MICHAEL J. TURMON

ASSESSING GENERALIZATION OF FEEDFORWARD NEURAL NETWORKS

A dissertation presented to the faculty
of the graduate school of

Cornell University

in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

August 1995

Assessing Generalization of Feedforward Neural Networks

Abstract

We address the question of how many training samples are required to ensure that the performance of a neural network of given complexity on its training data matches that obtained when fresh data is applied to the network. This desirable property may be termed 'reliable generalization.' Well-known results of Vapnik give conditions on the number of training samples sufficient for reliable generalization, but these are higher by orders of magnitude than practice indicates; other results in the mathematical literature involve unknown constants and are useless for our purposes.

We seek to narrow the gap between theory and practice by transforming the problem into one of determining the distribution of the supremum of a Gaussian random field in the space of weight vectors. This is addressed first by application of a tool recently proposed by D. Aldous called the Poisson clumping heuristic, and then by related probabilistic techniques. The idea underlying all the results is that mismatches between training set error and true error occur not for an isolated network but for a group or 'clump' of similar networks. In a few ideal situations—perceptrons learning halfspaces, machines learning axis-parallel rectangles, and networks with smoothly varying outputs—the clump size can be derived and asymptotically precise sample size estimates can be found via the heuristic.

In more practical situations, when formal knowledge of the data distribution is unavailable, the size of this group of equivalent networks can be related to the original neural network problem via a function of a correlation coefficient. Networks having prediction error correlated with that of a given network are said to be within the 'correlation volume' of the latter. Means of computing the correlation volume based on estimating such correlation coefficients using the training data are proposed and discussed. Two simulation studies are performed. In the cases we have examined, informative estimates of the sample size needed for reliable generalization are produced by the new method.

Vita

Michael Turmon was born in 1964 in Kansas City, Missouri, and he grew up in that city but not that state. From the time he sawed a telephone in half as a child, it was evident that he was meant to practice engineering of some sort—the more theoretical, the better. In 1987 he received Bachelor's degrees in Computer Science and in Electrical Engineering from Washington University in St. Louis, where he also had the good fortune to take classes from William Gass and Howard Nemerov.

Taking a summer off for a long bicycle tour, Michael returned to Washington University for graduate study, where he earned his Master's degree in Electrical Engineering in 1990. During this time he was supported by a fellowship from the National Science Foundation. His thesis concerned applications of constrained maximum-likelihood spectrum estimation to narrowband direction-finding, and uses of parallel processing to compute these estimates.

Feeling a new challenge was in order, Michael got into Cornell in 1990—and, perhaps the greater feat, lured his girlfriend to Ithaca. One major achievement of his time here was marrying Rebecca in June 1993. Another was completing a nice hard program in electrical engineering with emphasis on probability and statistics. Michael finished work on his dissertation in May 1995.

It is possible, possible, possible. It must
be possible. It must be that in time
The real will from its crude compoundings come,

Seeming, at first, a beast disgorged, unlike,
Warmed by a desperate milk. To find the real,
To be stripped of every fiction except one,

The fiction of an absolute.

–Wallace Stevens

Acknowledgments

Great thanks are due Professor T.L. Fine for the guidance and commitment he has given me throughout this work and for expecting the most of me. Terry was always there with a new idea when all leads seemed to be going nowhere, or a new direction when everything seemed settled. His attitude of pursuing research because it matters to the world rather than as an abstract exercise was a great influence. Finally, the time I spent making diagrams to accompany his yellow pads of analysis proved to me the untruth of the claim that right-handed people juggle symbols while left-handed people think geometrically.

I thank David Heath and Toby Berger for their help on my committee. Further thanks to Toby for the occasional word game, mathematical puzzle, or research question.

I would also like to thank P. Subrahmanya and Jamal Mohamad Youssef for a number of enlivening excursions via the whiteboard in 393 ETC. Thanks to Jim LeBlanc for asking good questions. Sayandev Mukherjee, Jen-Lun Yuan, and Wendy Wong introduced me to the rest of the world of neural networks, and James Chow and Luan Ling Lee to a world of engineering beyond that. Thanks to Srikant Jayaraman for organizing the weekly seminar.

I am glad to have been influenced along the way by the outstanding teaching of Don Snyder, Gennady Samorodnitsky, Harry Kesten, Venkat Anantharam, and Terrence Fine. Thanks to Michael Miller for his encouragement and enthusiasm.

I thank my parents and family for their continuing love and support. Deep thanks to Rebecca for her love and her dynamism, and for her good cheer as things got busy.

Contents

Tables

Figures

1 Introduction

IN THE PAPER by Le Cun et al. [22] we read of a nonlinear classifier, a neural network, used to recognize handwritten decimal digits. The inputs to the classifier are gray-scale images of 16×16 pixels, and the output is one of 10 codes representing the digits. The exact construction of the classifier is not of interest right now; what does matter is that its functional form is fixed at the outset of the process so that selection of a classifier means selecting values for $d = 9760$ real numbers, called the weight vector. No probabilistic model is assumed known for the digits. Instead, $n = 7291$ input/output pairs are used to find a weight vector approximately minimizing the squared error between the desired outputs and the classifier outputs on the known data.

In summary: based on 7291 samples, the 9760 parameters of a nonlinear model are to be estimated. It is not too surprising that a function can be selected from this huge family that agrees well with the training data. The surprise is rather that the mean squared error computed on a separate test set of handwritten characters agrees reasonably well with the error on the training set (.0180 and .0025 respectively for MSE normalized to lie between 0 and 1). The classifier has *generalized* from the training data.

This state of affairs is rather common for neural networks across a wide variety of application areas. In table 1.1 are several recent applications of neural networks, listed with the corresponding number of free parameters in the model and the number of input/output pairs used to select the model. One has an intuitive idea that for a given problem, good performance on the training set should imply good performance on the test set as long as the ratio n/d is large enough; general experience would indicate that this ratio should surely be greater than unity, but just how large is unclear. From the table, we see that the number of data points per parameter varies over more than three orders of magnitude.

One reason such a large range is seen in this table is that statistics has had little advice for practitioners about this problem. The most useful line of argument was initiated by Vapnik [53] which computes upper bounds on a satisfactory n/d: if the number of data per parameter is this high, accuracy of a given degree between test and training error is guaranteed with high probability; we say the architecture *reliably generalizes.* While Vapnik's result confirms intuition in the broad sense, the upper bounds have proven to be higher by orders of magnitude than practice indicates. We seek to narrow the chasm between statistical theory and neural network practice by finding reasonable estimates of the sample size at which the architecture reliably generalizes.

Table 1.1: Some recent neural network applications.

n	d	n/d	Application	Source
7291	2578	2.83	Digit Recognition	LeCun et al. [23]
4104	3040	1.35	Vowel Classification	Atlas et al. [21]
3190	2980	1.07	Gene Identification	Noordewier et al. [42]
2025*	2100	0.96	Medical Imaging	Nekovi [41]
7291	9760	0.75	Digit Recognition	LeCun et al. [22]
105	156	0.67	Commodity Trading	Collard [13]
150	360	0.42	Robot Control	Gullapalli [28]
200	1540	0.13	Image Classification	Delopoulos [15]
1200	36 600	1/30	Vehicle Control	Pomerleau [45]
3171	376 000	1/120	Protein Structure	Fredholm et al. [20]
20	8200	1/410	Signature Checking	Mighell et al. [40]
160	165 000	1/1000	Face Recognition	Cottrell, Metcalfe [14]

Shown are the number of samples used to train the network (n), the number of distinct weights (d), and the number of samples per weight. The starred entry is a conservative estimate of an equivalent number of independent samples; the training data in this application was highly correlated.

§1.1 Terms of the Problem

We formalize the problem in these terms:

- The inputs $x \in R^p$ and outputs $y \in R$ have joint probability distribution P which is unknown to the observer who only has the training set $\mathcal{T} := \{(x_i, y_i)\}_{i=1}^n$ of pairs drawn i.i.d. from P.
- Models are neural networks $\eta(x; w)$ where x is the input and $w \in \mathcal{W} \subseteq R^d$ parameterizes the network. The class of allowable nets is $\mathcal{N} = \{\eta(\cdot; w)\}_{w \in \mathcal{W}}$.
- The performance of a model may be measured by any loss function. We will consider

$$\nu_{\mathcal{T}}(w) := \frac{1}{n} \sum_{i=1}^{n} \big(\eta(x_i; w) - y_i\big)^2 \tag{1.1}$$

$$\mathcal{E}(w) := E\big(\eta(x; w) - y\big)^2 \quad ; \tag{1.2}$$

  the former is the *empirical error* and is accessible while the latter depends on the unknown P and is not. In the classification setting, inputs and outputs are binary, $\mathcal{E}(w)$ is error probability, and $\nu_{\mathcal{T}}(w)$ is error frequency.

- Two models are of special importance, they are

$$w^* := \arg\min_{w \in \mathcal{W}} \nu_{\mathcal{T}}(w) \tag{1.3}$$

$$w^0 := \arg\min_{w \in \mathcal{W}} \mathcal{E}(w) \tag{1.4}$$

  where either may not be unique. The goal of the training algorithm is to find w^*.

§1.2 A Criterion For Generalization

Since P is unknown, $\mathcal{E}(w^*)$ cannot be found directly. One measure is just $\nu_{\mathcal{T}}(w^*)$, but this is a biased estimate of $\mathcal{E}(w^*)$ because of the way w^* is selected. We treat this problem by finding an n such that

$$\sup_{w \in \mathcal{W}} |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \leq \epsilon \quad \text{with probability } \tau \tag{1.5}$$

for τ near one. The seeming overkill of including all weights in the supremum makes sense when one realizes that to limit the group of weights to be considered, one must take into account the algorithm used to find w^*. In particular its global properties seem needed because the issue of what the error surface looks like around the limit points of the algorithm must be dealt with. However, little is known about the global properties of any of the error-minimization algorithms currently in use—several variations of gradient descent, and conjugate gradient and Newton-Raphson methods.

Let us examine three implications of (1.5).

- Even for a training algorithm that does not minimize $\nu_{\mathcal{T}}$,

  $$|\nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^*)| \leq \epsilon \quad \text{w.p.}\tau \tag{1.6a}$$

  so that the ultimate performance of the selected model can be verified simply by its behavior on the training set. It is hard to overstate the importance of (1.6a) in the typical situation where the selected neural network has no interpretation based on a qualitative understanding of the data, i.e. the neural network is used as a black box. In the absence of a rationale for why the network models the data, statistical assurance that it does so becomes very important.

- Provided $\nu_{\mathcal{T}}(w) \leq \nu_{\mathcal{T}}(w^0)$,

  $$\mathcal{E}(w) - \mathcal{E}(w^0) \leq 2\epsilon \quad \text{w.p.}\tau$$

  and in particular,

  $$0 \leq \mathcal{E}(w^*) - \mathcal{E}(w^0) \leq 2\epsilon \quad \text{w.p.}\tau. \tag{1.6b}$$

  This follows from

  $$\begin{aligned} \mathcal{E}(w^*) - \mathcal{E}(w^0) &= \mathcal{E}(w^*) - \nu_{\mathcal{T}}(w^*) \ + \ \nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^0) \\ &\leq \mathcal{E}(w^*) - \nu_{\mathcal{T}}(w^*) \ + \ \nu_{\mathcal{T}}(w^0) - \mathcal{E}(w^0) \\ &\leq 2\epsilon \end{aligned}$$

  If this much confidence in the training algorithm is available, then w^* is close to w^0 in true squared error.

- Similarly, if $\nu_{\mathcal{T}}(w) \leq \nu_{\mathcal{T}}(w^0)$ then

  $$|\mathcal{E}(w^0) - \nu_{\mathcal{T}}(w)| \leq \epsilon \quad \text{w.p.}\tau,$$

  and in particular

  $$|\mathcal{E}(w^0) - \nu_{\mathcal{T}}(w^*)| \leq \epsilon \quad \text{w.p.}\tau. \tag{1.6c}$$

This is because

$$\mathcal{E}(w^0) \le \mathcal{E}(w^*) \le \nu_{\mathcal{T}}(w^*) + \epsilon$$
$$\mathcal{E}(w^0) \ge \nu_{\mathcal{T}}(w^0) - \epsilon \ge \nu_{\mathcal{T}}(w^*) - \epsilon \quad .$$

This gives information about how effective the family of nets is: if $\nu_{\mathcal{T}}(w^*)$ is much larger than the tolerance ϵ, no network in the architecture is performing well.

We contrast determining the generalization ability of an architecture by ensuring (1.5) with two other approaches. The simpler method uses a fraction, typically half, of the available input/output pairs to form say $\nu_{\mathcal{T}}^{(1)}(w)$ and select w^*. The remainder of the data is used to find an independent replica $\nu_{\mathcal{T}}^{(2)}(w)$ of $\nu_{\mathcal{T}}^{(1)}(w)$ by which estimates of the type (1.6a) are obtained. The powerful argument against this approach is its use of only half the available data to select w^*.

The cross-validation method (see [19]) avoids wasting data by holding out just one piece of data and training the network on the remainder. This leave-one-out procedure is repeated n times while noting the prediction error on the excluded point. The cross-validation estimate of generalization error is the average of the excluded-point errors. The advantages of this method lie in its simplicity and frugality, while drawbacks are that it is computationally intensive and difficult to analyze, so very little is known about the quality of the error estimates. More telling to us, such single-point analyses can never give information of a global nature such as (1.6b) and (1.6c) above. Using only cross-validation forces one into a point-by-point examination of the weight space when far more informative results are available.

§1.3 A Modified Criterion

We shall see that it may be preferable to establish conditions under which

$$\sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sigma(w)} \le \epsilon \quad \text{with probability } \tau \text{ near 1} \tag{1.7}$$

where $\sigma^2(w) := \mathrm{Var}(\nu_{\mathcal{T}}(w)) = \mathrm{Var}((y - \eta(x;w))^2)$. Normalization is useful because the weight largest in variance generally dominates the exceedance probability, and typically such networks are poor models. In binary classification for example, $\sigma^2(w) = \mathcal{E}(w)(1-\mathcal{E}(w))$ is maximized at $\mathcal{E}(w) = 1/2$.

Continuing in this classification context, we explore the implications of (1.7). These are regulated by $\sigma(w^*)$ and $\sigma(w^0)$. If we make the reasonable assumption that $\mathcal{E}(w^*) \le 1/2$, then by minimality of w^0, $\sigma(w^*) \ge \sigma(w^0)$. (Alternatively, if the architecture is closed under complementation then minimality of w^0 implies not only $\mathcal{E}(w^*) \ge \mathcal{E}(w^0)$, but also $\mathcal{E}(w^*) \le 1 - \mathcal{E}(w^0)$, so again $\sigma(w^*) \ge \sigma(w^0)$.) Knowing this allows the manipulations in §1.2 to be repeated, yielding

$$|\nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^*)| \le \epsilon\sqrt{\mathcal{E}(w^*)(1-\mathcal{E}(w^*))} \tag{1.8a}$$
$$0 \le \mathcal{E}(w^*) - \mathcal{E}(w^0) \le 2\epsilon\sqrt{\mathcal{E}(w^*)(1-\mathcal{E}(w^*))} \tag{1.8b}$$
$$|\mathcal{E}(w^0) - \nu_{\mathcal{T}}(w^*)| \le \epsilon\sqrt{\mathcal{E}(w^*)(1-\mathcal{E}(w^*))} \tag{1.8c}$$

which hold simultaneously with probability τ. To understand the essence of the new assertions, note that if $\nu_{\mathcal{T}}(w^*) = 0$, then the first condition says that $\mathcal{E}(w^*) \leq \epsilon^2/(1+\epsilon^2) \simeq \epsilon^2$. Now this allows us to conclude the second two errors are also of order ϵ^2 since $\sqrt{\mathcal{E}(w^*)(1-\mathcal{E}(w^*))} \leq \epsilon/(1+\epsilon^2)$. All three conclusions are tightened considerably.

In the general case, the following hold with probability τ:

$$|\nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^*)| \leq \epsilon\sigma(w^*) \tag{1.9a}$$

$$0 \leq \mathcal{E}(w^*) - \mathcal{E}(w^0) \leq \epsilon\big(\sigma(w^*) + \sigma(w^0)\big) \tag{1.9b}$$

$$|\mathcal{E}(w^0) - \nu_{\mathcal{T}}(w^*)| \leq \epsilon\big(\sigma(w^*) \vee \sigma(w^0)\big) \quad . \tag{1.9c}$$

We would expect $\sigma(w^0) \leq \sigma(w^*)$ which can be used to simplify the above expressions to depend only on $\sigma(w^*)$. Then $\sigma(w^*)$ can be estimated from the data as

$$\hat{\sigma}^2(w) = \sum_{i=1}^{n} \big((y_i - \eta(x_i; w))^2 - \nu_{\mathcal{T}}(w)\big)^2 \quad .$$

In any case, we would expect $\sigma(w^*)$ to be significantly smaller than the maximum variance, so that the assertions above are again stronger than the corresponding unnormalized ones.

§1.4 Related Areas of Research

Before considering the problem in greater detail, let us mention that tightly related work is going on under two other names. In probability and statistics, the random entity $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$ is known as an empirical process, and the supremum of this process is a generalized Kolmogorov-Smirnov statistic. We will return to this viewpoint later on. See the development of Pollard [43] or the survey of Gaenssler and Stute [26].

In theoretical computer science, the field of computational learning theory is concerned, as above, with selecting a model ('learning a concept') from a sequence of observed data or queries. Within this field, the idea of PAC (probably approximately correct) learning is very closely related to our formulation of the neural network problem. Computer scientists are also interested in algorithms for finding a near-optimal model in polynomial time, an issue we do not address. For an introduction see Kearns and Vazirani [33], Anthony and Biggs [9], or the original paper of Valiant [52].

§1.5 Contributions

After reviewing prior work on the problem of generalization in neural networks in chapter 2, we introduce a new tool from probability theory called the Poisson clumping heuristic in chapter 3. The idea is that mismatches between empirical error and true error occur not for an isolated network but for a 'clump' of similar networks, and computations of exceedance probability come down to obtaining the expected size of this clump. In chapter 4 we demonstrate the validity and appeal of the Poisson clumping technique by examining several examples of networks for which the mean clump size can be computed analytically.

An important feature of the new sample size estimates is that they depend on simple properties of the architecture and the data: this has the advantage of being tailored to a given problem but the potential disadvantage of our having to compute them. Since in general analytic information about the network is unavailable, in chapter 5 we develop ways to estimate the mean clump size using the training data. Some simulation studies in chapter 6 show the usefulness of the new sample size estimates.

The high points here are chapters 4, 5, and 6. The contributions of this research are:

- Introduction of the Poisson clumping view, which provides a means of visualizing the error process which is also amenable to analysis and empirical techniques.
- In §4.2 and §4.3 we give precise estimates of the sample size needed for reliable generalization for the problems of learning orthants and axis-oriented rectangles. In §4.4 we give similar estimates for the problem of learning for linear threshold units.
- In §4.5 we consider neural nets having twice differentiable activation functions, so that the error $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$ is smooth, yielding a local approximation which allows determination of the mean clump size. Again estimates of the sample size needed for reliable generalization are given.
- In §6.3, after having developed some more tools, we find estimates of the clump size under the relative distance criterion (1.7), which allows tight sample size estimates to be obtained for the problem of learning rectangles.
- Finally in chapters 5 and 6 a method for empirically finding the *correlation volume*, which is an estimate of the size of a group of equivalent networks, is outlined. In chapter 6 the method is tested for some sample architectures.

§1.6 Notation

With some exceptions, including the real numbers R, sets are denoted by script letters. The × symbol is Cartesian product. The indicator of a set $\mathcal{A}$ is $1_{\mathcal{A}}$. We use & and $\|$ for logical and and or, while $\wedge$ and $\vee$ denote the minimum and maximum. Equals by definition is $:=$ and $\overset{\mathcal{D}}{=}$ stands for equality in distribution. Generally $|\cdot|$ is absolute value and $\|\cdot\|_{\mathcal{W}}$ is the supremum of the given function over $\mathcal{W}$.

Context differentiates vectors from scalars except for $\mathbf{0}$ and ∞, which are vectors with all components equal to 0 and ∞ respectively. Vectors are columns, and a raised T is matrix transpose. A real function f has gradient ∇f which is a column vector, and Hessian matrix $\nabla\nabla f$. The determinant is denoted by $|\cdot|$. The volume of the unit sphere in R^d is $\kappa_d = 2\sqrt{\pi}^d / d\Gamma(d/2)$.

A standard normal random variable has density $\phi(x)$ and cdf $\Phi(x) = 1 - \bar{\Phi}(x)$. In appendix A, §A.2 shows that the asymptotic expansion $\bar{\Phi}(x) \simeq x^{-1}\phi(x)$ is accurate as an approximation even as low as $x \geq 1$. The same is true for the Stirling formula, §A.1.

One focus of this work is developing approximations for exceedance probabilities based on a heuristic method. The approximations we develop will be encapsulated and highlighted with the label 'Result' as distinct from a mathematically proper 'Theorem'.

2 Prior Work

§2.1 The Vapnik Bound

CONTEMPORARY INTEREST in the above formulation of the learning problem is largely due to the work of Vapnik and Chervonenkis [54] and Vapnik [53], which was first brought to the attention of the neural network community by Baum and Haussler [10]. We briefly outline the result.

2.1 Definition *The family of classifiers $\mathcal{N}$ is said to* shatter *a point set $\mathcal{S} \subset R^p$ if*

$$(\forall \mathcal{S}' \subseteq \mathcal{S})(\exists \eta \in \mathcal{N})(\forall x \in \mathcal{S})\, \eta(x) = 1 \iff x \in \mathcal{S}' \quad ,$$

i.e. $\mathcal{N}$ is rich enough to dichotomize $\mathcal{S}$ in any desired way.

2.2 Definition *The* Vapnik-Chervonenkis (VC) dimension *of $\mathcal{N}$ is the greatest integer v such that*

$$(\exists \mathcal{S} \subset R^p)(\mathrm{card}(\mathcal{S}) = v \;\&\; \mathcal{N} \textit{ shatters } \mathcal{S}) \quad .$$

If $\mathcal{N}$ shatters sets of arbitrary size, then say $v = \infty$.

If $v < \infty$, $\mathcal{N}$ shatters no set having more than v points. The results of Vapnik and Chervonenkis hinge on the surprising, purely combinatorial,

2.3 Lemma (Sauer) *For a given family of classifiers $\mathcal{N}$, either $v = \infty$ or, for all $n \geq v$, the number of dichotomies of any point set $\mathcal{S}$ of cardinality n that are generated by $\mathcal{N}$ is no more than $\sum_{i=0}^{v} \binom{n}{i} \leq 1.5n^v/v! \leq (en/v)^v$.*

Proof. See Sauer [46] for the first expression and Vapnik [53] for the bound. □

Sauer [46] points out that the class 'all point sets in R^p of cardinality v' has VC dimension v and achieves the first bound of the lemma. Table 2.1 lists some classifier architectures and their VC dimensions. We note that the VC dimension of an architecture having d independently adjusted real parameters is generally about d. We may now state

2.4 Theorem *[53, ch. 6, thm. A.2] Let the VC dimension of the binary classifiers $\mathcal{N}$ be $v < \infty$. Then*

$$P(\|\, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|\, \|_{\mathcal{W}} > \epsilon) \leq 6 \left(\frac{2en}{v}\right)^v \exp(-n\epsilon^2/4) \quad . \qquad (2.1)$$

Table 2.1: Sample Vapnik-Chervonenkis dimensions

Class	Representative	VC Dimension
Orthants	$\times_{i=1}^{p}(-\infty, w_i]$	p
Rectangles	$\times_{i=1}^{p}[w_{0i}, w_{1i}]$	$2p$
Halfspaces (I)	$\{x \,:\, w^{\mathsf{T}}x \geq 0\}$	p
Halfspaces (II)	$\{x \,:\, w^{\mathsf{T}}x \geq w_0\}$	$p+1$
Linear Space	$\{x \,:\, \sum_{k=1}^{d} w_k \phi_k(x) \geq 0\}$	d

In each case the classifier architecture consists of versions of the shown prototype, a subset of R^p, as parameters w are varied. Most of these results are proved by Wenocur and Dudley [56], although some of them are elementary.

