removing the logarithm; admitted)**.** *There is a universal constant κ such that for every hypothesis class $\mathcal{H}$ satisfying Assumption 4.1 with $\mathrm{VCdim}(\mathcal{H}) = d \ge 1$, every distribution $\mathcal{D}$ and every $n \ge 1$,*

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}} \big|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\big|\Big] \;\le\; \kappa\sqrt{\frac{d}{n}}.$$

Consequently $n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon,\delta) \le \kappa'(d+\log(1/\delta))/\varepsilon^2$ *and* $n_{\mathcal{H}}(\varepsilon,\delta) \le \kappa'(d+\log(1/\delta))/\varepsilon^2$ *for a universal* κ'.

The bound follows from Dudley's entropy integral [Dud67] combined with a bound on the covering numbers of VC classes; such bounds go back to Dudley [Dud78], and were sharpened by Haussler [Hau95]. Talagrand [Tal94] obtained the corresponding sharp tail bounds. A complete proof along the lines sketched above is [Ver18, Thm. 8.3.23], with a covering bound of the same form as Haussler's. The "consequently" part follows from the first by McDiarmid's inequality exactly as in Theorem 4.49. (The case $d = 0$ is excluded only because the right-hand side would vanish: for a single hypothesis h, $\mathbb{E}\left|L_{S_n}(h) - L_{\mathcal{D}}(h)\right|$ is positive in general and at most $1/(2\sqrt{n})$ by the Cauchy–Schwarz inequality.) We shall use Theorem 4.52 only in remarks: none of the theorems of this chapter depends on it. Since d/ε^2 is also a lower bound (Theorem 4.61), it shows that the agnostic sample complexity is $\Theta\big((d+\log(1/\delta))/\varepsilon^2\big)$.

Remark 4.53 (A class of VC dimension 1 cannot do better than $n^{-1/2}$). Could a very simple class enjoy faster uniform convergence? No. Take $\mathcal{H} = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\}$, $Y \equiv 0$ and X with a continuous distribution function F. Then $L_{S_n}(\mathbb{1}_{(-\infty,a]}) = F_n(a)$ is the empirical distribution function and $L_{\mathcal{D}}(\mathbb{1}_{(-\infty,a]}) = F(a)$, so that the uniform deviation is the Kolmogorov–Smirnov statistic $\sup_a |F_n(a) - F(a)|$. Its exact order is $n^{-1/2}$: by Kolmogorov's theorem $\sqrt{n}\sup_a|F_n - F|$ converges in law to a non-degenerate limit, and in a non-asymptotic form, $\mathbb{E}\sup_a|F_n(a)-F(a)| \ge \mathbb{E}\left|F_n(m) - \frac{1}{2}\right| \ge \frac{1}{2\sqrt{3n}}$ at a median m (Exercise 4.11). Conversely, the Dvoretzky–Kiefer–Wolfowitz inequality, with the optimal constant due to Massart [Mas90],

$$\mathbb{P}\Big(\sup_{a\in\mathbb{R}}\big|F_n(a) - F(a)\big| > t\Big) \;\le\; 2e^{-2nt^2}, \qquad t > 0,$$

(admitted; [DKW56; Mas90]) shows that for this class the uniform deviation is controlled *exactly* as for a single hypothesis, as if $|\mathcal{H}| = 1$, without the factor $\sqrt{\log n}$ given by Theorem 4.49. So the rate $n^{-1/2}$ is the truth for every non-trivial class, whatever its VC dimension, and the VC dimension only affects the constant – as Theorem 4.52 confirms.

Remark 4.54 (Measurability, once more). Assumption 4.1 is a convenient sufficient condition, but neither the weakest nor the most natural one. Dudley [Dud78] identified the *image admissible Suslin* classes, and Pollard [Pol84] the closely related *permissible* classes, for which all the suprema above are measurable (possibly for the completed measure) and the symmetrisation argument goes through; see [Dud99, Chap. 5]. Some condition is necessary, and the failure is not merely one of measurability. Under the continuum hypothesis, $[0,1]$ can be well ordered so that every initial segment is countable; the class $\mathcal{E}$ of these initial segments has VC dimension 1 (like half-lines, Example 4.17), each $E \in \mathcal{E}$ is countable, hence of Lebesgue measure 0, and yet every finite sample is contained in some $E \in \mathcal{E}$. For X uniform on $[0,1]$ this gives $\sup_{E\in\mathcal{E}}|P_n(E) - P(E)| = 1$ for every sample: a class of VC dimension 1 without the uniform convergence property [Dud99, Chap. 5]. An alternative, used in [BLM13], is to *define* the expected supremum over an arbitrary class as the supremum of the expected suprema over its finite subclasses. The bounds of this section on expected suprema remain true with this definition, with the same proofs; but the statements in probability, and the proofs about empirical risk minimisers, involve the supremum itself, about which this definition says nothing – and the example above shows that the high-probability statements can genuinely fail.

4.5.6 The realizable case: a fast rate through the double sample

In the realizable case, Proposition 4.12 gave the rate $\log|\mathcal{H}|/n$ instead of $\sqrt{\log|\mathcal{H}|/n}$. The analogue for VC classes is obtained by the original double-sample argument of Vapnik and Chervonenkis, which we now present. Unlike the Rademacher route, it uses the growth function at $2n$.

Theorem 4.55 (Consistent hypotheses in a VC class)**.** *Let $\mathcal{H}$ satisfy Assumption 4.1, let $\mathcal{D}$ be any distribution, and let $\varepsilon > 0$ and $n \geq 8/\varepsilon$. Then*

$$\mathbb{P}\Big(\exists h \in \mathcal{H} :\ L_{\mathcal{S}_n}(h) = 0 \text{ and } L_{\mathcal{D}}(h) > \varepsilon\Big) \ \leq\ 2\,\tau_{\mathcal{H}}(2n)\,2^{-n\varepsilon/2}.$$

Proof. Call E the event on the left. If $h \in \mathcal{H}$ satisfies $L_{\mathcal{S}_n}(h) = 0$ and $L_{\mathcal{D}}(h) > \varepsilon$, the approximating $h_j \in \mathcal{H}_0$ of Assumption 4.1 coincide with h on the sample for j large, and $L_{\mathcal{D}}(h_j) \to L_{\mathcal{D}}(h)$ by dominated convergence, so that $L_{\mathcal{S}_n}(h_j) = 0$ and $L_{\mathcal{D}}(h_j) > \varepsilon$ for j large. Hence E does not change if $\mathcal{H}$ is replaced by $\mathcal{H}_0$, and the same holds for the event E'' below. We may and do assume that $\mathcal{H}$ is countable.

Step 1: the ghost sample. Let $\mathcal{S}_n'$ be an independent copy of $\mathcal{S}_n$, and

$$E'' := \big\{\exists h \in \mathcal{H} :\ L_{\mathcal{S}_n}(h) = 0 \text{ and } L_{\mathcal{S}_n'}(h) \geq \varepsilon/2\big\}.$$

Enumerate $\mathcal{H}$ and, on E, let $h_{\mathcal{S}_n}$ be the first h with $L_{\mathcal{S}_n}(h) = 0$ and $L_{\mathcal{D}}(h) > \varepsilon$; it is a measurable function of $\mathcal{S}_n$. Conditionally on $\mathcal{S}_n$, on the event E, $nL_{\mathcal{S}_n'}(h_{\mathcal{S}_n})$ is binomial with parameters n and $p = L_{\mathcal{D}}(h_{\mathcal{S}_n}) > \varepsilon$, so by Chebyshev's inequality

$$\mathbb{P}\big(L_{\mathcal{S}_n'}(h_{\mathcal{S}_n}) < \tfrac{\varepsilon}{2} \,\big|\, \mathcal{S}_n\big) \leq \mathbb{P}\Big(\big|nL_{\mathcal{S}_n'}(h_{\mathcal{S}_n}) - np\big| > \tfrac{np}{2} \,\Big|\, \mathcal{S}_n\Big) \leq \frac{np(1-p)}{(np/2)^2} \leq \frac{4}{n\varepsilon} \leq \frac{1}{2}.$$

Hence $\mathbb{P}(E'') \geq \mathbb{E}\big[\mathbb{1}_E\,\mathbb{P}(L_{\mathcal{S}_n'}(h_{\mathcal{S}_n}) \geq \varepsilon/2 \mid \mathcal{S}_n)\big] \geq \tfrac{1}{2}\mathbb{P}(E)$.

Step 2: random swaps. Write $\mathcal{S}_n = (Z_1, \dots, Z_n)$ and $\mathcal{S}_n' = (Z_1', \dots, Z_n')$. For $r \in \{\pm 1\}^n$, exchanging Z_i and Z_i' for the indices with $r_i = -1$ preserves the law of the double sample (Step 2 of Theorem 4.41), hence $\mathbb{P}(E'') = \mathbb{E}\big[\mathbb{P}_\xi(E''_\xi)\big]$, where E''_ξ is the event E'' computed on the swapped samples and ξ is uniform on $\{\pm 1\}^n$, independent of the data. Fix the $2n$ points z_i, z_i', and fix h. Let q be the number of indices i such that h errs on at least one of z_i, z_i'. If h errs on both z_i and z_i' for some i, then after any swap it errs on the first sample, and h does not realise E''_ξ. Otherwise, h realises E''_ξ only if each of the q errors lands in the second sample – which happens with probability 2^{-q} – and if $q \geq n\varepsilon/2$. So the probability that this h realises E''_ξ is at most $2^{-n\varepsilon/2}$. This probability depends on h only through the error pattern of h on the $2n$ points, which is determined by the trace of h on $(x_1, \dots, x_n, x_1', \dots, x_n')$; there are at most $\tau_{\mathcal{H}}(2n)$ such traces. By the union bound over traces, $\mathbb{P}_\xi(E''_\xi) \leq \tau_{\mathcal{H}}(2n)2^{-n\varepsilon/2}$ for every realisation of the $2n$ points.

Combining the two steps, $\mathbb{P}(E) \leq 2\mathbb{P}(E'') \leq 2\tau_{\mathcal{H}}(2n)2^{-n\varepsilon/2}$. □

Note that the theorem holds for every $\mathcal{D}$, realizable or not: it bounds the probability that *some* hypothesis with large risk is consistent with the sample. Realizability is needed only to ensure that the empirical risk minimiser is consistent with the sample.

Corollary 4.56 (Realizable upper bound)**.** *Let $\mathcal{H}$ satisfy Assumption 4.1, with* $\mathrm{VCdim}(\mathcal{H}) = d \geq 1$*, and let $\mathcal{D}$ be realizable by $\mathcal{H}$. Then $\min_h L_{\mathcal{S}_n}(h) = 0$ almost surely, and for every $n \geq d$ and $\delta \in (0,1)$, with probability at least $1-\delta$, every $h \in \mathcal{H}$ with $L_{\mathcal{S}_n}(h) = 0$ – in particular every empirical risk minimiser – satisfies*

$$L_{\mathcal{D}}(h) \ \leq\ \max\Big\{\frac{2}{n}\Big(d\log_2\frac{2en}{d} + \log_2\frac{2}{\delta}\Big),\ \frac{8}{n}\Big\}.$$

Consequently,

$$n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) \;\le\; n^{\mathrm{re}}_{ERM}(\varepsilon,\delta) \;\le\; \left\lceil \frac{8}{\varepsilon\log 2}\left(d\log\frac{32}{\varepsilon}+\log\frac{2}{\delta}\right)\right\rceil.$$

Proof. If $\inf_h L_{\mathcal{D}}(h)=0$, pick h_j with $L_{\mathcal{D}}(h_j)\le 1/j$; then $\mathbb{P}(\min_h L_{\mathcal{S}_n}(h)>0)\le \mathbb{P}(L_{\mathcal{S}_n}(h_j)>0)\le n/j$ for every j. Next, let ε_0 denote the maximum in the display. Then $n\ge 8/\varepsilon_0$ and, by Corollary 4.31 with $2n\ge d$, $2\tau_{\mathcal{H}}(2n)2^{-n\varepsilon_0/2}\le 2(2en/d)^d 2^{-d\log_2(2en/d)-\log_2(2/\delta)}=\delta$. Theorem 4.55 gives $\mathbb{P}(\exists h: L_{\mathcal{S}_n}(h)=0,\ L_{\mathcal{D}}(h)>\varepsilon_0)\le\delta$. For the sample complexity, the event to control is $\{L_{\mathcal{D}}(\widehat{h}_n)\ge\varepsilon\}$; it is contained in $\{L_{\mathcal{D}}(\widehat{h}_n)>\varepsilon''\}$ for every $\varepsilon''<\varepsilon$, so by the first part it suffices that $n\varepsilon''\ge 8$ and $n\varepsilon''\ge 2d\log_2(2en/d)+2\log_2(2/\delta)$ for some $\varepsilon''<\varepsilon$, i.e. that $n\varepsilon>8$ and $n\varepsilon>2d\log_2(2en/d)+2\log_2(2/\delta)$. In terms of $u=2en/d$, the second condition reads $u>a\log u+b$ with $a=4e/(\varepsilon\log 2)\ge 1$ and $b=4e\log_2(2/\delta)/(d\varepsilon)$, and by Lemma 4.48(ii) it holds as soon as

$$n\;\ge\;\frac{8d}{\varepsilon\log 2}\log\frac{8e}{\varepsilon\log 2}+\frac{4\log(2/\delta)}{\varepsilon\log 2};$$

since $8e/\log 2<32$, the bound of the statement suffices. It also gives $n\ge d$ and $n\varepsilon\ge 8\log 32/\log 2=40>8$. □

The rate is now $d\log n/n$, the analogue of $\log|\mathcal{H}|/n$, and the sample complexity $O\big((d\log(1/\varepsilon)+\log(1/\delta))/\varepsilon\big)$ is the classical bound of Blumer, Ehrenfeucht, Haussler and Warmuth [Blu+89], who proved it with the constants $\max\{\frac{4}{\varepsilon}\log_2\frac{2}{\delta},\frac{8d}{\varepsilon}\log_2\frac{13}{\varepsilon}\}$.

Remark 4.57 (The optimal realizable sample complexity). Is the factor $\log(1/\varepsilon)$ necessary? For empirical risk minimisation in general, yes: there are classes of VC dimension d and consistent learners for them that need of the order of $(d\log(1/\varepsilon)+\log(1/\delta))/\varepsilon$ examples (see [Han16] and the references therein). But the sample complexity of the *class* is smaller. Haussler, Littlestone and Warmuth [HLW94] constructed, through the "one-inclusion graph" of the class, an (improper) algorithm with expected risk at most $d/(n+1)$; and Hanneke [Han16] proved, after a long series of partial results, that

$$n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta)\;\le\;\kappa\,\frac{d+\log(1/\delta)}{\varepsilon}$$

for a universal constant κ (admitted), with an algorithm that takes a majority vote of empirical risk minimisers trained on cleverly chosen subsamples. Together with Theorem 4.63, this determines the realizable sample complexity up to constants. Here, for once, the distinction between proper and improper learners of Remark 4.6 matters.

Remark 4.58 (Relative deviations). Theorems 4.49 and 4.55 are the two extremes of a single inequality, due to Vapnik and Chervonenkis [VC74] in the following form (admitted; for a proof with these constants see [AS93]): for every $t>0$,

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}}\frac{L_{\mathcal{D}}(h)-L_{\mathcal{S}_n}(h)}{\sqrt{L_{\mathcal{D}}(h)}}>t\right)\le 4\,\tau_{\mathcal{H}}(2n)\,e^{-nt^2/4}.$$

Solving for $L_{\mathcal{D}}(h)$ as in the proof of Theorem 2.67, it gives, with probability at least $1-\delta$, simultaneously for all $h\in\mathcal{H}$,

$$L_{\mathcal{D}}(h)\le L_{\mathcal{S}_n}(h)+2\sqrt{L_{\mathcal{S}_n}(h)\frac{\log\tau_{\mathcal{H}}(2n)+\log(4/\delta)}{n}}+4\,\frac{\log\tau_{\mathcal{H}}(2n)+\log(4/\delta)}{n},$$

the exact analogue for VC classes of the variance-adaptive bound (2.37): rate n^{-1} for consistent hypotheses, rate $n^{-1/2}$ with the factor $\sqrt{L_{\mathcal{S}_n}(h)}$ in general. The mechanism is the same as there – the variance of $\mathbb{1}\{h(X)\neq Y\}$ is at most its mean $L_{\mathcal{D}}(h)$ – and the modern way to exploit it systematically is the *localisation* of Rademacher complexities [Kol11].

4.6 The fundamental theorem of statistical learning

4.6.1 The qualitative statement

Theorem 4.59 (Fundamental theorem of statistical learning)**.** *Let $\mathcal{H}$ be a class of classifiers $\mathcal{X} \to \{0,1\}$ satisfying Assumption 4.1, with the 0–1 loss. The following are equivalent.*

(1) $\mathcal{H}$ has the uniform convergence property.

(2) Every (measurable) empirical risk minimiser agnostically PAC learns $\mathcal{H}$.

(3) $\mathcal{H}$ is agnostically PAC learnable.

(4) $\mathcal{H}$ is PAC learnable (in the realizable case).

(5) Every (measurable) empirical risk minimiser PAC learns $\mathcal{H}$ in the realizable case.

(6) $\mathrm{VCdim}(\mathcal{H}) < \infty$.

Proof. (1)⇒(2) is Corollary 4.10. (2)⇒(3) and (5)⇒(4) hold because a measurable empirical risk minimiser exists (Section 4.2.1). (2)⇒(5) and (3)⇒(4): an algorithm that agnostically PAC learns $\mathcal{H}$ also PAC learns it in the realizable case, since for a realizable $\mathcal{D}$ condition (4.1) reads $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \geq \varepsilon) \leq \delta$; thus $n^{\mathrm{re}}_{\mathcal{A}} \leq n_{\mathcal{A}}$. (4)⇒(6) is the contrapositive of Theorem 4.28. (6)⇒(1) is Theorem 4.49 if $\mathrm{VCdim}(\mathcal{H}) \geq 1$; if $\mathrm{VCdim}(\mathcal{H}) = 0$, the class has a single element and uniform convergence is Corollary 1.70. □

The logical structure deserves a comment. Five of the six properties are *analytic* or *algorithmic*: they involve distributions, samples and probabilities. The sixth is purely *combinatorial*: it involves only finite subsets of $\mathcal{X}$ and the labellings $\mathcal{H}$ induces on them, and can be checked with pencil and paper – as we did in Section 4.3.2. The circle of implications passes through both worlds: from combinatorics to analysis by the Sauer–Shelah lemma and the Rademacher machinery, and back by the no-free-lunch theorem. It also says that the simplest conceivable algorithm, empirical risk minimisation, is as good as any other *as far as learnability is concerned*: if anything learns $\mathcal{H}$, empirical risk minimisation does. (It is not always as good *quantitatively*, see Remark 4.57.) None of this survives intact outside binary classification: for multiclass problems with infinitely many labels, learnability is no longer characterised by uniform convergence, nor achieved by empirical risk minimisation, and the right combinatorial parameter is different [SB14, Chap. 29]: it is the dimension introduced by Daniely and Shalev-Shwartz, as shown only recently by Brukhim, Carmon, Dinur, Moran and Yehudayoff [Bru+22]. For online learning the VC dimension is replaced by the Littlestone dimension [SB14, Chap. 21].

4.6.2 Lower bounds

We now prove that the upper bounds of Section 4.5 are sharp, up to constants and the factor $\log(1/\varepsilon)$. All lower bounds hold for every learning algorithm, proper or not, and require no measurability assumption on $\mathcal{H}$. They rest on the minimax toolkit of Chapter 3, which we recall in the form used here. For probability measures P_0, P_1 on the same space and any test ϕ with values in $\{0,1\}$, Le Cam's lemma (Lemma 3.53) together with the Bretagnolle–Huber inequality (Proposition 3.32) gives

$$\max_{j\in\{0,1\}} P_j(\phi \neq j) \;\geq\; \frac{1 - \mathrm{TV}(P_0, P_1)}{2} \;\geq\; \frac{1}{4} e^{-\mathrm{KL}(P_0 \| P_1)}; \tag{4.9}$$

Assouad's lemma (Lemma 3.66) states that if $(P_b)_{b\in\{0,1\}^m}$ are probability measures and $\hat{b}$ is any estimator with values in $\{0,1\}^m$, then, with ρ_H the Hamming distance and $b^{(j)}$ the vector b with its j-th coordinate flipped,

$$\frac{1}{2^m}\sum_{b}\mathbb{E}_b\big[\rho_H(\hat{b},b)\big] \;\geq\; \frac{m}{2}\min_{b,\,j}\Big(1-\mathrm{TV}\big(P_b,P_{b^{(j)}}\big)\Big); \tag{4.10}$$

Pinsker's inequality is $\mathrm{TV}\le\sqrt{\mathrm{KL}/2}$ (Proposition 3.28), and relative entropy tensorises, $\mathrm{KL}(P^{\otimes n}\|Q^{\otimes n}) = n\,\mathrm{KL}(P\|Q)$ (Proposition 3.20). Finally, Lemma 3.71 – the elementary step at the end of the proof of the no-free-lunch theorem – says that a random variable $Z\in[0,1]$ with $\mathbb{E}Z\ge\frac14$ satisfies $\mathbb{P}(Z\ge\frac18)\ge\frac17$ (indeed $\frac14\le\mathbb{E}Z\le\frac18+\frac78\mathbb{P}(Z\ge\frac18)$).

We shall need the relative entropy between two coins with opposite biases.

Lemma 4.60 (Two opposite coins)**.** *For $\gamma\in[0,\frac12)$,*

$$\mathrm{kl}\big(\tfrac12+\gamma,\tfrac12-\gamma\big) = 2\gamma\log\frac{1+2\gamma}{1-2\gamma} \;\le\; \frac{8\gamma^2}{1-4\gamma^2}.$$

Proof. The identity is a direct computation from Notation 2.22. For $u\in[0,1)$, $\log\frac{1+u}{1-u} = 2\sum_{k\ge0}\frac{u^{2k+1}}{2k+1}\le 2u\sum_{k\ge0}u^{2k} = \frac{2u}{1-u^2}$; take $u=2\gamma$. (This is the computation made, with $\varepsilon=2\gamma$, in the proof of Proposition 3.60.) □

The bound is sharp as $\gamma\to0$, where both sides are equivalent to $8\gamma^2$. Its right-hand side is exactly half the χ^2-divergence $(2\gamma)^2/(\frac14-\gamma^2)$ between the two coins, in accordance with the general fact that the relative entropy between two nearby laws is asymptotically half their χ^2-divergence.

Theorem 4.61 (Agnostic lower bound)**.** *Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H})\ge d\ge1$ (d finite).*

(a) For $\varepsilon\in(0,\frac{1}{16}]$ and $\delta\in(0,\frac17)$,

$$n_{\mathcal{H}}(\varepsilon,\delta) \;\ge\; \frac{3d}{1024\,\varepsilon^2}.$$

(b) For $\varepsilon\in(0,\frac12]$ and $\delta\in(0,\frac14)$,

$$n_{\mathcal{H}}(\varepsilon,\delta) \;\ge\; \frac{3}{8\varepsilon^2}\log\frac{1}{4\delta}.$$

(c) Consequently, for $\varepsilon\in(0,\frac{1}{16}]$ and $\delta\in(0,\frac{1}{16}]$,

$$n_{\mathcal{H}}(\varepsilon,\delta) \;\ge\; \frac{3}{2048}\,\frac{d+\log(1/\delta)}{\varepsilon^2}.$$

Proof. Fix a learning algorithm $\mathcal{A}$ and a sample size n. As explained in Remark 4.3, it suffices to exhibit, when n is below the stated bound, a distribution $\mathcal{D}$ for which $\mathbb{P}(L_{\mathcal{D}}(\mathcal{A}_n(S_n))\ge\inf_h L_{\mathcal{D}}(h)+\varepsilon)>\delta$.

(a) Let $C=\{x_1,\dots,x_d\}$ be shattered by $\mathcal{H}$ and $\gamma=4\varepsilon\le\frac14$. For $b\in\{0,1\}^d$ let $\mathcal{D}_b$ be the law of (X,Y) where X is uniform on C and, given $X=x_j$, $Y\sim\mathrm{Ber}(\frac12+\gamma(2b_j-1))$: the label is b_j with probability $\frac12+\gamma$. Since C is shattered, there is $h_b\in\mathcal{H}$ with $h_b(x_j)=b_j$ for all j,

and h_b is a Bayes classifier of $\mathcal{D}_b$. For every classifier g, predicting t at x_j costs $\frac{1}{2}-\gamma$ if $t = b_j$ and $\frac{1}{2}+\gamma$ otherwise, so that

$$L_{\mathcal{D}_b}(g) - \inf_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h) = L_{\mathcal{D}_b}(g) - L_{\mathcal{D}_b}(h_b) = \frac{2\gamma}{d}\,\rho_H\big(\hat{b}(g), b\big) = 8\varepsilon\,\frac{\rho_H(\hat{b}(g), b)}{d},$$

where $\hat{b}(g) = (g(x_1), \dots, g(x_d))$ (this is also the excess-risk identity (1.6), since $\big|2\eta_b - 1\big| = 2\gamma$ on C). Let $P_b = \mathcal{D}_b^{\otimes n}$ and $\hat{b} = \hat{b}(\mathcal{A}_n(\mathcal{S}_n))$, an estimator of b. The laws $\mathcal{D}_b$ and $\mathcal{D}_{b^{(j)}}$ differ only in the conditional law of Y given $X = x_j$, an event of probability $1/d$, so that, by the chain rule (Proposition 3.19) with equal first marginals, $\mathrm{KL}(\mathcal{D}_b \| \mathcal{D}_{b^{(j)}}) = \frac{1}{d}\,\mathrm{kl}(\frac{1}{2}+\gamma, \frac{1}{2}-\gamma)$ (the two coins are interchanged, and $\mathrm{kl}(\frac{1}{2}-\gamma, \frac{1}{2}+\gamma) = \mathrm{kl}(\frac{1}{2}+\gamma, \frac{1}{2}-\gamma)$). By tensorisation and Lemma 4.60, using $1 - 4\gamma^2 = 1 - 64\varepsilon^2 \geq \frac{3}{4}$,

$$\mathrm{KL}\big(P_b \| P_{b^{(j)}}\big) \leq \frac{n}{d}\cdot\frac{8\gamma^2}{1-4\gamma^2} \leq \frac{512\,n\,\varepsilon^2}{3d}.$$

If $n \leq 3d/(1024\varepsilon^2)$, this is at most $\frac{1}{2}$, and Pinsker's inequality gives $\mathrm{TV}(P_b, P_{b^{(j)}}) \leq \frac{1}{2}$. Assouad's inequality (4.10) then yields $2^{-d}\sum_b \mathbb{E}_b[\rho_H(\hat{b}, b)/d] \geq \frac{1}{4}$, so there is b with $\mathbb{E}_b[Z] \geq \frac{1}{4}$, where $Z = \rho_H(\hat{b}, b)/d \in [0,1]$. By Lemma 3.71, $P_b(Z \geq \frac{1}{8}) \geq \frac{1}{7} > \delta$, and on the event $\{Z \geq \frac{1}{8}\}$ the excess risk is $8\varepsilon Z \geq \varepsilon$. Thus (4.1) fails for $\mathcal{D}_b$ at every $n \leq 3d/(1024\varepsilon^2)$, and $n_{\mathcal{A}}(\varepsilon, \delta) > 3d/(1024\varepsilon^2)$.

(b) Let x_1 be a point shattered by $\mathcal{H}$ and $\gamma = \varepsilon/2 \leq \frac{1}{4}$. For $t \in \{0,1\}$, let $\mathcal{D}_t$ be the law of (x_1, Y) with $Y \sim \mathrm{Ber}(\frac{1}{2} + \gamma(2t-1))$. The risk of g under $\mathcal{D}_t$ is $\frac{1}{2}-\gamma$ if $g(x_1) = t$ and $\frac{1}{2}+\gamma$ otherwise, and both values are attained in $\mathcal{H}$; hence the excess risk of g is $2\gamma\,\mathbb{1}\{g(x_1) \neq t\} = \varepsilon\,\mathbb{1}\{g(x_1) \neq t\}$, and (4.1) fails under $\mathcal{D}_t$ as soon as $P_t(\phi \neq t) > \delta$, where $\phi = \mathcal{A}_n(\mathcal{S}_n)(x_1)$ is a test between $P_0 = \mathcal{D}_0^{\otimes n}$ and $P_1 = \mathcal{D}_1^{\otimes n}$. By (4.9), tensorisation and Lemma 4.60, using $1 - 4\gamma^2 = 1 - \varepsilon^2 \geq \frac{3}{4}$,

$$\max_t P_t(\phi \neq t) \;\geq\; \frac{1}{4}\exp\Big(-n\frac{8\gamma^2}{1-4\gamma^2}\Big) \;\geq\; \frac{1}{4}\exp\Big(-\frac{8n\varepsilon^2}{3}\Big),$$

which exceeds δ when $n < \frac{3}{8\varepsilon^2}\log\frac{1}{4\delta}$. Hence $n_{\mathcal{A}}(\varepsilon, \delta) \geq \frac{3}{8\varepsilon^2}\log\frac{1}{4\delta}$.