Proof. We sketch the idea of Vapnik's proof. Standard symmetrization inequalities give

$$P(\|[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)]\|_{\mathcal{W}} > \epsilon) \leq 2P(\|[\nu_{\mathcal{T}}(w) - \nu_{\mathcal{T}'}(w)]\|_{\mathcal{W}} > \epsilon/2)$$

where $\nu_{\mathcal{T}'}(w)$ is the empirical error computed on a "phantom" training set $\mathcal{T}'$ which is independent of $\mathcal{T}$ but has the same distribution. While the bracketed quantity on the LHS depends continuously on w, the corresponding one on the RHS depends only on where the $2n$ random pairs in $\mathcal{T}$ and $\mathcal{T}'$ fall. By Sauer's lemma, the nets in $\mathcal{N}$ can act on these points in at most $((2n)e/v)^v$ ways, so there are effectively only this many classifiers in $\mathcal{N}$. The probability that a single such net exhibits a discrepancy is a large deviation captured by the exponential factor. The overall probability is then handled by a union bound, where the polynomial bounds the number of distinct nets and the exponential bounds the probability of a discrepancy.

This is the essence of the argument, but the difficulty to be overcome is that precisely which networks are in the group of 'differently-acting' classifiers depends on the (random) training set. Some ingenious conditioning and randomization techniques must be used in the proof. □

The bound (2.1) is a polynomial in n of fixed degree v multiplying an exponential which decays in n, so the probability may be made arbitrarily small by an appropriately large sample size n. It is worthwhile to appreciate some unusual features of this bound:

- There are no unknown constant prefactors.
- The bound does not depend on any characteristics of the unknown probability distribution P. We term this *uniformity across distributions.*
- The bound likewise is independent of the function y which is to be estimated. This is *uniformity across target functions.*
- The bound holds for all networks. As discussed in §1.2, this provides information about $\mathcal{E}(w^*)$ as well as the efficacy of the architecture and how close the selected net is to the optimal one. This is *uniformity across networks.*

To understand the predictions offered by (2.1), note that the exponential form of the bound implies that after it drops below unity, it heads to zero very quickly. It is therefore most useful to find the *critical sample size* at which the bound drops below unity. The calculation in §C.1 shows this critical size is very close to

$$n_c = \frac{9.2v}{\epsilon^2} \log \frac{8}{\epsilon} \quad . \tag{2.2}$$

For purposes of illustration take $\epsilon = .1$ and $v = 50$, for which $n_c = 202\,000$. A neural network with $v = 50$ has about 50 free parameters, so the recommendation is for 4000 training samples per parameter, disagreeing by at least three orders of magnitude with the experience of even conservative practitioners (compare table 1.1).

In the introduction we proposed to pin down the performance of a data model which is selected on the basis of a training set by finding a sample size for which with probability nearly one,

$$\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, \|_{\mathcal{W}} < \epsilon \quad . \tag{2.3}$$

The resulting estimate, while remarkable for its explicitness and universality, is far too large. Our principal concern will be to find ways of making a tighter estimate of (2.3).

One way to improve (2.1) is to note that an ingredient of the Vapnik bound is the pointwise Chernoff bound

$$\begin{aligned} P(\nu_{\mathcal{T}}(w) - \mathcal{E}(w) > \epsilon) &\leq \exp\bigl(-n\epsilon^2 / \bigl(2\mathcal{E}(w)(1 - \mathcal{E}(w))\bigr)\bigr) \\ &\leq \exp(-2n\epsilon^2) \end{aligned} \tag{2.4}$$

which has been weakened via $0 \leq \mathcal{E}(w) \leq 1$. However, since we anticipate $\mathcal{E}(w) \approx 0$ the second bound seems unwise: for the classifiers of interest it is a gross error. This is a reflection of the simple fact mentioned in section 1.3 that typically the maximum-variance point (here $\mathcal{E}(w) = 1/2$) dominates exceedance probabilities such as (2.3). (See e.g. [39, 50] and [36, ch. 3].) Resolution may be added to (2.1) by examining instead

$$P(\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, / \sqrt{\mathcal{E}(w)(1 - \mathcal{E}(w))} \, \|_{\mathcal{W}} > \epsilon) \quad . \tag{2.5}$$

Vapnik approximates this criterion and proves

2.5 Theorem *[53, ch. 6, thm. A.3] Let the VC dimension of the binary classifiers $\mathcal{N}$ be v. Then*

$$P(\bigl\| \bigl(\mathcal{E}(w) - \nu_{\mathcal{T}}(w)\bigr) / \sqrt{\mathcal{E}(w)} \, \|_{\mathcal{W}} > \epsilon) \leq 8 \Bigl(\frac{2en}{v}\Bigr)^v \exp\left(-n\epsilon^2/4\right) \quad . \tag{2.6}$$

This results in the critical sample size

$$n_c = \frac{9.2v}{\epsilon^2} \log \frac{8}{\epsilon} \quad , \tag{2.7}$$

above which with high probability

$$(\forall w \in \mathcal{W}) \, \frac{\mathcal{E}(w) - \nu_{\mathcal{T}}(w)}{\sqrt{\mathcal{E}(w)}} \leq \epsilon \quad .$$

The same conclusions as (1.8) are now possible. By way of illustration let us consider the first such conclusion which is the bound on $\nu_{\mathcal{T}}(w^*) - \mathcal{E}(w^*)$. If the net of interest w^* has $\nu_{\mathcal{T}}(w^*) = 0$ (for example, if the architecture is sufficiently rich) then we may essentially replace ϵ^2 by ϵ in (2.7):

$$n_c = \frac{4.6v}{\epsilon} \log \frac{64}{\epsilon} \tag{2.8}$$

samples are sufficient for $\mathcal{E}(w^*) < \epsilon$ with high probability. Using the same $v = 50$ and $\epsilon = 0.1$ yields a sample size sufficient for reliable generalization of about $n = 14\,900$, which is still unrealistically high.

§2.2 Further Developments

The VC tools and results were introduced to the theoretical computer science community by Blumer et al. [11]. In addition to examining methods of selecting a network $\eta(\cdot; w)$ on the basis of $\mathcal{T}$, VC methods are used to find

$$P(\big\| |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, 1_{\{0\}}(\nu_{\mathcal{T}}(w)) \big\|_{\mathcal{W}} > \epsilon) \leq 2\Big(\frac{2en}{v}\Big)^v 2^{-n\epsilon/2} \quad , \tag{2.9}$$

and as pointed out by Anthony and Biggs [9, thm. 8.4.1]

$$n_c = \frac{5.8v}{\epsilon} \log \frac{12}{\epsilon} \tag{2.10}$$

samples are enough to force this below unity. As in (2.8) we see the $O\big((v/\epsilon) \log 1/\epsilon\big)$ dependence when working near $\mathcal{E}(w) = 0$. By careful tuning of two parameters used in deriving (2.9), Shawe-Taylor et al. [48] find the sufficient condition

$$n_c = \frac{2v}{\epsilon(1 - \sqrt{\epsilon})} \log \frac{6}{\epsilon} \tag{2.11}$$

provided that only networks, if any, having $\nu_{\mathcal{T}}(w) = 0$ are used. Once more trying out $v = 50$, $\epsilon = 0.1$ gives $n = 6000$, which is the tightest estimate in the literature but still out of line with practice. The methods used to show (2.10) and (2.11) make strong use of the $\nu_{\mathcal{T}}(w) = 0$ restriction so it seems unlikely that they can be extended to the case of noisy data or imperfect models.

Haussler [30] (see also Pollard [44]) applies similar tools in a more general decision-theoretic setting. In this framework [25], a function $l(y, a) \geq 0$ captures the loss incurred by taking action (e.g. choosing the class) $a \in \mathcal{A}$ when the state of nature is y. Nets $\eta(\cdot; w)$ then become functions into $\mathcal{A}$, and the risk $r(w) := E\, l(y, \eta(x; w))$ is the generalization of probability of error. This is estimated by $\hat{r}(w) := n^{-1} \sum_{i=1}^n l(y_i, \eta(x_i; w))$, and the object of interest is

$$P(\| \rho(\hat{r}(w), r(w)) \|_{\mathcal{W}} > \epsilon) \tag{2.12}$$

where ρ is some distance metric. For instance, the formulation (2.1) has $l(y, \eta) = (y - \eta)^2$ and $\rho(r, s) = |r - s|$. Haussler uses the relative-distance metric

$$d_\nu(r, s) := \frac{|r - s|}{\nu + r + s} \quad \text{for } \nu > 0. \tag{2.13}$$

Letting $\nu = \epsilon$ and $\alpha = 1/2$ yields a normalized criterion similar to dividing by the standard deviation, but rather cruder.

Now suppose the loss function is bounded between 0 and 1, and for each y, $l(y,a)$ is monotone in a (perhaps increasing for some y and decreasing for others). Haussler finds [30, thm. 8]

$$P(\, \|d_\nu(\hat{r}(w), r(w))\|_{\mathcal{W}} \geq \alpha) \leq 8 \left(\frac{16e}{\alpha\nu} \log \frac{16e}{\alpha\nu} \right)^{v_p} e^{-\alpha^2 \nu n/8} \tag{2.14}$$

where v_p is the *pseudo dimension*[1] of the possibly real-valued functions in $\mathcal{N}$, which coincides with the VC dimension for $\{0,1\}$-valued functions. To force this bound below unity requires about

$$n_c = \frac{16 v_p}{\alpha^2 \nu} \log \frac{8e}{\alpha\nu} \tag{2.15}$$

samples. This is to date the formulation of the basic VC theory having the most generality, although again the numerical bounds offered are not tight enough.

§2.3 Applications to Empirical Processes

When Vapnik and Chervonenkis proved theorem 2.4, it was done as a generalization on the classical Glivenko-Cantelli theorem on uniform convergence of an empirical cumulative distribution function (cdf) to an actual one. To see the connection, define

$$D_n := \|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)\|_{\mathcal{W}} \tag{2.16}$$

and consider the case where $y \equiv 0$, x takes values in R, and $\eta(x;w) = 1_{(-\infty,w]}(x)$. Then $\{x : (\eta(x;w) - y)^2 = 1\} = (-\infty, w]$ and $\mathcal{E}(w) = F(w)$, the distribution of x.

2.6 Theorem (Glivenko-Cantelli)

$$D_n \to 0 \quad a.s. \tag{2.17}$$

Of course this is implied by the assertion of Vapnik above on noting (as in table 2.1) that the VC dimension of the functions $\eta(x;w)$ is one, whereby the exponential bound on $P(D_n > \epsilon)$ implies $\sum_{n=1}^{\infty} P(D_n > \epsilon) < \infty$ which the Borel-Cantelli lemma turns in to almost sure convergence.

It is then natural to ask if a rescaled version of D_n converges in distribution. Kolmogorov showed that it did and by direct methods found the limiting distribution

$$(\forall F \text{ cts.})(\forall b > 0)\, P(\sqrt{n} D_n > b) \to e^{-2b^2} \quad . \tag{2.18}$$

The less direct but richer path is to analyze the stochastic process

$$Z_n(w) := \sqrt{n}[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)] \quad . \tag{2.19}$$

Doob [17] made the observation that, by the ordinary central limit theorem, the limiting finite-dimensional distributions of this process are

[1] The pseudo dimension is defined as follows. For some training set $x_1, \ldots, x_n$ consider the cloud of points in R^n of the form $[\eta(x_1;w) \cdots \eta(x_n;w)]$ for $w \in \mathcal{W}$. Then v_p is the largest n for which there exists a training set and a center $p_0 \in R^n$ such that some piece of the cloud occupies all 2^n orthants around p_0.

Gaussian, with the same covariance function $R(w,v) = w \wedge v - wv$ as the Brownian bridge. His conjecture that the limit distribution of the supremum of the empirical process Z_n (which is relatively hard to find) equalled that of the supremum of the Brownian bridge was proved shortly thereafter [16].

The most immediate generalization of this empirical process setup is to vector random variables. Now $w, x \in R^d$, and $\eta(x; w) = 1_{(-\infty,w]}(x)$ where $(-\infty, w] := \times_{i=1}^{d}(-\infty, w_i] \subset R^d$ so that again $\mathcal{E}(w) = F(w)$. Kiefer [34] showed that for all $\delta > 0$ there exists $c = c(d, \delta)$ such that

$$(\forall n, b > 0, F)\, P(\sqrt{n} D_n > b) \leq c e^{-2(1-\delta)b^2} \tag{2.20}$$

Dudley has shown the equivalence for large n of the distribution of the supremum of the empirical process and the corresponding Gaussian process. Adler and Brown [3] have further shown that under mild conditions on F there exists a $c = c(F)$ such that for all $n > n(b)$,

$$c^{-1} b^{2(d-1)} e^{-2b^2} \leq P(\sqrt{n} D_n > b) \leq c\, b^{2(d-1)} e^{-2b^2} \quad , \tag{2.21}$$

thus capturing the polynomial factor. However, neither the constant factor nor the functional form for $n(b)$ is available, so this bound is not of use to us. Adler and Samorodnitsky [4] provide similar results for other classes of sets, e.g. rectangles in R^d and half-planes in R^2.

The results (2.18), (2.20), and (2.21) on the distribution of the supremum of an empirical process are derived as limits in n for fixed b. In a highly technical paper [51], Talagrand extends these results not only to apply to all VC classes, but also by finding a sample size at which the bound becomes valid. Talagrand's powerful bound (his thm. 6.6) is (now written in terms of (ϵ, n) rather than (b, n)):

$$P(\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, \|_{\mathcal{W}} > \epsilon) \leq K_1 \Big(\frac{K_2 n \epsilon^2}{v} \Big)^v e^{-2n\epsilon^2}, \tag{2.22}$$

for all $n \geq K_3 v/\epsilon^2$, where the three constants are universal. This gives the critical value of about

$$n_c = \frac{(K_3 \vee (1/2) \log K_2) v}{\epsilon^2} \quad . \tag{2.23}$$

Unfortunately for any application of this result, the constants are inaccessible and "the search of sharp numerical constants is better left to others with the talent and the taste for it" [51, p. 31]. It does, however, illustrate that the order of dependence (without restriction on $\nu_{\mathcal{T}}(w)$) is v/ϵ^2, without the extra logarithmic factor seen throughout §2.1, §2.2.

§2.4 The Expected Error

Instead of looking at the probability of a significant deviation of $\nu_{\mathcal{T}}(w^*)$ from $\mathcal{E}(w^*)$, some approaches examine $E\,\mathcal{E}(w^*)$. In doing this no information about the variability of $\mathcal{E}(w^*)$ is obtained unless $E\,\mathcal{E}(w^*) \approx \mathcal{E}(w^0)$, which implies $\mathcal{E}(w^*)$ is near $\mathcal{E}(w^0)$ with high probability. In this sense these methods are similar to classical statistical efforts to determine consistency and bias of estimators. Additionally, as remarked in §1.2, using this criterion precludes saying anything about the performance of the selected net relative to the best net in the architecture, or about the efficacy of the architecture. On the other hand, the results are interesting because they seek to incorporate information about the training method.

Such results are usually expressed in terms of *learning curves*, or values of $E\,\mathcal{E}(\eta^*)$ as a function of n (and perhaps another parameter representing complexity). This is somewhat analogous to the ϵ, n, and v of VC theory, although the relation between ϵ and $E\,\mathcal{E}(\eta^*)$ is indirect.

Haussler et al. [31] present an elegant and coherent analysis of this type. The authors assume that the target y is expressible as a deterministic function $\eta^0 : R^p \to \{0,1\}$, and that $\eta^0 \in \mathcal{N}$. Assuming knowledge of a prior π_0 on $\mathcal{N}$ which satisfies a mild nondegeneracy condition, the authors show that

$$E\,\mathcal{E}(w^*) \leq \frac{2v}{n} \tag{2.24}$$

when w^* is obtained by sampling from the posterior π given the n observations. In the more realistic case where no such prior is known, it is proved that

$$E\,\mathcal{E}(w^*) \leq (1 + o(1))\frac{v}{n}\log\frac{n}{v} \tag{2.25}$$

where $o(1) \to 0$ as $n/v \to \infty$ and now w^* is chosen from a posterior generated by an assumed prior π_*. (The bound is not a function of this prior except possibly in the remainder term.) Amari and Murata [8] obtain results similar to (2.24) via familiar statistical results like asymptotic normality of parameter estimates. In place of the VC dimension v is the trace of a certain product of asymptotic covariance matrices.

Work on this problem of a different character has also been done by researchers in statistical physics. Interpreting the training error $\nu_{\mathcal{T}}(w)$ as the "energy" of a system and the training algorithm as minimizing that energy allows the application of thermodynamic techniques. Some specific learning problems have been analyzed in detail (most notably the perceptron with binary weights treated in [49] and confirmed by [38]) and unexpected behaviors found, principally a sharp transition to near-zero error at certain values of n/d. Unfortunately the work in this area as published suffers from heavy use of physically motivated but mathematically unjustified approximations. For example, the 'annealed approximation' replaces the mean free energy $E \log Z(\beta)$ by $\log EZ(\beta)$ (the latter is an upper bound), and goes on to approximate $\partial/\partial\beta$ of the former *by differentiating the upper bound as if it were the original quantity.* When applied to physical systems such approximations have a verifiable interpretation; however, such intuitions are generally lacking in the neural network setting. Neural networks, after all, are mathematical objects and are not constrained by physical law in the same way a ferromagnet is. It remains to be seen if this work, summarized in [47, 55], can be formalized enough to be trustworthy.

§2.5 Empirical Studies

Some researchers have tried to determine the generalization error for example scenarios via simulation studies. Such studies are important to us as they will allow us to check the validity of the sample size estimates we find.

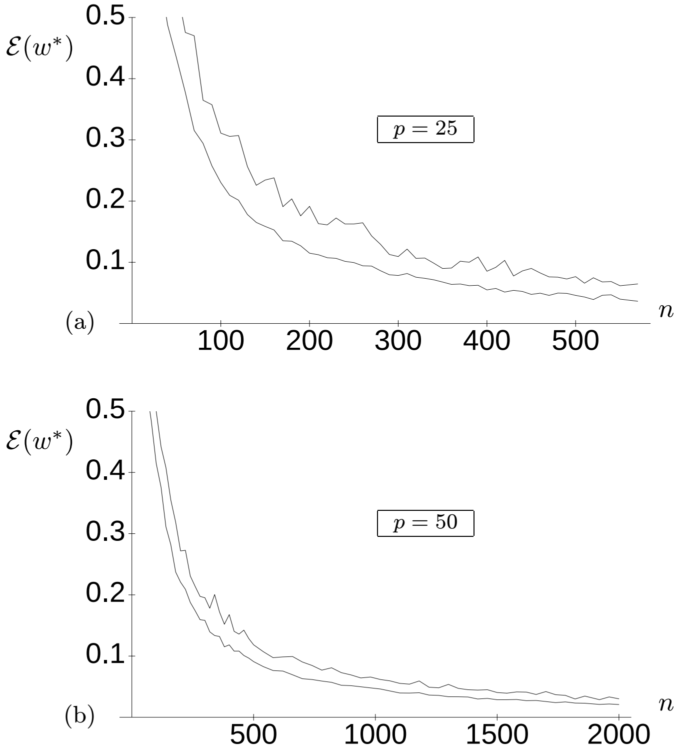

Figure 2.1: Cohn-Tesauro experiments on generalization

Shown are learning curves for the threshold function in two input dimensions. The lower curve in each panel is the average value of $\mathcal{E}(w^*)$ over about 40 independent runs. The upper curve is the largest value observed in these runs.

Cohn and Tesauro [12] have done a careful study examining how well neural networks can be trained to learn (among others) the 'threshold function' taking inputs in $[0, 1]^p$ and producing a binary output that is zero unless the sum of all inputs is larger than $p/2$. This is a linearly separable function. Two sizes $p = 25$ and $p = 50$ are chosen and the class of nets used to approximate is linear threshold units with p inputs.[2] The data distribution is uniform over the input space.

Nets are selected by the standard backpropagation algorithm, and their error computed on a separate test set of 8000 examples. Forty such training/test procedures are repeated to obtain independent estimates of $\mathcal{E}(w^*)$. Averaging these values gives an estimate of $E\mathcal{E}(w^*)$ as in §2.4, but for the reasons outlined there this is not our main interest; we are rather interested in the distribution of the discrepancy $\mathcal{E}(w^*) - \nu_{\mathcal{T}}(w^*)$. The differencing operation has little effect since in the trials $\nu_{\mathcal{T}}(w^*) \approx 0$ generally. We examine the distributional aspects by looking at, for a given function, p, and n, the sample mean of $\mathcal{E}(w^*) - \nu_{\mathcal{T}}(w^*)$ and

[2] Networks with continuously-varying outputs are used as a device to aid weight selection, but the final network from which empirical and "true" errors are computed from has outputs in $\{0, 1\}$.

the largest observed value in the 40 trials. These results are shown in figure 2.1. The lower curves (representing sample mean) have an excellent fit to $0.87\, p/n$, and the upper curves (extreme value) fit well to $1.3\, p/n$.

§2.6 Summary

Motivated by the strength of the results possible by knowing the distribution of the maximum deviation between empirical and true errors, we consider the Vapnik bound, which holds independent of target function and data distribution. The original form of this bound results in extreme overestimates of sample size, and making some assumptions about the selected network ($\nu_{\mathcal{T}}(w^*) = 0$) allows them to be reduced, but not enough to be practical. Work to this point in the neural net community on this formulation of the question of reliable generalization has focused exclusively on reworkings of the Vapnik ideas.

We propose to use rather different techniques—which are approximations rather than bounds—to estimate the same probability pursued in the Vapnik approach. In this approach, sample size estimates depend on the problem at hand through the target function and the data distribution. We will see that in some cases, these estimates are quite reasonable in the sense of being comparable with practice.

3 The Poisson Clumping Heuristic

NOW WE DESCRIBE the approach we take to the problem of generalization in neural networks. This is based on one familiar idea—a passage to a normal limit via generalized central limit theorems—and one not so familiar—finding the exceedances of a high level by a stochastic process using a new tool called the Poisson clumping heuristic. We transform the empirical process $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$ to a Gaussian process, and this into a *mosaic process* of scattered sets in weight space which represent regions of significant disagreement between $\mathcal{E}(w)$ and its estimate $\nu_{\mathcal{T}}(w)$.

§3.1 The Normal Approximation

For the large values of n we anticipate, the central limit theorem informs us that

$$Z_n(w) := \sqrt{n}\,[\nu_{\mathcal{T}}(w) - \mathcal{E}(w)] \tag{3.1}$$

has nearly the distribution of a zero-mean Gaussian random variable; the multivariate central limit theorem shows further that the collection $\{Z_n(w_1), \ldots, Z_n(w_k)\}$ has asymptotically a joint Gaussian distribution. The random variable of interest to us is $\|Z_n(w)\|_{\mathcal{W}}$ which depends on infinitely many points in weight space. To treat this type of convergence we need a functional central limit theorem (FCLT) written compactly

$$Z_n \Rightarrow Z \tag{3.2}$$

which means that for bounded continuous (in terms of the uniform distance metric $\rho(Z, Z') = \big\| \, |Z(w) - Z'(w)| \, \big\|_{\mathcal{W}}$) functionals f taking whole sample paths on $\mathcal{W}$ to R, the ordinary random variables

$$f(Z_n(\cdot)) \Rightarrow f(Z(\cdot)) \quad . \tag{3.3}$$

The supremum function $\|\cdot\|_{\mathcal{W}}$ for compact $\mathcal{W}$ is trivially such a bounded continuous function, and is the only one of interest here. FCLT's are well-known for classifiers of finite VC dimension: e.g. [43, ch. 7, thm. 21] and [36, thm. 14.13] are results ensuring that (3.3) holds for VC classes for any underlying distribution P. FCLT's also apply to neural network regressors having, say, bounded outputs and whose corresponding graphs[1] have finite VC dimension [7]. These theorems imply it is

[1] One of the several ways to extend the VC dimension to functions $f : R^p \to R$ is to find the ordinary VC dimension of the sets $\{(x, y) : 0 \le y \le f(x) \,\|\, f(x) \le y \le 0\}$ in R^{p+1}.

reasonable, for the moderately large n we envision, to approximate[2]

$$\begin{aligned} P(\big\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, \big\|_{\mathcal{W}} > \epsilon) &\simeq P(\big\| \, |Z(w)| \, \big\|_{\mathcal{W}} > \epsilon\sqrt{n}) \\ &\leq 2P(\|Z(w)\|_{\mathcal{W}} > \epsilon\sqrt{n}) \end{aligned}$$

where $Z(w)$ is the Gaussian process with mean zero and covariance

$$R(w,v) := EZ(w)Z(v) = \operatorname{Cov}\big((y - \eta(x;w))^2, \, (y - \eta(x;v))^2\big) \quad .$$

The problem about extrema of the original empirical process is equivalent to one about extrema of a corresponding Gaussian process.

A remark is in order about one aspect of the proposed approximation. While it is true that for fixed w

$$Z_n(w) \Rightarrow Z(w)$$

so that, since the limiting distribution is continuous,

$$\frac{P(Z_n(w) \leq \alpha)}{P(Z(w) \leq \alpha)} \underset{n \to \infty}{\rightarrow} 1 \quad ,$$

this is not generally true when $\alpha = \alpha(n) = \epsilon\sqrt{n}$; in fact, the fastest α can grow while maintaining the CLT is the much slower $\sqrt[6]{n}$, see [24, sec. XVI.7]. However, this conventional mathematical formulation is not what we desire. We only wish, for finite large n, the denominator to be a reasonable estimate of the numerator; moreover, we do not go into the tail of the normal distribution because we only desire to make $P(\|Z(w)\|_{\mathcal{W}} \geq b)$ of order perhaps .01. In other words, while we write $\alpha(n) = \epsilon\sqrt{n}$, we in effect *choose ϵ so that $\alpha(n)$ remains moderate.*

§3.2 Using the Poisson Clumping Heuristic

The Poisson clumping heuristic (PCH), introduced and developed in a remarkable book [6] by D. Aldous, provides a tool of wide applicability for estimating exceedance probabilities. Consider the excursions above a high level b of a sample path of a stochastic process $Z(w)$. As in figure 3.1a, the set $\{w : Z(w) \geq b\}$ can be visualized as a group of smallish *clumps* scattered sparsely in weight space $\mathcal{W}$. The PCH says that, provided Z has no long-range dependence and the level b is large, these clumps are generated independently of each other and thrown down at random (that is, centered on points of a Poisson process) on $\mathcal{W}$. Figure 3.1b illustrates the associated clump process. The vertical arrows illustrate two clump centers (points of the Poisson process); the clumps themselves are bounded by the bars surrounding the arrows.

Formally, such a so-called mosaic process consists of two stochastically independent mechanisms:

[2] Doob first proposed this idea for the class of indicator functions of intervals in R^1:

> We shall assume, until a contradiction frustrates our devotion to heuristic reasoning, that *in calculating asymptotic $x_n(t)$ process distributions when $n \to \infty$ we may simply replace $x_n(t)$ process by the $x(t)$ process.* It is clear that this cannot be done in all possible situations, but let the reader who has never used this sort of reasoning exhibit the first counter example. [17, p. 395]

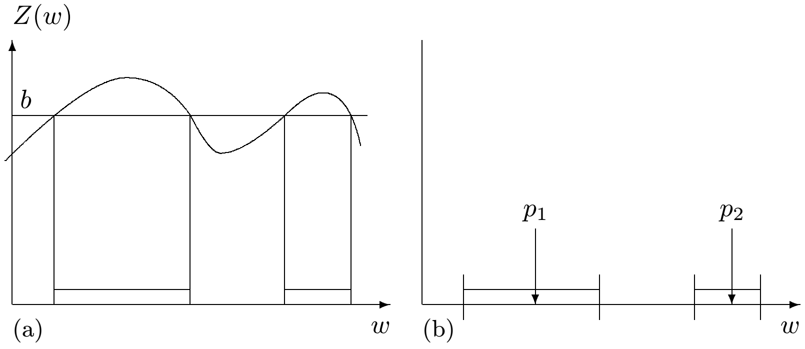

Figure 3.1: The Poisson clumping heuristic

The original process is on the left; the associated clump process is on the right.

- A Poisson process on $\mathcal{W}$ with intensity $\lambda_b(w)$ generating random points $\mathcal{P} = \{p\} \subset \mathcal{W}$. We assume throughout that $\int_{\mathcal{W}} \lambda_b(w)\, dw < \infty$ so that $\mathcal{P}$ is finite.
- For each $w \in \mathcal{W}$ there is a process choosing $\mathcal{C}_b(w) \subset \mathcal{W}$ from a distribution on sets, parameterized by b, which may vary across weight space. For example, $\mathcal{C}_b(w)$ might be chosen from a countable collection of sets according to probabilities that depend on w, or it might be a randomly scaled version of an elliptical exemplar having orientation depending on w and size inversely proportional to b.

According to this setup choose an independent random set $\mathcal{C}_b(p)$ for each Poisson point $p \in \mathcal{P}$. The mosaic process is

$$\mathcal{S}_b := \bigcup_{p \in \mathcal{P}} (p + \mathcal{C}_b(p)) \quad .$$

See [29] for more on mosaic processes.

The assertion of the PCH is that, for large b and Z having no long-range dependence,

$$1_{\mathcal{S}_b}(\cdot) \overset{\mathfrak{D}}{\approx} 1_{[b,\infty)}(Z(\cdot)) \tag{3.4}$$

in the sense of (3.2). This claim is not proved in general; instead

- the idea is justified in terms of its physical appeal.
- the Poisson approximation (3.4) is vindicated by rigorous proofs in certain special cases, e.g. for discrete- and continuous-time stationary processes on the real line [35].
- about 200 diverse examples are given in [6], in discrete, continuous, and multiparameter settings, for which the method both gives reasonable estimates and for which the estimates agree with known rigorous results.

Defining N_b as the total number of clumps in $\mathcal{S}_b$ and $N_b(w)$ as the number of clumps containing w gives the translation into a global equation and a local equation:

$$P(\|Z(w)\|_{\mathcal{W}} > b) = P(N_b > 0) = 1 - e^{-\int_{\mathcal{W}} \lambda_b(w)\,dw} \tag{3.5a}$$

$$p_b(w) := P(Z(w) > b) = P(N_b(w) > 0) \quad . \tag{3.5b}$$

The next result shows how to use $C_b(w) := \mathrm{vol}(\mathcal{C}_b(w))$ and the local equation to find the intensity $\lambda_b(w)$.

3.1 Lemma *$N_b(w)$ is Poisson distributed. If $\lambda_b(w)$ and the distribution of $\mathcal{C}_b(w)$ are nearly constant in a neighborhood of w, and if with high probability $w + \mathcal{C}_b(w)$ is contained within this neighborhood, then $EN_b(w) \simeq \lambda_b(w) EC_b(w)$.*

Proof. Note N_b is Poisson with mean $\Lambda_b = \int_{\mathcal{W}} \lambda_b(w)\,dw$. Drop the b subscripts.

$$\begin{aligned}
Ee^{iuN(w)} &= E\,E\left[e^{iuN(w)} \;\middle|\; N\right] \\
&= E\,E\left[\exp(iu\sum_{k=1}^{N} 1_{p_k+\mathcal{C}(p_k)}(w)) \;\middle|\; N\right] \\
&= E\prod_{k=1}^{N} E\exp(iu1_{p_k+\mathcal{C}(p_k)}(w)) \\
&= E\left(1-\rho_w+\rho_w e^{iu}\right)^N \\
&= \sum_{N=0}^{\infty} e^{-\Lambda}\frac{\Lambda^N}{N!}\left(1-\rho_w+\rho_w e^{iu}\right)^N \\
&= \exp(-\Lambda\rho_w(1-e^{iu})) \quad ,
\end{aligned}$$

with $\rho_w := P(w \in p + \mathcal{C}_b(p) \,|\, p \in \mathcal{W})$, the probability that a particular clump in $\mathcal{W}$ captures w. The characteristic function of $N_b(w)$ is that of a Poisson r.v. with mean $\Lambda_b\rho_w$, proving the first assertion. For the second, initially suppose the clump process is stationary so that $\lambda_b(w) = \lambda_b$ and all clumps have the distribution of $\mathcal{C}_b$. Then ρ_w is the fraction of trials in which a randomly-placed patch $\mathcal{C}_b$ intersects a given point w. Provided edge effects can be ignored ($C_b \ll \mathrm{vol}(\mathcal{W})$ with high probability) this is just $EC_b/\mathrm{vol}(\mathcal{W})$. In the nonstationary case, let $B_w \subset \mathcal{W}$ be a small ball containing w. Dropping subscripts,

$$\begin{aligned}
\rho_w &= P(w \in p + \mathcal{C}(p) \,|\, p \in \mathcal{W}) \\
&= P(p \in B_w \,|\, p \in \mathcal{W})\, P(w \in p + \mathcal{C}(p) \,|\, p \in B_w) + \\
&\quad P(p \in B_w^{\mathsf{C}} \,|\, p \in \mathcal{W})\, P(w \in p + \mathcal{C}(p) \,|\, p \in B_w^{\mathsf{C}}) \\
&\overset{(a)}{\simeq} P(p \in B_w \,|\, p \in \mathcal{W})\, P(w \in p + \mathcal{C}(p) \,|\, p \in B_w) \\
&\overset{(b)}{\simeq} P(p \in B_w)\, P(w \in p + \mathcal{C}(w) \,|\, p \in B_w) \\
&\overset{(c)}{\simeq} \frac{\int_{B_w} \lambda(w')\,dw'}{\Lambda} \cdot \frac{EC(w)}{\mathrm{vol}(B_w)} \\
&\overset{(d)}{\simeq} \frac{\lambda(w) EC(w)}{\Lambda}
\end{aligned} \tag{3.6}$$

where (a) is justified since B_w is large enough to contain all clumps hitting w, (b) by the local stationarity of $\mathcal{C}_b(w)$, (c) since again the clump size is small relative to B_w, and (d) by the local stationarity of the intensity. □

In our application, occurrence of a clump in weight space corresponds to existence of a large value of $Z(w)$, or a large discrepancy between $\mathcal{E}(w)$ and its estimate $\nu_{\mathcal{T}}(w)$. We therefore anticipate operating in a regime where $N_b = 0$ with high probability and equivalently $(\forall w) N_b(w) = 0$ with high probability, so that with lemma 3.1, the global/local equations (3.5) become

$$P(N_b > 0) = 1 - e^{-\int_{\mathcal{W}} \lambda_b(w)\,dw} \overset{\text{(a)}}{\simeq} \int_{\mathcal{W}} \lambda_b(w)\,dw \tag{3.7a}$$

$$P(N_b(w) > 0) = 1 - e^{-\lambda_b(w) EC_b(w)} \overset{\text{(b)}}{\simeq} \lambda_b(w) EC_b(w) \quad . \tag{3.7b}$$

To sum up, the heuristic calculation ends in the RHS of the upper equation, and this being low validates approximation (a), showing $P(N_b = 0)$ is near unity. *A fortiori* the LHS of lower equation is small, which validates approximation (b).

The first fundamental relation, which we treat as an equality, stems from the local equation above:

$$\boxed{p_b(w) = \lambda_b(w) EC_b(w)} \quad . \tag{3.8}$$

Letting $\bar{\Phi}(b) = P(N(0,1) > b)$ and $\sigma^2(w) = R(w,w)$, we have $p_b(w) = \bar{\Phi}(b/\sigma(w))$, and the second fundamental equation is (3.8) substituted into the global equation (3.7b):

$$\boxed{P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)}\,dw} \quad . \tag{3.9}$$

The idea behind the derivation is that the point exceedance probabilities are not additive, but the Poisson intensity is. Local properties of the random field $(p_b(w), EC_b(w))$ allow the intensity to be determined, and the PCH tells us how to combine the intensities to determine the overall probability. Loosely speaking, (3.9) says that the probability of an exceedance is the sum of all the pointwise exceedance probabilities, each diminished by a factor indicating the interdependence of exceedances at different points. The remaining difficulty is finding the mean clump size $EC_b(w)$ in terms of the network architecture and the statistics of (x, y).

§3.3 Summary

We have described the rationale and tools for approximating in distribution the random variable $\|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)\|$ in this two-stage fashion:

$$\begin{array}{ccccc} \text{Empirical Process} & \overset{\text{FCLT}}{\Longrightarrow} & \text{Gaussian Process} & \overset{\text{PCH}}{\Longrightarrow} & \text{Mosaic Process} \\ \nu_{\mathcal{T}}(w) - \mathcal{E}(w) & & Z(w),\ R(w,v) & & \lambda_b(w),\ C_b(w) \end{array}$$

4 Direct Poisson Clumping

IN THIS CHAPTER we discuss several situations in which the Poisson clumping method can be used without simplifying approximations to give conditions for reliable generalization. The first few results examine variants of the problem of learning axis-aligned rectangles in R^d. Later we develop a general result applying when the architecture is smooth as a function of w.

Finding these precise results is calculation-intensive, so before beginning we mention the interest each of these problems has for us. The problem of learning orthants is relevant to applied probability as the first-studied, and best-known, example of uniform convergence (the Glivenko-Cantelli theorem). Learning rectangles, closely related to learning orthants, has been examined several times in the PAC learning literature, e.g. in [33] as the problem of identifying men having medium build using their height and weight. (A natural decision rule is of the type: a man is of medium build if his height is between 1.7 and 1.8 meters and his weight is between 75 and 90 kilograms, which is a rectangle in R^2.) The problem of learning halfspaces, or training a linear threshold unit, is the best-studied problem in the neural network literature. The last section details learning smooth functions. The results here have the advantage that they apply universally to all such network architectures (e.g. networks of sigmoidal nonlinearities), and that the methods are transparent.

Here's what we expect to learn from these examples. First, we will understand what determines the mean clump size, and develop some expectations about its general form which will be important in our later efforts to approximate it. Second, we will see that, given sufficient knowledge about the process, the PCH approach generates tight sample size estimates of a reasonable functional form. Finally, a side-effect of our efforts will be the realization that, although exact PCH calculations can be carried out for some simple cases, in general the approach of performing such calculations seems of limited practical applicability. This will motivate our efforts in chapter 5 to approximate the clump size.

§4.1 Notation and Preliminaries

We establish straightforward notation for orthants and rectangles in R^d. For u, $v \in R^d$, write $u \leq v$ when the inequality is maintained in each coordinate, and write $[u, v]$ for $\{w : u \leq w \leq v\}$. Similarly $\wedge$ and $\vee$ are extended coordinatewise. Let $|u| := \mathrm{vol}([\mathbf{0}, u])$, which is zero if $u \not> \mathbf{0}$.

The empirical processes we will meet in the first few sections are best thought of in terms of a certain set-indexed Gaussian process. We introduce this process via some definitions which are intended to build intuition. Let μ be a positive measure μ on R^p.

4.1 Definition *The* μ-white noise $W(A)$ *is defined on Borel sets of finite* μ*-measure such that:*

$$\begin{aligned}&W(A) \overset{\mathcal{D}}{=} N(0,\mu(A))\\ &(\forall n)\,\{A_k\}_{k\le n} \textit{ disjoint } \implies \{W(A_k)\}_{k\le n} \textit{ independent}\\ &W(A)+W(B) = W(A\cup B)+W(A\cap B) \textit{ a.s.}\end{aligned}$$

(It is easy to verify that this process exists by checking that the covariance is nonnegative-definite.) $W(A)$ adds up a mass $\mu(A)$ of infinitesimal independent zero-mean "noises" that occur within the set A. To turn the set-indexed white noise into a random field, just parameterize some of the sets A by real vectors $w \in \mathcal{W}$. In particular,

4.2 Definition *The* μ-Brownian sheet *is*

$$W(w) := W((-\infty, w])$$

where $W(A)$ *is* μ*-white noise. To get* Brownian sheet, *take* μ *as Lebesgue measure on* $[0,1]^p$.

Brownian sheet is the p-dimensional analog of Brownian motion.

Returning to set-indexed processes, if μ is a probability measure we can define our main objective, the pinned Brownian sheet.

4.3 Definition *The* pinned set-indexed μ-Brownian sheet *is*

$$Z(A) := W(A) - \mu(A)W(R^p) \tag{4.1}$$

The pinned μ-Brownian sheet *is defined for* $w \in R^p$ *by*

$$Z(w) := Z((-\infty, w]) \tag{4.2}$$

where $Z(A)$ *is* μ*-Brownian sheet. To get* pinned Brownian sheet, *take* μ *as Lebesgue measure on the unit hypercube.*

The pinned Brownian sheet is a generalization of the Brownian bridge, and in statistics it occurs in the context of multidimensional Kolmogorov-Smirnov tests. The pinned set-indexed Brownian sheet inherits additivity from the associated white noise process:

$$\begin{aligned}Z(A)+Z(B) &= \big(W(A)+W(B)\big) - \big(\mu(A)+\mu(B)\big)W(R^p)\\ &= Z(A\cup B) + Z(A\cap B) \quad .\end{aligned} \tag{4.3}$$

Its covariance is

$$\begin{aligned}E\,Z(A)Z(B) &= \mu(A\cap B) - \mu(A)\mu(B)\\ &= 1/4 - \mu(A \bigtriangleup B)/2 \quad (\text{if } \mu(A)=1/2).\end{aligned} \tag{4.4}$$

To see the connection to the neural network classification problem, suppose the input data x is generated in R^p according to P, and y is deterministically based on x. Let A_w be the region where $\eta(\cdot; w) = 1$ and A_0 be that where $y = 1$. Then $B_w := A_0 \bigtriangleup A_w$ is the region of disagreement between the target and the network, where $(y-\eta(x;w))^2 = 1$. The covariance of the empirical process is

$$\begin{aligned} &nE(\nu_{\mathcal{T}}(w) - \mathcal{E}(w))(\nu_{\mathcal{T}}(w') - \mathcal{E}(w')) \\ &\qquad = \operatorname{Cov}(1_{B_w}(x), 1_{B_{w'}}(x)) = P(B_w \cap B_{w'}) - P(B_w)P(B_{w'}) \end{aligned} \tag{4.5}$$

which is the same as the pinned P-Brownian sheet indexed by the B_w. The limit Gaussian process discussed in chapter 3 is, for ‘noiseless’ classification (y a function of x), the pinned P-Brownian sheet of definition 4.3.

§4.2 Learning orthants

Suppose networks $\eta(x;w) = 1_{(-\boldsymbol{\infty},w]}(x)$ are used to learn a target function $y = y(x) = \eta(x;w^0)$. If x has a continuous distribution F then we may reduce the problem to a canonical form by noting

$$\begin{aligned} \|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)\|_{\mathcal{W}} &= \sup_{w \in \mathcal{W}} \Big[\frac{1}{n} \sum_{i=1}^{n} \big(\eta(x_i;w^0) - \eta(x_i;w)\big)^2 - \\ &\qquad\qquad E\big(\eta(x;w^0) - \eta(x;w)\big)^2 \Big] \\ &= \sup_{\tilde{w} \in [0,1]^d} \Big[\frac{1}{n} \sum_{i=1}^{n} \big(\eta(\tilde{x}_i;\tilde{w}^0) - \eta(\tilde{x}_i;\tilde{w})\big)^2 - \\ &\qquad\qquad E\big(\eta(\tilde{x};\tilde{w}^0) - \eta(\tilde{x};\tilde{w})\big)^2 \Big] \end{aligned} \tag{4.6}$$

where $\tilde{x} = \tilde{F}(x)$, $\tilde{w} = \tilde{F}(w)$, $\tilde{F}(x) = [F_1(x_1) \cdots F_d(x_d)]^{\mathsf{T}}$, and F_j is the (continuous) cdf of x_j. Continuity ensures that the marginals of $\tilde{x}$ are uniform on $[0,1]$, and by the construction, if the components of x are independent, then so are those of $\tilde{x}$ which must then be uniform on $[0,1]^p$. We will assume this in the sequel.

We can find the exceedance probability by first solving a simplified problem in which the target orthant[1] is empty ($y \equiv 0$) and it is desired to learn this with other orthants; this is also equivalent to the multidimensional Kolmogorov-Smirnov test. Then

$$P(\|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)\|_{\mathcal{W}} > \epsilon) \simeq P(\|Z(w)\|_{\mathcal{W}} > b) \tag{4.7}$$

where $Z(w)$ is the zero-mean Gaussian process with (from (4.5))

$$R(w,w') := EZ(w)Z(w') = |w \wedge w'| - |w||w'| \quad ;$$

this is the pinned Brownian sheet. Although the form of (4.7) as a function of b is known as in (2.21), the leading constant, so important for sample size estimates, is not, and finding it is the contribution of this section.

To set the problem up we follow Aldous [6, sec. J16]. The fundamental relation of chapter 3 is

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{EC_b(w)}\, dw \tag{4.8}$$

[1] We stretch the term ‘orthant’ to describe regions like $(-\boldsymbol{\infty}, w]$ because they are translated versions of the negative orthant $(-\boldsymbol{\infty}, \mathbf{0}]$ of points having all coordinates at most zero.

where $\sigma^2(w) = |w|(1 - |w|)$. The numerator is exponential in b while (see result 4.4) the denominator contributes a polynomial in b, so for large b the exponential dominates, and it is most significant where $\sigma(w)$ is largest. This surface, where $|w| = 1/2$, is denoted $\overline{\mathcal{W}}$.

We shall approximate the clump size near $\overline{\mathcal{W}}$ by its value on $\overline{\mathcal{W}}$; to find the latter we appeal to the fact (see below) that $EC_b(w)$ is determined by $R(w, w')$ for w' near w. To find this, take a small $\delta \in R^d$, let $w' = w + \delta$, and partition indices into $\mathcal{J}^+ = \{j \leq d : \delta_j > 0\}$ and $\mathcal{J}^-$. Then for $w \in \overline{\mathcal{W}}$, to terms of first order in δ,

$$\begin{aligned} R(w, w') &= \prod_{j \in \mathcal{J}^+} w_j \cdot \prod_{j \in \mathcal{J}^-} (w_j + \delta_j) - \frac{1}{2} \prod_{j \in \mathcal{J}^+} (w_j + \delta_j) \cdot \prod_{j \in \mathcal{J}^-} (w_j + \delta_j) \\ &= \frac{1}{2} + \sum_{j \in \mathcal{J}^-} \frac{1}{2} \frac{\delta_j}{w_j} - \frac{1}{2} \left(\frac{1}{2} + \sum_{1 \leq j \leq d} \frac{1}{2} \frac{\delta_j}{w_j} \right) + O(\delta^{\mathsf{T}} \delta) \\ &= \frac{1}{4} - \frac{1}{4} \sum_{1 \leq j \leq d} \frac{|\delta_j|}{w_j} + O(\delta^{\mathsf{T}} \delta) \end{aligned}$$

This covariance is locally the same as that of a process

$$\begin{aligned} &Y(w_1, \ldots, w_d) = d^{-1/2} \sum_{j=1}^{d} Y_j(w_j) \\ &EY_j(0) Y_j(t) = 1/4 - \gamma_j |t| \qquad EY_j(t) = 0 \quad . \end{aligned}$$

By appealing to this sort of expansion, Aldous suggests that the clump size decouples as shown below.