(c) For $\delta \leq \frac{1}{16}$, $\log\frac{1}{4\delta} \geq \frac{1}{2}\log\frac{1}{\delta}$. Hence $n_{\mathcal{H}}(\varepsilon, \delta) \geq \max\big(\frac{3d}{1024\varepsilon^2}, \frac{3\log(1/\delta)}{16\varepsilon^2}\big) \geq \frac{1}{2}\big(\frac{3d}{1024\varepsilon^2} + \frac{3\log(1/\delta)}{16\varepsilon^2}\big)$, which is at least the stated bound. □

Remark 4.62 (Where the lower bound comes from). The proof of (a) is a reduction to d independent coin-bias testing problems, one per point of the shattered set. Each point receives about n/d samples; a coin of bias $\frac{1}{2} \pm \gamma$ cannot be identified with fewer than about $1/\gamma^2$ tosses; so for $n \lesssim d/\gamma^2$ a constant fraction of the points are misclassified, each costing $2\gamma/d$ in excess risk. The shattering hypothesis is used exactly once: to guarantee that every one of the 2^d Bayes classifiers belongs to $\mathcal{H}$, so that the learner is compared with the Bayes risk and cannot hide behind the approximation error. The constant $3/1024$ is poor, and no effort has been made to improve it (the textbook proof of [SB14, Sec. 28.2.2], announced there as $n_{\mathcal{H}}(\varepsilon, \frac{1}{8}) \geq 8d/\varepsilon^2$ in our notation, actually yields $d/(512\varepsilon^2)$, less than (a)), but its order is right: by Theorem 4.52, d/ε^2 is also an upper bound. The labels of the hard distributions are *noisy*: the agnostic lower bound is a statement about learning in the presence of noise, and it is precisely the noise level $\frac{1}{2}-\gamma$, close to $\frac{1}{2}$, that makes the factor $1/\varepsilon^2$ appear. Under a margin condition (Section 2.7.2) the hard distributions above are excluded, and faster rates are possible [MN06].

Theorem 4.63 (Realizable lower bound). *Let $\mathcal{H}$ be a hypothesis class with* $\mathrm{VCdim}(\mathcal{H}) \geq d \geq 2$ *(d finite).*

(a) For $\varepsilon \in (0, \frac{1}{8}]$ and $\delta \in (0, \frac{1}{7})$,

$$n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon, \delta) \;\geq\; \frac{d-1}{16\,\varepsilon}.$$

(b) For $\varepsilon \in (0, \frac{1}{2}]$ and $\delta \in (0, \frac{1}{2})$,

$$n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon, \delta) \;\geq\; \frac{1}{2\varepsilon} \log \frac{1}{2\delta}.$$

(c) Consequently, for $\varepsilon \in (0, \frac{1}{8}]$ and $\delta \in (0, \frac{1}{8}]$,

$$n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon, \delta) \;\geq\; \frac{d + \log(1/\delta)}{64\,\varepsilon}.$$

Proof. Fix $\mathcal{A}$ and n.

(a) Let $\{x_0, x_1, \dots, x_m\}$, $m = d-1 \geq 1$, be shattered by $\mathcal{H}$. Let $\mathcal{D}_X$ put mass $1 - 8\varepsilon$ on x_0 and mass $8\varepsilon/m$ on each of $x_1, \dots, x_m$. For $b \in \{0,1\}^m$ let f_b be the labelling $f_b(x_0) = 0$, $f_b(x_j) = b_j$, and $\mathcal{D}_b$ the law of $(X, f_b(X))$ with $X \sim \mathcal{D}_X$. Since the set is shattered, some $h \in \mathcal{H}$ coincides with f_b on it, so $\mathcal{D}_b$ is realizable. For every classifier g,

$$L_{\mathcal{D}_b}(g) \;\geq\; \frac{8\varepsilon}{m} \sum_{j=1}^{m} \mathbb{1}\{g(x_j) \neq b_j\}.$$

Now draw b uniformly at random, independently of $X_1, \dots, X_n$ i.i.d. $\sim \mathcal{D}_X$, and let $\mathcal{S}_n = ((X_i, f_b(X_i)))_{i \leq n}$; conditionally on b, this is a sample from $\mathcal{D}_b$. Let $W \subseteq \{1, \dots, m\}$ be the (random) set of indices j such that x_j does not appear among $X_1, \dots, X_n$. The sample $\mathcal{S}_n$ is a function of $(X_1, \dots, X_n)$ and of $(b_j)_{j \notin W}$, so that, conditionally on these, the bits $(b_j)_{j \in W}$ are still i.i.d. uniform, while $\widehat{h}_n = \mathcal{A}_n(\mathcal{S}_n)$ is fixed. Hence $\mathbb{P}(\widehat{h}_n(x_j) \neq b_j \mid X_1, \dots, X_n, (b_i)_{i \notin W}) = \frac{1}{2}$ for $j \in W$, and the variable $Z := \frac{1}{m} \sum_{j \in W} \mathbb{1}\{\widehat{h}_n(x_j) \neq b_j\} \in [0,1]$ satisfies

$$\mathbb{E}Z = \frac{\mathbb{E}|W|}{2m} = \frac{1}{2}\Big(1 - \frac{8\varepsilon}{m}\Big)^n \;\geq\; \frac{1}{2}\Big(1 - \frac{8\varepsilon n}{m}\Big) \;\geq\; \frac{1}{4} \qquad \text{if } n \leq \frac{m}{16\varepsilon}.$$

Since $\mathbb{E}Z$ is the average over b of $\mathbb{E}_b Z$, there is a b with $\mathbb{E}_b Z \geq \frac{1}{4}$, and Lemma 3.71 gives $P_b(Z \geq \frac{1}{8}) \geq \frac{1}{7} > \delta$, where $P_b = \mathcal{D}_b^{\otimes n}$. On that event, $L_{\mathcal{D}_b}(\widehat{h}_n) \geq 8\varepsilon Z \geq \varepsilon$. Hence $n^{\mathrm{re}}_{\mathcal{A}}(\varepsilon, \delta) > m/(16\varepsilon)$.

(b) Let $\{x_0, x_1\}$ be shattered, and let $\mathcal{D}_X$ put mass $1 - \varepsilon$ on x_0 and ε on x_1. For $t \in \{0,1\}$ let $\mathcal{D}_t$ be the law of $(X, f_t(X))$, where $f_t(x_0) = 0$ and $f_t(x_1) = t$; both are realizable. If x_1 does not appear in the sample – an event E of probability $(1-\varepsilon)^n$ under both $\mathcal{D}_0$ and $\mathcal{D}_1$ – the samples drawn under $\mathcal{D}_0$ and under $\mathcal{D}_1$ coincide (couple them through the same $X_1, \dots, X_n$), hence so does $\widehat{h}_n(x_1)$, and $\widehat{h}_n(x_1) \neq t$ for exactly one t. In that case $L_{\mathcal{D}_t}(\widehat{h}_n) \geq \varepsilon$. Therefore, with $P_t = \mathcal{D}_t^{\otimes n}$,

$$P_0\big(L_{\mathcal{D}_0}(\widehat{h}_n) \geq \varepsilon\big) + P_1\big(L_{\mathcal{D}_1}(\widehat{h}_n) \geq \varepsilon\big) \;\geq\; \mathbb{P}(E) = (1-\varepsilon)^n,$$

and one of the two probabilities is at least $\frac{1}{2}(1-\varepsilon)^n$. Since $\log \frac{1}{1-\varepsilon} \leq \frac{\varepsilon}{1-\varepsilon} \leq 2\varepsilon$ for $\varepsilon \leq \frac{1}{2}$, this exceeds δ when $n < \frac{1}{2\varepsilon} \log \frac{1}{2\delta}$.

(c) For $d \geq 2$, $d-1 \geq d/2$, and for $\delta \leq \frac{1}{8}$, $\log \frac{1}{2\delta} \geq \frac{2}{3} \log \frac{1}{\delta}$. Hence $n^{\mathrm{re}}_{\mathcal{H}} \geq \max\big(\frac{d}{32\varepsilon}, \frac{\log(1/\delta)}{3\varepsilon}\big) \geq \frac{1}{2}\big(\frac{d}{32\varepsilon} + \frac{\log(1/\delta)}{3\varepsilon}\big) \geq \frac{d + \log(1/\delta)}{64\varepsilon}$. ∎

Remark 4.64 (On the realizable construction). Part (a) is due to Ehrenfeucht, Haussler, Kearns and Valiant [Ehr+89], who obtained the constant $\frac{1}{32}$ and the bound $\frac{1-\varepsilon}{\varepsilon}\log\frac{1}{\delta}$ for part (b). It is the no-free-lunch argument of Theorem 3.72 with one change: instead of spreading the mass uniformly over the shattered set, which would give an error of order 1 but only for $n \lesssim d$, we put almost all the mass on one "easy" point x_0 and spread a mass 8ε over the $d-1$ others. Each of these is then seen only after about d/ε examples, and as long as a constant fraction of them is unseen, the learner – who knows nothing about their labels – errs on half of them, which costs a constant fraction of 8ε. The hypothesis $d \geq 2$ cannot be dropped in (b): the class $\{0, 1\}$ of the two constant functions has VC dimension 1, and in the realizable case a single example reveals the target, whatever ε and δ. Part (a) holds for $d = 1$ in the trivial form $n^{\mathrm{re}} \geq 0$, and (b) holds for $d = 1$ as soon as $\mathcal{H}$ contains two functions that agree at one point and differ at another.

4.6.3 The quantitative statement

Theorem 4.65 (Fundamental theorem, quantitative form)**.** *Let $\mathcal{H}$ satisfy Assumption 4.1, with $2 \leq \mathrm{VCdim}(\mathcal{H}) = d < \infty$. For all $\varepsilon \in (0, \frac{1}{32})$ and $\delta \in (0, \frac{1}{16}]$:*

$$\begin{aligned}
\frac{3}{8192}\,\frac{d+\log(1/\delta)}{\varepsilon^2} \;\leq\; n^{\mathrm{UC}}_{\mathcal{H}}(\varepsilon,\delta) \;&\leq\; \left\lceil \frac{16\,d\log(8e/\varepsilon^2)+2\log(2/\delta)}{\varepsilon^2} \right\rceil,\\
\frac{3}{2048}\,\frac{d+\log(1/\delta)}{\varepsilon^2} \;\leq\; n_{\mathcal{H}}(\varepsilon,\delta) \;&\leq\; \left\lceil \frac{16\,d\log(8e/\varepsilon^2)+8\log(2/\delta)}{\varepsilon^2} \right\rceil,\\
\frac{d+\log(1/\delta)}{64\,\varepsilon} \;\leq\; n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) \;&\leq\; \left\lceil \frac{8}{\varepsilon\log 2}\left(d\log\frac{32}{\varepsilon}+\log\frac{2}{\delta}\right) \right\rceil.
\end{aligned}$$

The upper bounds for $n_{\mathcal{H}}$ and $n^{\mathrm{re}}_{\mathcal{H}}$ are achieved by any empirical risk minimiser.

Proof. The upper bounds are Theorems 4.49 and 4.50 and Corollary 4.56; the lower bounds for $n_{\mathcal{H}}$ and $n^{\mathrm{re}}_{\mathcal{H}}$ are Theorems 4.61 and 4.63. For $n^{\mathrm{UC}}_{\mathcal{H}}$, Corollary 4.10 gives $n_{\mathcal{H}}(\varepsilon',\delta) \leq n^{\mathrm{UC}}_{\mathcal{H}}(\varepsilon,\delta)$ for every $\varepsilon' > 2\varepsilon$; for $\varepsilon' \in (2\varepsilon, \frac{1}{16}]$, Theorem 4.61(c) gives $n^{\mathrm{UC}}_{\mathcal{H}}(\varepsilon,\delta) \geq \frac{3}{2048}(d+\log(1/\delta))/\varepsilon'^2$, and letting $\varepsilon' \downarrow 2\varepsilon$ yields the claim. □

Up to the constants and the factors $\log(1/\varepsilon)$, the three sample complexities are thus $(d+\log(1/\delta))/\varepsilon^2$, $(d+\log(1/\delta))/\varepsilon^2$ and $(d+\log(1/\delta))/\varepsilon$; the logarithmic factors can be removed by Theorem 4.52 in the first two cases and by Hanneke's theorem (Remark 4.57) in the third, so that, with universal constants $\kappa, \kappa' > 0$,

$$\begin{aligned}
\kappa\,\frac{d+\log(1/\delta)}{\varepsilon^2} \leq n_{\mathcal{H}}(\varepsilon,\delta) &\leq \kappa'\,\frac{d+\log(1/\delta)}{\varepsilon^2},\\
\kappa\,\frac{d+\log(1/\delta)}{\varepsilon} \leq n^{\mathrm{re}}_{\mathcal{H}}(\varepsilon,\delta) &\leq \kappa'\,\frac{d+\log(1/\delta)}{\varepsilon}.
\end{aligned}$$

The comparison with Proposition 4.12 is striking: everything is as if $\mathcal{H}$ were a finite class of cardinality e^d. Note, however, that for a finite class the two descriptions need not agree: $\mathrm{VCdim}(\mathcal{H}) \leq \log_2|\mathcal{H}|$ always, and the bound in terms of the VC dimension can be much better (Example 4.23); conversely, the factor $\log(1/\varepsilon)$ in our upper bounds makes $\log|\mathcal{H}|$ preferable when $|\mathcal{H}|$ is small.

Remark 4.66 (The constants). The constants of Theorem 4.65 are explicit but far from optimal, and no effort has been made to optimise them; the ratio between the upper and the lower bounds is of order $10^4\log(1/\varepsilon)$ in the agnostic case. They are nonetheless worth having, if only to check that nothing is hidden: in particular, the upper bounds involve no unspecified

constant, and could be used as they stand – with the caveat that they are worst-case over all distributions, and therefore very pessimistic for any particular one. For half-planes ($d = 3$) with $\varepsilon = 0.05$ and $\delta = 0.01$, for instance, the upper bound of Theorem 4.50 asks for about $1.9 \cdot 10^5$ examples, while the lower bounds of Theorem 4.61 only certify that 483 are needed (this is part (b); the combined form (c) gives 5). The truth, for the worst distribution, lies in between; both the constants of the upper bounds and the lower-bound constructions are crude. The data-dependent bound (4.5) is the practical answer to this pessimism.

4.7 Complements

4.7.1 A fixed distribution, almost sure convergence

The bounds of Theorem 4.49 are distribution-free. The intermediate steps, however, are not, and they give more.

Proposition 4.67 (VC entropy). *Let $\mathcal{H}$ satisfy Assumption 4.1, let $\mathcal{D}$ be a distribution, and call $\mathbb{E}\log|\mathcal{H}(X_1,\dots,X_n)|$ the* VC entropy *of $\mathcal{H}$ under $\mathcal{D}_X$. Then*

$$\mathbb{E}\Big[\sup_{h\in\mathcal{H}}\big(L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h)\big)\Big] \;\le\; \sqrt{\frac{2\,\mathbb{E}\log|\mathcal{H}(X_1,\dots,X_n)|}{n}}.$$

Proof. By Theorem 4.41, Lemma 4.40(v) and the first inequality of Corollary 4.45, the left-hand side is at most $2\mathbb{E}\sqrt{\log|\mathcal{H}(X_1,\dots,X_n)|\,/(2n)}$, and Jensen's inequality for the concave square root gives the claim. □

Since $\mathbb{E}\log|\mathcal{H}(X_1,\dots,X_n)| \le \log\tau_{\mathcal{H}}(n)$, this is never worse than Theorem 4.49, and it can be much better: if $\mathcal{D}_X$ is supported by m points, then $|\mathcal{H}(X_1,\dots,X_n)| \le 2^m$ and the bound is $\sqrt{2m\log 2/n}$ whatever $\mathrm{VCdim}(\mathcal{H})$. Vapnik and Chervonenkis [VC71] proved that the condition $\frac{1}{n}\mathbb{E}\log|\mathcal{H}(X_1,\dots,X_n)| \to 0$ is not only sufficient but also necessary for the uniform convergence of the empirical risks to the risks under the fixed distribution $\mathcal{D}$ (admitted). The distribution-free condition $\mathrm{VCdim}(\mathcal{H}) < \infty$ is the condition under which this holds uniformly in $\mathcal{D}$.

Corollary 4.68 (Uniform strong law of large numbers). *Let $\mathcal{H}$ satisfy Assumption 4.1 and have finite VC dimension. For every distribution $\mathcal{D}$, $\sup_{h\in\mathcal{H}}\left|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\right| \to 0$ almost surely as $n\to\infty$. In particular (Glivenko–Cantelli), the empirical distribution function of an i.i.d. real sample converges uniformly to the true one, almost surely.*

Proof. Let $d = \mathrm{VCdim}(\mathcal{H}) \ge 1$ (the case $d = 0$ is the strong law of large numbers) and apply (4.6) with $\delta = n^{-2}$: the event $\{\sup_h \left|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\right| > a_n\}$, with $a_n = \sqrt{2d\log(en/d)/n} + \sqrt{\log(2n^2)/(2n)} \to 0$, has probability at most n^{-2}, which is summable. By the Borel–Cantelli lemma, almost surely $\sup_h \left|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\right| \le a_n$ for all n large enough (compare Section 1.10.2). The Glivenko–Cantelli theorem is the case of half-lines with $Y \equiv 0$ (Example 4.17). □

4.7.2 Choosing the class: structural risk minimisation

In practice the class is not given: one has a nested family $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots$ of increasing complexity – polynomials of increasing degree, networks of increasing size – and must balance the approximation error $\inf_{h\in\mathcal{H}_j} L_{\mathcal{D}}(h)$, which decreases with j, against the estimation error, which increases with $\mathrm{VCdim}(\mathcal{H}_j)$. Since the bounds of this section hold with explicit constants, they can be used to make this choice from the data.

Theorem 4.69 (Structural risk minimisation). *Let $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots$ be classes satisfying Assumption 4.1, with $1 \le d_j = \mathrm{VCdim}(\mathcal{H}_j) < \infty$. For $\delta \in (0,1)$ let*

$$\mathrm{pen}_j := \sqrt{\frac{2d_j \log(en/d_j)}{n}} + \sqrt{\frac{\log(2j^2/\delta)}{2n}} \quad \text{if } n \ge d_j, \qquad \mathrm{pen}_j := +\infty \quad \text{otherwise.}$$

(i) With probability at least $1-\delta$, simultaneously for all j and all $h \in \mathcal{H}_j$, $L_{\mathcal{D}}(h) \le L_{\mathcal{S}_n}(h) + \mathrm{pen}_j$.

(ii) Assume $n \ge d_1$, and let $\widehat{h}_n$ minimise $L_{\mathcal{S}_n}(h) + \mathrm{pen}_j$ over all pairs (j,h) with $h \in \mathcal{H}_j$. Then for every fixed j, with probability at least $1-2\delta$,

$$L_{\mathcal{D}}(\widehat{h}_n) \;\le\; \inf_{h \in \mathcal{H}_j} L_{\mathcal{D}}(h) + \mathrm{pen}_j + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. (i) For each j with $n \ge d_j$, (4.3) and Corollary 4.45 applied at level $\delta_j = \delta/(2j^2)$ show that the inequality fails for some $h \in \mathcal{H}_j$ with probability at most δ_j; and $\sum_j \delta_j = \delta\pi^2/12 < \delta$. (ii) The minimiser exists: within each $\mathcal{H}_j$ the empirical risk takes finitely many values, and since $\mathrm{pen}_j \ge \sqrt{\log(2j^2/\delta)/(2n)} \to \infty$ while $L_{\mathcal{S}_n} \le 1$ and $\mathrm{pen}_1 < \infty$, only finitely many j compete. Let $\hat{\jmath}$ be the index selected. On the event of (i), for every $h \in \mathcal{H}_j$, $L_{\mathcal{D}}(\widehat{h}_n) \le L_{\mathcal{S}_n}(\widehat{h}_n) + \mathrm{pen}_{\hat{\jmath}} \le L_{\mathcal{S}_n}(h) + \mathrm{pen}_j$. Apply this to a fixed $h \in \mathcal{H}_j$ with $L_{\mathcal{D}}(h) \le \inf_{h' \in \mathcal{H}_j} L_{\mathcal{D}}(h') + \varepsilon'$, and bound $L_{\mathcal{S}_n}(h)$ by the one-sided Hoeffding inequality (first half of (1.28)) outside an event of probability δ; then let $\varepsilon' \downarrow 0$ as in the proof of Theorem 4.50. □

The rule performs, up to the term $\log(2j^2)$ in the penalty, as well as the bound (4.7) for empirical risk minimisation over the best class *chosen with the knowledge of* $\mathcal{D}$. Note what has been gained and what has been lost. A countable union $\bigcup_j \mathcal{H}_j$ of VC classes may have infinite VC dimension – think of polynomial threshold functions of all degrees – and is then not PAC learnable (Theorem 4.28); yet structural risk minimisation learns it in the weaker, *non-uniform* sense of Remark 4.4: the sample size needed to compete with a given $h \in \mathcal{H}_j$ depends on j, but not otherwise on $\mathcal{D}$ [SB14, Chap. 7]. Replacing the distribution-free penalty by the empirical Rademacher complexity (4.5) gives data-dependent model selection procedures [Kol01; BBL02].

4.8 Notes and references

The PAC model is due to Valiant [Val84]; its agnostic version, in which no assumption is made on the labels, to Haussler [Hau92] and Kearns, Schapire and Sellie [KSS94]. The connection with the much older work of Vapnik and Chervonenkis [VC71], who had proved the uniform law of large numbers for classes of finite VC dimension and the necessity of the condition for uniform convergence, was made by Blumer, Ehrenfeucht, Haussler and Warmuth [Blu+89], who proved the realizable half of the fundamental theorem (including Theorem 4.55 and Corollary 4.56); the realizable lower bound is [Ehr+89]. The name "fundamental theorem of statistical learning" and the six-fold statement of Theorem 4.59 are those of [SB14, Thms. 6.7–6.8], whose Chapters 2–6 and 26–28 cover the material of this chapter at a more elementary level, and whose proofs we have reorganised around Rademacher complexities and McDiarmid's inequality. Vapnik's own account is [Vap98]; [AB99] is a thorough treatment of the combinatorial side, including the agnostic lower bounds, for which see also [DL95] and [DGL96, Chap. 14]. The textbooks [MRT18], [Wai19, Chap. 4] and [Ver18, Chap. 8] present the same circle of ideas with different emphases.

The Sauer–Shelah lemma and its history are discussed in Remark 4.38; the presentation by induction follows [SB14, Sec. 6.5], and the entropy form of Lemma 4.34 is classical in information theory. Radon's theorem is [Rad21]; the use of vector spaces of functions to bound VC dimensions (Exercise 4.7) is due to Dudley [Dud78], see also [WD81].

Symmetrisation goes back to [VC71] in its double-sample form and to Giné and Zinn [GZ84] in its Rademacher form; [LT91, Chap. 4] is the reference for Rademacher processes, including the *contraction principle* (Lemma 6.1), which extends Lemma 4.40(v) to Lipschitz losses and makes Rademacher complexities the natural tool for margin-based and real-valued problems. Rademacher complexities were introduced as data-dependent penalties by Koltchinskii [Kol01], Koltchinskii and Panchenko [KP00], and Bartlett, Boucheron and Lugosi [BBL02], and systematised by Bartlett and Mendelson [BM02], from whom Theorem 4.46 is taken; the localised versions that yield fast rates are developed in [Kol11]. Massart's lemma is [Mas00, Lemma 5.2]; see also [SB14, Lemma 26.8]. The combination "McDiarmid for the fluctuations, symmetrisation and a maximal inequality for the mean" that structures Section 4.5 is the approach of the survey [BBL05].

Chaining is due to Dudley [Dud67]; its application to VC classes, which removes the logarithm in Theorem 4.49, rests on bounds on their covering numbers, due to Dudley [Dud78] and, in sharp form, to Haussler [Hau95]; the corresponding tail bounds are Talagrand's [Tal94]. The optimal realizable sample complexity was established by Hanneke [Han16]; a strikingly simple optimal learner – bagging of empirical risk minimisers – was later analysed by Larsen [Lar23]. The measurability questions of Remark 4.54 are treated in [Pol84] and [Dud99, Chap. 5].

Concluding the proof of uniform convergence with Markov's inequality instead of McDiarmid's would cost a factor $1/\delta^2$ instead of $\log(1/\delta)$ in the sample complexity (Remark 4.51). Our bound on the Rademacher complexity of a class of finite VC dimension is better by a factor 2 than the one usually quoted (for instance in [SB14, Sec. 28.1]), thanks to the centring in the proof of Corollary 4.45.

4.9 Exercises

Exercise 4.1 (⋆ Singletons)**.** Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{\mathbb{1}_{\{z\}} : z \in \mathbb{R}\} \cup \{0\}$.

(a) Show that $\mathrm{VCdim}(\mathcal{H}) = 1$ and compute $\tau_{\mathcal{H}}(n)$.

(b) In the realizable case, describe an empirical risk minimiser and show directly that its risk is at most $\log(1/\delta)/n$ with probability $1 - \delta$, for every realizable $\mathcal{D}$.

Exercise 4.2 (⋆ Two-sided thresholds: Sauer is not always tight)**.** Let $\mathcal{H} = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\} \cup \{\mathbb{1}_{[a,\infty)} : a \in \mathbb{R}\}$. Show that $\mathrm{VCdim}(\mathcal{H}) = 2$ and that $\tau_{\mathcal{H}}(n) = 2n$ for every $n \geq 1$. Compare with $\Phi_2(n)$, and conclude that the growth function of a class of VC dimension d can be much smaller than $\Phi_d(n)$.

Exercise 4.3 (⋆ Unions of intervals)**.** Let $N \geq 1$ and let $\mathcal{I}_N$ be the class of indicators of unions of at most N closed intervals of $\mathbb{R}$. Show that $\mathrm{VCdim}(\mathcal{I}_N) = 2N$. Compare with the bound given by Proposition 4.37.

Exercise 4.4 (⋆ The order of the quantifiers: finite sets)**.** Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{H} = \{\mathbb{1}_A : A \subseteq \mathbb{N} \text{ finite}\}$.

(a) Show that $\mathrm{VCdim}(\mathcal{H}) = \infty$, so that $\mathcal{H}$ is not PAC learnable.

(b) Let $\mathcal{D}$ be realizable by $\mathcal{H}$, with target $\mathbb{1}_A$, and let $\widehat{h}_n$ be the indicator of the set of sample points labelled 1. Show that $\mathbb{E}L_{\mathcal{D}}(\widehat{h}_n) \le |A|/(en)$. Why does this not contradict (a)?

Exercise 4.5 (⋆ Homogeneous half-spaces)**.** Let $\mathcal{H} = \{x \mapsto \mathbb{1}\{\langle w, x\rangle \ge 0\} : w \in \mathbb{R}^k\}$ be the class of homogeneous half-spaces of $\mathbb{R}^k$. Show that $\mathrm{VCdim}(\mathcal{H}) = k$. *Hint: for the upper bound, write a linear dependence $\sum_i \alpha_i x_i = 0$ between $k+1$ points and consider the labelling $\mathbb{1}\{\alpha_i > 0\}$, distinguishing whether some α_i is negative.*

Exercise 4.6 (⋆ Euclidean balls)**.** Let $\mathcal{H}$ be the class of indicators of closed Euclidean balls of $\mathbb{R}^k$.

(a) Using the map $x \mapsto (x, \|x\|^2) \in \mathbb{R}^{k+1}$ and Proposition 4.21, show that $\mathrm{VCdim}(\mathcal{H}) \le k+2$.

(b) Let C be a finite set and w, b such that $\langle w, x\rangle + b \ne 0$ for all $x \in C$. Show that the ball of centre $Rw/\|w\|$ and radius $\left(R^2 + 2Rb/\|w\|\right)^{1/2}$ induces on C the same labelling as the half-space $\{\langle w, x\rangle + b \ge 0\}$ for all R large enough. Deduce that $\mathrm{VCdim}(\mathcal{H}) \ge k+1$.

(In fact $\mathrm{VCdim}(\mathcal{H}) = k+1$ [Dud79]; the upper bound is more delicate, and we admit it.)

Exercise 4.7 (⋆ Classes defined by a vector space of functions)**.** Let V be a real vector space of functions $\mathcal{X} \to \mathbb{R}$, of finite dimension m, and $\mathcal{H} = \{x \mapsto \mathbb{1}\{g(x) \ge 0\} : g \in V\}$.