4.4 Result *[6, (J10k),(J16e)] If the process $Z(w)$ has a covariance of the form*

$$R(w, w') = \sigma^2 - \sum_{j=1}^{d} \gamma_j |w_j - w'_j| + O((w - w')^{\mathsf{T}} (w - w')) \quad ,$$

the clump size factors as

$$\begin{aligned} &EC(w) = \textstyle\prod_{j=1}^{d} EC_j(w) \\ &EC_j(w) = 1/((\gamma_j/\sigma^2)(b/\sigma)^2) \quad . \end{aligned}$$

Remark. Equation (J10k) of [6] has a typographical error, it is meant that the covariance decomposes as a sum and not a product. To get this result from (J10k), multiply the process in that equation by σ to match variances, and note that the rate of the original process crossing the level b/σ equals that of the scaled process crossing level b. Then use the fundamental relation (3.8) to get the clump size from the clump rate.

For $w \in \overline{\mathcal{W}}$, the factors of the clump size of the pinned Brownian sheet are therefore $EC_j(w) = w_j/4b^2$. Use this and (4.8) to find the probability as follows. Writing $\tilde{w} = [w_1 \cdots w_{d-1}]^\mathsf{T}$,

$$
\begin{aligned}
P(\|Z(w)\| > b) &\simeq \underset{[0,1]^d}{\int \cdots \int} \frac{\bar{\Phi}(b/\sigma)}{EC(w_1, \ldots, w_d)} \, dw_d \, d\tilde{w} \\
&\overset{(a)}{\simeq} \underset{|\tilde{w}|>1/2}{\int \cdots \int} \prod_{i=1}^{d-1} (EC_i(w_i))^{-1} \times \\
&\qquad\qquad \int_{1/2|\tilde{w}|}^{1} \frac{\bar{\Phi}(b/\sigma)}{EC_d(w_d)} \, dw_d \, d\tilde{w} \qquad (4.9) \\
&\overset{(b)}{\simeq} e^{-2b^2} \underset{|\tilde{w}|>1/2}{\int \cdots \int} \prod_{i=1}^{d-1} (EC_i(w_i))^{-1} \, d\tilde{w} \\
&= 4^{d-1} b^{2(d-1)} e^{-2b^2} \underset{|\tilde{w}|>1/2}{\int \cdots \int} \prod_{i=1}^{d-1} (1/w_i) \, d\tilde{w} \quad .
\end{aligned}
$$

In (a) we have restricted the region of integration to those $\tilde{w}$ for which there exists w_d large enough to make $|w| = 1/2$: these are weights with $|\tilde{w}| > 1/2$. At step (b) we have written $\bar{\Phi}$ via its asymptotic expansion (appendix §A.2) and used Laplace's method (§A.3, corollary A.3) with $\sigma_0 = 1/2$ and $-2H = 1/w_d^2$ on $\overline{\mathcal{W}}$.

It remains to find the constant factor

$$
\begin{aligned}
I_d &= \underset{\prod_{i=1}^{d-1} w_i > 1/2}{\int \cdots \int} \left(\textstyle\prod_{i=1}^{d-1} w_i{}^{-1} \right) dw_1 \cdots dw_{d-1} \\
&= \int_{1/2}^{1} \frac{1}{z} \left(\log \frac{1}{z} \right)^{d-2} \frac{dz}{(d-2)!} \qquad (4.10) \\
&= (\log 2)^{d-1} / (d-1)!
\end{aligned}
$$

using the volume element (see §C.2)

$$
\mathrm{vol}(\{w \in [0,1]^d \ : \ z \leq \textstyle\prod_1^d w_i \leq z + dz\}) = dz \, (\log 1/z)^{d-1} / (d-1)!
$$

Combining (4.9) and (4.10) yields

4.5 Result *The PCH estimate of exceedance probability for the pinned Brownian sheet is*

$$
P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \frac{(4 \log 2)^{d-1}}{(d-1)!} b^{2(d-1)} e^{-2b^2} \quad . \qquad (4.11)
$$

Remark. Adler and Brown [3] in theorem 2.1 prove that for large b, this probability is at least the above with 2 in place of $4 \log 2 \approx 2.7$, and in theorems 4.1 and 4.2 they find the same polynomial and exponent as above, although "the style of proof is such that it is impossible to closely monitor inequalities so as to estimate the constants"[3, p. 14]. The dependence on b is a factor of b^2 smaller than Talagrand's upper bound (2.22); this is because $\overline{\mathcal{W}}$ is of dimension $d-1$ rather than d.

For $d = 1$, (4.11) checks with the exact probability e^{-2b^2}, and for $d = 2$ it agrees with the asymptotic calculation of Hogan and Siegmund [32]. The asymptotic value is not known for $d > 2$. Our contribution is in carrying out the computation of the leading constant indicated by Aldous.

4.6 Corollary Result *Let the components of the d-dimensional input x be independent with continuous distribution, and let $y \equiv 0$. Suppose the family of orthants is used to learn y. Then*

$$n_c = \frac{d}{\epsilon^2} \quad .$$

samples are sufficient for reliable generalization.

Proof. At the critical value of n the probability (4.11) with $b = \epsilon\sqrt{n}$ drops below unity. Taking the log, applying Stirling's formula, and assuming $d \gg 1$ yields the criterion

$$\log(4e \log 2) + \log(b^2/d) - 2b^2/d \leq 0$$

which occurs at $b^2/d = 1.02 \approx 1$. □

4.7 Corollary Result *Let the components of the d-dimensional input x be independent with continuous distribution, and let $y = 1$ on some orthant. Suppose the family of orthants is used to learn y. Then*

$$n_c \leq 11\frac{d}{\epsilon^2}$$

samples are sufficient for reliable generalization.

Proof. Let Z be the pinned set-indexed Brownian sheet introduced in §4.1. A value of n slightly larger than n_c ensures that with high probability

$$(\forall w \in \mathcal{W})\, |Z([\mathbf{0}, w])| < b/\sqrt{11} \quad . \tag{4.12}$$

The function y to be learned corresponds, after the nonlinear mapping (4.6), to a fixed orthant $[\mathbf{0}, w_0]$, and the region of disagreement between y and $\eta(w; x)$ is now $[\mathbf{0}, w_0] \triangle [\mathbf{0}, w]$, which is not an orthant. However, by additivity of the set-indexed process,

$$Z([\mathbf{0}, w_0] \triangle [\mathbf{0}, w]) = Z([\mathbf{0}, w_0]) + Z([\mathbf{0}, w]) - 2Z([\mathbf{0}, w_0] \cap [\mathbf{0}, w]) \quad .$$

The sets indexing the RHS are all orthants, so (4.12) and the triangle inequality imply that with high probability the last two terms have magnitude less than $3b/\sqrt{11}$ uniformly in w.

The first term is a fixed normal random variable with variance at most 1/4, so

$$\begin{aligned} P\big(Z([\mathbf{0}, w_0]) > (1 - 3/\sqrt{11})b\big) &\leq \bar{\Phi}(2(1 - 3/\sqrt{11})b) \\ &\leq \bar{\Phi}(.19\sqrt{11d}) \\ &\leq \tfrac{1}{2}e^{-d/5} \end{aligned}$$

which is negligible for large d. So with high probability,

$$(\forall w \in \mathcal{W})\, |Z([\mathbf{0}, w_0] \bigtriangleup [\mathbf{0}, w])| \leq b \quad . \qquad \square$$

The VC dimension of the class of orthants in R^d is d (table 2.1), so results 4.6 and 4.7 are directly comparable to the Vapnik bound of (2.2): $n_c = (9.2v/\epsilon^2)\log 8/\epsilon$. The functional form of the new PCH-based result eliminates the undesirable $\log 1/\epsilon$ term, as well as yielding smaller sample size estimates (even using the imprecise shortcut of result 4.7).

§4.3 Learning rectangles

Now we solve the related problem of learning rectangles. Again first let $y \equiv 0$ but now classify the data using $\eta(x; w) = 1_{[u,v]}(x)$, where $u \leq v \in R^d$ and $w = [u^\mathsf{T} v^\mathsf{T}]^\mathsf{T} \in R^{2d}$. Supposing the cdf F of the random x is continuous and independent across coordinates allows us to assume without further loss of generality that x is uniform on $[0,1]^p$. Using (4.5), the derived process $Z(w)$ is zero-mean Gaussian with

$$\begin{aligned} R(w, w') &:= EZ(w)Z(w') \\ &= |(v \wedge v') - (u \vee u')| - |v - u||v' - u'| \quad ; \end{aligned}$$

Writing down (4.8), we see that the maximum-variance region $\overline{\mathcal{W}} := \{w : |v - u| = 1/2\}$ dominates the probability.

Again we estimate $EC_b(w)$ by examining the covariance locally about points in $\overline{\mathcal{W}}$. To do this write $v' = v + \delta^v$, $u' = u + \delta^u$, and partition the d indices into four sets $\mathcal{J}^{+-} = \{j : \delta_j^v \geq 0 > \delta_j^u\}$, etc. Proceeding as before,

$$\begin{aligned} R(w, w') &= |(v \wedge v') - (u \vee u')| - \frac{1}{2}|v' - u'| \\ &= \prod_{j \in \mathcal{J}^{++}} ((v_j - u_j) - \delta_j^u) \cdot \prod_{j \in \mathcal{J}^{--}} ((v_j - u_j) + \delta_j^v) \times \\ &\qquad \prod_{j \in \mathcal{J}^{+-}} (v_j - u_j) \cdot \prod_{j \in \mathcal{J}^{-+}} ((v_j - u_j) + (\delta_j^v - \delta_j^u)) \\ &\qquad - \frac{1}{2} \prod_{1 \leq j \leq d} ((v_j - u_j) + (\delta_j^v - \delta_j^u)) \\ &= \frac{1}{4} - \frac{1}{4} \sum_{1 \leq j \leq d} \frac{|\delta_j^v|}{v_j - u_j} - \frac{1}{4} \sum_{1 \leq j \leq d} \frac{|\delta_j^v|}{v_j - u_j} + O(\delta^\mathsf{T} \delta) \end{aligned}$$

Just as in the case of orthants, the covariance has a cusped shape at w and the clump size factors. By result 4.4, for $w \in \overline{\mathcal{W}}$,

$$EC_b(u, v) = \textstyle\prod_{i=1}^d EC_i^u(u, v) EC_i^v(u, v) \tag{4.13a}$$

$$EC_i^u(u, v) = EC_i^v(u, v) = (v_i - u_i)/4b^2 \quad . \tag{4.13b}$$

Let $\tilde{u} = [u_1 \cdots u_{d-1}]^\mathsf{T}$ and $\tilde{v} = [v_1 \cdots v_{d-1}]^\mathsf{T}$. The PCH procedure would have us find (suppressing some function arguments)

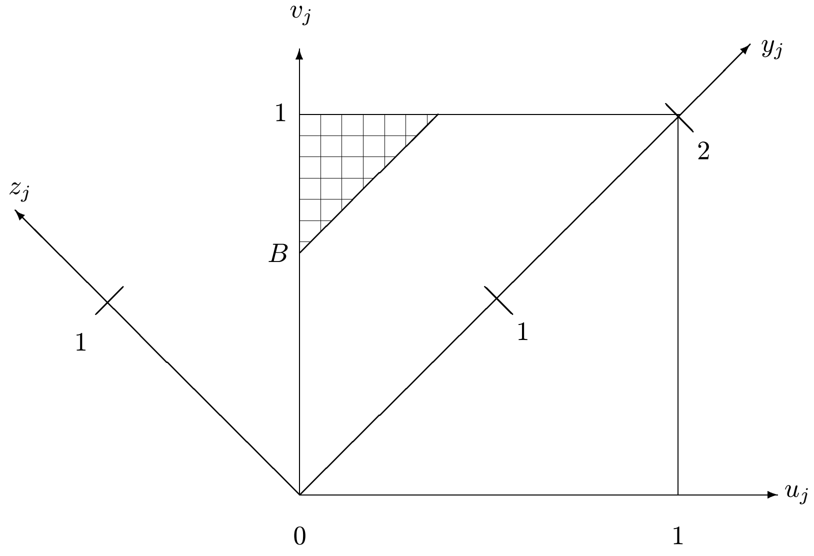

Figure 4.1: The region of integration in (4.14)

The boundary B is at the value $B = 1/\big(2\prod_{k=1}^{j-1}(v_k - u_k)\big)$.

$$
\begin{aligned}
P(\|Z(w)\| > b) &\simeq \int\cdots\int_{0\le u\le v\le 1} \frac{\bar{\Phi}(b/\sigma)}{EC(u_1, v_1, \ldots, u_d, v_d)} du_d\, dv_d \cdots du_1\, dv_1 \\
&\overset{(a)}{=} \int\cdots\int_{\substack{0\le\tilde{u}\le\tilde{v}\le 1\\ |\tilde{v}-\tilde{u}|\ge 1/2}} \prod_{j=1}^{d-1} \frac{du_j\, dv_j}{EC_j^u EC_j^v} \int_{1/2|\tilde{v}-\tilde{u}|}^{1} \frac{dv_d}{EC_d^v} \int \frac{\bar{\Phi}(b/\sigma)}{EC_d^u} du_d \\
&\overset{(b)}{\simeq} e^{-2b^2} \int\cdots\int_{\substack{0\le\tilde{u}\le\tilde{v}\le 1\\ |\tilde{v}-\tilde{u}|\ge 1/2}} \prod_{j=1}^{d-1} \frac{du_j\, dv_j}{EC_j^u EC_j^v} \int_{1/2|\tilde{v}-\tilde{u}|}^{1} \frac{dv_d}{EC_d^v} \\
&\overset{(c)}{\simeq} 2(4b^2)^{2d-1} e^{-2b^2} \int\cdots\int_{\substack{0\le\tilde{u}\le\tilde{v}\le 1\\ |\tilde{v}-\tilde{u}|\ge 1/2}} \prod_{j=1}^{d-1} \frac{du_j\, dv_j}{(v_j - u_j)} \int_{1/2|\tilde{v}-\tilde{u}|}^{1} dv_d
\end{aligned}
$$

where (a) follows from restricting the region of integration to $\overline{\mathcal{W}}$, or those $\tilde{u}$, $\tilde{v}$ for which a final component u_d, v_d exists for which $|v-u| = 1/2$. As in the case of orthants, (b) results after using the asymptotic expansion for $\bar{\Phi}$ and applying Laplace's method (corollary A.3) with $\sigma_0 = 1/2$ and $-2H = 1/(v_d - u_d)^2$ on $\overline{\mathcal{W}}$. Finally, relation (c) just uses the clump size of (4.13) and that $w \in \overline{\mathcal{W}}$.

The remaining factor is expressed as an iterated integral using the property that at stage j of integration, $v_j - u_j$ must be large enough to have $\prod_{k=1}^{j}(v_k - u_k) \ge 1/2$. As long as the product is this big, successful choices for the remaining variables indexed $j+1$ through d can be made.

The integral is

$$I_d = \iint\limits_{\substack{0\le u_1,\, v_1 \le 1 \\ v_1-u_1\ge 1/2}} \frac{du_1\,dv_1}{(v_1-u_1)} \iint\limits_{\substack{0\le u_2,\, v_2 \le 1 \\ v_2-u_2\ge 1/2(v_1-u_1)}} \frac{du_2\,dv_2}{(v_2-u_2)} \times\cdots\times$$

$$\iint\limits_{\substack{0\le u_{d-1},\, v_{d-1} \le 1 \\ v_{d-1}-u_{d-1}\ge \\ 1/2\prod_1^{d-2}(v_k-u_k)}} \frac{du_{d-1}\,dv_{d-1}}{(v_{d-1}-u_{d-1})} \int\limits_{\substack{v_d\le 1 \\ v_d\ge 1/2\prod_1^{d-1}(v_k-u_k)}} dv_d \tag{4.14}$$

$$= \int_{1/2}^{1} dz_1 \frac{1-z_1}{z_1} \int_{1/2z_1}^{1} dz_2 \frac{1-z_2}{z_2} \cdots \int_{1/2z_1\cdots z_{d-2}}^{1} dz_{d-1} \frac{1-z_{d-1}}{z_{d-1}} \int_{1/2z_1\cdots z_{d-1}}^{1} dz_d$$

since the $d-1$ integrals over $y_j = v_j + u_j$ (see figure 4.1) are easily performed. The dependence on $\prod(1-z_j)$ as well as $\prod z_j$ makes a simple transformation of variables impossible. To evaluate the constant we note that $I_d = I_d(1)$ under the recursive definition

$$I_1(z) := \int_{1/2z}^{1} dy \tag{4.15a}$$

$$I_d(z) := \int_{1/2z}^{1} \frac{1-y}{y} I_{d-1}(zy)\,dy \qquad (d>1). \tag{4.15b}$$

4.8 Lemma

$$I_d(z) = \frac{1}{2z}\frac{(\log 2z)^{2d-1}}{(2d-1)!} M(d, 2d; \log 2z)$$

where the confluent hypergeometric or Kummer function [1, ch. 13]

$$M(a,b;z) = \sum_{k=0}^{\infty} \frac{a^{\bar{k}}}{b^{\bar{k}}} \frac{z^k}{k!}$$

and $a^{\bar{k}} = a(a+1)\cdots(a+k-1)$ *is the* k*th rising power of* a. *Further,*

$$\sqrt{2} \le M(d, 2d; \log 2) \le 2 \quad .$$

Proof. The intricate calculation is in the appendix, §C.3. □

Combining these results yields

4.9 Result *The PCH estimate of exceedance probability for the rectangle indexed Brownian sheet is*

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq 2M(d,2d;\log 2)\frac{(4\log 2)^{2d-1}}{(2d-1)!} b^{2(2d-1)} e^{-2b^2}$$

where the Kummer function M *lies between* $\sqrt{2}$ *and* 2.

Remark. By the sandwich bounds of Adler and Samorodnitsky [4, ex. 3.2] the dependence on b above is correct. The VC dimension of rectangles is $2d$, so the result is of the same form as Talagrand's, although again with a smaller polynomial factor. This is an extension of Aldous's sketch for $d = 2$, although his equation (J17d) seems to be incorrect.

4.10 Corollary Result *Let the components of the d-dimensional input x be independent with continuous distribution, and let $y \equiv 0$. Suppose the family of rectangles is used to learn y. Then*

$$n_c = \frac{2d}{\epsilon^2}$$

samples are sufficient for reliable generalization.

Proof. We discard the leading factor of $2M(d, 2d; \log 2)$ since it is not important for d large. What remains is precisely the estimate of result 4.5 with $2d$ for d. □

4.11 Corollary Result *Let the components of the d-dimensional input x be independent with continuous distribution, and let $y = 1$ on some rectangle. Suppose the family of rectangles is used to learn y. Then*

$$n_c = 11\frac{2d}{\epsilon^2}$$

samples are sufficient for reliable generalization.

Proof. Similar to corollary 4.7. Let process Z be the pinned set-indexed Brownian sheet. If n is slightly larger than n_c then with high probability all rectangles have $|Z([u, v])| \leq b/\sqrt{11}$. Additivity of the pinned Brownian sheet implies that the mismatch process satisfies

$$Z([u_0, v_0] \triangle [u, v]) = Z([u_0, v_0]) + Z([u, v]) - 2Z([u_0, v_0] \cap [u, v]) \quad ,$$

and since the intersection of two rectangles is again a rectangle, the triangle inequality shows that the last two terms are uniformly bounded by $3b/\sqrt{11}$. A simple pointwise bound on $P\big(Z([u_0, v_0]) > (1-3/\sqrt{11})b\big)$ completes the argument. □

We note the satisfying result that as the VC dimension increases from d (orthants) to $2d$ (rectangles), the predicted sample size increases in the same way. As in the case of learning orthants, the PCH sample size estimates are of the order v/ϵ^2, without the extra $\log 1/\epsilon$, and the sample size estimates are notably smaller than the VC bound, especially in the first result which is more carefully calculated. Our benchmark example of $v = 50$, $\epsilon = 0.1$ gives in this case $n_c = 5000$ versus the VC prediction of 202 000.

§4.4 Learning hyperplanes

Now we analyze the most fundamental example of a neural network: the perceptron, or linear threshold unit. Suppose that the data distribution P is rotationally invariant, and that $\eta(x;w) = 1_{[0,\infty)}(w^{\mathsf{T}}x)$; this is a homogeneous linear threshold unit. Further, let $y = \eta(x;w^0)$. The networks are invariant to a positive scaling of the weights, so we assume $w^{\mathsf{T}}w = 1$, i.e. $\mathcal{W}$ is the surface of the unit ball in R^{d+1} and there are d free parameters. Without loss of generality we can take $w^0 = [1\,0\cdots0]^{\mathsf{T}}$; the problem is invariant with respect to rotations about the axis w^0. In this section we define

$$|w|_2 := (w^{\mathsf{T}}w)^{1/2} \quad ; \tag{4.16}$$

this is distinct from the earlier notation $|w|$.

As before we proceed by identifying the set of maximum-variance points and finding the clump size there. Letting A_w be the halfspace defined by the normal vector w, oriented to contain that vector, we see that

$$\sigma^2(w) = P(A_{w^0} \bigtriangleup A_w)\big(1 - P(A_{w^0} \bigtriangleup A_w)\big)$$

is maximized when $w^{\mathsf{T}}w^0 = 0$, which defines the set $\overline{\mathcal{W}}$. The clump size is constant on $\overline{\mathcal{W}}$ by rotational invariance. To find $EC_b(w)$, develop a local approximation for the covariance about any point $w \in \overline{\mathcal{W}}$ which we take as $w = [0\cdots0\,1]$ without loss of generality. Then $w' = [w'_1 \cdots w'_d \; \delta]$ for appropriate $\delta \approx 1$, and using (4.4), the covariance becomes

$$\begin{aligned} R(w,w') &= \tfrac{1}{4} - \tfrac{1}{2}P\big((A_w \bigtriangleup A_{w^0}) \bigtriangleup (A_{w'} \bigtriangleup A_{w^0})\big) \\ &= \tfrac{1}{4} - \tfrac{1}{2}P(A_w \bigtriangleup A_{w'}) \\ &= \tfrac{1}{4} - \tfrac{1}{2\pi}\cos^{-1}(w^{\mathsf{T}}w') \\ &= \tfrac{1}{4} - \tfrac{1}{2\pi}\cos^{-1}(\delta) \\ &= \tfrac{1}{4} - \tfrac{1}{2\pi}\Big(\sum_{k=1}^{d} {w'_k}^2\Big)^{1/2} + O({w'_1}^2 + \cdots + {w'_d}^2) \end{aligned} \tag{4.17}$$

where we have used the Taylor series for $\cos^{-1}$. This covariance is not of the same form as the Brownian bridge.

Aldous suggests

4.12 Result *[6, sec. J18] The clump size for a so-called* isotropic process *having covariance*

$$R(w,w') = \sigma^2 - \gamma\,|w-w'|_2 + o(|w-w'|_2)$$

for $w \in R^d$ is

$$EC_b(w) \simeq 1/(K_{d,1}\,(\gamma/\sigma^2)^d(b/\sigma)^{2d}) \tag{4.18}$$

$$d^{-d/2} \le K_{d,1} \le 1 \tag{4.19}$$

where $K_{d,1}$ is the isotropic constant.