(a) Show that $\mathrm{VCdim}(\mathcal{H}) \le m$. *Hint: if $x_1, \dots, x_{m+1}$ are given, the linear map $g \mapsto (g(x_1), \dots, g(x_{m+1}))$ is not onto.*

(b) Deduce again the upper bounds of Proposition 4.21 and of Exercise 4.6(a), and show that the class of sets $\{x \in \mathbb{R}^k : p(x) \ge 0\}$, p a polynomial of degree at most q, has VC dimension at most $\binom{k+q}{q}$.

Exercise 4.8 (Thresholds: no logarithm in the realizable case)**.** Let $\mathcal{H} = \{\mathbb{1}_{(-\infty,a]} : a \in \mathbb{R}\}$ and let $\mathcal{D}$ be realizable by $\mathcal{H}$ with target $\mathbb{1}_{(-\infty,a^*]}$, the marginal $\mathcal{D}_X$ having a continuous distribution function. Show that every $h \in \mathcal{H}$ consistent with the sample satisfies $L_{\mathcal{D}}(h) < \varepsilon$, except on an event of probability at most $2(1-\varepsilon)^n$, and hence that, uniformly over such distributions, $\lceil \log(2/\delta)/\varepsilon \rceil$ examples suffice to guarantee $\mathbb{P}(L_{\mathcal{D}}(\widehat{h}_n) \ge \varepsilon) \le \delta$ for every empirical risk minimiser. Compare with Corollary 4.56 and Remark 4.57.

Exercise 4.9 (⋆ Linear classes have dimension-free Rademacher complexity)**.** Let $R, B > 0$, let $x_1, \dots, x_n \in \mathbb{R}^k$ with $\|x_i\|_2 \le R$, and let $\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\}$. Show that

$$\widehat{\mathfrak{R}}_x(\mathcal{F}) \;\le\; \frac{BR}{\sqrt{n}}.$$

Comment on the absence of k, by comparison with Proposition 4.21 and Corollary 4.45.

Exercise 4.10 (⋆ Two-sided symmetrisation)**.** Under the assumptions of Theorem 4.41, show that

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|P_n f - Pf\right| \;\le\; 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(Z_i)\right| \;\le\; 4\,\mathfrak{R}_n(\mathcal{F}) + \frac{2}{\sqrt{n}}.$$

Deduce that, for a class of VC dimension $d \ge 1$ and $n \ge d$, $\mathbb{E}\sup_h \left|L_{\mathcal{S}_n}(h) - L_{\mathcal{D}}(h)\right| \le \sqrt{8d\log(en/d)/n} + 2/\sqrt{n}$. Compare with the bound $4\mathfrak{R}_n(\mathcal{F}) + \frac{1}{2\sqrt{n}}$ obtained after Proposition 4.43.

Exercise 4.11 (⋆ The rate $n^{-1/2}$ cannot be improved)**.** Let $\xi_1, \dots, \xi_n$ be independent Rademacher variables and $S = \sum_i \xi_i$.

(a) Compute $\mathbb{E}S^2$ and $\mathbb{E}S^4$, and deduce from Hölder's inequality that $\mathbb{E}|S| \geq \sqrt{n/3}$.

(b) Let $X_1, \dots, X_n$ be i.i.d. with a continuous distribution function F and empirical distribution function F_n. Show that $\mathbb{E}\sup_a |F_n(a) - F(a)| \geq 1/(2\sqrt{3n})$.

(c) More generally, show that if $\mathcal{H}$ contains a hypothesis h with $L_{\mathcal{D}}(h) = \frac{1}{2}$, then $\mathbb{E}\sup_{h' \in \mathcal{H}} \left|L_{S_n}(h') - L_{\mathcal{D}}(h')\right| \geq 1/(2\sqrt{3n})$.

(The optimal constant in (a) is $\mathbb{E}|S| \geq \sqrt{n/2}$, due to Szarek [Sza76].)

Exercise 4.12 (Shattered samples have maximal complexity)**.** (a) Show that if $\mathcal{H}$ shatters $\{x_1, \dots, x_n\}$ (distinct points), then $\widehat{\mathfrak{R}}_x(\mathcal{H}) = \frac{1}{2}$.

(b) Suppose $\mathrm{VCdim}(\mathcal{H}) = \infty$. Show that for all $N \geq n$ there is a distribution $\mathcal{D}$ with $\mathfrak{R}_n(\ell \circ \mathcal{H}) \geq \frac{1}{2} - \frac{n^2}{4N}$, and hence $\sup_{\mathcal{D}} \mathfrak{R}_n(\ell \circ \mathcal{H}) = \frac{1}{2}$ for every n.

(c) Deduce from Proposition 4.43 a direct proof that a class of infinite VC dimension (satisfying Assumption 4.1) does not have the uniform convergence property.

Exercise 4.13 (Sauer–Shelah by shifting)**.** Let C be finite and $\mathcal{V} \subseteq \{0,1\}^C$. For $c \in C$ and $f \in \mathcal{V}$, let $f^{c \to 0}$ be f with its value at c set to 0, and define $T_c(f) = f^{c \to 0}$ if $f(c) = 1$ and $f^{c \to 0} \notin \mathcal{V}$, and $T_c(f) = f$ otherwise. Let $S_c(\mathcal{V}) = \{T_c(f) : f \in \mathcal{V}\}$.

(a) Show that $|S_c(\mathcal{V})| = |\mathcal{V}|$.

(b) Show that every set shattered by $S_c(\mathcal{V})$ is shattered by $\mathcal{V}$.

(c) Show that after finitely many shifts one obtains a family $\mathcal{V}^*$ with $S_c(\mathcal{V}^*) = \mathcal{V}^*$ for every c, and that such a family is *hereditary*: if $f \in \mathcal{V}^*$ and $g \leq f$ pointwise, then $g \in \mathcal{V}^*$.

(d) Show that a hereditary family shatters the support of each of its elements, and deduce Theorem 4.30.

Chapter 5

Beyond Sums of Independent Variables

Summary. The inequalities of Chapter 2 concern sums of independent variables; the only exception was the bounded-difference inequality, which we used freely in Chapter 4 without discussing its limitations. This chapter is about what lies beyond: general functions of independent variables, and sums of dependent ones. The organising principle is the one of Chapter 2, applied twice: *control the variance first, then the exponential moments*. At the level of the variance, the Efron–Stein inequality bounds the fluctuations of any function of independent variables by the sum of its coordinate-wise fluctuations. At the exponential level, three tools are developed: negative association, which lets Chernoff's method treat certain dependent variables as if they were independent; the bounded-difference inequality, proved again slowly and from scratch; and the entropy method, in its simplest form, which yields Poisson-like concentration for self-bounding functions. Everything is put to work on one problem, the estimation of the *missing mass* – the probability that the next observation is a symbol never seen before – where Good's estimator, based on the number of symbols seen exactly once, is shown to be accurate to order $n^{-1/2}$ whatever the underlying distribution, by an argument that goes back to Chernoff's roots: a direct computation of the variance of the tilted law.

5.1 Introduction

Consider n independent draws $X_1, \dots, X_n$ from an unknown probability distribution P on a countable alphabet $\mathcal{X}$ – words of a language, species in an ecosystem, configurations of a physical system, settings of a cipher machine. After n observations, some symbols have been seen, most often very unequally, and others have not. The *missing mass*

$$M_n := P\big(\mathcal{X} \setminus \{X_1, \dots, X_n\}\big) = \sum_{x \in \mathcal{X}} p(x)\, \mathbb{1}\{x \notin \{X_1, \dots, X_n\}\}$$

is the probability that the next observation X_{n+1} is new. It is a quantity of obvious practical importance – it tells a linguist how much of the language the corpus has not yet covered, a cryptanalyst how many messages will use a key never seen before, a biologist how many species remain to be discovered – and it has a remarkable property: although M_n depends on the unknown P in a complicated way, it can be *estimated* with accuracy $n^{-1/2}$ *without any assumption* on P. The estimator, due to Turing and Good, is simply the proportion of observations that are symbols seen only once in the sample.

To prove this one needs two concentration results, neither of which is covered by the inequalities of Chapter 2. The estimator is a function of the independent variables $X_1, \dots, X_n$,

but it is not a sum: it is a count of *hapaxes*, symbols appearing exactly once, and changing one observation can create or destroy a hapax elsewhere. The missing mass itself *is* a sum, $\sum_x p(x)\mathbb{1}\{x \text{ unseen}\}$, but a sum of dependent terms: if one symbol has been seen, the others had fewer chances to be. These are the two situations announced in the title – non-linear functions of independent variables, and sums of dependent variables – and the missing mass will serve throughout as the example on which every tool is tried.

The chapter follows the method that Chapter 2 made systematic. A concentration inequality is a bound on the cumulant generating function, and a bound on the cumulant generating function is, by Proposition 2.6, a bound on the variance of the tilted laws. *Before* looking at exponential moments, one should therefore understand the variance itself: this is the content of Section 5.3, where we compute the variance of the missing mass by hand, discover that its terms are negatively correlated, and then prove the Efron–Stein inequality, which bounds the variance of *any* function of independent variables by a sum of coordinate-wise variances, without any structural assumption. Section 5.4 develops *negative association*, the general form of the negative correlation observed in the missing mass, and shows that Chernoff's method applies to negatively associated variables exactly as to independent ones. Section 5.5 returns to the bounded-difference inequality of Theorem 2.55 and proves it again, slowly, by an argument that freezes the past one coordinate at a time and uses nothing but Hoeffding's lemma and Fubini's theorem; the discussion that follows explains where independence is used, why the bound is sometimes very loose, and what a better bound would have to look like. Section 5.6 introduces the entropy method in its simplest form – tensorisation of the entropy and Herbst's argument – and derives from it the exponential inequality for self-bounding functions, which behave like Poisson variables. Section 5.7 then assembles all of this into the analysis of the missing mass: the Good–Turing estimator is handled by bounded differences, the missing mass by negative association and a direct computation of the variance of the tilted law, which gives sub-Gaussian tails with the *same* variance factor $1/(en)$ as the Chebyshev-level computation. Poissonisation and the error of histogram classification rules close the chapter.

5.2 The missing mass and the Good–Turing estimator

5.2.1 Setting and notation

Throughout the chapter, except in Sections 5.3.2, 5.5 and 5.6 where the X_i live in arbitrary measurable spaces, $\mathcal{X}$ is a finite or countably infinite set, the *alphabet*, P is a probability distribution on $\mathcal{X}$ with point masses

$$p(x) := P(\{x\}), \qquad x \in \mathcal{X}, \qquad \sum_{x \in \mathcal{X}} p(x) = 1,$$

and $X_1, \dots, X_n$ are independent with law P. (In the notation of Chapter 3, p is the density of P with respect to the counting measure.) For $x \in \mathcal{X}$ the *count* of x is

$$N_n(x) := \sum_{i=1}^{n} \mathbb{1}\{X_i = x\},$$

so that $N_n(x) \sim \mathrm{Bin}(n, p(x))$ and $\sum_x N_n(x) = n$. The sample is described, up to the order of the observations, by the vector of counts $(N_n(x))_{x \in \mathcal{X}}$, whose law is multinomial.

Definition 5.1 (Missing mass, hapaxes, Good–Turing estimator)**.** The *missing mass* of the sample is

$$M_n := \sum_{x \in \mathcal{X}} p(x)\, \mathbb{1}\{N_n(x) = 0\} = \mathbb{P}\big(X_{n+1} \notin \{X_1, \dots, X_n\} \;\big|\; X_1, \dots, X_n\big), \tag{5.1}$$

where X_{n+1} is a further independent draw from P. A *hapax* is a symbol seen exactly once in the sample; the number of hapaxes is

$$U_n := \sum_{x \in \mathcal{X}} \mathbb{1}\{N_n(x) = 1\} = \sum_{i=1}^{n} \mathbb{1}\big\{X_i \notin \{X_j : j \neq i\}\big\},$$

and the *Good–Turing estimator* of the missing mass is

$$\widehat{M}_n := \frac{U_n}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\big\{X_i \notin \{X_j : j \neq i\}\big\}. \tag{5.2}$$

We also write $D_n := \sum_{x \in \mathcal{X}} \mathbb{1}\{N_n(x) \geq 1\}$ for the number of distinct symbols in the sample.

The second expression of U_n follows from the first by grouping the observations by their value: a symbol with $N_n(x) = 1$ contributes one observation X_i which is not seen elsewhere, and a symbol with $N_n(x) \geq 2$ contributes none. The second expression of $\widehat{M}_n$ is the important one. It says that the Good–Turing estimator is a *leave-one-out* estimate of the missing mass: for each i, pretend that the sample consisted of the other $n-1$ observations and check whether X_i would have been "new" for that sample; then average over i. Since X_i is independent of the others, this is exactly the kind of experiment that defines the missing mass of a sample of size $n-1$ – which is the content of Good's lemma below. (The same idea, under the name *jackknife*, will reappear in Section 5.3.2; it is also the leave-one-out estimate of Section 2.7.6.)

The missing mass is a random variable with values in $[0, 1]$, a function of the sample, and it is *not observable*: computing it requires the knowledge of P. Its expectation is

$$\mathbb{E}M_n = \sum_{x \in \mathcal{X}} p(x)\big(1 - p(x)\big)^n, \tag{5.3}$$

a non-increasing function of n (as it should be: the missing mass of a sample decreases when the sample grows) which tends to 0 as $n \to \infty$ by dominated convergence, since each term does and the sum is dominated by $\sum_x p(x) = 1$.

Example 5.2 (Three regimes)**.** *Uniform law on k symbols.* Then $\mathbb{E}M_n = (1 - 1/k)^n \leq e^{-n/k}$: the missing mass is exponentially small as soon as $n \gg k$, but for $k = n$ it equals $(1-1/n)^n \to e^{-1}$, and $(1 - 1/n)^n \geq 1/4$ for all $n \geq 2$. So although $\mathbb{E}M_n \to 0$ for every fixed P, there is *no uniform rate*: $\sup_P \mathbb{E}M_n \geq 1/4$ for all n.

Power laws. Let $\mathcal{X} = \{1, 2, \dots\}$ and $p(x) = c_\alpha x^{-\alpha}$ with $\alpha > 1$ (Zipf's law, the typical shape of word frequencies). Comparing the sum (5.3) with the integral $\int_1^\infty c_\alpha x^{-\alpha} e^{-n c_\alpha x^{-\alpha}}\, \mathrm{d}x$ and substituting $x = n^{1/\alpha} y$ shows that $\mathbb{E}M_n \asymp n^{-(\alpha-1)/\alpha}$ (Exercise 5.1): the missing mass decays polynomially, with an exponent that is small when α is close to 1. This is the regime in which the estimation problem is interesting: the missing mass is large, it decreases slowly, and the number of symbols that would have to be seen to reduce it is astronomical.

Why $n^{-1/2}$ is the right scale. In the uniform case with $k \asymp n$, and in the Zipf case with $\alpha < 2$, the missing mass is of larger order than $n^{-1/2}$. The results of this chapter show that it can always be estimated to accuracy $n^{-1/2}$; in those regimes, the estimate is therefore informative.

5.2.2 Good's lemma: the bias of the estimator

The first result is an identity which, in the approximate form "the chance that the next animal sampled will belong to a new species is approximately n_1/N", is equation (9) of Good's paper [Goo53], where the idea is attributed to Turing; the exact form below is the variant of Good's theorem proved by McAllester and Schapire [MS00, Thm. 1]. It shows that $\widehat{M}_n$ estimates $\mathbb{E}M_n$ with a bias of order at most $1/n$, *whatever* P.

Proposition 5.3 (Good). *For every $n \geq 1$ and every distribution P on a countable alphabet,*

$$\mathbb{E}\big[\widehat{M}_n\big] = \mathbb{E}\big[M_{n-1}\big], \tag{5.4}$$

where M_{n-1} is the missing mass of the first $n-1$ observations (with $M_0 = 1$). Consequently

$$0 \;\leq\; \mathbb{E}\big[\widehat{M}_n\big] - \mathbb{E}\big[M_n\big] = \sum_{x\in\mathcal{X}} p(x)^2\big(1-p(x)\big)^{n-1} = \frac{1}{n}\,\mathbb{E}\Big[\sum_{x\in\mathcal{X}} p(x)\,\mathbb{1}\{N_n(x)=1\}\Big] \;\leq\; \frac{1}{n}. \tag{5.5}$$

Proof. By the leave-one-out expression (5.2) and the fact that the observations are identically distributed, $\mathbb{E}[\widehat{M}_n] = \mathbb{P}\big(X_n \notin \{X_1,\dots,X_{n-1}\}\big)$. Since X_n is independent of $(X_1,\dots,X_{n-1})$, conditioning on the latter gives $\mathbb{P}\big(X_n \notin \{X_1,\dots,X_{n-1}\}\big) = \mathbb{E}[M_{n-1}]$ by the second expression of the missing mass in (5.1), which is (5.4). Then, by (5.3),

$$\mathbb{E}[M_{n-1}] - \mathbb{E}[M_n] = \sum_x p(x)\big(1-p(x)\big)^{n-1}\big(1-(1-p(x))\big) = \sum_x p(x)^2\big(1-p(x)\big)^{n-1},$$

which is non-negative, and since $\mathbb{P}(N_n(x)=1) = np(x)(1-p(x))^{n-1}$ this is also

$$\frac{1}{n}\sum_x p(x)\mathbb{P}\big(N_n(x)=1\big) = \frac{1}{n}\,\mathbb{E}\Big[\sum_x p(x)\mathbb{1}\{N_n(x)=1\}\Big].$$

The last sum is the total mass of the hapaxes, at most 1. □

A direct computation, without the leave-one-out interpretation, gives the same result and is worth seeing once:

$$\begin{aligned}\mathbb{E}[\widehat{M}_n] - \mathbb{E}[M_n] &= \frac{1}{n}\sum_x \mathbb{P}\big(N_n(x)=1\big) - \sum_x p(x)\mathbb{P}\big(N_n(x)=0\big)\\ &= \sum_x p(x)\big(1-p(x)\big)^{n-1}\big[1-(1-p(x))\big].\end{aligned}$$

Remark 5.4 (Reading Good's lemma). Three comments. First, the bias bound $1/n$ is *distribution-free*: it holds for every P, finite or infinite, light-tailed or not. This is what makes the estimator useful in the regimes of Example 5.2, where $\mathbb{E}M_n$ itself has no distribution-free rate. Second, (5.5) says more than $1/n$: the bias is $\frac{1}{n}$ times the expected mass of the hapaxes, which is typically much smaller than 1; the elementary bound $p(1-p)^{n-1} \leq e^{-(n-1)p}p \leq 1/(e(n-1))$ (for $n \geq 2$) gives also $\mathbb{E}\widehat{M}_n - \mathbb{E}M_n \leq \sum_x p(x)^2 e^{-(n-1)p(x)} \leq 1/(e(n-1))$. Third, the estimator is biased *upwards*: it estimates the missing mass of a sample of size $n-1$, which is larger. The exchangeability argument also shows that $\mathbb{E}M_n = \mathbb{E}[U_{n+1}/(n+1)]$: with one more observation, $U_{n+1}/(n+1)$ would be an unbiased estimate of $\mathbb{E}M_n$. Robbins [Rob68] showed that $\mathbb{E}[(M_n - U_{n+1}/(n+1))^2] < 1/(n+1)$ for every P, a mean-square statement which already contains the $n^{-1/2}$ accuracy – for an estimator that uses one more observation than it estimates.

The identity (5.4) settles the bias. The accuracy of $\widehat{M}_n$ as an estimate of the *random* quantity M_n is then a matter of concentration: we need

$$\widehat{M}_n \approx \mathbb{E}\widehat{M}_n \qquad \text{and} \qquad M_n \approx \mathbb{E}M_n, \tag{5.6}$$

both with high probability. The first is a concentration statement for a function of n independent variables that is not a sum; the second is one for a sum of dependent variables. We begin, as announced, with the variance.

5.3 The variance first

5.3.1 The variance of the missing mass, by hand

The missing mass is a sum, $M_n = \sum_x p(x) I_x$ with $I_x := \mathbb{1}\{N_n(x) = 0\}$, and the variance of a sum of square-integrable variables is always

$$\mathbb{V}\mathrm{ar}\Big(\sum_x p(x) I_x\Big) = \sum_x p(x)^2\, \mathbb{V}\mathrm{ar}(I_x) + \sum_{x \neq y} p(x)p(y)\, \mathbb{C}\mathrm{ov}(I_x, I_y) \tag{5.7}$$

(all the series converge absolutely, since $\big|\mathbb{C}\mathrm{ov}(I_x, I_y)\big| \le 1$ and $\sum_{x,y} p(x)p(y) = 1$). The indicators I_x are dependent: if x has not been seen, the n observations fell into $\mathcal{X} \setminus \{x\}$, which makes it *less* likely that y was missed too. Quantitatively, with $q(x) := \mathbb{P}(N_n(x) = 0) = (1 - p(x))^n$,

$$\begin{aligned} \mathbb{C}\mathrm{ov}(I_x, I_y) &= \mathbb{P}\big(N_n(x) = 0,\, N_n(y) = 0\big) - q(x)q(y) \\ &= \big(1 - p(x) - p(y)\big)^n - \big(1 - p(x)\big)^n \big(1 - p(y)\big)^n \;\le\; 0, \end{aligned}$$

because $1 - p(x) - p(y) \le (1 - p(x))(1 - p(y)) = 1 - p(x) - p(y) + p(x)p(y)$. The indicators are *negatively correlated*, the cross terms in (5.7) only help, and

$$\mathbb{V}\mathrm{ar}(M_n) \;\le\; \sum_{x \in \mathcal{X}} p(x)^2\, q(x)\big(1 - q(x)\big) \;\le\; \sum_{x \in \mathcal{X}} p(x)^2 \big(1 - p(x)\big)^n \;\le\; \sum_{x \in \mathcal{X}} p(x)^2 e^{-np(x)} \;\le\; \frac{1}{en}. \tag{5.8}$$

The last step is the elementary inequality

$$u\, e^{-u} \le \frac{1}{e} \quad (u \ge 0), \qquad \text{so that} \qquad p(x)\, e^{-np(x)} = \frac{1}{n}\big(np(x)\big) e^{-np(x)} \le \frac{1}{en}, \tag{5.9}$$

applied to each term: $p(x)^2 e^{-np(x)} \le p(x)/(en)$, and $\sum_x p(x) = 1$. Note the mechanism, which will recur: a symbol of large mass $p(x)$ contributes a large weight $p(x)^2$, but it is almost surely seen, so that its indicator has a tiny variance $q(x) \le e^{-np(x)}$; a symbol of small mass may well be missed, but then it does not weigh much. The product $p(x)e^{-np(x)}$ is maximal for $p(x) = 1/n$, the symbols that are "seen about once".

Remark 5.5 (The bound (5.8) is sharp up to a constant). For the uniform law on n symbols, $\mathbb{V}\mathrm{ar}(M_n) \sim (e^{-1} - 2e^{-2})/n \approx 0.097/n$ (Exercise 5.2): the variance of the missing mass is genuinely of order $1/n$, and (5.8) loses only the constant $e^{-1} \approx 0.37$ against 0.097. For other laws the quantity

$$v_n := \sum_{x \in \mathcal{X}} p(x)^2 \big(1 - p(x)\big)^n,$$

which is the actual content of (5.8), can be much smaller than $1/(en)$ – for Zipf's law with exponent α it is of order $n^{-2+1/\alpha}$ (Exercise 5.1). We shall see that the *exponential* inequalities inherit this distribution-dependent variance term, at least on one side (Remark 5.46).

By Chebyshev's inequality (Corollary 1.46), with probability at least $1-\delta$,

$$|M_n - \mathbb{E}M_n| \;\le\; \frac{1}{\sqrt{en\delta}}.$$

This is a first, honest answer to the second half of (5.6). Its dependence on δ is poor – $1/\sqrt{\delta}$ where an exponential inequality would give $\sqrt{\log(1/\delta)}$ – but the dependence on n is the right one, and we have identified the variance proxy, $1/(en)$, that the exponential inequality of Section 5.7.2 will reproduce exactly.

5.3.2 The Efron–Stein inequality

The computation above used the special structure of M_n. The Good–Turing estimator $\widehat{M}_n = U_n/n$ has no such structure: U_n is a sum of indicators $\mathbb{1}\{N_n(x)=1\}$ which are neither independent nor, as far as one can see, negatively correlated. What is needed is a bound on the variance of an *arbitrary* function of independent variables, and the remarkable fact is that such a bound exists, and that it is tight for sums.

In this subsection and in Sections 5.5 and 5.6, $X_1,\dots,X_n$ are independent random variables, X_i taking its values in a measurable space $(\mathcal{X}_i,\mathcal{B}_i)$ with law P_i, $f:\mathcal{X}_1\times\cdots\times\mathcal{X}_n\to\mathbb{R}$ is measurable, and $Z:=f(X_1,\dots,X_n)$. We write $X=(X_1,\dots,X_n)$, $X^{(i)}:=(X_j)_{j\neq i}$ for the sample with the i-th variable removed, and, for a bounded or non-negative measurable function w of n variables,

$$\mathbb{E}^{(i)}w\,(x) \;:=\; \int_{\mathcal{X}_i} w(x_1,\dots,x_{i-1},y,x_{i+1},\dots,x_n)\;\mathrm{d}P_i(y), \tag{5.10}$$

the integral over the i-th coordinate alone, the others being frozen (and, more generally, for any w such that the integral exists for almost every $x^{(i)}$: by Fubini's theorem this is the case when $w(X)$ is integrable, and $\mathbb{E}^{(i)}[w^2]<\infty$ almost everywhere when $w(X)$ is square integrable). By Fubini's theorem $\mathbb{E}^{(i)}w$ is a measurable function of $x^{(i)}$, and $\mathbb{E}^{(i)}w(X)$ is a version of the conditional expectation $\mathbb{E}[w(X)\mid X^{(i)}]$; the operator $\mathbb{E}^{(i)}$ is linear, positive, and $\mathbb{E}[\mathbb{E}^{(i)}w(X)]=\mathbb{E}w(X)$. The *conditional variance with respect to the i-th coordinate* is

$$\mathbb{V}\mathrm{ar}^{(i)}\,w \;:=\; \mathbb{E}^{(i)}\big[(w-\mathbb{E}^{(i)}w)^2\big] = \mathbb{E}^{(i)}[w^2]-(\mathbb{E}^{(i)}w)^2,$$

again a function of $x^{(i)}$. Finally, $X'=(X'_1,\dots,X'_n)$ denotes an independent copy of X and

$$Z^{(i)} := f(X_1,\dots,X_{i-1},X'_i,X_{i+1},\dots,X_n)$$

the value of f when the i-th variable is *resampled*.

Theorem 5.6 (Efron–Stein inequality)**.** *Assume $\mathbb{E}[Z^2]<\infty$. Then*

$$\mathbb{V}\mathrm{ar}(Z) \;\le\; \sum_{i=1}^n \mathbb{E}\big[\mathbb{V}\mathrm{ar}^{(i)}\,f\,(X)\big] = \sum_{i=1}^n \mathbb{E}\Big[\big(Z-\mathbb{E}^{(i)}f(X)\big)^2\Big]. \tag{5.11}$$

Moreover, for each i,

$$\mathbb{E}\big[\mathbb{V}\mathrm{ar}^{(i)}\,f(X)\big] = \frac{1}{2}\,\mathbb{E}\big[(Z-Z^{(i)})^2\big] = \mathbb{E}\big[(Z-Z^{(i)})_+^2\big] = \inf_{Z_i}\mathbb{E}\big[(Z-Z_i)^2\big], \tag{5.12}$$

the infimum being over all square-integrable random variables of the form $Z_i=w_i(X^{(i)})$, w_i measurable on $\prod_{j\neq i}\mathcal{X}_j$.