Proof. Scale the process of section J18 by σ to match its variance to the one above. The rate of the original process crossing the level b/σ equals that of the scaled process crossing level b. The fundamental relation (3.8) then allows the clump size to be found from the clump rate. □

In such a case the covariance does not decouple so the clump size does not factor. For the process at hand, for $w \in \overline{\mathcal{W}}$,

$$(EC_b(w))^{-1} = (8/\pi)^d \, K_{d,1} \, b^{2d} \quad . \tag{4.20}$$

The exceedance probability is then easily written down via (3.9) as an integral over the effectively d-dimensional weight space

$$\begin{aligned} P(\|Z(w)\|_{\mathcal{W}} > b) &\simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)} \, dw \\ &\overset{\text{(a)}}{=} \int_{-1}^{1} \big[\kappa_d \, d \, (1 - w_1^2)^{(d-1)/2}\big] \frac{\bar{\Phi}(b/\sigma(w_1))}{EC_b(w_1)} \, dw_1 \\ &\overset{\text{(b)}}{\simeq} [\kappa_d \, d] \frac{1}{EC_b(w_1 = 0)} \frac{\pi}{8b^2} \exp^{-2b^2} \\ &\simeq \frac{\pi}{4} \, \frac{8^d}{\pi^{d/2}\Gamma(d/2)} K_{d,1} \, b^{2d-2} e^{-2b^2} \quad . \end{aligned} \tag{4.21}$$

The original integrand is a function of w_1 only, and at step (a) we have performed the integral over the other $d-1$ coordinates resulting in the bracketed factor, the surface area of a sphere of squared radius $1 - w_1^2$ embedded in R^d. (Recall κ_d is the volume of the unit sphere in R^d.) Relation (b) follows on writing $\bar{\Phi}(b/\sigma)$ via its asymptotic expansion and applying Laplace's method (corollary A.3) to the remaining one-dimensional integral with $-H/2 = 1/\pi^2$ and $\sigma_0 = 1/2$.

Remark. The VC dimension of this classifier architecture is $p = d + 1$, so the bound of Talagrand has the same exponent but its polynomial is too large by a factor of b^4. This method of calculation and the constant are new results.

The probability approximation leads directly to

4.13 Corollary Result *Let x have a radially symmetric distribution. Suppose we desire to learn a halfspace y with a perceptron architecture. Then the critical sample size satisfies*

$$\frac{1.3d}{\epsilon^2} \le n_c \le \frac{(1.36 + (1/3)\log d)d}{\epsilon^2} \quad . \tag{4.22}$$

Proof. Use the bounds for $K_{d,1}$ in result 4.12. See §C.4 for details. □

Again we see that the sufficient sample size is a constant, quite close to unity, times the number of parameters divided by ϵ^2. (The term $1/3 \log d$ does not play a significant role as it lies between 1.36 and 2.72 for $60 \le d \le 3500$.)

§4.5 Learning smooth functions

Now we pass to a qualitatively different regime in which $Z(w)$ is smooth enough to admit a local approximation, allowing direct computation of the clump size. By smooth, we mean that $Z(w)$ should have two derivatives, at least in the mean-square sense, in w. In neural net regression (as distinct from classification), this is often assured since the network is typically made of sigmoidal functions

$$\sigma(x; w) = \frac{1}{1 + e^{-w^{\mathsf{T}} x}} \tag{4.23}$$

that have bounded derivatives of all orders. In such a case the original process $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$ has two almost sure derivatives if the second moments of x and y exist finite.

We can then write a quadratic approximation to Z in the vicinity of a point w_0:

$$Z(w) \simeq Z_0 + (w - w_0)^{\mathsf{T}} G + \frac{1}{2}(w - w_0)^{\mathsf{T}} \mathbf{H}(w - w_0) \tag{4.24}$$

where the gradient $G = \nabla Z(w)$ and Hessian matrix $\mathbf{H} = \nabla\nabla Z(w)$ are both evaluated at w_0. One pictures a downward-turning parabola peaking near w_0 which attains height at least b; the clump size is the volume V of the ellipsoid in R^d formed by the intersection of the parabola with the level b. Provided $Z_0 \geq b$ and $\mathbf{H} < 0$, simple computations reveal that

$$V = \kappa_d \, \frac{(2(Z_0 - b) - G^{\mathsf{T}} \mathbf{H}^{-1} G)^{d/2}}{|{-\mathbf{H}}|^{1/2}} \quad . \tag{4.25}$$

We then wish to approximate

$$EC_b(w_0) \simeq E[V \mid Z(w_0) > b] \quad . \tag{4.26}$$

There are two issues here. The evident one is that since $C_b(w_0)$ is defined on a different probability space than V (which is derived from $Z(\cdot)$), these two random variables can only be equal in distribution. The subtle distinction is that the condition $Z(w_0) > b$ is not precisely equivalent to occurrence of a *clump center* at b, which conditions the event on the left above; in fact, the latter implies the former. However, it is apparent that the two events are closely related so that the approximation is reasonable. (We shall have more to say on the tightness of this approximation in chapter 5.)

The conditional mean of V is computed as follows. The same argument used to show that $Z(w)$ is approximately normal shows that G and $\mathbf{H}$ are approximately normal too. In fact,

$$\begin{aligned} E[\mathbf{H} \mid Z(w_0) = z] &= \frac{-z}{\sigma^2(w_0)} \Lambda_{02}(w_0) \\ \Lambda_{02}(w_0) &:= -E\, Z(w_0)\mathbf{H} \\ &= -\nabla_w \nabla_w R(w_0, w)\big|_{w=w_0} \end{aligned}$$

so that, since b (and hence z) is large, the second term in the numerator of (4.25) may be neglected. (The notation Λ_{02} is mnemonic for expectation of the product of zeroth and second derivatives of $Z(w)$.) We take the additional step of replacing the random Hessian by its mean, leaving

$$E[V \mid Z(w_0) = z] \simeq \kappa_d \sqrt{2}^d \left| \frac{\Lambda_{02}(w_0)}{\sigma^2(w_0)} \right|^{-1/2} \left(\frac{z - b}{z} \right)^{d/2} \quad .$$

The exceedance volume is then found by integrating out on z:

$$E[V \mid Z(w_0) > b] \simeq \kappa_d \sqrt{2}^d \left| \frac{\Lambda_{02}(w_0)}{\sigma^2(w_0)} \right|^{-1/2} I$$

where the remaining factor is (abbreviating $\sigma = \sigma(w_0)$)

$$\begin{aligned} I &:= \frac{1}{\sigma\,\bar{\Phi}(b/\sigma)} \int_b^\infty \left(\frac{z-b}{z}\right)^{d/2} e^{-z^2/2\sigma^2}\,dz \\ &\simeq \frac{b}{\sigma^2} \int_0^\infty \left(\frac{z}{z+b}\right)^{d/2} e^{-zb/\sigma^2} e^{-z^2/2\sigma^2}\,dz \\ &= \int_0^\infty \left(\frac{x}{x+b^2/\sigma^2}\right)^{d/2} e^{-x} e^{-x^2\sigma^2/2b^2}\,dx \\ &= \left(\frac{\sigma}{b}\right)^d \int_0^\infty \left[\left(1+\frac{\sigma^2 x}{b^2}\right)^{-d/2} e^{-x^2\sigma^2/2b^2}\right] x^{d/2} e^{-x}\,dx \quad . \end{aligned}$$

We again use $b \gg \sigma$ to justify the asymptotic expansion for $\bar{\Phi}$. The bracketed quantity is monotone increasing in b and has unity as its pointwise limit, so dominated convergence yields

$$EC_b(w_0) \simeq \kappa_d \sqrt{2}^d \left|\frac{\Lambda_{02}(w_0)}{\sigma^2(w_0)}\right|^{-1/2} \left(\frac{\sigma}{b}\right)^d \Gamma(d/2+1) \qquad (4.27)$$

where the RHS is both the asymptotic value and an upper bound. This is what we need for

4.14 Result *Let the network activation functions be twice continuously differentiable, and let $b \gg \sigma(w)$. Then provided $\Lambda_{02}(w) > 0$,*

$$EC_b(w) \simeq (2\pi)^{d/2} \frac{(\sigma(w)/b)^d}{|\Lambda_{02}(w)/\sigma^2(w)|^{1/2}} \quad .$$

The RHS is both the asymptotic value and an approximate upper bound.

Proof. Substitute κ_d into (4.27) and use $\Gamma(d/2+1) = (d/2)\Gamma(d/2)$. □

Remark. The relationship of this clump size to those in results 4.4 and 4.12 can be seen by expanding the covariance around w_0:

$$\begin{aligned} &R(w_0+w, w_0+w') = \\ &\quad \sigma_0^2 - \frac{1}{2}\begin{bmatrix} w^\mathsf{T} & w'^\mathsf{T} \end{bmatrix} \begin{bmatrix} \Lambda_{02} & -\Lambda_{11} \\ -\Lambda_{11} & \Lambda_{02} \end{bmatrix} \begin{bmatrix} w \\ w' \end{bmatrix} + o(w^\mathsf{T} w + w'^\mathsf{T} w') \qquad (4.28) \end{aligned}$$

where

$$\Lambda_{11} := E\,GG^\mathsf{T} = \nabla_w \nabla_{w'} R(w, w')|_{w=w'=w_0} \geq 0 \qquad (4.29)$$

Of course $R(w_0, w_0+w) = \sigma_0^2 - \frac{1}{2} w^\mathsf{T} \Lambda_{02}\, w$ so again it is the behavior of the covariance about w_0, in this case captured by Λ_{02}, that determines the clump size. In this case the sample paths are more regular so the clumps are larger, of order $1/b^d$ rather than this squared as with the nondifferentiable processes in previous sections.

Remark. At a local variance maximum, it is easy to see from (4.28) that $\Lambda_{02}(w) > 0$, so the expression for clump size is well-defined.

Substituting into the fundamental equation (3.9) yields

$$\begin{aligned} P(\|Z(w)\| > b) &\simeq (2\pi)^{-\frac{d}{2}} \int\limits_{\mathcal{W}} \left|\frac{\Lambda_{02}(w)}{\sigma^2(w)}\right|^{1/2} \left(\frac{b}{\sigma(w)}\right)^{d} \bar{\Phi}\left(\frac{b}{\sigma(w)}\right) dw \\ &\simeq (2\pi)^{-\frac{d+1}{2}} \int\limits_{\mathcal{W}} \left|\frac{\Lambda_{02}(w)}{\sigma^2(w)}\right|^{1/2} \left(\frac{b}{\sigma(w)}\right)^{d-1} e^{-\frac{b^2}{2\sigma^2(w)}} dw \end{aligned} \tag{4.30}$$

where use of the asymptotic expansion $\bar{\Phi}(z) \simeq (z\sqrt{2\pi})^{-1}\exp(-z^2/2)$ is justified since $(\forall w) b \gg \sigma(w)$ is necessary to have each individual probability $P(Z(w) \geq b)$ low—let alone the supremum. To proceed further, we need some information about the variance $\sigma^2(w)$ of $(y - \eta(x;w))^2$. In general this must come from the problem at hand, but suppose for example the process has a unique variance maximum $\bar{\sigma}^2$ at $\bar{w}$. Then the d-dimensional integral can be approximated, yielding

4.15 Result *Let the network activation functions be twice continuously differentiable. Let the variance have a unique maximum $\bar{\sigma}$ at $\bar{w}$ in the interior of $\mathcal{W}$ and the level $b \gg \bar{\sigma}$. Then the PCH estimate of exceedance probability is*

$$\begin{aligned} P(\|Z(w)\|_{\mathcal{W}} > b) &\simeq \frac{|\Lambda_{02}(\bar{w})|^{1/2}}{|\Lambda_{02}(\bar{w}) - \Lambda_{11}(\bar{w})|^{1/2}} \frac{\bar{\sigma}/b}{\sqrt{2\pi}} e^{-b^2/2\bar{\sigma}^2} \\ &\simeq \frac{|\Lambda_{02}(\bar{w})|^{1/2}}{|\Lambda_{02}(\bar{w}) - \Lambda_{11}(\bar{w})|^{1/2}} \bar{\Phi}(b/\bar{\sigma}) \quad . \end{aligned} \tag{4.31}$$

Furthermore, $\Lambda_{02} - \Lambda_{11}$ is positive-definite at $\bar{w}$; it is $-1/2$ the Hessian of $\sigma^2(w)$. The leading constant thus strictly exceeds unity.

Proof. Applying Laplace's method (corollary A.3) to the integral (4.30) results in (4.31) with a leading factor $|\Lambda_{02}(\bar{w})|^{1/2}/|-\nabla\nabla\sigma^2(\bar{w})/2|^{1/2}$. The indicated derivative is

$$\begin{aligned} \nabla\nabla E\, Z(w)^2 &= E\, \nabla\nabla Z(w)^2 \\ &= 2E\, \nabla Z(w) \nabla^{\mathsf{T}} Z(w) + 2E\, Z(w) \nabla\nabla Z(w) \\ &= 2(\Lambda_{11}(w) - \Lambda_{02}(w)) \quad . \end{aligned}$$

At a strict maximum of variance, this must be negative-definite. □

The above probability is just $P(Z(\bar{w}) > b)$ multiplied by a factor to account for the other random variables in the supremum.

The estimate of exceedance probability is easily turned into a sample size estimate:

4.16 Corollary Result *Let the network activation functions be twice continuously differentiable. Let* $\mathrm{Var}\big((y - \eta(x;w))^2\big)$ *have a unique maximum $\bar{\sigma}^2$ at $\bar{w}$ in the interior of $\mathcal{W}$ and the level $b \gg \bar{\sigma}$. Then the critical sample size for reliable generalization is*

$$n_c = \frac{d\bar{\sigma}^2 \log K}{\epsilon^2}$$

$$K^d = |\Lambda_{02}(\bar{w})| \,/\, |\Lambda_{02}(\bar{w}) - \Lambda_{11}(\bar{w})| = \left|\mathbf{I}_d - \Lambda_{02}(\bar{w})^{-1}\Lambda_{11}(\bar{w})\right|^{-1} \quad .$$

Proof. For large d, the factor of $\bar{\sigma}/b\sqrt{2\pi}$ in result 4.15 is negligible. The remainder drops below unity at the given value of $b^2 = n\epsilon^2$. □

Again we see the $O(d/\epsilon^2)$ dependence. The constant depends rather weakly on the distribution and architecture, via the logarithm of K, the reciprocal of the geometric mean of the eigenvalues of the matrix $\mathbf{I}_d - \Lambda_{02}(\bar{w})^{-1}\Lambda_{11}(\bar{w})$.

In the complementary situation where the process is normalized by its standard deviation $\sigma(w)$, the variance is unity and all networks make a contribution to the exceedance probability. When the clump size is put into the fundamental equation (3.9), we find

4.17 Result *Let the network activation functions be twice continuously differentiable, and the level $b \gg 1$. Then the PCH estimate of exceedance probability is*

$$P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \simeq (2\pi)^{d/2}\, b^d\; \bar{\Phi}(b) \int_{\mathcal{W}} \left|\frac{\Lambda_{11}}{\sigma^2(w)}\right|^{1/2} dw \quad . \qquad (4.32)$$

The probability estimate on the right is correctly invariant to scaling both $Z(\cdot)$ and the index set $\mathcal{W}$.

Proof. By result 4.14, the clump size at some point w_0 is a function of the second derivative of the covariance at that point; call the corresponding quantity for the normalized process $\tilde{\Lambda}_{02}$. It is given by

$$\begin{aligned}\tilde{\Lambda}_{02} &= E\,\frac{Z(w_0)}{\sigma(w_0)}\,\nabla\nabla\frac{Z(w)}{\sigma(w)}\Big|_{w=w_0}\\ &= \nabla\nabla\frac{R(w_0,w)}{\sigma(w_0)\sigma(w)}\Big|_{w=w_0} \quad .\end{aligned}$$

The expansion (4.28) tells us

$$\begin{aligned}\sigma(w_0 + w) &\simeq \sigma_0\big(1 - \tfrac{1}{2}w^{\mathsf{T}}(\Lambda_{02} - \Lambda_{11})w/\sigma_0^2\big)\\ R(w_0, w) &\simeq \sigma_0^2\big(1 - \tfrac{1}{2}w^{\mathsf{T}}\Lambda_{02}\, w/\sigma_0^2\big)\end{aligned}$$

and after some routine algebra we find $\tilde{\Lambda}_{02} = \Lambda_{11}/\sigma_0^2 \geq 0$. The result is obtained by substituting the clump size into the fundamental equation. □

In place of the estimate of a constant multiple of $\bar{\Phi}(b)$ in the case of a unique variance maximum we now have a factor of b^d to account for the other, now more prominent, variables. The corresponding sample size estimate is

4.18 Corollary Result *Let the network activation functions be twice continuously differentiable. Then the critical sample size for reliable generalization of the normalized process is*

$$\begin{aligned}n_c &= \frac{d(3.2 + 1.2\log d + 2.4\log K)}{\epsilon^2}\\ K^d &:= \int_{\mathcal{W}} \left|\frac{\Lambda_{11}(w)}{\sigma^2(w)}\right|^{1/2} dw \quad .\end{aligned}$$

Above this sample size, with high probability,

$$(\forall w \in \mathcal{W})\, |Z(w)| \leq \epsilon\sigma(w) \quad .$$

Proof. See §C.5. □

If $d > 20$, say, then in order for K to make a significant contribution to the sample size, it must be above about 17, or the integral must be on the order of 20^d. So although K cannot be calculated directly in any but trivial problems, the required sample size is not highly sensitive to changes in the data distribution or the architecture. We obtain a sample size on the order of $d(3.2 + 1.2 \log d)/\epsilon^2$ over a fairly broad class of problems.

§4.6 Summary and Conclusions

In this chapter we have demonstrated how to use the PCH in an idealized setting to estimate sample sizes needed for reliable generalization. If analytic information about the process of exceedances is available, which it may be if the architecture and the target function are simple, the maximum-variance points can be characterized and the mean clump size computed there. This size is typically proportional to $1/b^{2d/\alpha}$ where $\alpha = 1$ for a rough (nondifferentiable) process and $\alpha = 2$ for a smooth one. In the most general terms, the constant of proportionality depends on the data distribution and the architecture, and for a given problem the constant is a function of the location of the clump in weight space. More specifically, the constants depend on the local behavior of the covariance $R(w_0, w_0 + w)$ about the point of interest. To reinforce this idea we summarize results 4.4, 4.12, and 4.14:

Name	$1 - R(w_0, w_0 + w)$	$EC_b(w_0)$
Cusped	$\sum_{j=1}^{d} \gamma_j \lvert w_j \rvert$	$1/\big(\prod_{j=1}^{d} \gamma_j b^2\big)$
Rough Isotropic	$\big(\sum_{j=1}^{d} (\gamma\, w_j)^2\big)^{1/2}$	$1/(K_{d,1}\gamma^d b^{2d})$
Smooth	$w^\mathsf{T} \Lambda_{02}\, w/2$	$(2\pi)^{d/2}/(\lvert\Lambda_{02}\rvert^{1/2}\, b^d)$

The exceedance probability is calculated from the variance and clump size information in a way that is generally complex if exact results are desired. However, the probability is for our purposes fairly closely determined by qualitative factors like the maximum variance of the process (which defines the exponential factor $\bar{\Phi}(b/\bar{\sigma})$), whether the process is smooth or rough (determining the exponent of the leading polynomial in b), and how many weights there are (which controls the exponent of any leading constant factors). We saw this especially in corollaries 4.13 and 4.18, where the unknown or loosely-bounded constant factors had a small effect on the sample size.

The results we have found (all except one for unnormalized processes) agree in functional form with known bounds in the empirical process literature; agreement also holds in the few low-dimensional cases for which precise asymptotics are known. On the other hand, in the context of the neural network literature where the Vapnik bound is the only tool allowing explicit estimates, then new sample size approximations have eliminated a wasteful factor of $\log 1/\epsilon$ as well as coming equipped with tight constants. The results for orthants, rectangles, and linear threshold units are quite satisfying.

Of course, if they are to be useful in practice, sample size estimates cannot rely on detailed knowledge of the data distribution and the target function. Having in this chapter shown that the PCH method gives informative sample size bounds, we turn in the second part of this work to ways of making it practical.

5 Approximating Clump Size

WE HAVE SEEN that the Poisson clumping technique allows computation of the sample size needed for reliable generalization when sufficient analytic information is known about the process to approximate it locally and thus compute its clump size. In practice such detailed information is not available, and as a remedy, in this chapter we work in steps towards a tractable approximation to clump size. First we introduce a tight upper bound to the clump size, called the bundle size, and show the additional role it plays in rigorous lower bounds to exceedance probabilities independent of the PCH context. We then show how to compute the bundle size from the covariance information of the Gaussian process. Finally, we define a related 'correlation volume' which can be estimated in a robust way from the training data.

§5.1 The Mean Bundle Size

We start by defining a simpler analog of the clump size.

5.1 Definition *The* unconditioned bundle size *of the process Z is*

$$D_b := \int_{\mathcal{W}} 1_{(b,\infty)}(Z(w'))\, dw' \quad .$$

This is simply the volume of weight space where $Z(w)$ climbs above level b. With high probability $D_b = 0$ because no exceedance is observed, motivating the more useful

5.2 Definition *The* mean bundle size *of Z is*

$$E[D_b \,|\, Z(w) > b] \quad . \tag{5.1}$$

We briefly alter the overall level of this discussion to remark on two technical issues. First, if the sample paths $Z(w)$ are continuous as functions $\mathcal{W} \to R$, then D_b is well-defined since it is the measure of the open set $Z^{-1}\big((b,\infty)\big)$. Second, continuity of Z also ensures that D_b is a random variable by the following argument. Clearly $D_b > x_0$ only if there is an open $G \subseteq \mathcal{W}$, of Lebesgue measure $x > x_0$, over which $Z(w) > b$. Since finite unions of open rectangles having rational endpoints approximate open sets, G must contain such a finite union having measure greater than x_0. On the other hand, if such a union of rectangles with measure more than x_0 is contained in G, it must be that $D_b > x_0$. Formally,

$$\{D_b > x_0\} = \bigcup_{x > x_0} \bigcup_{N \geq 1} \bigcup_{\substack{\hat{G} = \cup_{i \leq N} R_i \\ \mathrm{vol}(\hat{G}) = x}} \bigcap_{w \in \hat{G}} \{Z(w) > b\}$$

where x is rational, the R_i are rectangles having rational endpoints, and the components of w are rational, so all operations are countable and the left-hand-side is measurable.

As the name suggests, the bundle size is different from the clump size because the former includes all exceedances of the level b, not just the region corresponding to a given clump center. The bundle size is therefore an overestimate when the total number of clumps $N_b \sim \mathrm{Pois}(\Lambda_b)$ exceeds one, but recall that we are in a regime where b (or equivalently the number of samples n) is large enough so that $\Lambda_b \ll 1$ and

$$\frac{P(N_b > 1)}{P(N_b = 1)} = \frac{1 - e^{-\Lambda_b} - \Lambda_b e^{-\Lambda_b}}{\Lambda_b e^{-\Lambda_b}} \leq \frac{1}{2}\Lambda_b \, e^{\Lambda_b} \ll 1 \quad .$$

The overestimate of $EC_b(w)$ by $E[D_b \,|\, Z(w) > b]$ due to inclusion of multiple clumps is negligible. We will call the assumption that at most one clump occurs in $\mathcal{W}$ the “single-clump condition.”

There is, however, another source of error due to biased sampling. To make this evident, fix b and suppose clumps are generated homogeneously at rate λ (with $\lambda \, \mathrm{vol}(\mathcal{W}) \ll 1$) and from the same distribution across the weight space; say that with equal probability clumps are cubes of volume either α or β, with $\alpha < \beta$. Of course $EC_b(w) = (\alpha + \beta)/2$. On the other hand, since with high probability there is either zero or one clump, if $Z(w) > b$ then D_b equals either α or β, but not with equal probability. For suppose p is the Poisson point (clump center) whose corresponding clump has captured w. Then D_b equals α if p fell within the cubical region of volume α placed around w, and β if p landed in the larger box of volume β. Given that w is covered, then, a large clump is more likely to have done it, and, assuming the single-clump condition, $E[D_b \,|\, Z(w) \!>\! b] = (\alpha^2 + \beta^2)/(\alpha + \beta)$ (For $\alpha \ll \beta$, the mean bundle size is about twice the mean clump size.)

It is easy to quantify this sampling effect via

5.3 Proposition *[6, sec. A6] Let f_C be the density of $C_b(w)$, and assume (as in lemma 3.1) that the rate $\lambda_b(w)$ is nearly constant in some ball B_w about w, that the distribution of $C_b(w)$ is essentially constant in B_w, and that $C_b(w) \ll \mathrm{vol}(B_w)$ with high probability. Then under the single-clump condition, the conditional density of D_b given $Z(w) > b$ is*

$$f_D(v) = \frac{v f_C(v)}{EC_b(w)} \quad .$$

Proof. The proportion of B_w covered by clumps having volume in $(v, v + dv)$ is $v \lambda f_C(v) dv$; this is the volume of such clumps times their rate of occurrence. The average proportion of B_w covered by any clump is $\lambda EC_b(w)$ which is again the rate of occurrence times the average size. The ratio is the probability that a covered point in B_w is covered by a clump of volume in $(v, v + dv)$; this is also $f_D(v)dv$. □

A nice way to express the overestimate is

$$\frac{E[D_b \,|\, Z(w) \!>\! b]}{EC_b(w)} = \frac{EC_b(w)^2}{(EC_b(w))^2} \geq 1 \quad . \tag{5.2}$$

As in the example, the variability of the clump size controls the accuracy of its bound via the bundle size.

§5.2 Harmonic Mean Inequalities

Now we show how the above picture of the bundle size as an approximation to clump size can be strengthened considerably: it turns out that the bundle size is also a tool for obtaining rigorous lower bounds to exceedance probability, without appeal to the PCH. (See also appendix B.) First we present a modified union bound for discrete unions [5] to develop intuition.

5.4 Proposition *Let $S_1, S_2, \ldots$ be measurable sets and N be the number of sets occurring. If $N < \infty$ a.s. then*

$$P\Big(\bigcup_{i=1}^{\infty} S_i\Big) = \sum_{i=1}^{\infty} P(S_i)\, E[N^{-1} \,|\, S_i]$$

Since $N \geq 1$ on S_i, the union bound follows easily.

Proof. $N_i := N^{-1}$ on S_i and zero otherwise is well-defined by hypothesis, and the simple equivalence

$$\begin{aligned} 1_{\bigcup S_i} &= \sum_{i=1}^{\infty} N_i 1_{S_i} \quad \text{a.s.} \\ \implies P\Big(\bigcup_{i=1}^{\infty} S_i\Big) &= \sum_{i=1}^{\infty} E\, N_i 1_{S_i} \\ &= \sum E\, E[N_i 1_{S_i} \,|\, 1_{S_i}] \\ &= \sum E\, 1_{S_i} E[N_i \,|\, 1_{S_i}] \\ &= \sum E\, 1_{S_i} E[N_i \,|\, S_i] \\ &= \sum P(S_i)\, E[N^{-1} \,|\, S_i] \end{aligned}$$

since $N_i = N^{-1}$ on S_i. □

The next proposition is an extension to uncountable index sets $\mathcal{W} \subseteq R^d$. For clarity we make explicit the dependence on the experimental outcome $\underline{\omega} \in \Omega$ (further distinguished from $w \in \mathcal{W}$ by the underbar).

5.5 Proposition *Let $S_w \subseteq \Omega$ be measurable sets for each $w \in \mathcal{W}$, and let θ be a measure on $\mathcal{W}$. Assume that*

$$D = D(\underline{\omega}) := \theta(\{w \in \mathcal{W} \,:\, \underline{\omega} \in S_w\}) \tag{5.3}$$

is a well-defined random variable. If the regularity conditions $D(\underline{\omega}) < \infty$ a.s. and $\underline{\omega} \in S_w \implies D(\underline{\omega}) > 0$ are met, then

$$P\Big(\bigcup_{w \in \mathcal{W}} S_w\Big) = \int_{\mathcal{W}} P(S_w)\, E[D^{-1} \,|\, S_w]\, \theta(dw) \quad .$$

In order for $D(\underline{\omega})$ to make sense, for each fixed $\underline{\omega}$ the θ-measure must be defined. Then, the resulting function $\Omega \to R$ must be a random variable.

Proof. The regularity conditions on $D(\underline{\omega})$ allow us to define $D_w(\underline{\omega}) = 1/D(\underline{\omega})$ for $\underline{\omega} \in S_w$ and zero otherwise. Proceeding as before,

$$\begin{aligned}
1_{\bigcup S_w} \quad \text{a.s.} &= \int_{\mathcal{W}} D_w 1_{S_w}\,\theta(dw) \\
\implies P\big(\bigcup_{w\in\mathcal{W}} S_w\big) &= \int_{\mathcal{W}} E\,D_w 1_{S_w}\,\theta(dw) \\
&= \int_{\mathcal{W}} E\,E[D_w 1_{S_w} \,|\, 1_{S_w}]\,\theta(dw) \\
&= \int_{\mathcal{W}} E\,1_{S_w} E[D_w \,|\, 1_{S_w}]\,\theta(dw) \\
&= \int_{\mathcal{W}} E\,1_{S_w} E[D_w \,|\, S_w]\,\theta(dw) \\
&= \int_{\mathcal{W}} P(S_w)\,E[D^{-1} \,|\, S_w]\,\theta(dw)
\end{aligned}$$

since $D_w(\underline{\omega}) = D^{-1}(\underline{\omega})$ if $\underline{\omega} \in S_w$. □

5.6 Corollary *If $Z(w)$ is continuous and $D_b < \infty$ a.s.,*

$$\begin{aligned}
P(\|Z(w)\|_{\mathcal{W}} > b) &= \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{E[D_b^{-1} \,|\, Z(w)\!>\!b]^{-1}}\,dw \\
&\geq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{E[D_b \,|\, Z(w)\!>\!b]}\,dw
\end{aligned}$$

Proof. Take θ as Lebesgue measure and $S_w = \{\underline{\omega} \in \Omega : Z(w) > b\}$ in the proposition. Then $D(\underline{\omega}) = D_b$. As shown after definition 5.1, continuity of $Z(w)$ as a function $\mathcal{W} \to R$ ensures that D_b is a well-defined random variable. In fact, continuity tells us that the preimage $Z^{-1}\big((b,\infty)\big) \subseteq \mathcal{W}$ is open a.s., so if $Z(w_0) > b$ the preimage is also nonempty and its Lebesgue measure is positive. The second assertion is a consequence of the harmonic mean inequality: $f > 0 \implies (Ef^{-1})^{-1} \leq Ef$. □

We note that the analytic part of the PCH can be obtained, using the single-clump condition to express the conditional density of D_b, by substituting

$$\begin{aligned}
E[D_b^{-1} \,|\, Z(w) > b]^{-1} &= \left(\int_0^\infty v^{-1} f_D(v)\,dv\right)^{-1} \\
&= \left(\int_0^\infty \frac{1}{EC_b(w)} f_C(v)\,dv\right)^{-1} \\
&= EC_b(w)
\end{aligned} \tag{5.4}$$

into the corollary.

§5.3 Computing Bundle Size

The bundle size approximation replaces the difficult-to-find $EC_b(w)$ with the more transparent

$$E[D_b \,|\, Z(w)\!>\!b] = \int_{\mathcal{W}} P(Z(w') > b \,|\, Z(w) > b)\,dw' \quad . \tag{5.5}$$

From the clump sizes presented in §4.6, we expect the estimate (5.5) to be, as a function of b, of the form (const $\sigma/b)^k$ for, say, $k = d$ or $k = 2d$. In particular, we do not anticipate the exponentially small clump size at w resulting from $(\forall w')\, P(Z(w') > b \,|\, Z(w) > b) \approx \bar\Phi(b/\sigma')$. To achieve these polynomial sizes, the $Z(w')$ and $Z(w)$ must be highly correlated, which we expect to happen for w' in a neighborhood of w. As with clump size, the behavior of the covariance in a neighborhood of w is the key to finding the bundle size at w.

Since $Z(w)$ and $Z(w')$ are jointly normal, abbreviate

$$\begin{aligned}
\sigma &= \sigma(w)\\
\sigma' &= \sigma(w')\\
\rho &= \rho(w,w') = R(w,w')/(\sigma\sigma')\\
\zeta &= \zeta(w,w') := (\sigma/\sigma')\frac{1-\rho\sigma'/\sigma}{\sqrt{1-\rho^2}} && (5.6a)\\
&= \left(\frac{1-\rho}{1+\rho}\right)^{1/2} \quad \text{(if } \sigma \text{ constant)} \quad . && (5.6b)
\end{aligned}$$

To show how to compute (5.5) we need

5.7 Lemma *For $\rho \ge 0$, $\zeta \ge 0$ and $\rho\sigma/\sigma' \le 1$,*

$$\bar\Phi\big((b/\sigma)\,\zeta\big) \le P(Z' > b \,|\, Z > b) \le \bar\Phi\big((b/\sigma)\sqrt{1+\zeta^2}\big)/\,\bar\Phi(b/\sigma) \quad .$$

The lower bound to P holds even for $\zeta < 0$.

Proof. For independent standard normal (X, Y),

$$\begin{bmatrix} Z \\ Z' \end{bmatrix} \overset{\mathcal{D}}{=} \begin{bmatrix} \sigma & 0 \\ \sigma'\rho & \sigma'\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} \quad ,$$

so that

$$\begin{aligned}
P(Z' > b,\, Z > b) &= P(\sigma X > b,\, \sigma'\rho X + \sigma'\sqrt{1-\rho^2}Y > b)\\
&= P(X \ge b/\sigma,\, Y \ge \alpha b/\sigma - \beta X)
\end{aligned}$$

where $\alpha = (\sigma/\sigma')/\sqrt{1-\rho^2} > 0$ and $\beta = \rho/\sqrt{1-\rho^2} \ge 0$. The corresponding region L is shown in figure 5.1; L is bounded by the lines

$$\begin{aligned}
\ell &: x = b/\sigma\\
\ell' &: y = \alpha b/\sigma - \beta x
\end{aligned}$$

which intersect at $p_\cap = (b/\sigma, \zeta\, b/\sigma)$ since $\alpha - \beta = \zeta$.

The lower bound is easy. No matter what ζ is, line ℓ' slopes downward implying that L contains the rectangular region $x \ge b/\sigma$, $y \ge \zeta\, b/\sigma$. Independence of X and Y implies the probability of the latter set is $\bar\Phi(b/\sigma)\,\bar\Phi(\zeta\, b/\sigma)$.

For the upper bound, we seek the closest approach p_* of L to the origin, which clearly occurs either on the vertical boundary, on the oblique boundary, or at $p_\cap$. The first case is precluded by the condition $\zeta \ge 0$. Observe that $p_* = p_\cap$ if the point on ℓ' nearest to $\mathbf{0}$ falls to the left

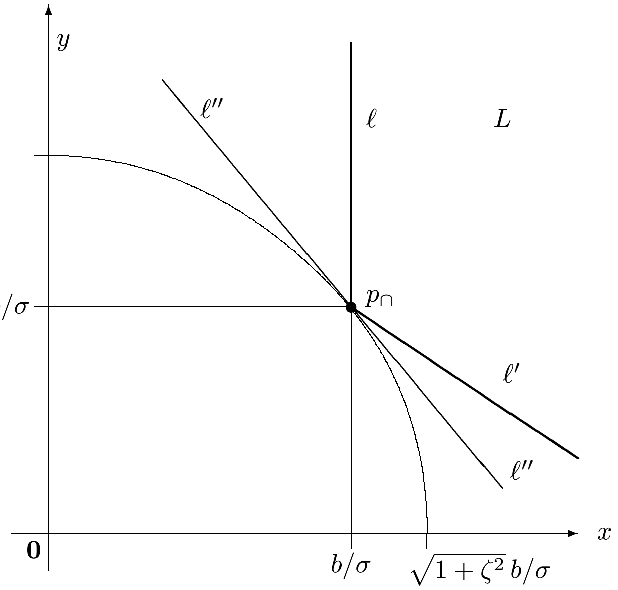

Figure 5.1: Finding a bivariate normal probability

This shows the geometry of the Gaussian integral in lemma 5.7.

of ℓ. Differentiation and some algebra shows that this occurs provided $\rho\sigma/\sigma' \leq 1$.

Thus let ℓ'' be the line passing through $p_\cap$ which is perpendicular to the line connecting $\mathbf{0}$ and $p_\cap$. Let L'' be the halfspace above and to the right of ℓ''. Then $L \subseteq L''$ so $P(L) \leq P(L'')$. But by rotational invariance of (X, Y), $P(L'') = P(X \geq (b/\sigma)\sqrt{1+\zeta^2})$. □

Now it is easy to show

5.8 Result *If* $b/\sigma \gg 1$ *and* $\rho\sigma/\sigma' \leq 1$,

$$E[D_b \,|\, Z(w) > b] \simeq \int_{\mathcal{W}} \bar{\Phi}((b/\sigma)\,\zeta)\, dw' \quad . \tag{5.7}$$

Proof. The integrand of (5.5) is $P(Z(w') > b \,|\, Z(w) > b)$. Fix some w' and suppose for the moment $\zeta \geq 0$. Apply the lemma, and then note that since $b \gg \sigma$, the asymptotic expansion for the upper bound part of the lemma yields (for, say, $b \geq 3\sigma$; see §A.2)

$$\begin{aligned} \bar{\Phi}\big((b/\sigma)\,\zeta\big) \leq P(Z' > b \,|\, Z > b) &\leq \frac{1.1}{\sqrt{1+\zeta^2}} \exp\big(-\tfrac{1}{2}((b/\sigma)\,\zeta)^2\big) \\ &\leq 1.1 \exp\big(-\tfrac{1}{2}((b/\sigma)\,\zeta)^2\big) \quad . \end{aligned} \tag{5.8}$$

On the other hand, if $\zeta < 0$, the lower bound of the lemma and the trivial upper bound show

$$\bar{\Phi}\big((b/\sigma)\,\zeta\big) \leq P(Z' > b \,|\, Z > b) \leq 1 \quad . \tag{5.9}$$

Integrating the conditional probability to find $E[D_b \,|\, Z(w) > b]$, we see the only w'-dependence is through ζ. For $\zeta < 0$, (5.9) shows the lower bound is off by at most a factor of two. For positive ζ near zero, (5.8) shows the lower bound is again within small constant factors of the upper bound. (Here we are not in the tails of $\bar{\Phi}$.) Finally, for ζ far from zero, say greater than unity, the upper bound differs from the lower bound by a small constant multiple of $\zeta(b/\sigma)$. A small disagreement in this regime is unimportant since this part of the integral does not contribute much to the bundle size—the integrand is exponentially small. □

The bundle size estimate of clump size will be used in the fundamental equation (3.9) to find

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{\int_{\mathcal{W}} \bar{\Phi}\left((b/\sigma)\,\zeta\right)\, dw'}\, dw \quad . \tag{5.10}$$

We have remarked that the bundle size represents those weights w' for which $Z(w')$ is correlated with $Z(w)$. To see how ζ captures this idea, note that the probability estimate in the integral is determined by weights where $\sigma(w)$ is maximized. (This includes the case where $\sigma(w)$ is constant.) Then $\rho\sigma'/\sigma \leq 1$ so $\zeta \geq 0$. In order to have ζ small, it is necessary to have $\rho \approx 1$, or in other words $w \approx w'$. On the other hand, weights w' not highly correlated with w have ζ so large that $\bar{\Phi}((b/\sigma)\zeta)$ does not make a significant contribution to a bundle size of order $(\sigma/b)^k$.

To sum up, we have presented and developed, in the context of Poisson clumping, an upper bound $E[D_b \,|\, Z(w) > b]$ to mean clump size whose final expression is lemma 5.8. When used in place of $EC_b(w)$ in the fundamental equation

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{EC_b(w)} dw \quad , \tag{5.11}$$

corollary 5.6 shows that a lower bound to the exceedance probability results, notwithstanding the validity of the PCH. The tightness of this lower bound is given by (5.2) or the harmonic mean inequality; it involves the variability of the clump size.

§5.4 Bundle Size Examples

We can use result 5.8 to check our belief that the mean bundle size is a good estimate of the clump size. For instance, suppose the covariance is of the 'cusped' form like the Brownian bridge of result 4.4,

$$R(w, w+v) = \sigma^2 - \sum_{j=1}^{d} \gamma_j |v_j| + O(v^{\mathsf{T}} v) \quad , \tag{5.12}$$

and the variance is either constant because of normalization, or of a smoothly varying form like $\sigma^2(w+v) = \sigma^2 - \alpha v^{\mathsf{T}} v$. Then the condition $\rho\sigma/\sigma' \leq 1$ is met, and in fact

$$\rho(w, w+v) \simeq 1 - \sum_{j=1}^{d} \frac{\gamma_j}{\sigma^2} |v_j| \tag{5.13a}$$

$$\zeta(w, w+v) \simeq \frac{1}{\sqrt{2}} \Big(\sum_{j=1}^{d} \frac{\gamma_j}{\sigma^2}\, |v_j| \Big)^{1/2} \tag{5.13b}$$

The estimate of $EC_b(w)$ is

$$\begin{aligned}
E[D_b \,|\, Z(w) > b] &\simeq \int \bar{\Phi}((b/\sigma)\,\zeta)\,dv \\
&\simeq \int \bar{\Phi}\Big(\frac{b}{\sigma\sqrt{2}}\Big(\sum_{j=1}^{d} \frac{\gamma_j}{\sigma^2}\,|v_j|\Big)^{1/2}\Big)\,dv \\
&\overset{\text{(a)}}{=} \prod_{j=1}^{d} \frac{\gamma_j}{\sigma^2} \cdot \Big(\frac{2\sigma^2}{b^2}\Big)^d \int \bar{\Phi}\Big(\Big(\sum_{j=1}^{d}|u_j|\Big)^{1/2}\Big)\,du \\
&\overset{\text{(b)}}{=} \frac{2^d}{(d-1)!} \cdot \prod_{j=1}^{d} \frac{\gamma_j}{\sigma^2} \cdot \Big(\frac{2\sigma^2}{b^2}\Big)^d \int_0^\infty z^{d-1}\,\bar{\Phi}\big(\sqrt{z}\,\big)\,dz \\
&\overset{\text{(c)}}{=} 2^{d-1}\binom{2d}{d} \cdot \prod_{j=1}^{d} \frac{\gamma_j}{\sigma^2} \cdot \big(\tfrac{\sigma}{b}\big)^{2d} \\
&\overset{\text{(d)}}{=} 2^{d-1}\binom{2d}{d}\,EC_b(w)
\end{aligned} \tag{5.14}$$

Relation (a) follows by a change of variables $u = \mathbf{M}v$ for $\mathbf{M}$ a diagonal matrix; step (b) is the substitution $z = \sum_{j=1}^{d}|u_j|$ where the volume element is

$$\mathrm{vol}(\{u \in R^d \,:\, \textstyle\sum|u_j| \in (z, z+dz)\}) = \tfrac{2^d}{d!}\big((z+dz)^d - z^d\big) \quad .$$

The integral at (c) is done by parts; see §C.6. Finally (d) just uses the clump size from result 4.4.

The overestimate is about $2^d \cdot 2^{2d} = 8^d$, or a factor of eight in each dimension of the weight space; also note that the dependence on the level b and the local expansion γ_j is captured. Because of the exponential behavior of the exceedance probability for large b, such constant factors do not influence the sample size much. In a similar fashion $E[D_b \,|\, Z(w) > b]$ can be computed (§C.6) for the other process models we have come across, with results summarized below:

Name	$1 - R(w_0, w_0 + w)$	$E[D_b \,\|\, Z(w) > b]/EC_b(w_0)$
Cusped	$\sum_{j=1}^{d} \gamma_j \|w_j\|$	8^d
Rough Isotropic	$\big(\sum_{j=1}^{d}(\gamma\, w_j)^2\big)^{1/2}$	6^d to $14^d(d/2)!$
Smooth	$w^{\mathsf{T}}\Lambda_{02}\,w/2$	$\leq 2^d$

In each case the factor introduced in substituting $E[D_b \,|\, Z(w) > b]$ for $EC_b(w)$ is a small constant per weight.

§5.5 The Correlation Volume

The bundle size can be used directly in a lower bound to exceedance probability if one has information about the covariance of Z; we shall see an example of the use of such information to find $E[D_b \,|\, Z(w) > b]$ in chapter 6. However, since we ultimately desire an estimate which is computable from the training set, in this section we make one more simplification, which also reveals an important property of the integral of result 5.8 for $E[D_b \,|\, Z(w) > b]$.

We introduce

5.9 Definition *For $\tau \le 1$ let*

$$\mathcal{V}_\tau(w) := \{w' \in \mathcal{W} \,:\, \zeta(w, w') \le \tau\} \quad ,$$

then the correlation volume *of the process Z is*

$$V_\tau(w) := \mathrm{vol}(\mathcal{V}_\tau(w)) \quad .$$

As mentioned below result 5.8, the major contribution to the bundle size comes for $\zeta(w, w') \approx 0$, and in standard situations this means $\rho(w, w') \approx 1$. Thus $\mathcal{V}_\tau(w)$ and its volume $V_\tau(w)$ represent those weight vectors w' whose errors $Z(w')$ are highly correlated with $Z(w)$.

From the monotonicity of $\bar{\Phi}$

$$E[D_b \,|\, Z(w) > b] \ge \int_{\mathcal{V}_\tau(w)} \bar{\Phi}((b/\sigma)\zeta)\, dw' \ge V_\tau(w)\, \bar{\Phi}((b/\sigma)\tau) \quad . \tag{5.15}$$

The bound being used is the simple $\bar{\Phi}(\beta z) \ge \bar{\Phi}(\beta\tau) 1_{(-\infty,\tau]}(z)$. In this case it is quite accurate, as we shall see.

In some cases the correlation volume is convex or nearly so. Consider the important special case of a process with constant variance (e.g. the self-normalized case). Because the function $\rho \mapsto \big((1-\rho)/(1+\rho)\big)^{1/2}$ is monotone decreasing, $w' \in \mathcal{V}_\tau(w)$ is equivalent to $\rho(w, w') \ge (1-\tau^2)/(1+\tau^2)$. So if two points w', $w'' \in \mathcal{V}_\tau(w)$ then

$$\beta\rho(w, w') + (1-\beta)\rho(w, w'') \ge (1-\tau^2)/(1+\tau^2)$$

If the correlation $\rho(w, \cdot)$ is concave in its second argument then $\rho(w, \beta w' + (1-\beta)w'')$ is at least as large as the LHS and the convex combination is also in $\mathcal{V}_\tau(w)$. For our purposes $\rho(w, w')$ need not be concave for all $w' \in \mathcal{W}$; concavity in the neighborhood of w for which $\rho(w, w') \ge (1-\tau^2)/(1+\tau^2)$ is enough. In particular, either of

$$\rho(w, w+v) \simeq 1 - v^{\mathsf{T}} \mathbf{M}\, v + o(v^{\mathsf{T}} v)$$

$$\rho(w, w+v) \simeq 1 - \sum_{j=1}^{d} \gamma_j |v_j| + O(v^{\mathsf{T}} v)$$

are approximately concave if τ is small.