Proof. Step 1: a martingale decomposition. For $0 \le i \le n$ let

$$g_i(x_1,\dots,x_i) := \int f(x_1,\dots,x_i,y_{i+1},\dots,y_n)\,\mathrm{d}P_{i+1}(y_{i+1})\cdots\mathrm{d}P_n(y_n),$$

the integral of f over the last $n-i$ coordinates ($g_n = f$, $g_0 = \mathbb{E}Z$). By Fubini's theorem these are measurable, and integrating over one more coordinate gives

$$g_{i-1}(x_1,\dots,x_{i-1}) = \int g_i(x_1,\dots,x_{i-1},y)\,\mathrm{d}P_i(y). \tag{5.13}$$

Set $\Delta_i := g_i(X_1,\dots,X_i) - g_{i-1}(X_1,\dots,X_{i-1})$; then $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$. The Δ_i are square integrable (by Jensen, $\mathbb{E}[g_i(X_1,\dots,X_i)^2] \le \mathbb{E}[Z^2]$) and *orthogonal*: for $i<j$, integrating first over X_j with the other variables frozen, (5.13) gives

$$\int \big(g_j(x_1,\dots,x_{j-1},y) - g_{j-1}(x_1,\dots,x_{j-1})\big)\,\mathrm{d}P_j(y) = 0,$$

while Δ_i does not depend on x_j; hence $\mathbb{E}[\Delta_i\Delta_j]=0$ by Fubini. Therefore

$$\mathbb{V}\mathrm{ar}(Z) = \mathbb{E}\Big[\Big(\sum_{i=1}^n \Delta_i\Big)^2\Big] = \sum_{i=1}^n \mathbb{E}\big[\Delta_i^2\big]. \tag{5.14}$$

Step 2: each increment is a conditional expectation of a centred variable. Let $\mathbb{E}_i$ denote integration over the coordinates $i+1,\dots,n$, so that $g_i = \mathbb{E}_i f$. Since the coordinate i is integrated out in g_{i-1} and the order of integration is immaterial (Fubini),

$$g_{i-1} = \mathbb{E}_i\big[\mathbb{E}^{(i)}f\big], \qquad \text{hence} \qquad \Delta_i = \mathbb{E}_i\big[f - \mathbb{E}^{(i)}f\big](X_1,\dots,X_i).$$

By Jensen's inequality for the probability measure $P_{i+1}\otimes\cdots\otimes P_n$,

$$\Delta_i^2 \le \mathbb{E}_i\big[(f-\mathbb{E}^{(i)}f)^2\big](X_1,\dots,X_i), \qquad \text{hence} \qquad \mathbb{E}[\Delta_i^2] \le \mathbb{E}\big[(f-\mathbb{E}^{(i)}f)^2(X)\big] = \mathbb{E}[\mathbb{V}\mathrm{ar}^{(i)}f(X)]$$

after taking expectations. With (5.14) this is (5.11).

Step 3: the three forms of the conditional variance. Fix $x^{(i)}$ and write

$$\varphi(y) := f(x_1,\dots,x_{i-1},y,x_{i+1},\dots,x_n), \qquad y\in\mathcal{X}_i,$$

a square-integrable function under P_i (for almost every $x^{(i)}$). If Y,Y' are independent with law P_i, then $\mathbb{E}[(\varphi(Y)-\varphi(Y'))^2] = 2\,\mathbb{V}\mathrm{ar}\,\varphi(Y)$, which is the first equality of (5.12) once integrated over $x^{(i)}$; and

$$\big(\varphi(Y)-\varphi(Y')\big)^2 = \big(\varphi(Y)-\varphi(Y')\big)_+^2 + \big(\varphi(Y')-\varphi(Y)\big)_+^2,$$

the two terms having the same expectation by exchange of Y and Y', which gives the second. Finally, for every constant c, $\mathbb{E}[(\varphi(Y)-c)^2] = \mathbb{V}\mathrm{ar}\,\varphi(Y) + (\mathbb{E}\varphi(Y)-c)^2 \ge \mathbb{V}\mathrm{ar}\,\varphi(Y)$; applied with $c = w_i(x^{(i)})$ and integrated over $x^{(i)}$, this gives $\mathbb{E}[(Z-Z_i)^2] \ge \mathbb{E}[\mathbb{V}\mathrm{ar}^{(i)}f(X)]$, with equality for $w_i = \mathbb{E}^{(i)}f$. □

Remark 5.7 (Tensorisation of the variance). Theorem 5.6 says that the variance is *subadditive* with respect to the product structure: the variance of $f(X)$ under the product measure $P_1\otimes\cdots\otimes P_n$ is at most the sum of the expected variances of f along each coordinate. For a sum $f(x)=\sum_i x_i$ one has $\mathbb{V}\mathrm{ar}^{(i)}f = \mathbb{V}\mathrm{ar}(X_i)$ and (5.11) is an *equality*: the Efron–Stein

inequality is exact for sums of independent variables, and the loss in general is entirely in Jensen's inequality of Step 2. This is the first instance of a pattern that organises the chapter: a one-dimensional fact (here, nothing at all – the conditional variance is simply left as it is) is "tensorised" along the martingale $g_0, g_1, \dots, g_n$ of Step 1. The bounded-difference inequality (Section 5.5) tensorises Hoeffding's lemma in the same way; the entropy method (Section 5.6) tensorises an inequality for the entropy instead of the variance.

Remark 5.8 (History: the jackknife). Efron and Stein [ES81] proved (5.11) for *symmetric* functions of $n+1$ i.i.d. variables, in the following form. Let $S(X_1, \dots, X_{n+1})$ be symmetric and let $S_{(i)} := S(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+1})$ be the statistic computed on the sample with the i-th observation deleted. The *jackknife estimate of the variance* of $S(X_1, \dots, X_n)$, proposed by Quenouille and Tukey, is $\sum_{i \le n+1}(S_{(i)} - \bar{S})^2$ with $\bar{S} = \frac{1}{n+1}\sum_i S_{(i)}$; Efron and Stein showed that its expectation is at least $\mathbb{V}\mathrm{ar}\, S(X_1, \dots, X_n)$ – the jackknife is conservative. Since $\sum_{i \le n+1}(S_{(i)} - \bar{S})^2 = \frac{1}{2(n+1)}\sum_{i,j}(S_{(i)} - S_{(j)})^2$ and, by exchangeability, $(S_{(n+1)}, S_{(i)})$ has the law of $(Z, Z^{(i)})$ for $Z = S(X_1, \dots, X_n)$ and $i \le n$, the jackknife quantity has expectation $\frac{1}{2}\sum_{i \le n} \mathbb{E}[(Z - Z^{(i)})^2]$: Efron and Stein's inequality is exactly the first form of (5.12). (The cruder bound $\sum_i \mathbb{E}[(Z - Z_i)^2]$ with the leave-one-out $Z_i = S_{(i)}$, given by the last form, is twice as large.) Steele [Ste86] removed the symmetry assumption, and the formulation with $\mathbb{V}\mathrm{ar}^{(i)}$ is from [BLM13], whose Chapter 3 is the reference for this section. The leave-one-out Good–Turing estimator (5.2) is a jackknife in exactly this sense, which is why the method of this section is so well suited to it.

5.3.3 First consequences

Corollary 5.9 (Bounded differences, variance version)**.** *If f has the bounded-difference property (2.32) with constants $c_1, \dots, c_n$, then $\mathbb{V}\mathrm{ar}(Z) \le \frac{1}{4}\sum_{i=1}^n c_i^2$.*

Proof. For frozen $x^{(i)}$, the function $y \mapsto f(x_1, \dots, y, \dots, x_n)$ takes its values in an interval of length at most c_i, so its variance under P_i is at most $c_i^2/4$ by Popoviciu's inequality (Exercise 1.5); that is, $\mathbb{V}\mathrm{ar}^{(i)} f \le c_i^2/4$ identically. □

This is the variance-level counterpart of Theorem 2.55, with the same proxy $\frac{1}{4}\sum_i c_i^2$; it suffers from the same defect, discussed in Remark 2.59: the c_i are worst-case constants. The next notion captures a common situation in which the variance is instead of the order of the *mean*.

Definition 5.10 (Self-bounding functions)**.** A measurable function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to [0, \infty)$ is *self-bounding* if there exist measurable functions $f_i : \prod_{j \ne i} \mathcal{X}_j \to \mathbb{R}$, $1 \le i \le n$, such that for all x,

$$0 \;\le\; f(x) - f_i(x^{(i)}) \;\le\; 1 \qquad \text{and} \qquad \sum_{i=1}^n \big(f(x) - f_i(x^{(i)})\big) \;\le\; f(x). \tag{5.15}$$

Typically f_i is the same statistic computed on the sample without its i-th observation; the first condition says that removing one observation can only decrease f, and by at most one; the second, that the total decrease caused by the n removals, performed one at a time from the full sample, is at most f itself.

Corollary 5.11 (Variance of a self-bounding function)**.** *If f is self-bounding and $Z = f(X)$ is square integrable, then $\mathbb{V}\mathrm{ar}(Z) \le \mathbb{E}Z$.*

Proof. With $Z_i = f_i(X^{(i)})$, (5.12) and (5.15) give $\mathbb{V}\mathrm{ar}(Z) \le \sum_i \mathbb{E}[(Z - Z_i)^2] \le \sum_i \mathbb{E}[Z - Z_i] \le \mathbb{E}Z$, using $(Z - Z_i)^2 \le Z - Z_i$ since $0 \le Z - Z_i \le 1$ (the Z_i are square integrable because $Z - 1 \le Z_i \le Z$). □

Example 5.12 (The number of distinct symbols is self-bounding)**.** Let $f(x_1, \dots, x_n) = |\{x_1, \dots, x_n\}|$ be the number of distinct values in the sample, so that $f(X) = D_n$, and let $f_i(x^{(i)})$ be the number of distinct values among the other $n-1$. Removing x_i decreases the count by 1 if x_i is a hapax of x and by 0 otherwise, so $\sum_i (f(x) - f_i(x^{(i)}))$ is the number of hapaxes, which is at most the number of distinct values. Hence D_n is self-bounding, and $\mathbb{V}\mathrm{ar}(D_n) \le \mathbb{E}D_n = \sum_x \big(1 - (1-p(x))^n\big)$. Note that $\sum_x I_x = |\mathcal{X}| - D_n$ when $\mathcal{X}$ is finite: the number of *unseen* symbols has the same variance. For the uniform law on k symbols, $\mathbb{E}D_n \le \min(k, n)$, and the bound $\mathbb{V}\mathrm{ar}(D_n) \le \mathbb{E}D_n$ is of the right order (Exercise 5.4); compare with Corollary 5.9, which only gives $\mathbb{V}\mathrm{ar}(D_n) \le n/4$.

Example 5.13 (Configuration functions and suprema)**.** Let $\mathcal{G}$ be a countable class of measurable functions $g : \mathcal{X} \to [0,1]$ and $f(x) = \sup_{g\in\mathcal{G}} \sum_{i=1}^n g(x_i)$, with $f_i(x^{(i)}) = \sup_{g\in\mathcal{G}} \sum_{j\neq i} g(x_j)$. Then f is self-bounding: clearly $f_i(x^{(i)}) \le f(x)$, and if $g^* \in \mathcal{G}$ satisfies $\sum_i g^*(x_i) \ge f(x) - \varepsilon$, then $f_i(x^{(i)}) \ge \sum_{j\neq i} g^*(x_j) \ge f(x) - \varepsilon - g^*(x_i)$, so $f(x) - f_i(x^{(i)}) \le g^*(x_i) + \varepsilon \le 1 + \varepsilon$ and $\sum_i (f(x) - f_i(x^{(i)})) \le \sum_i g^*(x_i) + n\varepsilon \le f(x) + n\varepsilon$; let $\varepsilon \downarrow 0$. Suprema of empirical processes indexed by classes of non-negative functions are therefore self-bounding; Exercise 5.10 shows that the logarithm of the number of labellings of the sample by a class of classifiers, the "random VC entropy" of Chapter 4, is self-bounding too.

Example 5.14 (The missing mass and its estimator, by Efron–Stein)**.** *The estimator.* Changing one observation x_i into x_i' changes the number of hapaxes by at most 2: removing x_i changes U_n by -1 if x_i was a hapax, by $+1$ if x_i was seen exactly twice, by 0 otherwise; adding x_i' likewise changes it by at most 1 in absolute value. So U_n has bounded differences with $c_i = 2$, and Corollary 5.9 gives

$$\mathbb{V}\mathrm{ar}\big(\widehat{M}_n\big) = \frac{\mathbb{V}\mathrm{ar}(U_n)}{n^2} \le \frac{4n/4}{n^2} = \frac{1}{n}. \tag{5.16}$$

The hapax count is *not* self-bounding (removing an observation can increase it), and (5.16) is of the right order in the worst case (Exercise 5.3).

The missing mass, without using negative correlation. Let us compute the Efron–Stein bound for $M_n = f(X)$ in the form $\sum_i \mathbb{E}[(Z - Z^{(i)})_+^2]$. The resampled missing mass $M_n^{(i)}$ is smaller than M_n only if the new observation X_i' is a symbol unseen in $X_1, \dots, X_n$, and then $M_n - M_n^{(i)} \le p(X_i')$; so $(M_n - M_n^{(i)})_+ \le p(X_i')\mathbb{1}\{X_i' \notin \{X_1, \dots, X_n\}\}$ and

$$\mathbb{E}\big[(M_n - M_n^{(i)})_+^2\big] \le \sum_x p(x)\cdot p(x)^2\big(1-p(x)\big)^n \le \frac{1}{n^2}\sum_x p(x) \sup_{u\ge 0} u^2 e^{-u} = \frac{4}{e^2 n^2}.$$

Summing over i, $\mathbb{V}\mathrm{ar}(M_n) \le 4/(e^2 n) \approx 0.54/n$. This is weaker than (5.8) by the factor $4/e \approx 1.47$, but it is obtained by a general method that did not need to notice that the terms of M_n are negatively correlated: the Efron–Stein inequality sees M_n as a function of the independent X_i, not as a sum of dependent indicators. The two points of view will be compared again at the exponential level, where only the second one survives (Remark 5.33).

Putting (5.16), (5.8) and Proposition 5.3 together with Chebyshev's inequality, we obtain a first, complete, distribution-free guarantee for the Good–Turing estimator: for every P and every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\big|\widehat{M}_n - M_n\big| \le \frac{1}{n} + \Big(1 + \frac{1}{\sqrt{e}}\Big)\sqrt{\frac{2}{n\delta}} \le \frac{1}{n} + \frac{2.3}{\sqrt{n\delta}}. \tag{5.17}$$

(Apply Chebyshev to each of the two deviations of (5.6) with confidence $\delta/2$, and use (5.5) for the bias.) The rest of the chapter improves the factor $1/\sqrt{\delta}$ to $\sqrt{\log(1/\delta)}$, which is what a confidence region valid simultaneously for many n, or a union bound over many estimators, requires. Doing so needs exponential moments, and exponential moments need the structural tools of the next three sections.

5.4 Negative association

The missing mass is a sum of indicators which are negatively correlated, and Section 5.3.1 showed how useful this is at the level of the variance. For Chernoff's method one needs more than covariances: one needs the exponential moment of the sum, $\mathbb{E}e^{\lambda \sum_x p(x) I_x}$, to be at most the product of the exponential moments, as it would be under independence. The notion that makes this true is *negative association*, introduced by Alam and Saxena [AS81] and studied systematically by Joag-Dev and Proschan [JP83]. It is a strong form of negative dependence, stable under the operations that arise in practice, and satisfied by the models we care about: sampling without replacement, random permutations, and the occupancy counts of balls thrown into bins – which is to say, the counts $N_n(x)$ of our sample.

5.4.1 Definition and first properties

A function $F : \mathbb{R}^I \to \mathbb{R}$ is *non-decreasing* if $F(x) \le F(x')$ whenever $x_i \le x'_i$ for every $i \in I$, and *non-increasing* if $-F$ is non-decreasing. For $I \subseteq \{1, \dots, n\}$ we write $X_I = (X_i)_{i \in I}$.

Definition 5.15 (Negative association)**.** Real random variables $X_1, \dots, X_n$ are *negatively associated* (NA) if for every pair of disjoint sets $I, J \subseteq \{1, \dots, n\}$ and every pair of bounded measurable functions $F : \mathbb{R}^I \to \mathbb{R}$, $G : \mathbb{R}^J \to \mathbb{R}$ which are both non-decreasing or both non-increasing,

$$\mathbb{E}\big[F(X_I)\, G(X_J)\big] \;\le\; \mathbb{E}\big[F(X_I)\big]\, \mathbb{E}\big[G(X_J)\big]. \tag{5.18}$$

A countable family of random variables is NA if each of its finite subfamilies is.

Only the values of F and G on the supports of X_I and X_J matter, so it is enough that they be bounded and monotone on a product set containing the support: a monotone function on $D = \prod_{i \in I} D_i$ with $D_i \subseteq \mathbb{R}$ extends to a monotone function on $\mathbb{R}^I$ (for $D_i = \{0, 1\}$ or $D_i = \mathbb{N}$, compose with $u \mapsto 0 \vee \lfloor u \rfloor \wedge \max D_i$ coordinatewise). We use this without further comment, e.g. for $u \mapsto \mathbb{1}\{u = 0\}$ on $\mathbb{N}$ or $u \mapsto e^{\lambda u}$ on $\{0, 1\}$.

In words: if a monotone function of some of the variables is large, a monotone function of the *other* variables tends to be small. Independent variables are NA, with equality in (5.18). The requirement that I and J be disjoint is essential: for $I = J$ and $F = G$, (5.18) would say $\mathbb{V}\mathrm{ar}\, F(X_I) \le 0$.

By truncation and dominated convergence, (5.18) extends from bounded to square-integrable monotone functions (the truncation $(F \wedge K) \vee (-K)$ is monotone, and the products are dominated by $\frac{1}{2}(F^2 + G^2)$). We shall use this freely. The following elementary fact about a *single* real variable will be used repeatedly: a single variable is "positively associated with itself".

Lemma 5.16 (Chebyshev's correlation inequality)**.** *Let S be a real random variable and $\bar{F}, \bar{G} : \mathbb{R} \to \mathbb{R}$ bounded, $\bar{F}$ non-decreasing and $\bar{G}$ non-increasing. Then $\mathbb{E}[\bar{F}(S)\bar{G}(S)] \le \mathbb{E}[\bar{F}(S)]\,\mathbb{E}[\bar{G}(S)]$. If $\bar{F}$ and $\bar{G}$ are both non-decreasing, the inequality is reversed.*

Proof. Let S' be an independent copy of S. For all s, s', $(\bar{F}(s) - \bar{F}(s'))(\bar{G}(s) - \bar{G}(s')) \leq 0$, since the two factors have opposite signs or vanish. Taking expectations and expanding, $2\mathbb{E}[\bar{F}(S)\bar{G}(S)] - 2\mathbb{E}[\bar{F}(S)]\mathbb{E}[\bar{G}(S)] \leq 0$. The other case is identical. □

Proposition 5.17 (First properties). *Let $X_1, \dots, X_n$ be NA.*

(i) Every subfamily $(X_i)_{i\in I}$ is NA.

(ii) If the X_i are square integrable, $\mathbb{C}\mathrm{ov}(X_i, X_j) \leq 0$ for all $i \neq j$.

(iii) If $I_1, \dots, I_k$ are pairwise disjoint and $F_1, \dots, F_k$ are bounded, non-negative, *measurable functions of $X_{I_1}, \dots, X_{I_k}$ respectively, all non-decreasing or all non-increasing, then $\mathbb{E}\big[\prod_{j=1}^k F_j\big] \leq \prod_{j=1}^k \mathbb{E}[F_j]$.*

(iv) For all reals $r_1, \dots, r_n$,

$$\mathbb{P}\big(X_i \geq r_i \text{ for all } i\big) \leq \prod_{i=1}^n \mathbb{P}(X_i \geq r_i) \qquad \text{and} \qquad \mathbb{P}\big(X_i \leq r_i \text{ for all } i\big) \leq \prod_{i=1}^n \mathbb{P}(X_i \leq r_i).$$

Proof. (i) and (ii) are immediate from the definition (for (ii), take $F(x_i) = x_i$, $G(x_j) = x_j$, square integrable). For (iii), induct on k: the product $\prod_{j\geq 2} F_j$ is a bounded, non-negative function of $X_{I_2\cup\cdots\cup I_k}$ which is non-decreasing (resp. non-increasing) as a product of non-negative functions which all are, so (5.18) applied to F_1 and $\prod_{j\geq 2} F_j$ gives $\mathbb{E}[\prod_j F_j] \leq \mathbb{E}[F_1]\mathbb{E}[\prod_{j\geq 2} F_j]$, and the induction hypothesis finishes. (iv) is (iii) with the indicators $\mathbb{1}\{x_i \geq r_i\}$, which are non-decreasing, or $\mathbb{1}\{x_i \leq r_i\}$, which are non-increasing. □

The two inequalities of (iv) are called *negative upper* and *negative lower orthant dependence*; they are the multivariate form [JP83, Def. 2.3] of Lehmann's quadrant dependence [Leh66], and they are what one gets from negative association when only events of the form "all the variables are large" or "all are small" are of interest. Property (iii) is the one that matters for Chernoff's method.

5.4.2 Two closure properties

Negative association would be of little use if it had to be verified from the definition in each case. Its value lies in the following two stability properties, which reduce most verifications to a handful of elementary models.

Proposition 5.18 (Union of independent NA families). *If $(X_i)_{i\in I}$ and $(Y_j)_{j\in J}$ are two NA families which are independent of each other, then the family $(X_i, Y_j)_{i\in I, j\in J}$ is NA.*

Proof. Let A, B be disjoint sets of indices of the joint family, and write $A = A_1 \sqcup A_2$, $B = B_1 \sqcup B_2$ according to whether an index refers to an X or to a Y. Let F be a bounded non-decreasing function of (X_{A_1}, Y_{A_2}) and G one of (X_{B_1}, Y_{B_2}) (the non-increasing case is identical). By independence and Fubini's theorem, $\mathbb{E}[FG] = \int \mathbb{E}\big[F(X_{A_1}, y_{A_2})G(X_{B_1}, y_{B_2})\big]\, \mathrm{d}P_Y(y)$, where P_Y is the law of Y. For each fixed y, the two functions inside are bounded non-decreasing functions of X_{A_1} and X_{B_1}, with $A_1 \cap B_1 = \emptyset$, so by negative association of X the integrand is at most $\bar{F}(y_{A_2})\bar{G}(y_{B_2})$ with $\bar{F}(y_{A_2}) := \mathbb{E}[F(X_{A_1}, y_{A_2})]$ and $\bar{G}(y_{B_2}) := \mathbb{E}[G(X_{B_1}, y_{B_2})]$. These are bounded non-decreasing functions of y_{A_2} and y_{B_2} (measurable by Fubini), and $A_2 \cap B_2 = \emptyset$; by negative association of Y, $\int \bar{F}\bar{G}\, \mathrm{d}P_Y \leq \int \bar{F}\, \mathrm{d}P_Y \int \bar{G}\, \mathrm{d}P_Y = \mathbb{E}[F]\,\mathbb{E}[G]$. □

Proposition 5.19 (Concordant monotone transformations). *Let $X_1, \dots, X_n$ be NA, let $I_1, \dots, I_k$ be pairwise disjoint, and let $F_j : \mathbb{R}^{I_j} \to \mathbb{R}$, $1 \leq j \leq k$, be measurable functions which are all non-decreasing, or all non-increasing. Then $F_1(X_{I_1}), \dots, F_k(X_{I_k})$ are NA.*

Proof. Let $A, B \subseteq \{1, \dots, k\}$ be disjoint and let Φ, Ψ be bounded functions of $(F_j)_{j\in A}$ and $(F_j)_{j\in B}$, both non-decreasing (say). If the F_j are non-decreasing, the composite $x \mapsto \Phi\big((F_j(x_{I_j}))_{j\in A}\big)$ is a bounded non-decreasing function of $X_{\bigcup_{j\in A} I_j}$, and likewise for Ψ; if the F_j are non-increasing, both composites are non-increasing. In both cases the two composites are concordant monotone functions of disjoint blocks of the X_i, and (5.18) for X gives (5.18) for the $F_j(X_{I_j})$. □

In particular, sums of disjoint blocks of NA variables are NA; so are $(-X_i)$, $(\mathbb{1}\{X_i \geq r_i\})$, $(\mathbb{1}\{X_i = 0\})$ for non-negative X_i (the map $u \mapsto \mathbb{1}\{u = 0\}$ is non-increasing on $[0, \infty)$), and so on. What is *not* allowed is to mix non-decreasing and non-increasing maps, or to apply a non-monotone map such as $u \mapsto \mathbb{1}\{u = 1\}$: the hapax indicators $\mathbb{1}\{N_n(x) = 1\}$ are not obtained from the counts by a monotone map, and nothing in this section applies to them.

5.4.3 The basic examples

Lemma 5.20 (The zero–one lemma)**.** *Let $Y_1, \dots, Y_m$ be $\{0, 1\}$-valued random variables with $\sum_{j=1}^m Y_j \leq 1$ almost surely. Then $Y_1, \dots, Y_m$ are NA.*

Proof. Let I, J be disjoint and F, G bounded, both non-decreasing or both non-increasing. Set $\tilde{F}(y_I) := F(y_I) - F(0)$ and $\tilde{G}(y_J) := G(y_J) - G(0)$. Since at most one of the Y_j is non-zero, almost surely $Y_I = 0$ or $Y_J = 0$, so that $\tilde{F}(Y_I)\tilde{G}(Y_J) = 0$ almost surely. On the other hand $\tilde{F}(Y_I)$ and $\tilde{G}(Y_J)$ are both ≥ 0 (non-decreasing case) or both ≤ 0 (non-increasing case), so $\mathbb{E}[\tilde{F}(Y_I)]\mathbb{E}[\tilde{G}(Y_J)] \geq 0$. Expanding the products,

$$\mathbb{E}[F(Y_I)G(Y_J)] - \mathbb{E}[F(Y_I)]\mathbb{E}[G(Y_J)] = \mathbb{E}[\tilde{F}(Y_I)\tilde{G}(Y_J)] - \mathbb{E}[\tilde{F}(Y_I)]\mathbb{E}[\tilde{G}(Y_J)] = -\mathbb{E}[\tilde{F}(Y_I)]\mathbb{E}[\tilde{G}(Y_J)] \leq 0.$$

□

This elementary lemma, together with the two closure properties, already gives the model that matters most for us.

Theorem 5.21 (Occupancy counts are negatively associated)**.** *Let $X_1, \dots, X_n$ be independent random variables with values in a countable set $\mathcal{X}$, with arbitrary (not necessarily identical) laws, and let $N_n(x) = \sum_{i=1}^n \mathbb{1}\{X_i = x\}$. Then the family of counts $(N_n(x))_{x\in\mathcal{X}}$ is NA. Consequently, the families $(\mathbb{1}\{N_n(x) = 0\})_{x\in\mathcal{X}}$ and $(\mathbb{1}\{N_n(x) \geq 1\})_{x\in\mathcal{X}}$ are NA.*

Proof. Fix a finite $A \subseteq \mathcal{X}$. For each i, the variables $(\mathbb{1}\{X_i = x\})_{x\in A}$ are $\{0, 1\}$-valued with sum at most 1, hence NA by Lemma 5.20. These n families are independent of each other, so by Proposition 5.18 (applied $n - 1$ times) the whole family $(\mathbb{1}\{X_i = x\})_{1\leq i\leq n,\, x\in A}$ is NA. The counts $N_n(x) = \sum_i \mathbb{1}\{X_i = x\}$, $x \in A$, are non-decreasing functions of the disjoint blocks $(\mathbb{1}\{X_i = x\})_{i\leq n}$, hence NA by Proposition 5.19. Finally $u \mapsto \mathbb{1}\{u = 0\}$ is non-increasing and $u \mapsto \mathbb{1}\{u \geq 1\}$ non-decreasing on $\{0, 1, 2, \dots\}$, and Proposition 5.19 applies again. Since A was arbitrary, the countable families are NA. □

The classical formulation is in terms of *balls and bins*: n balls are thrown independently into bins indexed by $\mathcal{X}$, ball i landing in bin x with probability $\mathbb{P}(X_i = x)$, and $N_n(x)$ is the number of balls in bin x. Dubhashi and Ranjan [DR98] gave this proof; the zero–one lemma is their Lemma 8 (for a sum equal to 1, with a proof through the FKG inequality; the direct proof above is the elementary one they attribute to McDiarmid), and the observation that a single multinomial draw is NA "trivially from the definition" is already in [JP83, Sec. 3.1], where the multinomial case (identical laws) is obtained from it by the closure properties, exactly as above, with a second proof by conditioning (Remark 5.25).

The second basic example is sampling without replacement, or more generally random permutations. The proof is longer, but it is instructive: it rests on Chebyshev's correlation inequality and on the stochastic monotonicity of uniform random subsets.

Theorem 5.22 (Permutation distributions are negatively associated)**.** *Let $a_1, \dots, a_n$ be real numbers, not necessarily distinct, let π be a uniformly distributed random permutation of $\{1, \dots, n\}$, and let $X_i := a_{\pi(i)}$. Then $X_1, \dots, X_n$ are NA.*

Proof. We use two facts about uniform random subsets. Let E be a finite set and, for $0 \le s \le |E|$, let Σ_s be a uniformly distributed random s-subset of E, identified with its indicator vector $(\mathbb{1}\{e \in \Sigma_s\})_{e \in E}$.

(α) *Monotonicity in the size.* For every bounded non-decreasing Φ of the indicator vector, $s \mapsto \mathbb{E}\Phi(\Sigma_s)$ is non-decreasing. Indeed, if Σ_{s+1} is obtained from Σ_s by adding a uniformly chosen element of $E \setminus \Sigma_s$, then the law of Σ_{s+1} is invariant under the permutations of E and supported by the $(s+1)$-subsets, so Σ_{s+1} is uniform among them; and $\Sigma_s \subseteq \Sigma_{s+1}$ gives $\Phi(\Sigma_s) \le \Phi(\Sigma_{s+1})$.