Let us show that the bound (5.15) gives reasonable results in our stock of examples. For the cusped covariance of (5.12), we found

$$\zeta(w, w+v) \simeq \frac{1}{\sqrt{2}} \Big(\sum_{j=1}^{d} \frac{\gamma_j}{\sigma^2}\, |v_j| \Big)^{1/2} \tag{5.16}$$

so that

$$\begin{aligned}
V_\tau(w) &\simeq \operatorname{vol}\bigl(\{v \in R^d \, : \, \sum_{j=1}^{d} \frac{\gamma_j}{\sigma^2}\, |v_j| \le 2\tau^2\}\bigr) \\
&= 2^d \cdot \prod_{j=1}^{d} \frac{\sigma^2}{\gamma_j} \cdot \tau^{2d} \operatorname{vol}\bigl(\{u \in R^d \, : \, \sum_{j=1}^{d} |u_j| \le 1\}\bigr) \qquad (5.17) \\
&= \frac{4^d}{d!} \cdot \prod_{j=1}^{d} \frac{\sigma^2}{\gamma_j} \cdot \tau^{2d} \quad .
\end{aligned}$$

The approximate lower bound to clump size is $\bar{\Phi}((b/\sigma)\tau)V_\tau(w)$, where the parameter $\tau \le 1$ is our choice. Using the asymptotic expansion for $\bar{\Phi}$ and differentiating yields the threshold

$$\tau^2 \frac{b^2}{\sigma^2} = 2d - 1 \simeq 2d \quad ;$$

in retrospect a large d (say $d \ge 5$) justifies use of the asymptotic expansion. The approximate lower bound is

$$\begin{aligned}
E[D_b \,|\, Z(w) > b] &\ge \bar{\Phi}(\sqrt{2d})\frac{4^d}{d!}\left(2d\frac{\sigma^2}{b^2}\right)^d \prod_{j=1}^{d} \frac{\sigma^2}{\gamma_j} \\
&\simeq \frac{1}{2\sqrt{2\pi d}} 8^d \cdot \prod_{j=1}^{d} \frac{\sigma^2}{\gamma_j} \cdot \left(\frac{\sigma}{b}\right)^{2d}
\end{aligned} \qquad (5.18)$$

where we have used the Stirling approximation and the asymptotic expansion for $\bar{\Phi}$. This differs by a factor of essentially only $\sqrt{d}$ from the actual integral as computed in (5.14).

It is similarly possible (§C.7) to find the correlation volumes for the isotropic covariance and the smooth process. The results are as below.

Name	$V_\tau(w)$	Ratio
Cusped	$(4^d/d!) \cdot \prod_{j=1}^{d} (\sigma^2/\gamma_j) \cdot \tau^{2d}$	$\sqrt{2\pi d}$
Rough Isotropic	$2^d \kappa_d (\sigma^2/\gamma)^d \tau^{2d}$	$\sqrt{2\pi d}$
Smooth	$\le 2^d \, \kappa_d \, \lvert \Lambda_{02}/\sigma^2 \rvert^{-1/2} \, \tau^d$	$\approx \sqrt{\pi d}$

In the table

$$\text{Ratio} := \frac{E[D_b \,|\, Z(w) > b]}{V_\tau(w)\, \bar{\Phi}((b/\sigma)\tau)} \quad .$$

The correlation volume given for the smooth process is an upper bound which is achieved if the variance is constant, and the corresponding ratio is also exact in that case.

The central conclusion to be gained from this table is that the correlation volume approximation to bundle size is quite accurate—by our undemanding standard it is as good as an equality. The quality of the approximation is due to the large dimension of the weight space: most of the mass in the correlation volume is found near the boundary of $\mathcal{V}_\tau(w)$, and this is where the constant approximation to $\bar{\Phi}((b/\sigma)\,\zeta)$ is tightest. To estimate the bundle size at a point, rather than needing to find $\bar{\Phi}((b/\sigma)\,\zeta)$ across all $w' \in \mathcal{W}$, we need only find the boundary of $\mathcal{V}_\tau(w)$, the set of significant ζ.

§5.6 Summary

We have approximated the overall exceedance probability by a construction like

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{EC_b(w)}\, dw \quad . \tag{5.19}$$

This is a sum of the point exceedance probabilities—the numerator—each weighted according to how many other points are stochastically related to it. A large weighting factor indicates strong correlation of the network in question to neighboring nets, so its contribution to the exceedance probability is decreased; smaller factors indicate a more jagged process and give a larger contribution.

In this chapter, motivated by practical difficulty in finding the mean clump size, we have introduced several other measures of the idea of 'stochastically related':

$$\begin{aligned} EC_b(w) &\overset{\text{(a)}}{\simeq} E[D_b^{-1} \mid Z(w) > b]^{-1} \\ &\overset{\text{(b)}}{\leq} E[D_b \mid Z(w) > b] \\ &\overset{\text{(c)}}{\simeq} \int_{\mathcal{W}} \bar{\Phi}((b/\sigma)\,\zeta) dw' \\ &\overset{\text{(d)}}{\simeq} V_\tau(w)\, \bar{\Phi}((b/\sigma)\,\tau) \quad . \end{aligned} \tag{5.20}$$

The first two sections of the chapter show that (a) holds under the single-clump condition, discuss the cause and size of inequality (b), and show that $E[D_b \mid Z(w) > b]$ can be used in a PCH-style lower bound without appeal to the PCH. In §5.3 we prove (c), which uses ζ to link the weighting factor to the process correlation. This also allows computation of bundle sizes by direct integration; comparisons to the corresponding clump sizes show that the dependence on process parameters is preserved, but some multiplicative constants, which are not very significant for our purposes, are introduced. These constants represent the gap between the true asymptotic value of the probability as given by PCH and the lower bound as found via the mean bundle size. Finally, §5.5 introduces (d) and finds correlation volumes by elementary calculations. The result is that negligible inaccuracy is introduced in the final approximation.

There is an interesting link between the VC dimension (definition 2.2, lemma 2.3) and the several recastings, in terms of the stochastic process $Z(w)$, of the idea of inter-weight dependence. The counterpart of (5.19) for self-normalized $Z(w)$ and, say, the correlation volume, is

$$\begin{aligned} P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) &\simeq \bar{\Phi}(b) \int_{\mathcal{W}} \frac{1}{V_\tau(w)\, \bar{\Phi}(b\tau)}\, dw \\ &= \bar{\Phi}(b) E_U\, \frac{\text{vol}(\mathcal{W})}{V_\tau(w)\, \bar{\Phi}(b\tau)} \end{aligned} \tag{5.21}$$

where the expectation is taken with respect to a uniform distribution on $\mathcal{W}$. This probability is the worst-case point exceedance probability times a number representing degrees of freedom, or essentially independent networks. It is the average (across $\mathcal{W}$) of the number of network-clusters at correlation τ present within $\mathcal{W}$. This number of degrees of freedom plays the same role as the Sauer bound on distinct dichotomies does in the VC setup (see especially the proof sketch of the Vapnik bound, theorem 2.4), except that it retains dependence on the underlying data distribution P, the target function, and the network architecture. In addition, as we shall soon see, the number of degrees of freedom as defined in this chapter can be estimated from the training data.

6 Empirical Estimates of Generalization

In a search for a usable adjustment factor in the integral for the exceedance probability, we have progressed from the mean clump size (intractable in practice), through the mean bundle size, to the correlation volume, which is determined by the covariance function of Z. Now we show how to estimate the correlation volume of a given weight using the training set. Then we present an algorithm which uses such estimates to find approximations for the probability of reliable generalization in the absence of analytical information about the unknown P and the potentially complex network architecture $\mathcal{N}$. Two simulation studies, one for the problem of learning orthants and the other for learning halfspaces, show that the proposed empirical method works.

§6.1 Estimating Correlation Volume

Until now we have obtained the correlation volume via elementary considerations involving the process covariance; here is how to estimate it empirically. Fix a weight $w \in \mathcal{W}$. A (noisy) oracle for determining if some w' is in $\mathcal{V}_\tau(w)$ uses the training set to compute first

$$\begin{aligned}&(y_1 - \eta(x_1; w))^2, \ldots, (y_n - \eta(x_n; w))^2\\&(y_1 - \eta(x_1; w'))^2, \ldots, (y_n - \eta(x_n; w'))^2 \quad .\end{aligned}$$

and then successively

$$\begin{aligned}
\hat{\sigma}^2(w) &= \frac{1}{n}\sum_{i=1}^{n}\big[(y_i - \eta(x_i; w))^2 - \nu_{\mathcal{T}}(w)\big]^2\\
\hat{\sigma}^2(w') &= \frac{1}{n}\sum_{i=1}^{n}\big[(y_i - \eta(x_i; w'))^2 - \nu_{\mathcal{T}}(w')\big]^2\\
\hat{R}(w, w') &= \frac{1}{n}\sum_{i=1}^{n}\big[(y_i - \eta(x_i; w))^2 - \nu_{\mathcal{T}}(w)\big] \times\\
&\qquad\qquad \big[(y_i - \eta(x_i; w'))^2 - \nu_{\mathcal{T}}(w')\big]\\
\hat{\rho}(w, w') &= \hat{R}(w, w')/\big(\hat{\sigma}(w)\,\hat{\sigma}(w')\big)\\
\hat{\zeta}(w, w') &= (\hat{\sigma}/\hat{\sigma}')\frac{1 - \hat{\rho}\,\hat{\sigma}'/\hat{\sigma}}{\sqrt{1 - \hat{\rho}^2}}\\
&= \left(\frac{1 - \hat{\rho}}{1 + \hat{\rho}}\right)^{1/2} \qquad \text{(if } \sigma \text{ constant)}
\end{aligned}$$

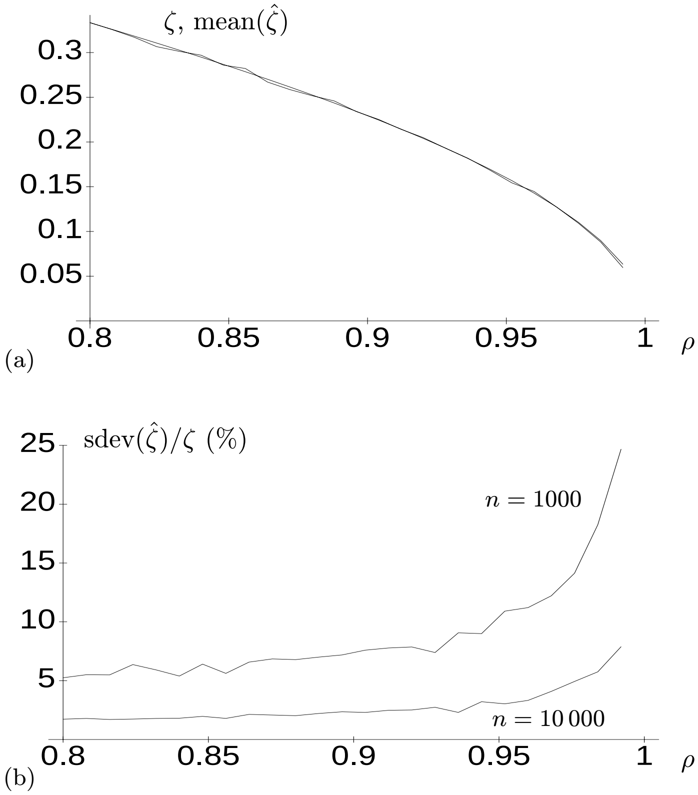

Figure 6.1: Estimating ζ for binary classification

In the upper plot, the curves nearly coincide. The ratio in the lower plot is shown in percent.

Then $\hat{\zeta}(w, w')$ is compared to τ to see if $w' \in \mathcal{V}_\tau(w)$ or not.

It is possible to reliably estimate ζ in this way, even when $\rho \approx 1$. Figure 6.1 illustrates this for the problem of binary classification. The error $(y - \eta(w; x))^2$ is a Bernoulli random variable, and $\hat{\zeta}$ is formed from n i.i.d. pairs (for w and w') of such variables having a given variance and correlation. Choosing $\sigma = \sigma' = 1/2$, the correlation is then varied from 0.8 to nearly unity, resulting in ζ dropping from about 1/3 to quite small values. (The estimator $\hat{\zeta}$ is forced to find the variances even though they are the same in this example.) The upper panel shows ζ and the sample mean of 100 independent $\hat{\zeta}$ estimates, each of which is computed on the basis of $n = 1000$ pieces of data. This plot shows the scale of ζ and demonstrates that $\hat{\zeta}$ is essentially unbiased, at least for n moderately large. The lower panel shows the ratio of standard deviation of $\hat{\zeta}$ to ζ, expressed as a percentage, for $n = 1000$ (upper curve) and $n = 10\,000$ (lower curve). Only for quite low values of ζ does the variance become significant. However, this variability at very low ζ does not influence estimates of $V_\tau(w)$ as long as the threshold τ is moderate, which in this simulation would mean greater than 1/20 or so.

A natural way to estimate $V_\tau(w)$ is to select a set $\mathcal{A}$ around w large enough to enclose $\mathcal{V}_\tau(w)$. Sample uniformly in that set M times, obtaining m hits on $\mathcal{V}_\tau(w)$, and form the Monte Carlo estimate

$$V_\tau(w) \approx \mathrm{vol}(\mathcal{A}) \, \frac{m}{M} \quad .$$

This scheme has a problem. If there is an overlap of just a factor of two in the width of $\mathcal{A}$ relative to $\mathcal{V}_\tau(w)$, then on the average $M = 2^d$ is necessary to expect even one hit—and this is far too many samples if $d = 50$. This is true even using importance sampling drawing from, say, a two-sided exponential distribution. Any uncertainty of the 'characteristic length' of the sampling distribution (e.g. the $1/e$ width of the exponential or the standard deviation of a Gaussian) is magnified tremendously by the dimension of the weight space.

In fact, the problem of finding a good (having relative error of size less than d^{Kd}) approximation to the volume of a convex set by a deterministic algorithm using calls to a membership oracle is NP-hard [37, sec. 4.1]. The recent discovery of a randomized polynomial-time algorithm [18] is of no practical use since the running time is of order greater than d^{16} [37, sec. 4.2].

In our problem, we can use certain reasonable properties of $\mathcal{V}_\tau(w)$ to help estimate the volume. For example, we know $\mathcal{V}_\tau(w)$ is centered about w, and can assume it extends symmetrically away from w. The simplest technique is to let $w' = w$ except in one coordinate and sample along each coordinate axis, stopping when $w' \notin \mathcal{V}_\tau(w)$. If $\mathcal{V}_\tau(w)$ is at least approximately convex (see the remarks after its definition), it is approximately contained in the cube defined by these intercepts, and in turn contains the simplex defined by them. The ratio of these two estimates is unfortunately $d!$. If the shape of $\mathcal{V}_\tau(w)$ is known, one can do better than this.

§6.2 An Algorithm

The basic relation is the lower bound and approximation

$$\begin{aligned} P(\|Z(w)\|_{\mathcal{W}} > b) &\geq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{E[D_b \mid Z(w) > b]} \, dw \\ &\simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{\int_{\mathcal{W}} \bar{\Phi}((b/\sigma)\,\zeta) dw'} \, dw \\ &=: \underline{P} \quad , \end{aligned} \tag{6.1}$$

where the latter holds for $\rho\sigma/\sigma' \leq 1$. We have seen that $\underline{P}$ is in turn upper-bounded as

$$\underline{P} \leq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma)}{V_\tau(w)\, \bar{\Phi}((b/\sigma)\tau)} \, dw \quad , \tag{6.2}$$

where the bound becomes a very tight approximation if τ is chosen correctly. We wish to use the estimate of correlation volume to find the integral.

Before going ahead, we take the important step of normalizing the process by its standard deviation $\sigma(w)$, which is desirable for reasons other than analytic simplicity. Primarily, normalization avoids the undesirable property (remarked on in §1.3, §4.6) of the sample size estimate n_c

being dominated, via the exponential factor $\bar{\Phi}(b/\sigma)$, by a small group of networks which are unlikely to be chosen as models. Such a dependence on the maximum-variance networks becomes especially problematic in the empirical setting, where the variance is not known exactly—finding networks of maximum variance becomes an optimization problem on the order of network selection itself. As a desirable side-effect of normalization, the condition $\rho\sigma/\sigma' \leq 1$ holds, and (6.2) simplifies to

$$\underline{P} \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\tau)} \int_{\mathcal{W}} V_\tau(w)^{-1}\, dw \quad . \tag{6.3}$$

This can be estimated by a Monte Carlo integral, using the method of §6.1 to find the integrand $V_\tau(w)$.

The method we propose is outlined in pseudocode form in figure 6.2. The procedure contains an outer loop over randomly drawn $w \in \mathcal{W}$ and an inner loop over the d coordinates of a weight vector. Some parts of the code are left ambiguous, such as the precise method of adjusting δ to find a boundary, the termination criterion in the loop for δ, and the method of finding the volume of the piece of $\mathcal{V}_\tau(w)$ that the loop on δ locates. These parts could all be different depending on the application. (For example, architectures with smoothly varying outputs can use a root-finding method or approximate slope of ζ for locating the boundary of $\mathcal{V}_\tau(w)$.)

The remaining difficulty is the choice of τ, which as we saw in §5.5 depends on b, which is unknown to us at the outset. Recomputing the integral for many different τ or b values can be avoided by making the reasonable assumption that

$$V_\tau(w) = K(w)\, \tau^{2d/\alpha} \tag{6.4}$$

with $\alpha = 1$ (rough process) or 2 (smooth process); see the table in §5.5. Thus α depends only on known qualitative aspects of the architecture. The coefficients may change as w varies but the basic form of the correlation does not.

Once the integral is computed for a reference τ_0, it can be scaled to a desired $\tau \ll 1$ via

$$\underline{P} \leq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\tau)\tau^{2d/\alpha}} \left(\tau_0^{2d/\alpha} \int_{\mathcal{W}} V_{\tau_0}(w)^{-1}\, dw \right) \quad . \tag{6.5}$$

The only τ-dependence in the lower bound is in the denominator of the leading factor; we wish to maximize it:

$$\begin{aligned} &\frac{d}{d\tau}\, \bar{\Phi}(b\tau)\tau^{2d/\alpha} \simeq \frac{1}{b}\frac{d}{d\tau}\phi(b\tau)\tau^{2d/\alpha-1} = 0 \\ &\implies \tau^2 b^2 = 2d/\alpha - 1 \simeq 2d/\alpha \quad . \end{aligned}$$

We have seen several times (e.g. the table at the end of §5.5) that this choice of τ causes the inequality (6.5) to become very tight. With this

```
$sum = 0$
for $N = 1$ to $N_{\text{weight}}$
    $w = \text{random}(\mathcal{W})$        $[*]$
    for $j = 1$ to $d$
        $\delta = \delta_0$
        while $(0 < \delta)$ & $(w + \delta e_j \in \mathcal{W})$
            $w' = w + \delta e_j$
            if $\hat{\zeta}((w, w')) \leq \tau$
                increase $\delta$
            else
                decrease $\delta$
            check termination
        endwhile
        if $\delta$ has both increased and decreased
            $boundary(j) = \delta$
        else
            go to $[*]$
    endfor
    $V_\tau(w) = \text{volume}(boundary)$
    $sum = sum + 1/V_\tau(w)$
endfor
$integral = \text{vol}(\mathcal{W}) \times sum/N_{\text{weight}}$
```

Figure 6.2: An algorithm for estimating generalization error

The symbol e_j denotes a vector in $\mathcal{W}$ with a one in coordinate j and all others zero.

in mind we write

$$\begin{aligned} P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) &\geq \underline{P} \\ &\simeq \left(\frac{b^2}{d}\right)^{d/\alpha} \bar{\Phi}(b) \left[\frac{\tau_0^{2d/\alpha} \int_{\mathcal{W}} V_{\tau_0}(w)^{-1}\, dw}{(2/\alpha)^{d/\alpha}\, \bar{\Phi}(\sqrt{2d/\alpha}\,)}\right] \qquad (6.6) \\ &= \exp(dQ) \left(\frac{b^2}{d}\right)^{d/\alpha} \bar{\Phi}(b) \end{aligned}$$

where the final line defines Q.

§6.3 Simulation: Learning Orthants

Here we consider again the problem of learning orthants by performing a simulation using the method outlined above to estimate exceedance probability. As in section 4.2, nets are indicator functions of translated negative orthants in R^d. The output $y \equiv 0$ and input vectors x are distributed uniformly on $[0,1]^d$. Networks are discontinuous so $Z(w)$ is ‘rough’ with $\alpha = 1$.

Before stating the results, we remark on one difficulty with binary classification problems such as this one. The method for determining

Table 6.1: Estimates of correlation volume, learning orthants

Empirical $V_\tau(w)$	Predicted $V_\tau(w)$	Log Ratio
2.39×10^{-11}	2.02×10^{-11}	0.16
1.47×10^{-11}	9.26×10^{-12}	0.46
2.56×10^{-11}	6.66×10^{-12}	1.34
6.41×10^{-14}	5.67×10^{-13}	−2.18
1.51×10^{-11}	1.57×10^{-11}	−0.03
1.55×10^{-11}	8.63×10^{-12}	0.58
3.88×10^{-12}	5.89×10^{-13}	1.88
4.98×10^{-12}	3.40×10^{-12}	0.38
5.41×10^{-12}	1.00×10^{-11}	−0.62
8.89×10^{-13}	2.38×10^{-12}	−0.98
2.12×10^{-14}	8.19×10^{-13}	−3.65
1.18×10^{-13}	1.76×10^{-12}	−2.70

correlation volume is based on varying w' about a chosen w to see when $w' \in \mathcal{V}_\tau(w)$. For an orthant classifier w having $|w| \ll 1$, no matter how much one coordinate w_j is altered, the classifier often acts identically on the training set. For example, suppose w has all coordinates less than $1/2$, and d is large. Then with high probability, no x in $\mathcal{T}$ is classified positively by $\eta(\cdot; w)$ since $P(x \leq w) = |w| \leq 2^{-d}$. Setting w' by increasing say w_1 to unity is unlikely to change the classification of any point since then $P(x \leq w') \leq 2^{-d+1}$; similarly decreasing w_1 to zero has no effect. The algorithm sees this as $\hat{\sigma}^2 = \hat{\sigma}'^2$ and $\hat{\rho} = 1$ so $\hat{\zeta}$ is indeterminate for any w'. Such points w must be discarded, which effectively restricts the algorithm to sampling regions where $|w|$ is not exponentially small. Interestingly, as we shall see, the resulting bias is not large enough to affect the sample size estimates greatly since all that is needed is order-of-magnitude estimates of the correlation volume. This difficulty would not arise if the network output varied continuously with w, for then a differential change in w would produce a corresponding change in output; the problem is an extreme example of why smoothly varying networks are often used in practice.

Table 6.1 shows a dozen randomly chosen points $w \in [0,1]^d$, for $d = 6$, and their associated correlation volumes. These volumes are found by the method described in this chapter, and by the theoretical prediction of correlation volume (see §C.8)

$$V_\tau(w) \simeq \frac{8^d}{d!} |w|(1 - |w|)^d \tau^{2d} \quad . \tag{6.7}$$

The threshold used is $\tau = 0.1$ and $n = 8192$, large enough to ensure that the difficulty mentioned above is unlikely to happen. For the empirical volume estimates, the extreme points of $\mathcal{V}_\tau(w)$ are found by varying just one coordinate of w, and the correlation volume is taken to be the size of the smallest 'diamond-shaped' region containing those points. In this way we exploit to some degree our knowledge of the Brownian sheet covariance (see (5.17)) to find the most accurate correlation volumes for comparison to theoretical predictions. In the rightmost column is the

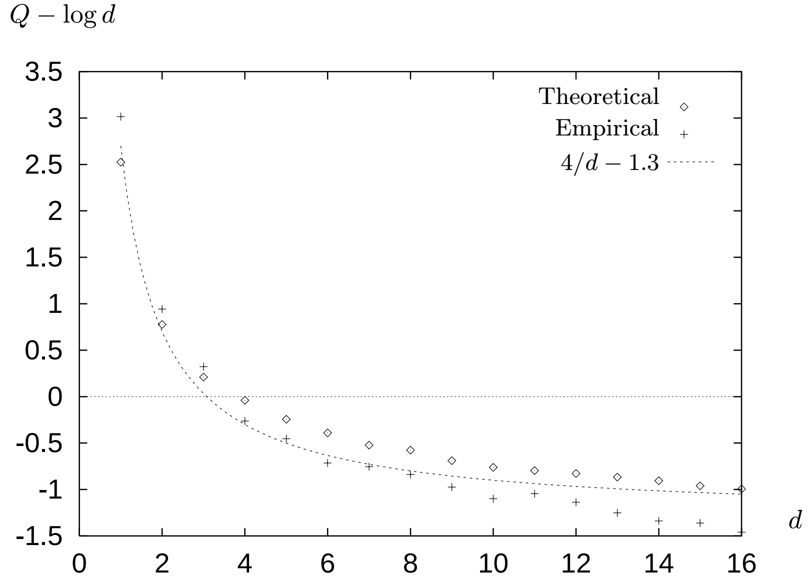

Figure 6.3: Empirical estimate of the leading constant in the case of learning orthants.