(β) *Conditioning on a trace.* Let Σ be a uniform k-subset of E and $I \subseteq E$. Conditionally on $\Sigma \cap I = b$, the set $\Sigma \setminus I$ is a uniform $(k - |b|)$-subset of $E \setminus I$; and conditionally on $|\Sigma \cap I| = s$, the set $\Sigma \cap I$ is a uniform s-subset of I. (Both because the conditional laws are uniform on the corresponding sets of configurations, by symmetry.)

The proof is by induction on the number m of distinct values among the a_i. If $m = 1$ there is nothing to prove. Throughout, I, J are disjoint and F, G bounded non-decreasing; the non-increasing case follows by applying the result to $-F, -G$, which does not change (5.18).

The case $m = 2$. Let $a < b$ be the two values, b appearing k times. Then $X_i = a + (b-a)B_i$ where $B = (B_i)_{i \le n}$ is the indicator vector of the set of positions carrying the value b, which is a uniform k-subset of $\{1, \dots, n\}$. By Proposition 5.19 it suffices to show that B is NA. Let $S := \sum_{i \in I} B_i$. By (β), conditionally on $B_I = b_I$ the vector B_J is the trace on J of a uniform $(k - S)$-subset of $\{1, \dots, n\} \setminus I$; hence $\mathbb{E}[G(B_J) \mid B_I] = \bar{G}(S)$, where $\bar{G}(s)$ is the expectation of G at the trace on J of a uniform $(k-s)$-subset of $\{1, \dots, n\} \setminus I$, a non-increasing function of s by (α). Likewise, conditionally on $S = s$ the vector B_I is a uniform s-subset of I, so $\mathbb{E}[F(B_I) \mid S] = \bar{F}(S)$ with $\bar{F}$ non-decreasing by (α). Therefore, by Lemma 5.16,

$$\mathbb{E}[F(B_I)G(B_J)] = \mathbb{E}\big[F(B_I)\bar{G}(S)\big] = \mathbb{E}\big[\bar{F}(S)\bar{G}(S)\big] \le \mathbb{E}[\bar{F}(S)]\,\mathbb{E}[\bar{G}(S)] = \mathbb{E}[F(B_I)]\,\mathbb{E}[G(B_J)].$$

The induction step. Let $m \ge 3$, let a^* be the largest value, appearing k times, and let $B = (B_i)$ with $B_i = \mathbb{1}\{X_i = a^*\}$: the set of positions of a^* is a uniform k-subset of $\{1, \dots, n\}$, so B is NA by the case $m = 2$. Conditionally on $B = b$, the vector $(X_i)_{i \notin b}$ is a uniform random arrangement, in the positions outside b, of the multiset of the $n - k$ remaining values, which has $m - 1$ distinct values; and $X_i = a^*$ for $i \in b$. Given $B = b$, $F(X_I)$ is therefore a bounded non-decreasing function of $(X_i)_{i \in I \setminus b}$ and $G(X_J)$ one of $(X_i)_{i \in J \setminus b}$, two disjoint sets of positions outside b. By the induction hypothesis,

$$\mathbb{E}[F(X_I)G(X_J) \mid B = b] \le \bar{F}(b)\,\bar{G}(b), \qquad \bar{F}(b) := \mathbb{E}[F(X_I) \mid B = b], \quad \bar{G}(b) := \mathbb{E}[G(X_J) \mid B = b].$$

Now $\bar{F}(b)$ depends on b only through $b_I = (b_i)_{i \in I}$: given $B = b$, the coordinates X_i, $i \in I \cap b$, equal a^* and the coordinates X_i, $i \in I \setminus b$, form a uniform sample without replacement of size $|I \setminus b|$ from the remaining multiset, whose law depends only on which positions of I lie in b. Moreover $\bar{F}$ is *non-decreasing* in b_I: if $b_I \le b'_I$ coordinatewise, take the sample $(V_i)_{i \in I \setminus b}$ of the first configuration; its sub-vector $(V_i)_{i \in I \setminus b'}$ is a uniform sample without replacement of the smaller size, so the vector equal to a^* on $I \cap b'$ and to V_i on $I \setminus b'$ has the conditional law

of X_I given $B_I = b'_I$, and it dominates coordinatewise the vector equal to a^* on $I \cap b$ and to V_i on $I \setminus b$, since $V_i \le a^*$. As F is non-decreasing, $\bar{F}(b_I) \le \bar{F}(b'_I)$. The same holds for $\bar{G}$ as a function of b_J. Taking expectations over B in the conditional inequality, and then applying the case $m = 2$ to the NA vector B with the non-decreasing functions $\bar{F}(b_I)$, $\bar{G}(b_J)$ of the disjoint blocks B_I, B_J,

$$\mathbb{E}[F(X_I)G(X_J)] \le \mathbb{E}\big[\bar{F}(B_I)\bar{G}(B_J)\big] \le \mathbb{E}[\bar{F}(B_I)]\,\mathbb{E}[\bar{G}(B_J)] = \mathbb{E}[F(X_I)]\,\mathbb{E}[G(X_J)]. \qquad \square$$

Corollary 5.23 (Sampling without replacement)**.** *Let $a_1, \dots, a_N$ be real numbers and let $X_1, \dots, X_n$, $n \le N$, be drawn from $\{a_1, \dots, a_N\}$ without replacement (each unordered subset of n indices being equally likely, in a uniformly random order). Then $X_1, \dots, X_n$ are NA.*

Proof. $(X_1, \dots, X_n)$ has the law of the first n coordinates of $(a_{\pi(1)}, \dots, a_{\pi(N)})$ for a uniform permutation π of $\{1, \dots, N\}$; apply Theorem 5.22 and Proposition 5.17(i). $\square$

Remark 5.24 (Negative regression dependence). A related notion, used by Dubhashi and Ranjan [DR98] and natural for discrete models, is *negative regression dependence*: the family (X_i) is NRD if for all disjoint I, J and all bounded non-decreasing G, the conditional expectation $\mathbb{E}[G(X_J) \mid X_I = x_I]$ is a non-increasing function of x_I – knowing that some variables are large makes the others stochastically smaller. The multinomial counts are NRD: with X_i i.i.d. of law P and $I \subseteq \mathcal{X}$ finite, conditionally on $(N_n(x))_{x \in I} = a$ the counts $(N_n(x))_{x \notin I}$ are the occupancy counts of $n - \sum_{x \in I} a_x$ balls thrown independently into the bins $\mathcal{X} \setminus I$ with probabilities $p(x)/(1 - P(I))$; a configuration with one ball less is obtained by deleting a ball, which can only decrease every count, so this conditional law is stochastically non-increasing in $\sum_x a_x$, hence in a. Negative regression dependence is the property that was used, in the guise of (α)–(β), in the proof of Theorem 5.22; the relations between the two notions are delicate – negative regression dependence is not preserved by monotone transformations [DR98, Rem. 22], and the orthant inequalities of Proposition 5.17(iv), which both notions imply, do not imply NA [JP83, Rem. 2.5] – and we shall only need NA.

Remark 5.25 (Joag-Dev and Proschan's route). Joag-Dev and Proschan [JP83] also give a second proof that multinomial vectors are NA: they show that if $Y_1, \dots, Y_k$ are independent with log-concave discrete laws (the Poisson law is an example), then conditionally on $\sum_j Y_j = s$ the vector $(Y_1, \dots, Y_k)$ is NA; and independent Poisson variables with parameters $np(x)$, conditioned on their sum being n, are exactly multinomial counts. This is the *Poissonisation* identity, which we shall prove and use in Section 5.7.5 for a different purpose.

5.4.4 Chernoff's method for negatively associated variables

Theorem 5.26 (NA variables can be treated as independent by Chernoff's method)**.** *Let $X_1, \dots, X_n$ be NA and $S_n = \sum_{i=1}^n X_i$. Then for every $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\lambda S_n}\big] \;\le\; \prod_{i=1}^{n} \mathbb{E}\big[e^{\lambda X_i}\big] \tag{5.19}$$

(both sides possibly infinite). Consequently every inequality of Chapters 1 and 2 whose proof uses independence only through the product formula (1.23) for exponential moments remains valid for NA summands: Hoeffding's inequality (Theorem 1.64) if $a_i \le X_i \le b_i$, Bennett's and Bernstein's inequalities (Theorems 2.40 and 2.47) if the X_i are centred with $X_i \le b$, and the sub-Gaussian and sub-gamma tail bounds of Sections 2.4 and 2.5.

Proof. For $\lambda > 0$ the functions $x_i \mapsto e^{\lambda x_i} \wedge K$ are non-negative, bounded and non-decreasing; Proposition 5.17(iii) gives $\mathbb{E}\prod_i (e^{\lambda X_i} \wedge K) \le \prod_i \mathbb{E}[e^{\lambda X_i} \wedge K]$, and $K \to \infty$ with monotone convergence gives (5.19). For $\lambda < 0$ the functions are non-increasing, and the same argument applies. The proofs of Theorems 1.64, 2.40 and 2.47 and of the additivity of sub-Gaussian and sub-gamma variables bound $\mathbb{E}e^{\lambda S_n}$ by $\prod_i \mathbb{E}e^{\lambda X_i}$ and then each factor separately; they use the equality (1.23) only as the inequality (5.19). Centring preserves negative association (Proposition 5.19 with the maps $x_i \mapsto x_i - \mathbb{E}X_i$). □

This is the whole point of negative association for us: *for Chernoff's method, NA variables can be treated as if they were independent*. Note that (5.19) is needed for λ of both signs if both tails are wanted, and that this is why the definition requires (5.18) for non-increasing as well as for non-decreasing functions.

Example 5.27 (Empty bins)**.** Throw n balls independently and uniformly into k bins, and let $E := \sum_{x=1}^{k} \mathbb{1}\{N_n(x) = 0\}$ be the number of empty bins, a sum of NA indicators (Theorem 5.21) with $\mathbb{E}E = k(1-1/k)^n \approx ke^{-n/k}$. Bernstein's inequality for NA variables gives, with probability at least $1-\delta$,

$$|E - \mathbb{E}E| \le \sqrt{2\,k(1-1/k)^n \log(2/\delta)} + \tfrac{1}{3}\log(2/\delta)\,:$$

the number of empty bins is $ke^{-n/k}$ up to fluctuations of order $\sqrt{ke^{-n/k}}$, much better than the $\sqrt{k}$ of Hoeffding's inequality when $n \gg k$ (Exercise 5.7). The same conclusion follows from the self-bounding structure of the number of occupied bins $D_n = k - E$ (Example 5.12) and Theorem 5.39 below.

5.4.5 Sampling without replacement: Hoeffding's comparison

Remark 1.68 of Chapter 1 announced that Hoeffding's inequality holds for sampling without replacement, because the exponential moments are then smaller than with replacement. We can now prove this in two ways. The first is immediate: by Corollary 5.23 the sample is NA, so Theorem 5.26 applies. The second, which is Hoeffding's own [Hoe63, Thm. 4], is a more precise statement – a comparison of the two sums for *every* convex function, not only for exponentials – whose proof is a nice coupling argument.

Theorem 5.28 (Hoeffding's comparison theorem)**.** *Let $a_1, \dots, a_N$ be real numbers. Let $X_1, \dots, X_n$ ($n \le N$) be drawn from $\{a_1, \dots, a_N\}$ without replacement, and let $Y_1, \dots, Y_n$ be i.i.d. uniform on $\{a_1, \dots, a_N\}$, that is, drawn with replacement. Then for every convex function $\varphi : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\,\varphi\Big(\sum_{i=1}^{n} X_i\Big) \;\le\; \mathbb{E}\,\varphi\Big(\sum_{i=1}^{n} Y_i\Big).$$

Proof. Realise the sample with replacement as $Y_i = a_{\iota_i}$ with $\iota_1, \dots, \iota_n$ i.i.d. uniform on $\{1, \dots, N\}$, and let $m_j := \big|\{i : \iota_i = j\}\big|$ be the number of times the index j was drawn, so that $\sum_i Y_i = \sum_{j=1}^{N} m_j a_j$ and $\sum_j m_j = n$. Let $\Sigma := \{j : m_j \ge 1\}$ be the set of distinct indices drawn, $|\Sigma| \le n$, and complete it into a set Σ' of exactly n indices by adding $n - |\Sigma|$ indices chosen uniformly at random, without replacement, from $\{1, \dots, N\} \setminus \Sigma$, independently of everything else given Σ.

Claim 1: Σ' is a uniform random n-subset of $\{1, \dots, N\}$. The law of the vector $(m_j)_{j \le N}$ is invariant under permutations of the indices (the ι_i are uniform), and the completion procedure commutes with permutations; hence the joint law of (m, Σ') is invariant under

the permutations σ of $\{1,\dots,N\}$ acting by $(m,\Sigma') \mapsto ((m_{\sigma^{-1}(j)})_j, \sigma(\Sigma'))$. In particular the law of Σ' is permutation invariant and supported by the n-subsets, so it is uniform on them. Consequently $\sum_{j\in\Sigma'} a_j$ has the law of $\sum_i X_i$.

Claim 2: $\mathbb{E}\big[\sum_j m_j a_j \;\big|\; \Sigma'\big] = \sum_{j\in\Sigma'} a_j$. Fix an n-subset A with $\mathbb{P}(\Sigma' = A) > 0$ and a permutation σ with $\sigma(A) = A$. By the invariance of Claim 1, for every vector c, $\mathbb{P}(m = c, \Sigma' = A) = \mathbb{P}(m = \sigma\cdot c, \Sigma' = \sigma(A)) = \mathbb{P}(m = \sigma\cdot c, \Sigma' = A)$: conditionally on $\Sigma' = A$, the vector $(m_j)_{j\in A}$ is exchangeable. Since $m_j = 0$ for $j \notin \Sigma'$ and $\sum_{j\in A} m_j = n = |A|$, we get $\mathbb{E}[m_j \mid \Sigma' = A] = 1$ for every $j \in A$, which is the claim.

By Claim 2 and the conditional Jensen inequality, $\varphi\big(\sum_{j\in\Sigma'} a_j\big) = \varphi\big(\mathbb{E}[\sum_j m_j a_j \mid \Sigma']\big) \le \mathbb{E}\big[\varphi(\sum_j m_j a_j) \;\big|\; \Sigma'\big]$; taking expectations and using Claim 1, $\mathbb{E}\varphi(\sum_i X_i) \le \mathbb{E}\varphi(\sum_j m_j a_j) = \mathbb{E}\varphi(\sum_i Y_i)$. □

Corollary 5.29 (Concentration without replacement)**.** *In the setting of Theorem 5.28, let* $\mu = \frac{1}{N}\sum_j a_j$ *and* $\sigma^2 = \frac{1}{N}\sum_j (a_j-\mu)^2$ *be the mean and variance of the population, and let* $\bar X_n = \frac{1}{n}\sum_i X_i$*. Then* $\psi_{\sum_i X_i}(\lambda) \le n\,\psi_{Y_1}(\lambda)$ *for all* λ*, and consequently Hoeffding's inequality (1.28) (if* $a \le a_j \le b$*), Bennett's and Bernstein's inequalities (2.24)–(2.30) (with* $v = n\sigma^2$ *and* $b = \max_j a_j - \mu$*) hold for* $\bar X_n$ *exactly as for the mean of* n *i.i.d. draws. Moreover* $\mathbb{V}\mathrm{ar}(\sum_i X_i) \le n\sigma^2$*; the exact value is* $n\sigma^2\frac{N-n}{N-1}$ *(Exercise 5.6).*

Proof. Apply Theorem 5.28 to $\varphi(u) = e^{\lambda u}$, then to $\varphi(u) = (u - n\mu)^2$; the inequalities follow from the cgf bound as in Chapter 2. □

The variance bound shows the direction of the effect, but it is coarse: sampling without replacement reduces the variance by the factor $(N-n)/(N-1)$, which is small when n is close to N (and zero for $n = N$, when the sample is the whole population and $\bar X_n = \mu$). Exponential inequalities capturing this factor exist [Ser74; BM15]; they require more than the comparison with the i.i.d. case.

5.5 The bounded-difference inequality, revisited

Theorem 2.55 was proved in Section 2.6 through the Doob martingale and a conditional version of Hoeffding's lemma, and it was used there and in Chapter 4 as a black box. The inequality deserves a second look, for two reasons. The first is pedagogical: the martingale proof hides, behind conditional expectations and essential suprema, an argument which is entirely elementary – freeze the past, apply Hoeffding's lemma to one coordinate, integrate, repeat – and which is the exact template of every tensorisation argument of this chapter. The second is practical: the inequality is sometimes very loose, the missing mass being a case in point, and one should understand precisely where the loss occurs in order to know what a better inequality must look like.

Theorem 5.30 (Bounded differences, restated)**.** *Let* $X_1,\dots,X_n$ *be independent,* X_i *with values in* $\mathcal{X}_i$*, and let* $f : \mathcal{X}_1\times\cdots\times\mathcal{X}_n \to \mathbb{R}$ *have the bounded-difference property (2.32) with constants* $c_1,\dots,c_n$*. Then* $Z = f(X_1,\dots,X_n)$ *satisfies*

$$\psi_Z(\lambda) = \log\mathbb{E}\big[e^{\lambda(Z-\mathbb{E}Z)}\big] \;\le\; \frac{\lambda^2}{8}\sum_{i=1}^n c_i^2 \qquad \text{for all } \lambda\in\mathbb{R}, \tag{5.20}$$

and consequently, for every $t > 0$, $\mathbb{P}(Z - \mathbb{E}Z \ge t) \le \exp\big(-2t^2/\sum_i c_i^2\big)$ *and* $\mathbb{P}(Z - \mathbb{E}Z \le -t) \le \exp\big(-2t^2/\sum_i c_i^2\big)$.

Second proof. We keep the notation g_i of the proof of Theorem 5.6: $g_i(x_1,\dots,x_i)$ is the integral of f over the last $n-i$ coordinates, $g_n = f$, $g_0 = \mathbb{E}Z$, and (5.13) holds. Note that f is bounded, by $\bigl|f(x)-f(y)\bigr| \le \sum_i c_i$, so that all integrals exist.

Step 1: a telescoping sum. Define the *increment functions*

$$d_i(x_1,\dots,x_i) \;:=\; g_i(x_1,\dots,x_i) - g_{i-1}(x_1,\dots,x_{i-1}), \qquad 1 \le i \le n.$$

Then, for every x, $f(x) - \mathbb{E}Z = g_n(x) - g_0 = \sum_{i=1}^n d_i(x_1,\dots,x_i)$. The i-th increment depends only on the first i coordinates: on the "past" $x_1,\dots,x_{i-1}$ and the "present" x_i.

Step 2: one coordinate at a time. Fix i and fix the past $x_1,\dots,x_{i-1}$. Consider the function of the present alone,

$$D(y) \;:=\; d_i(x_1,\dots,x_{i-1},y), \qquad y \in \mathcal{X}_i.$$

It has two properties. *(a) It is centred:* by (5.13), $\int D(y)\,\mathrm{d}P_i(y) = \int g_i(x_1,\dots,x_{i-1},y)\,\mathrm{d}P_i(y) - g_{i-1}(x_1,\dots,x_{i-1}) = 0$. *(b) Its oscillation is at most c_i:* for $y, y' \in \mathcal{X}_i$,

$$\begin{aligned} D(y) - D(y') &= g_i(x_1,\dots,x_{i-1},y) - g_i(x_1,\dots,x_{i-1},y') \\ &= \int \bigl[f(x_1,\dots,y,\dots) - f(x_1,\dots,y',\dots)\bigr]\,\mathrm{d}P_{i+1}\cdots\mathrm{d}P_n \;\le\; c_i, \end{aligned}$$

by integrating (2.32) over the future coordinates. Thus $D(X_i)$ is a centred random variable taking its values in an interval of length at most c_i (namely $[\inf D, \sup D]$, both finite), and *Hoeffding's lemma* (Lemma 1.62) gives

$$\int_{\mathcal{X}_i} e^{\lambda D(y)}\,\mathrm{d}P_i(y) \;\le\; e^{\lambda^2 c_i^2/8} \qquad \text{for every } \lambda \in \mathbb{R}, \tag{5.21}$$

whatever the frozen past $x_1,\dots,x_{i-1}$.

Step 3: peel off the last coordinate, then the next, and so on. By Step 1 and Fubini's theorem,

$$\mathbb{E}\bigl[e^{\lambda(Z-\mathbb{E}Z)}\bigr] = \int\cdots\int \exp\Bigl(\lambda\sum_{i=1}^n d_i(x_1,\dots,x_i)\Bigr)\,\mathrm{d}P_1(x_1)\cdots\mathrm{d}P_n(x_n).$$

Integrate over x_n first, the other coordinates being frozen. The factor $\exp(\lambda\sum_{i<n} d_i)$ does not involve x_n, and the remaining integral $\int e^{\lambda d_n(x_1,\dots,x_{n-1},y)}\,\mathrm{d}P_n(y)$ is at most $e^{\lambda^2 c_n^2/8}$ by (5.21). Hence

$$\mathbb{E}\bigl[e^{\lambda(Z-\mathbb{E}Z)}\bigr] \;\le\; e^{\lambda^2 c_n^2/8}\int\cdots\int \exp\Bigl(\lambda\sum_{i=1}^{n-1} d_i(x_1,\dots,x_i)\Bigr)\,\mathrm{d}P_1(x_1)\cdots\mathrm{d}P_{n-1}(x_{n-1}),$$

an expression of the same form with one coordinate less. Repeating the operation for $x_{n-1}, x_{n-2},\dots,x_1$ – the last step being $\int e^{\lambda d_1(y)}\,\mathrm{d}P_1(y) \le e^{\lambda^2 c_1^2/8}$ – yields (5.20).

Step 4: Chernoff. By Theorem 1.57 and (5.20), $\mathbb{P}(Z - \mathbb{E}Z \ge t) \le \exp(-\lambda t + \lambda^2\sum_i c_i^2/8)$ for every $\lambda > 0$; the choice $\lambda = 4t/\sum_i c_i^2$ gives the upper tail, and the lower tail follows by applying the result to $-f$. □

Let us now take the proof apart.

Remark 5.31 (Where independence is used, and what happens without it). Independence enters through the product form $P_1 \otimes \cdots \otimes P_n$ of the law of X, used three times: in the identity (5.13) behind the centring (a), in the oscillation bound (b) where the future is

integrated with respect to $P_{i+1} \otimes \cdots \otimes P_n$ independently of the past, and in the peeling of Step 3. Without independence the statement is false. Let $\varepsilon \sim \mathrm{Ber}(1/2)$ and $X_1 = \cdots = X_n = \varepsilon$, and let $f(x) = \frac{1}{n}\sum_i x_i$, which has bounded differences with $c_i = 1/n$. Then $Z = \varepsilon$, so $\mathbb{P}(|Z - \mathbb{E}Z| \geq \frac{1}{2}) = 1$, whereas (2.33) would give $2e^{-n/2}$. What goes wrong is Step 2(a): with the past frozen at x_1, the "present" X_i is no longer distributed according to P_i – it equals x_1 – and the increment is not centred. In the martingale language of Section 2.6, the Doob martingale $\mathbb{E}[Z \mid X_1, \ldots, X_i]$ jumps from $\frac{1}{2}$ to ε at $i = 1$: the conditional range of the first increment is 1, not $c_1 = 1/n$. This is the right way to think about the hypothesis of Theorem 2.55: what the proof needs is a bound on the *conditional* oscillation of the increments of $\mathbb{E}[Z \mid X_1, \ldots, X_i]$, and independence is what converts the deterministic bounded-difference property of f into such a bound (Theorem 2.56 is the statement with the hypothesis in its natural form).

Remark 5.32 (Two proofs, one mechanism). The martingale proof of Section 2.6 and the one above are the same argument written in two languages: the functions $d_i(X_1, \ldots, X_i)$ of Step 1 are the increments Δ_i of the Doob martingale, and (5.21) is the conditional Hoeffding lemma, with the conditioning replaced by the freezing of explicit variables. The second version is the one that generalises: it shows that the only ingredients are a *one-dimensional* inequality – Hoeffding's lemma, cgf at most $\lambda^2(\text{range})^2/8$ – and a *tensorisation* step, which multiplies the one-dimensional bounds along the coordinates. Replacing Hoeffding's lemma by Popoviciu's inequality (variance at most $(\text{range})^2/4$) gives Corollary 5.9, with the same proxy $\frac{1}{4}\sum_i c_i^2$: *the bounded-difference inequality is the exponential version of the Efron–Stein bound for bounded differences, and it costs nothing in the constants*. Replacing the variance by the entropy gives the method of the next section. The martingale version has its own advantages – it needs no independence, only conditionally bounded increments, and it accommodates random bounds $A_i \leq \Delta_i \leq B_i$ (Theorem 2.56) – and the two should be regarded as complementary. Finally, note that the ordering of the coordinates plays no role in the bound, although it does in the proof; and that, exactly as in Remark 2.58, Step 2(b) only requires the oscillation bound for $P_1 \otimes \cdots \otimes P_{i-1}$-almost every past and $P_i \otimes P_i$-almost every pair (y, y'), so that the almost-everywhere version of the hypothesis suffices.

Remark 5.33 (The loss, on the example of the missing mass). Consider $f(x) = \sum_y p(y)\mathbb{1}\{y \notin \{x_1, \ldots, x_n\}\}$, so that $Z = f(X) = M_n$. Changing one observation reveals at most one symbol and hides at most one, so f has bounded differences with $c_i = p_{\max} := \max_x p(x)$, and Theorem 5.30 gives $\mathbb{P}(|M_n - \mathbb{E}M_n| \geq t) \leq 2\exp(-2t^2/(np_{\max}^2))$. This is useless unless $p_{\max} \ll n^{-1/2}$: for a law with one symbol of mass $1/2$ and the rest spread thinly, the bound is $2e^{-8t^2/n}$, vacuous for $t \leq 0.29\sqrt{n}$, hence for every relevant $t \leq 1$ as soon as $n \geq 12$, while $\mathbb{V}\mathrm{ar}(M_n) \leq 1/(en)$ by (5.8). The reason is that the constants c_i are *worst-case*: the symbol of mass $1/2$ is almost never missing, so the configuration in which changing x_i moves M_n by $1/2$ has probability about 2^{-n}, but (2.32) must hold for it all the same. The Efron–Stein computation of Example 5.14 shows what the right quantity is: the *expected* squared increment $\sum_i \mathbb{E}[(Z - Z^{(i)})_+^2] \leq 4/(e^2 n)$, which is tiny. What one would like is an exponential inequality whose variance proxy is a quantity of this type – an average, or a self-bounding quantity – rather than $\sum_i c_i^2$. Such inequalities exist; the simplest of them is the self-bounding inequality of Theorem 5.39 below, and more general "exponential Efron–Stein" inequalities are given in [BLM03]. For the missing mass, however, the route through negative association (Section 5.7.2) is both more elementary and sharper, because it exploits the structure of M_n as a sum.

5.6 The entropy method, in its simplest form

The variance is not the only functional that tensorises. The *entropy* of a non-negative random variable does too, and when it is applied to $V = e^{\lambda Z}$ it yields a differential inequality for the cumulant generating function of Z whose solution is a concentration inequality. This is the *entropy method*, developed by Ledoux and by Boucheron, Lugosi and Massart after Talagrand's work; [BLM13] is its reference. We present the minimum needed to prove one inequality, the exponential inequality for self-bounding functions, and to explain why it has the shape it has. The relative entropy of Chapter 3, which appeared there as the cost of a change of measure, is the same object seen from another side.

5.6.1 Entropy of a non-negative variable and its two variational formulas

As in Section 3.3, let $\Phi(u) := u \log u$ for $u > 0$ and $\Phi(0) = 0$, a convex continuous function on $[0, \infty)$.

Definition 5.34 (Entropy)**.** For a bounded non-negative random variable V,

$$\operatorname{Ent}(V) \;:=\; \mathbb{E}\big[\Phi(V)\big] - \Phi\big(\mathbb{E}V\big) = \mathbb{E}[V \log V] - \mathbb{E}V \log \mathbb{E}V.$$

By Jensen's inequality $\operatorname{Ent}(V) \geq 0$, with equality if and only if V is almost surely constant; and $\operatorname{Ent}(cV) = c \operatorname{Ent}(V)$ for $c \geq 0$. If $\mathbb{E}V > 0$, the probability measure P_V with density $V/\mathbb{E}V$ with respect to $\mathbb{P}$ satisfies

$$\operatorname{Ent}(V) = \mathbb{E}V \cdot \operatorname{KL}\big(P_V \,\|\, \mathbb{P}\big), \tag{5.22}$$

so that Ent is a relative entropy in disguise, and the two formulas below are the Gibbs variational principle (Proposition 3.35) in disguise.

Proposition 5.35 (Variational formulas for the entropy)**.** *Let V be a bounded non-negative random variable.*

(i) For every constant $u > 0$, $\operatorname{Ent}(V) \leq \mathbb{E}\big[V \log V - V \log u - V + u\big]$, with equality for $u = \mathbb{E}V$ (when $\mathbb{E}V > 0$).

(ii) For every bounded measurable g with $\mathbb{E}[e^g] \leq 1$, $\mathbb{E}[Vg] \leq \operatorname{Ent}(V)$; if $V \geq c > 0$ almost surely, equality holds for $g = \log(V/\mathbb{E}V)$.

Proof. (i) The difference between the right-hand side and $\operatorname{Ent}(V)$ is $\mathbb{E}V\big[\log(\mathbb{E}V/u) - 1 + u/\mathbb{E}V\big]$, which is non-negative because $\log r \geq 1 - 1/r$ for $r = \mathbb{E}V/u > 0$, with equality iff $r = 1$. (If $\mathbb{E}V = 0$, both sides vanish.) (ii) For $u \geq 0$ and $v \in \mathbb{R}$, $uv \leq u \log u - u + e^v$: for $u > 0$ the function $v \mapsto uv - e^v$ is maximal at $v = \log u$, where it equals $u \log u - u$; for $u = 0$ the inequality reads $0 \leq e^v$. Apply it pointwise with $u = V/\mathbb{E}V$ and $v = g$ (if $\mathbb{E}V = 0$ there is nothing to prove), and take expectations:

$$\frac{\mathbb{E}[Vg]}{\mathbb{E}V} \;\leq\; \mathbb{E}\Big[\frac{V}{\mathbb{E}V} \log \frac{V}{\mathbb{E}V}\Big] - 1 + \mathbb{E}[e^g] \;\leq\; \frac{\operatorname{Ent}(V)}{\mathbb{E}V}.$$

If $V \geq c > 0$ and $g = \log(V/\mathbb{E}V)$, then $\mathbb{E}e^g = 1$ and $\mathbb{E}[Vg] = \operatorname{Ent}(V)$ by definition. □

5.6.2 Tensorisation of the entropy

We return to the product setting of Section 5.3.2: $X_1, \dots, X_n$ independent with laws P_i, $\mathbb{E}^{(i)}$ the integration over the i-th coordinate (5.10), and, for a bounded non-negative measurable function w of n variables,

$$\operatorname{Ent}^{(i)} w \;:=\; \mathbb{E}^{(i)}\big[\Phi(w)\big] - \Phi\big(\mathbb{E}^{(i)} w\big),$$

the entropy of w with respect to the i-th coordinate, a function of $x^{(i)}$.

Theorem 5.36 (Tensorisation of the entropy)**.** *For every bounded measurable $w \geq 0$ and $V = w(X_1, \dots, X_n)$,*

$$\operatorname{Ent}(V) \;\leq\; \sum_{i=1}^{n} \mathbb{E}\big[\operatorname{Ent}^{(i)} w\,(X)\big]. \tag{5.23}$$

Proof. Assume first that $w \geq c > 0$. Set $g := \log(V/\mathbb{E}V)$, a bounded variable with $\mathbb{E}e^{g} = 1$ and $\mathbb{E}[Vg] = \operatorname{Ent}(V)$. For $1 \leq i \leq n+1$ let

$$h_i(x_i, \dots, x_n) \;:=\; \frac{1}{\mathbb{E}V} \int w(y_1, \dots, y_{i-1}, x_i, \dots, x_n)\, \mathrm{d}P_1(y_1) \cdots \mathrm{d}P_{i-1}(y_{i-1}),$$

the average of $w/\mathbb{E}V$ over the first $i-1$ coordinates: $h_1 = w/\mathbb{E}V = e^{g}$, $h_{n+1} = 1$, each h_i is bounded below by $c/\mathbb{E}V > 0$, and by Fubini's theorem

$$\int h_i(y, x_{i+1}, \dots, x_n)\, \mathrm{d}P_i(y) = h_{i+1}(x_{i+1}, \dots, x_n). \tag{5.24}$$

Define $g_i := \log h_i - \log h_{i+1}$, a bounded measurable function of $(x_i, \dots, x_n)$. Then $\sum_{i=1}^{n} g_i = \log h_1 - \log h_{n+1} = g$, and by (5.24), since h_{i+1} does not depend on x_i,

$$\mathbb{E}^{(i)}\big[e^{g_i}\big] = \frac{\mathbb{E}^{(i)}[h_i]}{h_{i+1}} = 1.$$

Fix $x^{(i)}$. The functions $y \mapsto w(x_1, \dots, y, \dots, x_n)$ and $y \mapsto g_i(x_1, \dots, y, \dots, x_n)$ are, respectively, a bounded non-negative random variable and a bounded test function with $\int e^{g_i}\, \mathrm{d}P_i \leq 1$ on the probability space $(\mathcal{X}_i, P_i)$, so Proposition 5.35(ii) gives $\mathbb{E}^{(i)}[wg_i] \leq \operatorname{Ent}^{(i)} w$ pointwise in $x^{(i)}$. Therefore

$$\operatorname{Ent}(V) = \mathbb{E}[Vg] = \sum_{i=1}^{n} \mathbb{E}[Vg_i] = \sum_{i=1}^{n} \mathbb{E}\big[\mathbb{E}^{(i)}[wg_i](X)\big] \leq \sum_{i=1}^{n} \mathbb{E}\big[\operatorname{Ent}^{(i)} w(X)\big].$$

For a general bounded $w \geq 0$, apply this to $w + \varepsilon$ and let $\varepsilon \downarrow 0$: both sides converge by bounded convergence, Φ being continuous and bounded on bounded intervals. □

The proof is the entropy analogue of Step 1 of Theorem 5.6: the test function g is decomposed into pieces g_i, each of which is a legitimate test function for the i-th coordinate, and the duality formula is applied coordinate by coordinate. Compare with Proposition 3.20, the tensorisation of the relative entropy of Chapter 3: there the inequality is an equality for product measures, and the content is different.

5.6.3 From entropy to cumulant generating functions

The entropy method applies Theorem 5.36 to $V = e^{\lambda Z}$. The right-hand side is then controlled by the one-dimensional formula (i), and the left-hand side is expressed through the cumulant generating function.

Lemma 5.37 (The modified logarithmic Sobolev inequality)**.** *Let f be bounded, $Z = f(X)$, and let $f_1, \dots, f_n$ be bounded measurable functions, f_i of $x^{(i)}$, with $Z_i := f_i(X^{(i)})$. Then for every $\lambda \in \mathbb{R}$, with $\varphi(u) = e^{u} - u - 1$,*

$$\operatorname{Ent}\big(e^{\lambda Z}\big) \;\leq\; \sum_{i=1}^{n} \mathbb{E}\Big[e^{\lambda Z}\, \varphi\big(-\lambda(Z - Z_i)\big)\Big]. \tag{5.25}$$

Proof. Fix i and $x^{(i)}$, and apply Proposition 5.35(i) on $(\mathcal{X}_i, P_i)$ to the variable $e^{\lambda f}$ with the constant $u = e^{\lambda f_i}$ (which does not depend on the i-th coordinate):

$$\begin{aligned}\operatorname{Ent}^{(i)}\bigl(e^{\lambda f}\bigr) &\le \mathbb{E}^{(i)}\Bigl[e^{\lambda f}\bigl(\lambda f - \lambda f_i\bigr) - e^{\lambda f} + e^{\lambda f_i}\Bigr] \\ &= \mathbb{E}^{(i)}\Bigl[e^{\lambda f}\Bigl(\lambda(f - f_i) - 1 + e^{-\lambda(f-f_i)}\Bigr)\Bigr] = \mathbb{E}^{(i)}\Bigl[e^{\lambda f}\varphi\bigl(-\lambda(f - f_i)\bigr)\Bigr].\end{aligned}$$

Now sum over i, take expectations and use Theorem 5.36. □

The left-hand side of (5.25) is a derivative in disguise. If $F(\lambda) := \mathbb{E}e^{\lambda Z}$ and $L := \log F$ (for Z bounded, F is smooth, $F'(\lambda) = \mathbb{E}[Ze^{\lambda Z}]$, $L(0) = 0$ and $L'(0) = \mathbb{E}Z$), then

$$\operatorname{Ent}\bigl(e^{\lambda Z}\bigr) = \lambda\mathbb{E}\bigl[Ze^{\lambda Z}\bigr] - \mathbb{E}\bigl[e^{\lambda Z}\bigr]\log\mathbb{E}\bigl[e^{\lambda Z}\bigr] = \lambda F'(\lambda) - F(\lambda)L(\lambda) = F(\lambda)\bigl(\lambda L'(\lambda) - L(\lambda)\bigr). \tag{5.26}$$

A bound on the entropy of $e^{\lambda Z}$ in terms of F is therefore a differential inequality for L, and solving it is what is known as *Herbst's argument*. The simplest case is the sub-Gaussian one.

Lemma 5.38 (Herbst's argument)**.** *Let Z be bounded and $v \ge 0$. If $\operatorname{Ent}(e^{\lambda Z}) \le \frac{\lambda^2 v}{2}\mathbb{E}[e^{\lambda Z}]$ for all $\lambda \in \mathbb{R}$, then $\psi_Z(\lambda) \le \lambda^2 v/2$ for all λ: $Z - \mathbb{E}Z$ is sub-Gaussian with variance factor v, and $\mathbb{P}(|Z - \mathbb{E}Z| \ge t) \le 2e^{-t^2/(2v)}$.*

Proof. By (5.26) the hypothesis reads $\lambda L' - L \le \lambda^2 v/2$, that is, $(L(\lambda)/\lambda)' \le v/2$ for $\lambda \ne 0$. Since $L(\lambda)/\lambda \to L'(0) = \mathbb{E}Z$ as $\lambda \to 0$, integrating between 0 and $\lambda > 0$ gives $L(\lambda)/\lambda \le \mathbb{E}Z + \lambda v/2$, that is, $\psi_Z(\lambda) = L(\lambda) - \lambda\mathbb{E}Z \le \lambda^2 v/2$. For $\lambda < 0$, integrating between λ and 0 gives $\mathbb{E}Z - L(\lambda)/\lambda \le -\lambda v/2$, and multiplying by $\lambda < 0$ yields the same conclusion. □

Exercise 5.9 shows how Lemmas 5.37 and 5.38 recover the bounded-difference inequality (with a worse constant, $\sum_i c_i^2$ in place of $\frac{1}{4}\sum_i c_i^2$; the sharp constant can also be obtained by the entropy method, with more care [BLM13, Thm. 6.2]). The interesting applications are those where (5.25) produces a proxy which is *not* a sum of worst-case constants.

5.6.4 Self-bounding functions behave like Poisson variables

Theorem 5.39 (Exponential inequality for self-bounding functions)**.** *Let f be a bounded self-bounding function (Definition 5.10) and $Z = f(X_1, \dots, X_n)$. Then*

$$\psi_Z(\lambda) \;\le\; \varphi(\lambda)\,\mathbb{E}Z = \bigl(e^\lambda - \lambda - 1\bigr)\mathbb{E}Z \qquad \text{for all } \lambda \in \mathbb{R}. \tag{5.27}$$

Consequently, if $\mathbb{E}Z > 0$ (otherwise $Z = 0$ almost surely), with $h(u) = (1+u)\log(1+u) - u$ as in (2.11), for every $t > 0$,

$$\mathbb{P}\bigl(Z \ge \mathbb{E}Z + t\bigr) \le \exp\Bigl(-\mathbb{E}Z\, h\Bigl(\frac{t}{\mathbb{E}Z}\Bigr)\Bigr) \le \exp\Bigl(-\frac{t^2}{2\bigl(\mathbb{E}Z + t/3\bigr)}\Bigr), \tag{5.28}$$

$$\mathbb{P}\bigl(Z \le \mathbb{E}Z - t\bigr) \le \exp\Bigl(-\frac{t^2}{2\,\mathbb{E}Z}\Bigr); \tag{5.29}$$

and for every $\delta \in (0,1)$, each of the two inequalities

$$Z \;>\; \mathbb{E}Z - \sqrt{2\,\mathbb{E}Z\log\tfrac{1}{\delta}}, \qquad Z \;<\; \mathbb{E}Z + \sqrt{2\,\mathbb{E}Z\log\tfrac{1}{\delta}} + \tfrac{1}{3}\log\tfrac{1}{\delta}$$

holds with probability at least $1 - \delta$.

In words: (5.27) is exactly the bound $\psi(\lambda) \le \sigma^2\varphi(\lambda)$ satisfied, with equality, by a centred Poisson variable of parameter $\sigma^2 = \mathbb{E}Z$ (Proposition 2.33). *A self-bounding function fluctuates at most like a Poisson variable with the same mean*: its upper tail is Bennett-like with variance factor $\mathbb{E}Z$ and scale 1, its lower tail is sub-Gaussian with variance factor $\mathbb{E}Z$. This is the exponential counterpart of Corollary 5.11, $\mathbb{V}\mathrm{ar}\, Z \le \mathbb{E}Z$, again at no cost in the constants.

Proof. Apply Lemma 5.37 with the f_i of Definition 5.10. For $u \in [0,1]$ and $\lambda \in \mathbb{R}$, the convexity of $s \mapsto \varphi(-\lambda s)$ and $\varphi(0) = 0$ give $\varphi(-\lambda u) \le u\,\varphi(-\lambda)$. Since $0 \le Z - Z_i \le 1$ and $\sum_i (Z - Z_i) \le Z$,

$$\mathrm{Ent}\big(e^{\lambda Z}\big) \le \varphi(-\lambda)\,\mathbb{E}\Big[e^{\lambda Z}\sum_{i=1}^{n}(Z - Z_i)\Big] \le \varphi(-\lambda)\,\mathbb{E}\big[Ze^{\lambda Z}\big] = \varphi(-\lambda)F'(\lambda),$$

with F, L as in (5.26). Using (5.26), $\lambda F' - FL \le \varphi(-\lambda)F'$, that is,

$$\big(1 - e^{-\lambda}\big)F'(\lambda) \;\le\; F(\lambda)L(\lambda), \tag{5.30}$$

because $\lambda - \varphi(-\lambda) = 1 - e^{-\lambda}$. Define $H(\lambda) := L(\lambda)/(e^\lambda - 1)$ for $\lambda \neq 0$ and $H(0) := L'(0) = \mathbb{E}Z$; H is continuous on $\mathbb{R}$ and differentiable off 0, with

$$H'(\lambda) = \frac{L'(\lambda)(e^\lambda - 1) - L(\lambda)e^\lambda}{(e^\lambda - 1)^2}.$$

For $\lambda > 0$, (5.30) divided by $F(\lambda)(1 - e^{-\lambda}) > 0$ gives $L' \le L/(1 - e^{-\lambda})$, hence $L'(e^\lambda - 1) \le L(e^\lambda - 1)/(1 - e^{-\lambda}) = Le^\lambda$ and $H' \le 0$. For $\lambda < 0$, dividing by $F(\lambda)(1 - e^{-\lambda}) < 0$ gives $L' \ge L/(1 - e^{-\lambda})$, and multiplying by $e^\lambda - 1 < 0$ gives again $L'(e^\lambda - 1) \le Le^\lambda$, so $H' \le 0$. Thus H is non-increasing on $\mathbb{R}$. For $\lambda > 0$ this gives $L(\lambda) \le (e^\lambda - 1)\mathbb{E}Z$; for $\lambda < 0$ it gives $H(\lambda) \ge \mathbb{E}Z$, i.e. $L(\lambda) \le (e^\lambda - 1)\mathbb{E}Z$ as well, since $e^\lambda - 1 < 0$. In both cases $\psi_Z(\lambda) = L(\lambda) - \lambda\mathbb{E}Z \le \varphi(\lambda)\mathbb{E}Z$, which is (5.27).

The tail bounds follow as for the Poisson law: Chernoff's bound with the cumulant generating function $\sigma^2\varphi(\lambda)$, $\sigma^2 = \mathbb{E}Z$, gives $\exp(-\sigma^2 h(t/\sigma^2))$ for the upper tail (Proposition 2.33), and $h(u) \ge u^2/(2(1 + u/3))$ ((2.17) in Proposition 2.33) gives the second form of (5.28); for the lower tail, $\varphi(\lambda) \le \lambda^2/2$ for $\lambda \le 0$ gives (5.29). The confidence statement is Corollary 2.34. □

Example 5.40 (Distinct symbols and empty bins)**.** By Example 5.12, the number D_n of distinct symbols in the sample satisfies, with probability at least $1 - \delta$,

$$|D_n - \mathbb{E}D_n| \;\le\; \sqrt{2\,\mathbb{E}D_n \log\tfrac{2}{\delta}} + \tfrac{1}{3}\log\tfrac{2}{\delta}.$$

Since $\mathbb{E}D_n \le \min(n, |\mathcal{X}|)$ and can be much smaller than n – for Zipf's law with exponent α it is of order $n^{1/\alpha}$ (Exercise 5.1) – this improves considerably on the bounded-difference inequality, which only gives $\sqrt{(n/2)\log(2/\delta)}$. For balls and bins (Example 5.27) it recovers the $\sqrt{ke^{-n/k}}$ fluctuation of the number of empty bins obtained there by negative association, this time without any use of negative association: the two structures – a sum of NA indicators, a self-bounding function – lead to the same conclusion by different routes.

Remark 5.41 (What the entropy method sees, and what it does not). The hypothesis of Theorem 5.39 is a statement about the increments $Z - Z_i$ that holds *pointwise*, like the bounded-difference property, but it is of a different nature: it bounds the sum of the increments $\sum_i (Z - Z_i)$, and therefore of their squares, by Z itself, a random quantity, instead of by a constant. This is the simplest instance of the general philosophy of the entropy

method, in which the variance proxy is a random variable (here Z, in [BLM03] the quantities $\sum_i (Z - Z^{(i)})_\pm^2$) and the differential inequality for L is solved accordingly. The missing mass does not fit: it is not self-bounding, and the Efron–Stein quantity $\sum_i (M_n - M_n^{(i)})_+^2$ is small only in expectation, not uniformly. The next section shows that its structure as a sum of negatively associated terms is the right handle, and that it leads to the *same* Poisson-like shape – a Bennett-type upper tail and a sub-Gaussian lower tail – by a direct computation of tilted variances.

5.7 Concentration of the missing mass

We now have the tools to prove (5.6) with exponential confidence. The two halves use two different tools: the bounded-difference inequality for the estimator, negative association together with the tilted variance representation of Chapter 2 for the missing mass itself.

5.7.1 The Good–Turing estimator: bounded differences

Proposition 5.42 (Concentration of the Good–Turing estimator)**.** *For every P, every $n \geq 1$ and every $t > 0$,*

$$\mathbb{P}\big(\widehat{M}_n - \mathbb{E}\widehat{M}_n \geq t\big) \leq e^{-nt^2/2}, \qquad \mathbb{P}\big(\widehat{M}_n - \mathbb{E}\widehat{M}_n \leq -t\big) \leq e^{-nt^2/2}.$$

Proof. As noted in Example 5.14, the hapax count $U_n = n\widehat{M}_n$ has bounded differences with constants 2, so $\widehat{M}_n$ has bounded differences with constants $c_i = 2/n$, $\sum_i c_i^2 = 4/n$, and Theorem 5.30 gives $\exp(-2t^2/(4/n)) = e^{-nt^2/2}$. □

The proxy $\frac{1}{4}\sum_i c_i^2 = 1/n$ is of the right order in the worst case (Exercise 5.3), so nothing is lost here that could be recovered by a finer argument – at least without assumptions on P. When $\mathbb{E}U_n \ll n$, however, the estimator fluctuates much less than $n^{-1/2}$, and the bounded differences cannot see it; Section 5.7.5 shows how Poissonisation does.

5.7.2 The missing mass: negative association and tilted variances

Recall $I_x = \mathbb{1}\{N_n(x) = 0\}$ and $q(x) = \mathbb{P}(I_x = 1) = (1 - p(x))^n$. The *independent surrogate* of the missing mass is

$$\widetilde{M}_n := \sum_{x \in \mathcal{X}} p(x) Z_x, \qquad Z_x \sim \mathrm{Ber}(q(x)) \text{ independent}, \tag{5.31}$$

a sum of independent weighted Bernoulli variables with the same marginals as the terms of M_n; the series converges absolutely since $\sum_x p(x) = 1$, and $\mathbb{E}\widetilde{M}_n = \mathbb{E}M_n$.

Lemma 5.43 (Comparison with the surrogate)**.** *For every $\lambda \in \mathbb{R}$, $\mathbb{E}\big[e^{\lambda M_n}\big] \leq \mathbb{E}\big[e^{\lambda \widetilde{M}_n}\big] = \prod_{x \in \mathcal{X}} \mathbb{E}\big[e^{\lambda p(x) Z_x}\big]$, the infinite product being the limit of the products over finite sets increasing to $\mathcal{X}$. Consequently $\psi_{M_n} \leq \psi_{\widetilde{M}_n}$ on $\mathbb{R}$.*

Proof. Let $A \subseteq \mathcal{X}$ be finite. By Theorem 5.21 the indicators $(I_x)_{x \in A}$ are NA, and $u \mapsto e^{\lambda p(x) u}$ is non-negative, bounded on $\{0, 1\}$, and monotone in the same direction for all x; by Proposition 5.17(iii) and the independence of the Z_x,

$$\mathbb{E}\Big[e^{\lambda \sum_{x \in A} p(x) I_x}\Big] \leq \prod_{x \in A} \mathbb{E}\big[e^{\lambda p(x) I_x}\big] = \prod_{x \in A} \mathbb{E}\big[e^{\lambda p(x) Z_x}\big] = \mathbb{E}\Big[e^{\lambda \sum_{x \in A} p(x) Z_x}\Big].$$

Let A_k be finite sets increasing to $\mathcal{X}$. The partial sums $\sum_{x\in A_k} p(x)I_x$ and $\sum_{x\in A_k} p(x)Z_x$ increase to M_n and $\widetilde{M}_n$, and the exponentials are bounded by $e^{|\lambda|}$; dominated convergence on both sides gives the inequality and identifies the product. The last claim follows from $\mathbb{E}M_n = \mathbb{E}\widetilde{M}_n$. □

By Chernoff's bound, every deviation bound for $\widetilde{M}_n$ is therefore a deviation bound for M_n. The surrogate is a sum of *independent* variables, but not an ordinary one: its terms have very different weights $p(x)$, and the heavy ones have very small variances. The classical inequalities fail (Remark 5.33; Bennett's inequality with $b = p_{\max}$ has the same defect), and the way out is to go back to Chapter 2 and compute the variance of the tilted law *by hand*.

Theorem 5.44 (Concentration of the missing mass)**.** *For every distribution P on a countable alphabet, every $n \geq 1$ and every $t > 0$, with $v_n = \sum_x p(x)^2(1-p(x))^n \leq 1/(en)$,*

$$\mathbb{P}\big(M_n \leq \mathbb{E}M_n - t\big) \;\leq\; \exp\Big(-\frac{t^2}{2v_n}\Big) \;\leq\; \exp\Big(-\frac{e\,n\,t^2}{2}\Big), \tag{5.32}$$

$$\mathbb{P}\big(M_n \geq \mathbb{E}M_n + t\big) \;\leq\; \exp\Big(-\frac{n}{e}\,\varphi(-et)\Big) \;\leq\; \exp\Big(-\frac{e\,n\,t^2}{2+et}\Big), \tag{5.33}$$

where $\varphi(u) = e^u - u - 1$. Moreover $\psi_{\widetilde{M}_n}(\lambda) \leq \frac{\lambda^2}{2en(1-\lambda/n)}$ for $0 < \lambda < n$, so that $M_n - \mathbb{E}M_n$ satisfies the sub-gamma tail bound (2.20) with variance factor $1/(en)$ and scale $1/n$: for every $y > 0$,

$$\mathbb{P}\Big(M_n \;\geq\; \mathbb{E}M_n + \sqrt{\frac{2y}{en}} + \frac{y}{n}\Big) \;\leq\; e^{-y}. \tag{5.34}$$

Both tails are sub-Gaussian with variance factor $1/(en)$ for small t: the exponent of (5.33) is $\frac{n}{e}\varphi(-et) = \frac{ent^2}{2} - \frac{e^2nt^3}{6} + \dots$, and the lower tail is exactly $e^{-ent^2/2}$. The variance factor is the one found at the Chebyshev level in (5.8): *the exponential inequality costs nothing*. Figure 5.2 compares the exponents.

Proof. Step 1: the cumulant generating function of the surrogate. For each x, let ψ_x be the centred cgf of the term $p(x)Z_x$,

$$\psi_x(\lambda) := \log \mathbb{E}\big[e^{\lambda p(x)(Z_x - q(x))}\big] = \log\big(1 - q(x) + q(x)e^{\lambda p(x)}\big) - \lambda p(x) q(x).$$

By Lemma 5.43, taking logarithms and using $\mathbb{E}\widetilde{M}_n = \sum_x p(x)q(x)$, $\psi_{\widetilde{M}_n}(\lambda) = \sum_x \psi_x(\lambda)$, the series being a limit of finite sums (and convergent: $0 \leq \psi_x(\lambda) \leq$ the bound (5.36) below, which is summable).

Step 2: the variance of the tilted Bernoulli term. Differentiating,

$$\psi_x'(\lambda) = p(x)\big(q_\lambda(x) - q(x)\big), \qquad q_\lambda(x) := \frac{q(x)e^{\lambda p(x)}}{1 - q(x) + q(x)e^{\lambda p(x)}},$$

and since $\lambda \mapsto q_\lambda(x)$ satisfies $\frac{\mathrm{d}}{\mathrm{d}\lambda} q_\lambda(x) = p(x) q_\lambda(x)(1 - q_\lambda(x))$,

$$\psi_x''(\lambda) = p(x)^2\, q_\lambda(x)\big(1 - q_\lambda(x)\big). \tag{5.35}$$

This is Theorem 2.3 in the simplest case: $q_\lambda(x)$ is the parameter of the λ-tilted law of Z_x (the tilt by $e^{\lambda p(x) Z_x}$ multiplies the odds of $\{Z_x = 1\}$ by $e^{\lambda p(x)}$), and (5.35) is the variance of $p(x)Z_x$ under that law. By Taylor's formula with integral remainder, since $\psi_x(0) = \psi_x'(0) = 0$,

$$\psi_x(\lambda) = \int_0^\lambda (\lambda - \ell)\, \psi_x''(\ell)\, \mathrm{d}\ell \;\leq\; \int_0^\lambda (\lambda - \ell)\, p(x)^2 q_\ell(x)\, \mathrm{d}\ell \tag{5.36}$$

(for $\lambda < 0$ the integral is $\int_\lambda^0 (\ell - \lambda)\psi_x''(\ell)\,\mathrm{d}\ell \geq 0$; in both cases the integrand is non-negative). Summing over x (Tonelli),

$$\psi_{\widetilde{M}_n}(\lambda) \;\leq\; \int_0^\lambda (\lambda - \ell)\,\varsigma^2(\ell)\,\mathrm{d}\ell, \qquad \varsigma^2(\ell) := \sum_{x\in\mathcal{X}} p(x)^2 q_\ell(x), \tag{5.37}$$

which is (2.2) for the surrogate, with the variance of the tilted law bounded by the sum of the tilted means $q_\ell(x)$ weighted by $p(x)^2$. Everything now rests on bounding $\varsigma^2(\ell)$.

Step 3: the lower tail. For $\ell \leq 0$, $e^{\ell p(x)} \leq 1$ and $q_\ell(x) \leq q(x)$ (the tilt decreases the odds), so

$$\varsigma^2(\ell) \;\leq\; \sum_x p(x)^2 q(x) = v_n \;\leq\; \sum_x p(x)^2 e^{-np(x)} \;\leq\; \frac{1}{en}$$

by (5.9), exactly as in (5.8). Hence $\psi_{\widetilde{M}_n}(\lambda) \leq \lambda^2 v_n/2$ for $\lambda \leq 0$, and Theorem 1.57 applied to $-\widetilde{M}_n$ gives $\mathbb{P}(\widetilde{M}_n \leq \mathbb{E}\widetilde{M}_n - t) \leq e^{-t^2/(2v_n)}$; by Lemma 5.43 (with $\lambda < 0$) the same holds for M_n, which is (5.32).

Step 4: the upper tail. For $0 \leq \ell < n$, $q_\ell(x) \leq q(x)e^{\ell p(x)} \leq e^{-np(x)}e^{\ell p(x)} = e^{-(n-\ell)p(x)}$, so by (5.9) with $n - \ell > 0$ in place of n,

$$\varsigma^2(\ell) \;\leq\; \sum_x p(x)^2 e^{-(n-\ell)p(x)} \;\leq\; \frac{1}{e(n-\ell)}.$$

The tilted variance blows up as $\ell \uparrow n$, and this is genuine: for tilts of order n the heavy symbols are no longer almost surely seen under the tilted law. Integrating,

$$\psi_{\widetilde{M}_n}(\lambda) \;\leq\; \frac{1}{e}\int_0^\lambda \frac{\lambda - \ell}{n - \ell}\,\mathrm{d}\ell = \frac{1}{e}\Big[\lambda - (n-\lambda)\log\frac{n}{n-\lambda}\Big] = \frac{n}{e}\,g\Big(\frac{\lambda}{n}\Big),$$

for $0 \leq \lambda < n$, where $g(a) := a + (1-a)\log(1-a)$, using $\frac{\lambda-\ell}{n-\ell} = 1 - \frac{n-\lambda}{n-\ell}$. Chernoff's bound then gives, for every $t > 0$,

$$\mathbb{P}\big(M_n \geq \mathbb{E}M_n + t\big) \leq \exp\Big(-\sup_{0<\lambda<n}\big\{\lambda t - \tfrac{n}{e}g(\lambda/n)\big\}\Big) = \exp\Big(-n\sup_{0<a<1}\big\{at - g(a)/e\big\}\Big).$$

Since $g'(a) = -\log(1-a)$, the supremum is attained at $1 - a = e^{-et} \in (0,1)$, where $g(a) = 1 - e^{-et} - et\,e^{-et}$, and its value is $t(1-e^{-et}) - \frac{1}{e}(1 - e^{-et} - ete^{-et}) = \frac{1}{e}\big(et - 1 + e^{-et}\big) = \frac{1}{e}\varphi(-et)$. This is the first bound of (5.33). The second follows from the elementary inequality

$$\varphi(-u) = e^{-u} - 1 + u \;\geq\; \frac{u^2}{2+u} \qquad (u \geq 0): \tag{5.38}$$

the function $r(u) = (2+u)\varphi(-u) - u^2 = (2+u)e^{-u} - 2 + u$ vanishes at 0 and $r'(u) = 1 - (1+u)e^{-u} \geq 0$. Finally, $g(a) = \sum_{k\geq 2}\frac{a^k}{k(k-1)}$ (expand $(1-a)\log(1-a)$), so $g(a) \leq \frac{1}{2}\sum_{k\geq 2} a^k = \frac{a^2}{2(1-a)}$ and $\psi_{\widetilde{M}_n}(\lambda) \leq \frac{\lambda^2}{2en(1-\lambda/n)}$ for $0 < \lambda < n$: in the notation of Definition 2.36, $\widetilde{M}_n - \mathbb{E}\widetilde{M}_n \in \Gamma_+(1/(en), 1/n)$, and Theorem 2.37 with Lemma 5.43 gives (5.34). □

Remark 5.45 (Reading the proof). The argument is Remark 2.7 in action, and it is worth saying in words what the variance computation found. Under the ℓ-tilted law, a symbol of mass $p(x)$ is missing with probability about $e^{-(n-\ell)p(x)}$: tilting by $\ell > 0$ *rewards* configurations with a large missing mass, and the effect is the same as reducing the sample size from n to $n - \ell$. The variance of the tilted surrogate is therefore at most what the Chebyshev-level computation

gives for a sample of size $n-\ell$, namely $1/(e(n-\ell))$, and the cumulant generating function is the double integral of this quantity. The lower tail is simpler because negative tilts make missing symbols rarer, and the untilted variance bound v_n serves for all $\ell \le 0$. No inequality of Chapter 2 could have found this, because each of them bounds the tilted variance by a quantity that depends on the range of the summands – $p_{\max}$ here – and the whole point is that the summands of large range have negligible variance.

Remark 5.46 (Distribution-dependent variance factors). Step 3 proves more than (5.32): the lower tail is sub-Gaussian with the *distribution-dependent* variance factor $v_n = \sum_x p(x)^2(1-p(x))^n$, which can be much smaller than $1/(en)$ (Remark 5.5). For the upper tail, Step 4 only used $q(x) \le e^{-np(x)}$ and the worst case over $p(x)$; Ben-Hamou, Boucheron and Ohannessian [BBO17] show that $M_n - \mathbb{E}M_n$ is sub-gamma on the right with scale $1/n$ and a variance factor of the same nature as v_n (namely $2\,\mathbb{E}[\text{number of symbols seen at least twice}]/n^2$), and that under a regular-variation assumption on the tail of P these variance factors are of the exact order of $\mathbb{V}\mathrm{ar}(M_n)$. The proof of Theorem 5.44 is a version of theirs and of McAllester and Ortiz's [MO03], whose Theorem 10 is the second form of (5.32) and whose Theorem 12, $\mathbb{P}(M_n \ge \mathbb{E}M_n + t) \le \exp(-ent^2/(2(1+et)))$, is a slightly weaker relaxation of (5.33) (they bound the tilted variance by its value at the largest tilt instead of integrating $1/(e(n-\ell))$); both are obtained by the same "Gibbs-variance lemma", which is our Proposition 2.11.

Remark 5.47 (Sharper constants by a sharper Bernoulli bound). The constants $e/2$ and 1 in the exponents are not the best known. McAllester and Ortiz [MO03, Thm. 16] and Berend and Kontorovich [BK13, Thm. 2.1] prove the cleaner upper-tail bound

$$\mathbb{P}\big(M_n \ge \mathbb{E}M_n + t\big) \le e^{-nt^2},$$

by replacing the tilted-variance computation with a sharp bound on the cumulant generating function of a Bernoulli variable, due to Kearns and Saul [KS98] (with a complete proof in [BK13]), which we state without proof: if $Z \sim \mathrm{Ber}(q)$ then for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\big[e^{\lambda(Z-q)}\big] \;\le\; \frac{1-2q}{4\log\frac{1-q}{q}}\,\lambda^2 \;\le\; \frac{\lambda^2}{4\log(1/q)},$$

the middle term being read as $\lambda^2/8$ for $q = \frac12$ (Hoeffding's lemma, which this inequality refines for $q \ne \frac12$). Applied to the surrogate, $\psi_{\widetilde{M}_n}(\lambda) \le \lambda^2 \sum_x p(x)^2/(4\log(1/q(x))) \le \lambda^2 \sum_x p(x)^2/(4np(x)) = \lambda^2/(4n)$ since $\log(1/q(x)) = -n\log(1-p(x)) \ge np(x)$; this is sub-Gaussianity with variance factor $1/(2n)$, on both sides, hence e^{-nt^2} for both tails. For the lower tail Berend and Kontorovich obtain $e^{-1.92\,nt^2}$ by a finer version of the same idea. The tilted-variance route is less sharp but more transparent, and it is the one that extends to the Bernstein-type statements of Remark 5.46 and to the histogram rules of Section 5.7.6.

5.7.3 The confidence region

Theorem 5.48 (Distribution-free estimation of the missing mass)**.** *For every distribution P on a countable alphabet, every $n \ge 1$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\widehat{M}_n - \frac{1}{n} - \Big(\sqrt{2}+\sqrt{\tfrac{2}{e}}\Big)\sqrt{\frac{\log(4/\delta)}{n}} \;\le\; M_n \;\le\; \widehat{M}_n + \Big(\sqrt{2}+\sqrt{\tfrac{2}{e}}\Big)\sqrt{\frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n}. \tag{5.39}$$

In particular $\big|\widehat{M}_n - M_n\big| \le 2.28\sqrt{\log(4/\delta)/n} + (1+\log(4/\delta))/n$.

Proof. Write $y = \log(4/\delta)$. Consider the four events

$$\widehat{M}_n - \mathbb{E}\widehat{M}_n \geq \sqrt{2y/n}, \qquad \widehat{M}_n - \mathbb{E}\widehat{M}_n \leq -\sqrt{2y/n},$$
$$M_n - \mathbb{E}M_n \leq -\sqrt{2y/(en)}, \qquad M_n - \mathbb{E}M_n \geq \sqrt{2y/(en)} + y/n.$$

Each has probability at most $e^{-y} = \delta/4$: the first two by Proposition 5.42, the third by (5.32), the fourth by (5.34). Outside their union, of probability at most δ, Proposition 5.3 gives

$$M_n < \mathbb{E}M_n + \sqrt{\tfrac{2y}{en}} + \tfrac{y}{n} \leq \mathbb{E}\widehat{M}_n + \sqrt{\tfrac{2y}{en}} + \tfrac{y}{n} < \widehat{M}_n + \sqrt{\tfrac{2y}{n}} + \sqrt{\tfrac{2y}{en}} + \tfrac{y}{n},$$

and symmetrically $M_n > \mathbb{E}M_n - \sqrt{2y/(en)} \geq \mathbb{E}\widehat{M}_n - \frac{1}{n} - \sqrt{2y/(en)} > \widehat{M}_n - \sqrt{2y/n} - \frac{1}{n} - \sqrt{2y/(en)}$. □

Remark 5.49 (What the theorem says). *(a)* Nothing is assumed on P: not its support, not its tail, not even that $\mathcal{X}$ is finite. The width of the region, of order $\sqrt{\log(1/\delta)/n}$, is the same for all P, whereas M_n itself ranges from exponentially small (finite support) to $n^{-(\alpha-1)/\alpha}$ (Zipf) – the estimate is informative exactly when $M_n \gg n^{-1/2}$, which is the case in the regimes of Example 5.2 where the problem is interesting (Figure 5.1 shows the region on a Zipf law). *(b)* The estimator's own fluctuation dominates: the term $\sqrt{2}$ comes from Proposition 5.42, the term $\sqrt{2/e} \approx 0.86$ from the missing mass. The bias $1/n$ and the sub-gamma correction $\log(4/\delta)/n$ are lower order. *(c)* The first bound of this type is Theorem 9 of McAllester and Schapire [MS00], who proved the upper bound $M_n \leq \widehat{M}_n + (2\sqrt{2} + \sqrt{3})\sqrt{\log(3/\delta)/n}$ and left the lower one open; it was settled by McAllester and Ortiz [MO03]. Combining the e^{-nt^2} tails of Remark 5.47 with Proposition 5.42 gives the slightly weaker constant $1 + \sqrt{2}$ in place of $\sqrt{2} + \sqrt{2/e}$, but without the $\log(4/\delta)/n$ term.

5.7.4 How sharp is this?

Three questions. Is the fluctuation of M_n really of order $n^{-1/2}$? Is the accuracy $n^{-1/2}$ of the estimator the best possible? And is the additive accuracy the right thing to ask for?

The fluctuations of M_n. For the uniform law on n symbols, $\mathbb{V}\mathrm{ar}(M_n) \sim (e^{-1} - 2e^{-2})/n \approx 0.097/n$ (Exercise 5.2), so M_n does fluctuate at scale $n^{-1/2}$, and the variance factor $1/(en)$ of Theorem 5.44 is sharp up to the constant $0.37/0.097 \approx 3.8$ – the bound $v_n \leq 1/(en)$ ignores the factor $(1 - q(x))$ and the negative covariances of (5.7), and this is where the constant is lost. For other laws $\mathbb{V}\mathrm{ar}(M_n)$ is smaller, often by orders of magnitude (Remark 5.5), and the distribution-free bound is then pessimistic; but a *distribution-free* statement cannot do better, since it must cover the uniform case.

The accuracy of the estimator. Here the right measure is the minimax mean squared error

$$R_n^* := \inf_{\widehat{M}} \sup_{P} \mathbb{E}\big[(\widehat{M} - M_n)^2\big],$$

the infimum being over all estimators $\widehat{M} = \widehat{M}(X_1, \ldots, X_n)$ and the supremum over all distributions on a countable alphabet. We admit the following results, due to Rajaraman, Thangaraj and Suresh [RTS17] and to Acharya, Bao, Kang and Sun [Ach+18]: as $n \to \infty$,

$$\frac{0.570}{n} + o\Big(\frac{1}{n}\Big) \;\leq\; R_n^* \;\leq\; \sup_{P} \mathbb{E}\big[(\widehat{M}_n - M_n)^2\big] = \frac{\alpha_{\mathrm{GT}}}{n} + o\Big(\frac{1}{n}\Big), \tag{5.40}$$

where $\alpha_{\mathrm{GT}} = \max_{0 \leq u \leq 1} u(1 - u - \log u) \approx 0.608$, the worst case for the Good–Turing estimator being attained, asymptotically, by uniform laws on $k \asymp n$ symbols. Thus the minimax rate is

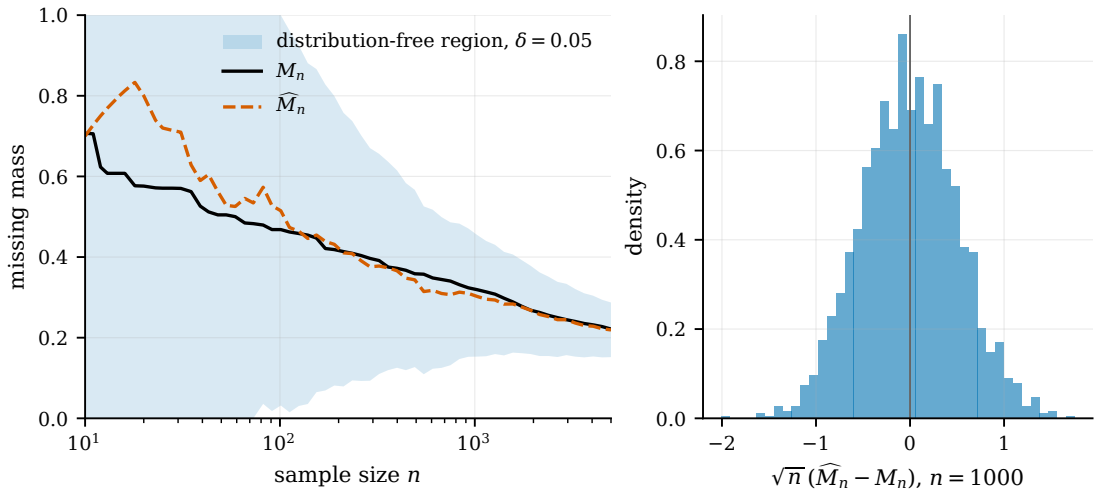


Figure 5.1: Left: one sample path of the missing mass M_n (solid) and of the Good–Turing estimator $\widehat{M}_n$ (dashed), for a Zipf law with exponent 1.2 truncated at 10^6 symbols, together with the confidence region (5.39) at level $\delta = 0.05$ (shaded). Right: histogram of $\sqrt{n}\,(\widehat{M}_n - M_n)$ over 2000 independent samples of size $n = 1000$, for the same law; the fluctuations are of order $n^{-1/2}$, but far inside the distribution-free region.

exactly $1/n$ for the squared error, the Good–Turing estimator attains it, and it is within 7% of the minimax constant; whether it is exactly minimax is open. In terms of Theorem 5.48: the width $n^{-1/2}$ of the confidence region cannot be improved for any estimator, uniformly over P (compare Exercise 5.2, where the mean squared error of $\widehat{M}_n$ is computed explicitly in the uniform case; the lower bound of (5.40) is a Bayesian argument restricted to uniform laws on k symbols, with a prior on k of order n under which the Bayes estimator is asymptotically the maximum-likelihood estimator of the missing mass [Ach+18, Thm. 2]; a two-point reduction to the estimation of a Bernoulli parameter, in the spirit of Section 3.5, only gives $0.25/n$ [RTS17, Thm. 11]).

Additive versus multiplicative accuracy. When $\mathbb{E}M_n \ll n^{-1/2}$ – finite alphabets, light tails – Theorem 5.48 says nothing useful: the missing mass is smaller than the error bar. What one would like then is a *relative* guarantee, $\widehat{M}_n/M_n \to 1$. This cannot hold in general: for geometric laws the Good–Turing estimator is not even consistent in this sense [OD12, Prop. 2]. Under a regular-variation assumption on the tail of P – the mass of the symbols of probability at least u behaves like $u^{-\alpha}$ times a slowly varying function, $\alpha \in (0,1)$, as for Zipf's law – Ohannessian and Dahleh [OD12] and Ben-Hamou, Boucheron and Ohannessian [BBO17] prove multiplicative concentration, $M_n/\mathbb{E}M_n \to 1$ with exponential bounds, and the consistency of $\widehat{M}_n/M_n$; the distribution-dependent variance factors of Remark 5.46 are the tool. This is the natural continuation of the subject, and we leave it there.

5.7.5 Poissonisation

The dependence between the counts $N_n(x)$ comes from the constraint $\sum_x N_n(x) = n$. A classical device removes it: draw the sample size itself at random, from a Poisson law. The counts then become independent, every functional of the counts becomes a functional of independent variables, and the price is a comparison between the two models.

Proposition 5.50 (Poissonisation). *Let $(X_i)_{i\geq 1}$ be i.i.d. with law P and let $T \sim \mathrm{Poi}(n)$ be independent of the sequence. Set $N_T(x) := \sum_{i=1}^{T} \mathbb{1}\{X_i = x\}$. Then the random variables $(N_T(x))_{x\in\mathcal{X}}$*

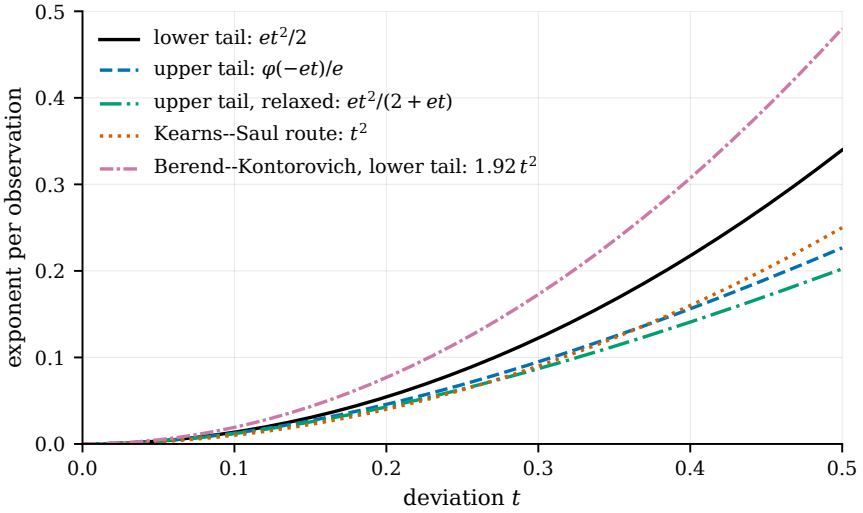


Figure 5.2: The exponents of Theorem 5.44, per observation: the lower tail $et^2/2$ (5.32), the upper tail $\varphi(-et)/e$ (5.33) and its sub-gamma relaxation $et^2/(2+et)$, compared with the t^2 of the Kearns–Saul route (Remark 5.47) and the $1.92\,t^2$ of Berend and Kontorovich for the lower tail. All coincide with $et^2/2$ to first order for small t, except the last two.

are independent, with $N_T(x) \sim \mathrm{Poi}(np(x))$.

Proof. Assume first $\mathcal{X} = \{x_1, \dots, x_k\}$ finite. For non-negative integers $m_1, \dots, m_k$ with $m = \sum_j m_j$, conditioning on $T = m$, the counts are multinomial, so

$$\mathbb{P}\big(N_T(x_j) = m_j,\ 1 \le j \le k\big) = e^{-n}\frac{n^m}{m!} \cdot \frac{m!}{m_1! \cdots m_k!} \prod_{j=1}^{k} p(x_j)^{m_j} = \prod_{j=1}^{k} e^{-np(x_j)} \frac{(np(x_j))^{m_j}}{m_j!},$$

using $\sum_j p(x_j) = 1$: the joint law is the product of the announced Poisson laws. For a countable $\mathcal{X}$ and a finite $A \subseteq \mathcal{X}$, apply this to the lumped alphabet $A \cup \{*\}$, where $*$ stands for $\mathcal{X} \setminus A$ (the lumped sequence is i.i.d. on a finite alphabet, and its counts are those of the original one on A): the family $(N_T(x))_{x \in A}$ is independent with the right marginals, for every finite A. □

Proposition 5.51 (Depoissonisation)**.** *In the setting of Proposition 5.50, for every measurable* $\Psi : \mathbb{N}^{\mathcal{X}} \to [0, \infty)$,

$$\mathbb{E}\big[\Psi\big((N_n(x))_{x \in \mathcal{X}}\big)\big] \ \le\ e\sqrt{n}\ \mathbb{E}\big[\Psi\big((N_T(x))_{x \in \mathcal{X}}\big)\big];$$

in particular, every event concerning the counts which has probability at most δ *under Poisson sampling has probability at most* $e\sqrt{n}\,\delta$ *under sampling of fixed size* n.

Proof. Conditioning on T, $\mathbb{E}\Psi(N_T) = \sum_{m \ge 0} \mathbb{P}(T = m)\mathbb{E}\Psi(N_m) \ge \mathbb{P}(T = n)\mathbb{E}\Psi(N_n)$, and $\mathbb{P}(T = n) = e^{-n}n^n/n!$. It remains to see that $n! \le e\sqrt{n}\,n^n e^{-n}$. Since log is concave, the trapezoidal rule underestimates its integral: $\frac{1}{2}(\log k + \log(k+1)) \le \int_k^{k+1} \log u \ \mathrm{d}u$ for $k \ge 1$. Summing for $k = 1, \dots, n-1$ gives $\log n! - \frac{1}{2}\log n \le \int_1^n \log u \ \mathrm{d}u = n\log n - n + 1$, which is the claim. □

The factor $e\sqrt{n}$ looks expensive, but it only costs an additive $\log(e\sqrt{n})$ in an exponent, which is negligible compared with the nt^2 of the inequalities we are after – as long as t is not too small. It can be reduced to 2 for events whose probability is monotone in the sample size [MU05, Thm. 5.10], at the cost of a fact about the median of the Poisson law that we shall not prove.

Proposition 5.52 (The Poissonised missing mass and its estimator)**.** *Let*

$$M_T := \sum_{x\in\mathcal{X}} p(x)\mathbb{1}\{N_T(x)=0\} \qquad \text{and} \qquad U_T := \sum_{x\in\mathcal{X}} \mathbb{1}\{N_T(x)=1\}$$

be the missing mass and the hapax count of the Poissonised sample. Then:

(i) $\mathbb{E}M_T = \mathbb{E}[U_T/n] = \sum_x p(x)e^{-np(x)}$*, and* $0 \le \mathbb{E}M_T - \mathbb{E}M_n \le 4/(e^2 n)$*;*

(ii) M_T *satisfies* (5.32)–(5.34) *verbatim, with* $\mathbb{E}M_T$ *in place of* $\mathbb{E}M_n$ *and* $\sum_x p(x)^2 e^{-np(x)}$ *(still at most* $1/(en)$*) in place of* v_n*;*

(iii) U_T *is a sum of independent indicators with* $\mathbb{V}\mathrm{ar}(U_T) \le \mathbb{E}U_T$*, so that for every* $t > 0$*,*

$$\mathbb{P}\big(|U_T - \mathbb{E}U_T| \ge t\big) \le 2\exp\Big(-\frac{t^2}{2(\mathbb{E}U_T + t/3)}\Big).$$

Proof. (i) $\mathbb{P}(N_T(x)=0) = e^{-np(x)}$ and $\mathbb{P}(N_T(x)=1) = np(x)e^{-np(x)}$ give the two expectations, which are equal: *under Poisson sampling the Good–Turing estimator is exactly unbiased*. Next, $e^{-np} - (1-p)^n \ge 0$ because $1-p \le e^{-p}$, and since $e^{-p}(1-p^2) \le 1-p$ (that is, $e^p \ge 1+p$) one has $(1-p)^n \ge e^{-np}(1-p^2)^n \ge e^{-np}(1-np^2)$, so that $\mathbb{E}M_T - \mathbb{E}M_n = \sum_x p(x)\big(e^{-np(x)} - (1-p(x))^n\big) \le \sum_x p(x)\, np(x)^2 e^{-np(x)} \le \frac{1}{n}\sum_x p(x) \sup_{u\ge 0} u^2 e^{-u} = \frac{4}{e^2 n}$. (ii) By Proposition 5.50, M_T *is* an independent surrogate of the form (5.31), with $q(x) = e^{-np(x)}$ in place of $(1-p(x))^n$; the proof of Theorem 5.44 used $q(x)$ only through $q(x) \le e^{-np(x)}$, and applies verbatim (Lemma 5.43 is not even needed). (iii) The indicators $\mathbb{1}\{N_T(x)=1\}$ are independent, so $\mathbb{V}\mathrm{ar}(U_T) = \sum_x \mathbb{P}(N_T(x)=1)(1-\mathbb{P}(N_T(x)=1)) \le \mathbb{E}U_T$, and Bernstein's inequality (2.28) applies with $b=1$ and $v = \mathbb{E}U_T$ (the countable sum is handled as in Lemma 5.43). □

The interesting item is (iii). The bounded-difference inequality sees the hapax count as a function of n variables with constants 2, and gives fluctuations of order $\sqrt{n}$ whatever P; Poissonisation sees it as a sum of independent indicators and gives fluctuations of order $\sqrt{\mathbb{E}U_T}$, which is much smaller when few symbols are hapaxes – for Zipf's law with exponent α, $\mathbb{E}U_n \asymp n^{1/\alpha}$ (Exercise 5.1). Transferring the bound to the fixed-n sample by Proposition 5.51 costs the factor $e\sqrt{n}$, and the comparison of the means costs $4/(e^2 n)$; the result is the following variance-adaptive guarantee for the Good–Turing estimator, which should be compared with Proposition 5.42.

Corollary 5.53 (A variance-adaptive bound for the Good–Turing estimator)**.** *For every* P*,* $n \ge 1$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\Big|\widehat{M}_n - \sum_x p(x)e^{-np(x)}\Big| \le \frac{\sqrt{2\,\mathbb{E}U_T\, y} + y/3}{n}, \qquad y := \log\frac{2e\sqrt{n}}{\delta},$$

with $\mathbb{E}U_T = n\sum_x p(x)e^{-np(x)}$*; and* $\sum_x p(x)e^{-np(x)}$ *differs from* $\mathbb{E}M_n$ *by at most* $4/(e^2 n)$*.*

Proof. By (iii) and Theorem 2.37, under Poisson sampling $|U_T - \mathbb{E}U_T| < \sqrt{2\mathbb{E}U_T y} + y/3$ except on an event of probability at most $2e^{-y} = \delta/(e\sqrt{n})$; this event concerns the counts only, so by Proposition 5.51 its fixed-n counterpart has probability at most δ. Divide by n and use (i). □

For Zipf's law with exponent α the bound is of order $n^{1/(2\alpha)-1}\sqrt{\log n}$, against $n^{-1/2}$ in Theorem 5.48: a substantial gain for $\alpha > 1$, obtained at the price of a $\sqrt{\log n}$ and of a bound

that depends on P through $\mathbb{E}U_T$ – which, however, can be estimated by U_n itself, a point we do not pursue ([BBO17, Sec. 5] does). Poissonisation is thus complementary to negative association: it applies to functionals, like the hapax count, that have no monotone structure, and it yields variance-sensitive bounds; negative association keeps the sample size fixed and loses nothing in the exponent.

5.7.6 Histogram classification rules

We close with a classification problem in which the missing mass appears naturally. Return to the framework of Chapter 1: (X, Y) has law $\mathcal{D}$ on $\mathcal{X} \times \{0, 1\}$, where $\mathcal{X}$ is now an arbitrary measurable space, and $(X_i, Y_i)_{i \le n}$ is an i.i.d. sample. Fix a countable measurable partition $(C_j)_{j \in J}$ of $\mathcal{X}$ into *cells*, and write

$$p_j := \mathcal{D}_{\mathcal{X}}(C_j), \qquad \eta_j := \mathbb{P}(Y = 1 \mid X \in C_j) \ \ (\text{if } p_j > 0),$$

$$u_j := \left|2\eta_j - 1\right|, \qquad \rho_j := 1 - 2\sqrt{\eta_j(1-\eta_j)}$$

(and, by convention, $\eta_j = \frac{1}{2}$, $u_j = \rho_j = 0$ when $p_j = 0$: such cells carry no mass and play no role). The *histogram rule* is the majority vote within each cell. Let $n_j^y := \left|\{i : X_i \in C_j,\ Y_i = y\}\right|$ be the number of observations of cell j with label y; for $x \in C_j$,

$$\widehat{h}_n(x) := \begin{cases} 1 & \text{if } n_j^1 > n_j^0, \\ 0 & \text{if } n_j^1 < n_j^0, \\ \xi_j & \text{if } n_j^1 = n_j^0 \text{ (including the empty cell)}, \end{cases}$$

where the ξ_j are independent fair coin tosses, independent of the sample. This randomised tie-breaking is not a cosmetic choice: it is what makes the error probability of the vote in a cell a *monotone* function of the number of observations in the cell (Lemma 5.54), and the analysis rests on this monotonicity (Exercise 5.12).

Among the classifiers which are constant on each cell, the best one is $\bar{h}(x) := \mathbb{1}\{\eta_j \ge \frac{1}{2}\}$ for $x \in C_j$: for any such h, conditioning on the cell gives $L_{\mathcal{D}}(h) = \sum_j p_j \big[\eta_j \mathbb{1}\{h = 0 \text{ on } C_j\} + (1-\eta_j)\mathbb{1}\{h = 1 \text{ on } C_j\}\big]$, minimised cell by cell by $\bar{h}$, with $L_{\mathcal{D}}(\bar{h}) = \sum_j p_j \min(\eta_j, 1-\eta_j)$. The same formula applied to the (random, cell-constant) classifier $\widehat{h}_n$ gives the *excess risk of the histogram rule over the best histogram rule*,

$$W_n \ := \ L_{\mathcal{D}}(\widehat{h}_n) - L_{\mathcal{D}}(\bar{h}) = \sum_{j \in J} p_j \, u_j \, E_j, \qquad E_j := \mathbb{1}\big\{\widehat{h}_n \ne \bar{h} \text{ on } C_j\big\}, \tag{5.41}$$

because a wrong vote in cell j costs $\left|\eta_j - (1-\eta_j)\right| = u_j$ per unit of mass. (Here $L_{\mathcal{D}}(\widehat{h}_n)$ is the risk conditionally on the sample and the coins; the difference $L_{\mathcal{D}}(\bar{h}) - L_{\mathcal{D}}^*$ is the approximation error of the partition, which concentration has nothing to say about.) The structure is that of the missing mass: a sum over cells of a weight times an indicator of failure, and the failure of cell j is likely only when the cell contains few observations. Indeed an empty cell fails with probability $\frac{1}{2}$, so $W_n \ge \frac{1}{2}\sum_j p_j u_j \mathbb{1}\{\text{cell } j \text{ empty}\}$: the missing mass of the partition, weighted by the u_j, is part of the excess risk.

Lemma 5.54 (Majority votes: the error is monotone in the number of voters)**.** *Let $\eta \in [0, 1]$, $m \ge 0$, let $B \sim \mathrm{Bin}(m, \eta)$ be the number of votes for 1 among m independent $\mathrm{Ber}(\eta)$ votes, and*

let $r_\eta(m)$ be the probability that the majority vote, with fair randomised tie-breaking, differs from $\mathbb{1}\{\eta \ge \frac12\}$. Then

$$r_\eta(m) \;\le\; \big(2\sqrt{\eta(1-\eta)}\big)^m = (1-\rho)^m, \qquad \rho = 1-2\sqrt{\eta(1-\eta)} = \big(\sqrt{\eta}-\sqrt{1-\eta}\big)^2,$$

and $m \mapsto r_\eta(m)$ is non-increasing.

Proof. By symmetry ($\eta \leftrightarrow 1-\eta$, votes relabelled) we may assume $\eta \ge \frac12$, so that the vote is wrong when $B < m/2$, and with probability $\frac12$ when $B = m/2$: $r_\eta(m) = \mathbb{P}(B < m/2) + \frac12\mathbb{P}(B = m/2) \le \mathbb{P}(B \le m/2)$. Chernoff's bound for the lower tail of a binomial variable gives, for $\lambda > 0$,

$$\mathbb{P}(B \le m/2) \le e^{\lambda m/2}\mathbb{E}e^{-\lambda B} = \Big[e^{\lambda/2}\big(1-\eta+\eta e^{-\lambda}\big)\Big]^m,$$

and the choice $e^{-\lambda} = (1-\eta)/\eta$ (legitimate for $\eta > \frac12$; for $\eta = \frac12$ the bound 1 is trivial) makes the bracket equal to $\sqrt{\eta/(1-\eta)}\cdot 2(1-\eta) = 2\sqrt{\eta(1-\eta)}$.

For the monotonicity, write B_m for the number of votes for 1 among the first m votes of an infinite sequence, and $k \ge 1$. *Even sizes change nothing:* $r_\eta(2k) = r_\eta(2k-1)$. Indeed, with $\pi_s := \mathbb{P}(B_{2k-1} = s)$,

$$r_\eta(2k) = \mathbb{P}(B_{2k-1} \le k-2) + \pi_{k-1}\big[(1-\eta)+\tfrac12\eta\big] + \tfrac12\pi_k(1-\eta), \qquad r_\eta(2k-1) = \mathbb{P}(B_{2k-1} \le k-2) + \pi_{k-1},$$

so $r_\eta(2k) - r_\eta(2k-1) = -\frac{\eta}{2}\pi_{k-1} + \frac{1-\eta}{2}\pi_k = 0$, because $\pi_k/\pi_{k-1} = \eta/(1-\eta)$ (the binomial coefficients $\binom{2k-1}{k}$ and $\binom{2k-1}{k-1}$ are equal). *Odd sizes help:* $r_\eta(2k+1) \le r_\eta(2k-1)$. Adding two votes to $2k-1$, the majority of $2k+1$ is wrong if $B_{2k-1} \le k-2$, or if $B_{2k-1} = k-1$ and the two new votes are not both 1, or if $B_{2k-1} = k$ and both are 0:

$$r_\eta(2k+1) - r_\eta(2k-1) = -\eta^2\pi_{k-1} + (1-\eta)^2\pi_k = \pi_{k-1}\big(-\eta^2 + \eta(1-\eta)\big) = \eta(1-2\eta)\pi_{k-1} \le 0.$$

Finally $r_\eta(0) = \frac12 \ge 1-\eta = r_\eta(1)$. Together these give the monotonicity on all of $\{0,1,2,\dots\}$. □

Lemma 5.55 (Randomised monotone functions of NA variables)**.** *Let $N_1,\dots,N_k$ be NA and, conditionally on $(N_j)_j$, let $E_1,\dots,E_k$ be independent $\{0,1\}$-valued variables with $\mathbb{P}(E_j = 1 \mid N_1,\dots,N_k) = r_j(N_j)$, where each r_j is non-increasing (or each non-decreasing). Then $E_1,\dots,E_k$ are NA.*

Proof. Let I, J be disjoint and F, G bounded non-decreasing. By conditional independence, $\mathbb{E}[F(E_I)G(E_J) \mid N] = \bar F(N_I)\bar G(N_J)$, where $\bar F(N_I) := \mathbb{E}[F(E_I) \mid N]$ is a non-decreasing function of the parameters $(r_i(N_i))_{i\in I}$ (increasing the parameter of an independent Bernoulli variable increases it stochastically), hence a non-increasing function of N_I when the r_i are non-increasing; likewise for $\bar G$. Negative association of N gives $\mathbb{E}[\bar F(N_I)\bar G(N_J)] \le \mathbb{E}[\bar F(N_I)]\mathbb{E}[\bar G(N_J)] = \mathbb{E}[F(E_I)]\mathbb{E}[G(E_J)]$. □

Theorem 5.56 (Concentration of the error of the histogram rule)**.** *For every law $\mathcal{D}$, every countable partition and every $n \ge 1$, the excess risk (5.41) satisfies $\mathbb{V}\mathrm{ar}(W_n) \le 2/(en)$ and, for every $t > 0$,*

$$\mathbb{P}\big(W_n \le \mathbb{E}W_n - t\big) \le \exp\Big(-\frac{e\,n\,t^2}{4}\Big), \qquad \mathbb{P}\big(W_n \ge \mathbb{E}W_n + t\big) \le \exp\Big(-\frac{n\,t^2}{21}\Big).$$

Moreover $\mathbb{E}W_n \le (en)^{-1/2}\sum_j \sqrt{p_j}$, which is at most $\sqrt{k/(en)}$ for a partition into k cells.

Proof. Step 1: structure. Let $N_j := n_j^0 + n_j^1$ be the number of observations in cell j. The pairs (cell of X_i, Y_i) are i.i.d., so conditionally on the cell indices the labels are independent, $Y_i \sim \mathrm{Ber}(\eta_j)$ if $X_i \in C_j$; together with the independence of the coins, this shows that conditionally on $(N_j)_j$ the indicators E_j are independent with $\mathbb{P}(E_j = 1 \mid N) = r_{\eta_j}(N_j)$, r as in Lemma 5.54. The counts $(N_j)_j$ are the occupancy counts of the cells, NA by Theorem 5.21, and r_{η_j} is non-increasing; by Lemma 5.55, the $(E_j)_j$ are NA (for every finite set of cells). Set $w_j := p_j u_j$ and $\bar\rho_j := \mathbb{P}(E_j = 1) = \mathbb{E}[r_{\eta_j}(N_j)]$. Since $N_j \sim \mathrm{Bin}(n, p_j)$, Lemma 5.54 and the binomial generating function give

$$\bar\rho_j \le \mathbb{E}\big[(1-\rho_j)^{N_j}\big] = (1 - p_j\rho_j)^n \le e^{-np_j\rho_j} \le e^{-np_j u_j^2/2}, \tag{5.42}$$

where the last step is $u^2 = 1 - 4\eta(1-\eta) \le 2 - 4\sqrt{\eta(1-\eta)} = 2\rho$, i.e. $(2\sqrt{\eta(1-\eta)} - 1)^2 \ge 0$.

Step 2: the mean and the variance. With $s = p_j u_j^2$, $\mathbb{E}W_n = \sum_j w_j\bar\rho_j \le \sum_j \sqrt{p_j}\sqrt{s}\, e^{-ns/2} \le \sum_j \sqrt{p_j}\sup_{s\ge 0}\sqrt{s}e^{-ns/2} = (en)^{-1/2}\sum_j\sqrt{p_j}$, and $\sum_j\sqrt{p_j} \le \sqrt{k}$ by the Cauchy–Schwarz inequality when there are k cells. By negative association the covariances are non-positive, so

$$\mathbb{V}\mathrm{ar}(W_n) \le \sum_j w_j^2\bar\rho_j \le \sum_j p_j^2 u_j^2 e^{-np_ju_j^2/2} = \sum_j p_j\,\big(p_ju_j^2\big)e^{-np_ju_j^2/2} \le \sum_j p_j\cdot\frac{2}{en} = \frac{2}{en},$$

by (5.9) applied to $u = np_ju_j^2/2$.

Step 3: the lower tail. As in Lemma 5.43, negative association gives $\mathbb{E}e^{\lambda W_n} \le \mathbb{E}e^{\lambda\widetilde W_n}$ for all λ, where $\widetilde W_n = \sum_j w_j\widetilde E_j$ with independent $\widetilde E_j \sim \mathrm{Ber}(\bar\rho_j)$. The proof of Theorem 5.44, Steps 1–3, applies to $\widetilde W_n$ with w_j in place of $p(x)$ and $\bar\rho_j$ in place of $q(x)$: for $\ell \le 0$ the tilted variance is at most $\sum_j w_j^2\bar\rho_j \le 2/(en)$, whence the lower tail $\exp(-ent^2/4)$.

Step 4: the upper tail. Fix $\theta > 0$ and split the cells according to whether $u_j \le \theta$ or $u_j > \theta$: $W_n = W_n^{\le} + W_n^{>}$. The first part is deterministically small: $0 \le W_n^{\le} \le \sum_j p_ju_j\mathbb{1}\{u_j \le \theta\} \le \theta$, so $W_n^{\le} - \mathbb{E}W_n^{\le} \le \theta$, and therefore

$$\mathbb{P}\big(W_n \ge \mathbb{E}W_n + t + \theta\big) \le \mathbb{P}\big(W_n^{>} \ge \mathbb{E}W_n^{>} + t\big). \tag{5.43}$$

For the second part, the cells with $u_j > \theta$ form a sub-family, still NA, with surrogate $\widetilde W_n^{>}$, and for $0 \le \ell \le \frac{2}{5}n\theta$ the tilted parameters satisfy, by (5.42) and $u_j > \theta$,

$$\bar\rho_{j,\ell} \le \bar\rho_j e^{\ell w_j} \le \exp\Big(-np_ju_j^2/2 + \tfrac{2}{5}n\theta p_ju_j\Big) \le \exp\Big(-np_ju_j^2/2 + \tfrac{2}{5}np_ju_j^2\Big) = e^{-np_ju_j^2/10},$$

so that the tilted variance is at most $\sum_j p_j^2u_j^2e^{-np_ju_j^2/10} \le \sum_j p_j\cdot\frac{10}{en} = \frac{10}{en}$ (again (5.9)), and $\psi_{\widetilde W_n^{>}}(\lambda) \le 5\lambda^2/(en)$ for $0 \le \lambda \le \frac{2}{5}n\theta$. Chernoff's bound with $\lambda = ent/10$, which is admissible if $\theta \ge et/4$, gives $\mathbb{P}(W_n^{>} \ge \mathbb{E}W_n^{>} + t) \le \exp(-ent^2/10 + ent^2/20) = \exp(-ent^2/20)$. Choosing $\theta = et/4$ in (5.43) and setting $s = t(1 + e/4)$,

$$\mathbb{P}\big(W_n \ge \mathbb{E}W_n + s\big) \le \exp\Big(-\frac{e\,n\,s^2}{20(1+e/4)^2}\Big) \le \exp\Big(-\frac{ns^2}{21}\Big),$$

since $e/(20(1+e/4)^2) \approx 0.048 > 1/21$. □

Remark 5.57 (Comments). *(a)* The theorem is distribution-free and partition-free: the fluctuations of the error of a histogram rule are of order $n^{-1/2}$ whatever the partition, even

one with infinitely many cells, and the bounded-difference inequality could not have shown it (changing one observation can change the vote of a cell of mass p_j, so the constants c_i are of order $\max_j p_j$, exactly as in Remark 5.33). The expected excess risk, on the other hand, depends on the partition through $\sum_j \sqrt{p_j} \le \sqrt{k}$: this is the usual estimation error of a k-cell histogram, and the balance between it and the approximation error $L_{\mathcal{D}}(\bar{h}) - L^*_{\mathcal{D}}$ – which decreases when the cells shrink – is the bias–variance trade-off of histogram rules [DGL96, Chap. 9]. *(b)* The result and the method are those of McAllester and Ortiz [MO03, Thm. 17], who obtain the exponents $nt^2/7$ and $nt^2/9$ in a more general setting (an arbitrary bounded loss and more than two labels) by combining negative association with the Kearns–Saul bound of Remark 5.47; Lemma 5.55 is their Lemma 8. The constant $1/21$ of the upper tail reflects the crude splitting of Step 4 and no attempt to optimise it; the lower tail, where the tilted variance is controlled by the untilted one, comes out with the natural constant. *(c)* The cells with $u_j \le \theta$ are those in which the two labels are nearly equally likely. Their votes are nearly coin tosses and never become reliable, however many observations they receive; but they cost little, since a wrong vote there is nearly as good as a right one. This is why they can be set aside in Step 4, and it is the one place where the structure of the problem differs from the missing mass, in which every symbol becomes reliably seen once $np(x) \gg 1$.

5.8 Notes and references

The missing mass. The estimator U_n/n was devised by Turing at Bletchley Park during the Second World War, in the cryptanalysis of the Enigma traffic (the problem being to estimate how often the Kriegsmarine would use a key setting never seen before), and published after the war by Good [Goo53], who attributes the idea and "an intuitive demonstration" to Turing; the paper treats the general problem of estimating the total probability of the species seen r times, with the formula $(r+1)n_{r+1}/N$ (Exercise 5.5), and its sequel [GT56] predicts the number of new species in a larger sample. The exact identity (5.4) and its r-count version are in [MS00]; Robbins [Rob68] noted the mean-square accuracy of the one-step-ahead version. Good–Turing smoothing became a standard tool of natural language processing in the 1980s [GS95], where the unseen bigrams of a corpus are the missing mass. The concentration of $\widehat{M}_n$ by the bounded-difference inequality and the first one-sided bound on M_n are in [MS00]; both tails of M_n, by negative association and what they call the Gibbs-variance lemma – the variance representation (2.4) of Chapter 2 – are in [MO03], which is the source of Section 5.7.2 and Section 5.7.6. The sharper constants of Remark 5.47 are from [BK13]; the Kearns–Saul inequality is from [KS98], with a complete proof in [BK13]. Distribution-dependent and multiplicative concentration, under regular variation, are developed in [OD12] and [BBO17]; the latter also contains the Poissonised analysis of the Good–Turing estimator of Section 5.7.5 and a wealth of related results on occupancy counts. The minimax results (5.40) are from [RTS17] and [Ach+18]. Poissonisation as a tool for discrete distribution estimation is used systematically in [Ach+13]; the depoissonisation inequalities are [MU05, Sec. 5.4]. The missing mass also arises in sequential problems: when K sources of items are available and the goal is to discover as many distinct items as possible, the Good–Turing estimate of the missing mass of each source is the natural index on which to base the choice of the next source [BEG13].

The variance. The Efron–Stein inequality [ES81] was proved for symmetric functions, in connection with the jackknife estimate of variance of Quenouille and Tukey [Que56; Tuk58]; Steele [Ste86] gave the general form, and the presentation through $\mathbb{V}\mathrm{ar}^{(i)}$ and the martingale decomposition follows [BLM13, Chap. 3]. Self-bounding functions were

introduced, without a name, in [BLM00], where configuration functions (Example 5.13) and the random VC entropy (Exercise 5.10) are the motivating examples; the terminology and the (a, b)-generalisation are from [BLM03; BLM09].

Negative association. The notion was introduced by Alam and Saxena [AS81]; Joag-Dev and Proschan [JP83] proved the closure properties, the negative association of permutation distributions (Theorem 5.22, their Theorem 2.11, by a different argument) and of multinomial vectors, and the general theorem on independent log-concave variables conditioned on their sum (Remark 5.25). The bivariate notions – quadrant and regression dependence – are Lehmann's [Leh66]. The balls-and-bins treatment of Section 5.4.3, with the zero–one lemma and the remark that Chernoff's method goes through unchanged, is [DR98]; see also [DP09, Chap. 3] and Wajc's notes [Waj17], which collect the properties used here. Hoeffding's comparison theorem is [Hoe63, Thm. 4]; the proof by completion of the set of distinct indices given in Section 5.4.5 is a rewriting of his. Refinements capturing the finite-population correction are in [Ser74] and, with a martingale approach and matching lower bounds, in [BM15].

Bounded differences and the entropy method. The bounded-difference inequality is [McD89]; the second proof of Section 5.5 is folklore and is, for instance, the proof given in [BLM13, Thm. 6.2] up to the replacement of Hoeffding's lemma by the entropy method. The entropy method originates in Ledoux's proof of Talagrand's concentration inequalities for product measures [Led97; Tal95], through logarithmic Sobolev inequalities and Herbst's argument; its development into a general tool is due to Boucheron, Lugosi and Massart [BLM00; BLM03] and Massart [Mas07]. Theorem 5.36 is the subadditivity of entropy, proved here by the duality formula as in [BLM13, Thm. 4.10]; Lemma 5.37 is the modified logarithmic Sobolev inequality of [BLM13, Thm. 6.6], and Theorem 5.39 is [BLM13, Thm. 6.12], originally [BLM00, Thm. 1] (the lower tail being new there). What we have not covered is the heart of the method: the "exponential Efron–Stein" inequalities [BLM03], which bound the cgf of Z through the random variables $\sum_i (Z - Z^{(i)})_+^2$, the Gaussian and Poisson logarithmic Sobolev inequalities, and the convex distance inequality of Talagrand [Tal95]. Chapters 5–8 of [BLM13] and [Led01] are the references.

Histogram rules. The consistency and the rates of histogram classification rules are the subject of [DGL96, Chap. 9]; the concentration of their error around its mean, in the form of Theorem 5.56, is [MO03, Thm. 17]. The monotonicity of the error of majority votes in the number of voters (Lemma 5.54) is the content of Condorcet's jury theorem.

5.9 Exercises

Exercise 5.1 (⋆ Zipf's law)**.** Let $\mathcal{X} = \{1, 2, \dots\}$ and $p(x) = c_\alpha x^{-\alpha}$ with $\alpha > 1$ and $c_\alpha = 1/\sum_{x \geq 1} x^{-\alpha}$.

(a) Show that if $f : [1, \infty) \to [0, \infty)$ is non-decreasing then non-increasing (unimodal), then $\left| \sum_{x \geq 1} f(x) - \int_1^\infty f(u)\, \mathrm{d}u \right| \leq 2 \sup f$.

(b) Show that $\mathbb{E} M_n \sim c_\alpha^{1/\alpha} I_\alpha \, n^{-(\alpha-1)/\alpha}$ with $I_\alpha = \int_0^\infty y^{-\alpha} e^{-y^{-\alpha}}\, \mathrm{d}y < \infty$, and that $v_n = \sum_x p(x)^2 (1 - p(x))^n \asymp n^{-2+1/\alpha}$.

(c) Show that $\mathbb{E} D_n \asymp n^{1/\alpha}$ and $\mathbb{E} U_n \asymp n^{1/\alpha}$.

Exercise 5.2 (⋆ The uniform law)**.** Let P be uniform on k symbols and $q = (1 - 1/k)^n$.

(a) Show that $\mathbb{E} M_n = q$ and $\mathbb{V}\mathrm{ar}(M_n) = \frac{1}{k} q (1 - q) + \left(1 - \frac{1}{k}\right) \left[(1 - \frac{2}{k})^n - q^2 \right]$.

(b) Show that for $k = n$, $n\,\mathbb{V}\mathrm{ar}(M_n) \to e^{-1} - 2e^{-2} \approx 0.097$.

Exercise 5.3 (★ The hapax count). (a) Show that the bounded-difference constant 2 of the hapax count U_n (Example 5.14) cannot be replaced by a smaller one.

(b) Show that for the uniform law on n symbols, $\mathbb{V}\mathrm{ar}(\widehat{M}_n) \sim (e^{-1} - e^{-2})/n \approx 0.23/n$: the proxy $1/n$ of (5.16) is of the right order.

Exercise 5.4 (★ Distinct symbols). (a) Show that $\mathbb{E}D_n = \sum_x (1 - (1 - p(x))^n)$.

(b) For the uniform law on k symbols, compute $\mathbb{V}\mathrm{ar}(D_n)$ and show that $\mathbb{V}\mathrm{ar}(D_n) \sim (e^{-1} - 2e^{-2})n$ when $k = n$. Relate this to Exercise 5.2.

(c) Compare, when $k = n$ and when $k \ll n$, the three variance bounds available for D_n: $\mathbb{E}D_n$ (Corollary 5.11), $n/4$ (Corollary 5.9) and the truth.

Exercise 5.5 (Good's formula for the r-counts). For $r \geq 0$ let $M_{n,r} := \sum_x p(x)\mathbb{1}\{N_n(x) = r\}$ be the total mass of the symbols seen exactly r times ($M_{n,0} = M_n$) and $U_{n,r} := \sum_x \mathbb{1}\{N_n(x) = r\}$ their number ($U_{n,1} = U_n$). Show that for $0 \leq r < n$,

$$\mathbb{E}M_{n,r} = \frac{r+1}{n-r}\,\mathbb{E}\big[U_{n,r+1} - M_{n,r+1}\big], \qquad \text{hence} \qquad \left|\mathbb{E}M_{n,r} - \frac{r+1}{n-r}\mathbb{E}U_{n,r+1}\right| \leq \frac{r+1}{n-r}.$$

Interpret the case $r = 0$ and Good's estimate $(r+1)U_{n,r+1}/n$ of $M_{n,r}$.

Exercise 5.6 (★ Finite population correction). In the setting of Corollary 5.29, show that $\mathbb{C}\mathrm{ov}(X_i, X_j) = -\sigma^2/(N-1)$ for $i \neq j$ and deduce $\mathbb{V}\mathrm{ar}(\sum_{i \leq n} X_i) = n\sigma^2 \frac{N-n}{N-1}$. *Hint: the sum of all N draws is constant.*

Exercise 5.7 (★ Empty bins). In Example 5.27, show that $\mathbb{V}\mathrm{ar}(E) \leq \mathbb{E}E$, prove the Bernstein bound stated there, and compare it with Hoeffding's inequality when $n = k \log k$.

Exercise 5.8 (Negative correlation is weaker than negative association). (a) Let (X_1, X_2) be a pair of random variables with $\mathbb{P}(X_1 = X_2 = 0) = 0.3$, $\mathbb{P}(X_1 = X_2 = 1) = 0.4$, and $\mathbb{P}(X_1 = 0, X_2 = 2) = \mathbb{P}(X_1 = 2, X_2 = 0) = 0.15$. Show that $\mathbb{C}\mathrm{ov}(X_1, X_2) < 0$ but that (X_1, X_2) is not NA.

(b) Show that a pair (X_1, X_2) is NA if and only if it is negatively quadrant dependent: $\mathbb{P}(X_1 \leq s, X_2 \leq t) \leq \mathbb{P}(X_1 \leq s)\mathbb{P}(X_2 \leq t)$ for all s, t. *Hint: for bounded U, V and an independent copy (U', V'), write $2\,\mathbb{C}\mathrm{ov}(U, V) = \mathbb{E}[(U - U')(V - V')]$ and expand $U - U' = \int (\mathbb{1}\{a < U\} - \mathbb{1}\{a < U'\})\,\mathrm{d}a$.*

(c) Show that the hapax indicators $(\mathbb{1}\{N_n(x) = 1\})_{x \in \mathcal{X}}$ are not NA in general.

Exercise 5.9 (★ Bounded differences by the entropy method). Let f have bounded differences with constants c_i, the $\mathcal{X}_i$ being countable (or, more generally, the infima and suprema below being measurable functions of $x^{(i)}$). Using Lemma 5.37 with $f_i(x^{(i)}) = \inf_y f(x_1, \ldots, y, \ldots, x_n)$ for $\lambda \geq 0$ and with the supremum for $\lambda < 0$, and then Lemma 5.38, show that $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq 2\exp\big(-t^2/(2\sum_i c_i^2)\big)$. Where is the factor 4 lost with respect to Theorem 5.30?

Exercise 5.10 (★ The random VC entropy is self-bounding). For a random variable Y with finitely many values, let $H(Y) = -\sum_y \mathbb{P}(Y = y)\log \mathbb{P}(Y = y)$ be its Shannon entropy, and $H(Y \mid Y') = \sum_{y'} \mathbb{P}(Y' = y')H(Y \mid Y' = y')$ the conditional entropy.

(a) Prove the chain rule $H(Y, Y') = H(Y') + H(Y \mid Y')$, that $H(Y \mid Y') \le H(Y)$, and that $H(Y) \le \log(\text{number of values})$.

(b) (Han's inequality.) For $Y = (Y_1, \dots, Y_n)$ with finitely many values, show that $\sum_{i=1}^{n} \big[H(Y) - H(Y^{(i)})\big] \le H(Y)$.

(c) Deduce that for a finite set $V \subseteq \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with projections $V^{(i)}$ (forget the i-th coordinate), $\sum_i \log\big(|V| \,/\big|V^{(i)}\big|\big) \le \log|V|$.

(d) Let $\mathcal{H}$ be a class of binary classifiers on $\mathcal{X}$ and, with the notation of Definition 4.13, let $Z = \log_2 |\mathcal{H}(X_1, \dots, X_n)|$ be the logarithm of the number of labellings of the sample by $\mathcal{H}$. Show that Z is a self-bounding function of $X_1, \dots, X_n$, and state what Corollary 5.11 and Theorem 5.39 give for it.

Exercise 5.11 (Poissonisation: what is lost)**.** (a) Compute $\mathbb{V}\mathrm{ar}(M_T)$ exactly.

(b) For the uniform law on n symbols, show that $\mathbb{E}M_T - \mathbb{E}M_n \sim e^{-1}/(2n)$: the bound $4/(e^2 n)$ of Proposition 5.52 is of the right order.

(c) Deduce from Propositions 5.51 and 5.52 a lower-tail bound for M_n, and compare it with (5.32).

Exercise 5.12 (⋆ Ties in the histogram rule)**.** Suppose that ties (and empty cells) are resolved deterministically in favour of the label 0.

(a) For $\eta > \frac{1}{2}$, compute the error probabilities $r_\eta(1)$ and $r_\eta(2)$ of the vote and show that $r_\eta(2) > r_\eta(1)$.

(b) Show that the bound $r_\eta(m) \le (2\sqrt{\eta(1-\eta)})^m$ still holds.

(c) Which step of the proof of Theorem 5.56 fails, and which conclusions survive?

Exercise 5.13 (Hypergeometric tails)**.** A population of size N contains K ones and $N - K$ zeros, $\mu = K/N$, and $\bar{X}_n$ is the mean of n draws without replacement.

(a) Show that $\mathbb{P}(\bar{X}_n \ge \mu + t) \le e^{-n\,\mathrm{kl}(\mu+t,\mu)}$ for $0 < t \le 1 - \mu$, with kl as in Notation 2.22.

(b) Show on the case $n = N$ that this bound can be very far from the truth, and explain why no bound obtained from Theorem 5.28 can capture the effect of n/N.

Exercise 5.14 (Suprema of empirical processes)**.** Let $\mathcal{G}$ be a countable class of functions with values in $[0,1]$ and $Z = \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} g(X_i)$. Write down the bounds of Theorem 5.39 for Z and compare them with the bounded-difference inequality. When is the self-bounding bound better? Why does this not apply directly to the uniform deviation $\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_{\mathcal{S}_n}(h))$ of Chapter 4?