Crosses are estimates of Q obtained using the training set; diamonds are theoretical predictions. The dotted line is an approximate curve fitted to the points.

log ratio between the two volumes; this is never even as large as d. The table verifies that, in principle, it is possible to generate quite accurate estimates of correlation volume using our empirical method.

In figure 6.3 is a graph of Q versus d for this scenario. At each d, 50 independent estimates of Q are averaged to get an idea of average performance.[1] Each such Q is found via a Monte Carlo integral using 100 points w, as described above, with correlation volumes determined from a training set of size $100d$. (Points w for which no correlation volume could be found were discarded and a new choice made.) The reference level was $\tau_0 = 0.1$.

Corresponding theoretical predictions of the scale factor for the probability are also shown; they are based on choosing 100 points w at random and performing a Monte Carlo integration just as for the empirical correlation volume. In this case the correlation volume used is (6.7) instead the empirical estimate. The agreement of this theoretical integral with the empirical curve is quite close. The theoretical curve has larger Q values because it does not reject the weights w for which $|w|$ is extremely small; such weights also have very small correlation volumes which increases Q.

Also on the plot is the fitted curve $Q - \log d \approx (4/d) - 1.3$; the

[1] This averaging would not be done in practice but is used here to get an idea of typical Q estimates. The sample standard deviation of the empirical Q decreases smoothly from 0.5 at $d = 1$ to 0.2 at $d = 16$. That of the theoretical Q decreases from 0.5 to 0.15.

conclusion is that

$$\begin{aligned}\exp(dQ) &\simeq \exp\bigl(d(4/d - 1.3 + \log d)\bigr) \\ &= e^4\, e^{-1.3\, d}\, d^d \quad .\end{aligned} \tag{6.8}$$

Using this fitted curve, we find (see §C.9) that a critical sample size of about

$$n_c = \frac{2.5\, d \log d}{\epsilon^2} \tag{6.9}$$

is enough to ensure that with high probability,

$$\sup_{w \in \mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sqrt{\mathcal{E}(w)(1-\mathcal{E}(w))}} < \epsilon \quad .$$

As remarked in §1.3, under the proviso that the selected network has $\nu_{\mathcal{T}}(w^*) = 0$, we may in essence replace ϵ^2 by ϵ

§6.4 Simulation: Perceptrons

As another application of the proposed algorithm, consider the following example of a perceptron. Nets are $\eta(x; w) = 1_{[0,\infty)}(w^T x)$ for $w \in \mathcal{W} = R^d$, and data x is uniform on $[-1/2, 1/2]^d$. Suppose $y = \eta(x; w^0)$ and $w^0 = [1 \cdots 1]^{\mathsf{T}}$. Nets are discontinuous so $Z(w)$ is 'rough' with $\alpha = 1$. We will call this version P-Emp meaning the perceptron using empirical analysis.

We have seen two versions of this problem before. In §2.5 are 'learning curves', somewhat analogous to our tradeoff between n, d, and ϵ, for a similar problem. In that problem data is uniform on $[0, 1]^d$ and the target function equals unity if the sum of inputs exceeds $d/2$; this is just a spatial translation of the problem we consider. The difference is that the networks used as models have continuously-varying outputs. Nonetheless we would expect some rough agreement between the estimates of generalization ability. We will call this setup P-CT after the authors of that study.

In §4.4 exact PCH methods were applied to a problem like this one, except that the data distribution was rotationally invariant. Also, the ordinary distance of $\nu_{\mathcal{T}}(w)$ from $\mathcal{E}(w)$, not relative distance, was used, and this tends to give pessimistic estimates of generalization ability. We will call this setup P-An for the perceptron network with analytical knowledge of the data model.

For the perceptron, as demonstrated in §4.4, the set of inputs on which a network w' disagrees with w is a wedge-shaped slice of $[-1/2, 1/2]^d$. By varying just one coordinate of w to get w', this region eventually is made to have appreciable probability measure. The problem that arose with learning orthants does not come up here.

In figure 6.4 is the empirically determined Q versus d for the threshold function. At each d twenty independent estimates of Q are averaged. Each estimate of Q is found via a Monte Carlo integral using 50 points sampled at random from $\mathcal{W}$, as described above, with correlation volumes determined from a training set of size $n = 100d$.

Over the range, say, $7 \le d \le 50$, we see $Q \approx 1$ and from (6.6),

$$P\Bigl(\Bigl\| \tfrac{Z(w)}{\sigma(w)} \Bigr\|_{\mathcal{W}} > b\Bigr) \approx e^d\, (b^2/d)^d\ \bar{\Phi}(b) \tag{6.10}$$

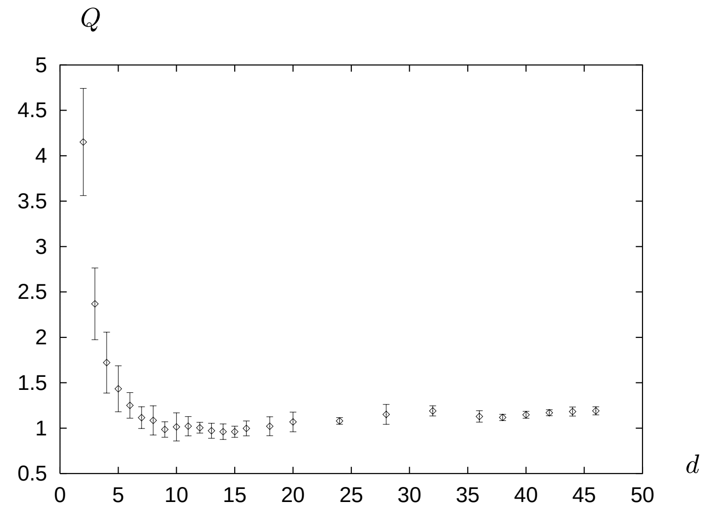

Figure 6.4: Empirical estimate of the leading constant for a perceptron architecture

Error bars span one sample standard deviation in each direction from the sample mean.

$$(1/d)\log P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \approx 1 + \log(b^2/d) - (1/2)(b^2/d) \qquad (6.11)$$

This falls below zero at $b^2/d = 5.4$ implying

$$\begin{aligned} \frac{b^2}{d} &= \frac{n_c\epsilon^2}{d} \simeq 5.4 \\ n_c &= \frac{5.4d}{\epsilon^2} \end{aligned} \qquad (6.12)$$

samples are enough to ensure that with high probability,

$$\sup_{w\in\mathcal{W}} \frac{|\nu_{\mathcal{T}}(w) - \mathcal{E}(w)|}{\sqrt{\mathcal{E}(w)(1-\mathcal{E}(w))}} < \epsilon \quad .$$

In particular the relations (1.8) hold with high probability for n a bit larger than n_c.

Now, following the discussion of the relative distance criterion in §1.3, if the selected network has $\nu_{\mathcal{T}}(w^*) = 0$, we may essentially replace ϵ^2 above by ϵ. Under this proviso, we see that

$$n_c = \frac{5.4d}{\epsilon} \qquad \text{[P-Emp]} \qquad (6.13)$$

samples are enough for reliable generalization, and in particular for $\mathcal{E}(w^*) \le \epsilon$ with high probability.

This can most easily be compared with the estimate in result 4.13, for the P-An setup:

$$\frac{1.3d}{\epsilon^2} \le n_c \le \frac{d(1.36 + (1/3)\log d)}{\epsilon^2} \quad . \qquad \text{[P-An]} \qquad (6.14)$$

The relative distance criterion has improved the sample size estimate considerably, especially for small ϵ, as expected.

Finally, in the P-CT problem, the true model y is exactly representable in the class of neural networks, so $\nu_{\mathcal{T}}(w^*) = 0$ can be attained. In fact most realizations used in computing the sample averages satisfy that restriction, and almost all have $\nu_{\mathcal{T}}(w^*) \leq 0.01$. As discussed in §2.5, the values of $E\mathcal{E}(w^*)$ making up the learning curve do not directly correspond to our ϵ; we proposed to examine instead the *largest* of the values observed in 40 independent trials. Cohn and Tesauro (see figure 2.1) find that generally

$$\mathcal{E}(w^*) \leq \frac{1.3d}{n} \quad . \qquad \text{[P-CT]} \tag{6.15}$$

The estimate (6.13), based on our more stringent conception (uniform over networks) of reliable generalization, predicts that $\mathcal{E}(w^*) \leq 5.4d/n$ with high probability. The sample sizes disagree by about a factor of four.

§6.5 Summary

The correlation volume, defined in terms of the covariance function of Z, is easy to estimate because the covariances can be estimated using the training set. Using the correlation volume to analyze the self-normalized process leads to a simple integral over the weight space of the reciprocal of correlation volume, which can be done by Monte Carlo integration.

This procedure was tried on two examples. The first, that of learning orthants in the relative distance formulation, is one that can be analyzed directly, so it provides a test of the empirical approach. The two methods give very similar answers.

The second example, more interesting from the neural network point of view, is learning halfspaces. While this problem cannot be analyzed directly, the resulting estimates of generalization can be compared to experiments in training real neural networks. The empirical approach gives estimates of sample size needed for reliable generalization that are within a factor of four of the experimental values.

7 Conclusions

We are motivated by the recent success of applications of neural network methods to diverse problems in classification and estimation. These applications are typically distinguished by the following characteristics:

- the problem at hand is poorly understood statistically;
- nonlinear models of high complexity are used as estimators or as classifiers;
- the data models are selected on the basis of performance on a given training set.

In such a situation, prior knowledge of the characteristics of an optimal or near-optimal solution is vague and statistical assurance that the selected model has good performance is essential. Turning this reasoning around, the frequent success enjoyed by applications of neural networks indicates that some principle of generalization is at work.

There are several ways to make the notion of generalization precise. We have argued that 'reliable generalization' can usefully be taken to mean that sample size is large enough so that for some small ϵ

$$\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, \|_{\mathcal{W}} < \epsilon \text{ with high probability} \quad . \tag{7.1}$$

This clearly ensures that the true error of the chosen network agrees with its observed error. Furthermore, the criterion (7.1) allows conclusions of a global nature to be drawn about the performance of the chosen network relative to the best network in the class. Such information can be used to adjust the architecture in a principled way.

Existing results, pioneered by Vapnik, provide bounds that differ by orders of magnitude from the experience of neural network practitioners. Research to date on global measures of generalization such as (7.1) has been based on refinements of the original Vapnik tools. Much of this work applies only to the special case where the empirical error can be forced to zero by the training algorithm. These results imply a number of samples of order $(v/\epsilon^2)\log(1/\epsilon)$, or $(v/\epsilon)\log(1/\epsilon)$ if the empirical error is driven to zero.

We pursue a new approach to the problem. Starting with the idea that the difference process $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$, when properly scaled, is approximately normal, we introduce the Poisson clumping heuristic as a means to analyze the regions of weight space where a significant discrepancy between $\mathcal{E}(w)$ and its estimate $\nu_{\mathcal{T}}(w)$ exists. The heuristic tells us that the overall exceedance probability is the sum of the point exceedance probabilities diminished by a factor having to do with the typical area of such a region:

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \int_{\mathcal{W}} \frac{P(Z(w) > b)}{EC_b(w)}\, dw = \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{EC_b(w)}\, dw \quad . \tag{7.2}$$

The numerator is well-understood. Expressions for the denominator depend on having a good knowledge of the Z process, and our subsequent efforts are directed at calculating or approximating this weighting factor.

In a few cases the clump size can be found exactly; in chapter 4 we work out some examples to provide a reference point and to demonstrate the calculations involved. We find (results 4.6 and 4.10) that for the problem of learning with orthants in R^d and of learning with rectangles in R^d, d/ϵ^2 and $2\,d/\epsilon^2$ samples are needed, respectively. For the rather similar problem of learning halfspaces (the perceptron architecture), about $1.3\,d/\epsilon^2$ samples are needed (result 4.13). In the qualitatively different regime where the neural net outputs vary smoothly with changes in the weights, we again (result 4.16) find sample sizes of order d/ϵ^2, although this time the constant is given in terms of a hard-to-evaluate covariance matrix. These figures eliminate the $\log 1/\epsilon$ factor in the Vapnik bounds as well as showing that the constants are in some cases rather small, implying that for a few typical problems anyway, the Vapnik bounds are very loose. Thus the problem posed by the criterion (7.1) does *not* in itself impose unreasonable sample sizes.

We find that there are two ways to proceed: one is to work with (7.2) directly, and the other is to use a self-normalized version

$$P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \simeq \int_{\mathcal{W}} \frac{\bar{\Phi}(b)}{EC_b(w)}\, dw \tag{7.3}$$

where the clump size is now that of the normalized process. Using this criterion later in chapter 4 (result 4.18), we find that the number of samples needed for reliable generalization can be reduced to the order of $d(3.2 + 1.2 \log d + K)/\epsilon$, where the constant again depends on the architecture and distribution. Since the dependence on the unknown constant is rather weak, over a rather wide class of problems we find a dependence of small constants times $(d \log d)/\epsilon$.

We view these results as encouraging, but as suggesting a more pragmatic approach. In chapter 5 we develop two more notions of an appropriate 'weighting factor', the bundle size and the correlation volume. We show (corollary 5.6) that the bundle size figures into the rigorous lower bound

$$P(\|Z(w)\|_{\mathcal{W}} > b) \geq \int_{\mathcal{W}} \frac{\bar{\Phi}(b/\sigma(w))}{E[D_b \,|\, Z(w) > b]}\, dw \quad .$$

Through result 5.8, $E[D_b \,|\, Z(w) > b]$ is linked to the process covariance, a connection which is exploited via the introduction of the correlation volume $V_\tau(w)$. The latter is a rigorous lower bound to the bundle size, but by working a series of examples we discover that, when the parameter τ is chosen correctly, the bound becomes very tight indeed.

Finally in chapter 6 we recommend a method for estimating the correlation volume at a given point using the training data itself. Exceedance probability estimates are then given by

$$P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \simeq \frac{\bar{\Phi}(b)}{\bar{\Phi}(b\,\tau)} \int_{\mathcal{W}} V_\tau(w)^{-1}\,dw \quad .$$

The proposed method uses a Monte Carlo integral to estimate the outer integral, and searches locally about w to find the boundaries of $\mathcal{V}_\tau(w)$. The procedure is tested on two simple examples familiar from chapter 4: learning orthants and learning halfspaces. In the latter case, estimates of the number of samples sufficient to learn a function differ by only a factor of four from experimental results using the common backpropagation algorithm. The simulations we have done indicate that for some problems the new method can provide useful guidance to practice.

To balance these conclusions we discuss some limitations of this research. Two related objections concern the heuristic basis of our work. First, in determining rates of strong convergence of the empirical process $\nu_{\mathcal{T}}(w) - \mathcal{E}(w)$ to zero, we use a central limit theorem (for which no convergence rate is known) to model the process of error discrepancies. Second, while we do find asymptotic values for exceedance probabilities of the limiting process, it is not known how quickly the asymptote is approached. Both of these difficult questions are still the active concern of probability theory, but practically they are probably not the largest hole in our arguments. We note that the central limit theorem has very broad applicability so it is reasonable to expect it to hold in the relatively straightforward framework (independent summands which for binary classification are 0/1 random variables) considered here. Similarly, as we have noted in chapter 4, for those situations in which the exact (non-asymptotic) exceedance probability is known (primarily in low dimensions) the asymptotic value is approached quite rapidly. Simulations of the pinned Brownian sheet in $d = 2$, for example, show that the asymptotic formula of result 4.5 is quite close to correct even for $b = 1.25$; such results are not unusual [2, p. 135].

I believe there are two principal gaps in the story presented here. The first is the upper bound (which we use as an estimate) of mean clump size by $E[D_b \,|\, Z(w) > b]$. Although it is a physically reasonable approximation, and in the examples we consider it lowers the probability estimate by unimportant constant factors, adoption of the bundle size is a source of error which does not vanish as the level b becomes large. Furthermore, in contrast with the central limit theorem, little is known about how universal an approximation this is. The second gap is the estimate of correlation volume when certain boundaries of the corresponding set are known. Even if the correlation volume is convex and the estimates of the boundary locations are exact, there is an uncertainty of as much as $d!$ in the resulting volume estimate. Unless something is known about the geometry of the correlation volume, this error remains a problem.

Finally, we return to an incompleteness of this work connected with the stringent requirement (7.1). We have seen indications that in certain problems—the perceptron for example—there is little penalty (perhaps about a factor of four in sample size) in requiring simultaneous agreement of empirical and true error across all weight space as opposed to at w^* only. For discrete weight spaces (e.g. $\mathcal{W} = \{0,1\}^d$), this may no longer be true; as mentioned in §2.4, in certain experiments with 'discrete perceptrons' there is a critical sample size above which the generalization error of the chosen network drops suddenly to zero. While the framework we have presented gives reason to believe that $E\mathcal{E}(w^*) \simeq O(d/n)$ provided $(\exists w^0 \in \mathcal{W})\, \mathcal{E}(w^0) = 0$, these experiments show more complex behavior. We conjecture that the discrete space allows a sudden transition to perfect generalization by eliminating the possiblity of infinitesimal 'jitter' about a given point in weight space. If so, global agreement between empirical and true error may not always be of interest.

There are evidently more subtleties to the phenomenon of reliable generalization than are captured by any theory to date. It is anticipated that the ideas and methods presented here are a usable contribution to this, as yet unknown, theory.

A Asymptotic Expansions

§A.1 Stirling's formula

Recall the well-known estimate

$$d! \simeq \sqrt{2\pi d}\, d^d e^{-d} \approx \left(\frac{d}{e}\right)^d \quad . \tag{A.1}$$

On several occasions we will desire compact expressions for probabilities like

$$\begin{aligned} \frac{1}{d}\log\left(\frac{b^{2d}}{d!}\,\bar{\Phi}(b)\right) &\simeq 2\log b - \frac{\log(d!)}{d} - \frac{b^2}{2d} \\ &\simeq 2\log b - \log d/e - \frac{b^2}{2d} \\ &\simeq \log(b^2/d) + 1 - b^2/d \end{aligned}$$

The exponent we arrive at is compared to zero to see if b is large enough to force the probability low. It is this type of manipulation that forces us to consider the accuracy of the two representations (A.1) of the Stirling formula. The first, which we preferred not to use in the example above, has very high accuracy (even at $d = 1$) as the standard plot of relative error in figure A.1a shows. The logarithmic error of the simpler bound is shown in figure A.1b. For $d \geq 5$, this error is less than 1/3, which in the above example is negligible relative to the constant unity.

§A.2 The normal tail

Let $\phi(z)$ be the density of $N(0,1)$ and $1 - \bar{\Phi}(z)$ its cdf. The classical asymptotic expansion for $\bar{\Phi}$ is

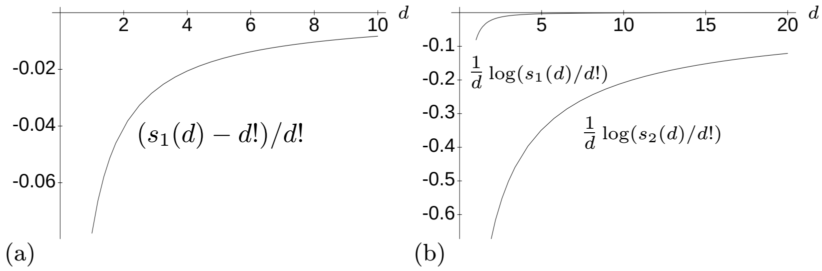

Figure A.1: Stirling's asymptotic expansion

Relative error of two forms of Stirling's approximation to the factorial, $s_1(d) = \sqrt{2\pi d}\, d^d e^{-d}$ and $s_2(d) = (d/e)^d$.

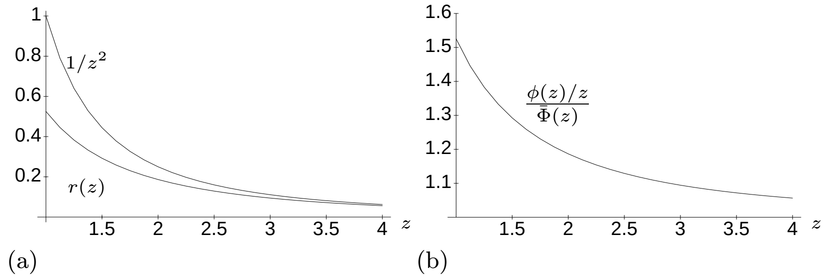

Figure A.2: The normal tail estimate

In the left panel, $r(z) = \left(\phi(z)/z - \bar\Phi(z)\right)/\,\bar\Phi(z)$.

A.1 Lemma *For $z > 0$,*

$$\frac{\phi(z)/z - \bar\Phi(z)}{\bar\Phi(z)} \leq 1/z^2 \quad .$$

Proof. Integrate by parts.

$$\begin{aligned}\bar\Phi(z) &= \int_z^\infty x\phi(x)\frac{dx}{x}\\ &= \phi(z)/z - \int_z^\infty \frac{\phi(x)}{x^2}\,dx\\ &\geq \phi(z)/z - \bar\Phi(z)/z^2 \quad \square\end{aligned}$$

Thus $\phi(z)/z \geq \bar\Phi(z)$, with the relative error going to zero as the argument squared. See figure A.2a. Figure A.2b shows that, for $z \geq 1$, the ratio between the two is less than about 1.5, and for $z \geq 3$, the ratio is less than 1.1, which is more than adequate for our purposes.

§A.3 Laplace's method

Laplace's method finds asymptotic expansions for integrals

$$\int_{\mathcal{W}} g(w)e^{-\lambda f(w)}\,dw$$

as $\lambda \to \infty$. The precise result is

A.2 Theorem *Let $f(w)$ be twice continuously differentiable with a unique positive minimum at w_0 in the interior of $\mathcal{W} \subseteq R^d$, and $g(w)$ be continuous at w_0. Write $K = \nabla\nabla f(w)|_{w_0}$ for the Hessian of f. Then, provided it converges absolutely for λ large enough,*

$$\int_{\mathcal{W}} g(w)e^{-\lambda f(w)}\,dw \simeq \sqrt{2\pi}^d g(w_0)\,|\lambda K|^{-1/2}\,e^{-\lambda f(w_0)}$$

in the sense that the ratio of the two sides goes to unity as $\lambda \to \infty$.

Proof. A proof is in [57, sec. IX.5], but the idea is simple. As $\lambda \to \infty$, the exponential peaks more sharply and only the behavior of f about the minimum matters, so expand f about w_0 and substitute into the exponential:

$$\int_{\mathcal{W}} g(w) e^{-\lambda f(w_0)} e^{-(w-w_0)^{\mathsf{T}}[\lambda K](w-w_0)/2}\, dw \\ \simeq g(w_0) e^{-\lambda f(w_0)} \int_{\mathcal{W}} e^{-(w-w_0)^{\mathsf{T}}[\lambda K](w-w_0)/2}\, dw$$

where In the second line we have used that g is changing slowly relative to the exponential. The integral is expanded to all of R^d—it is negligible away from w_0—and is easily performed. □

Letting $\lambda = b^2/2$ and $f(w) = 1/\sigma^2(w)$ above yields the more convenient

A.3 Corollary *Let $\sigma^2(w)$ be twice continuously differentiable with a unique maximum at w_0 in the interior of $\mathcal{W} \subseteq R^d$, and $g(w)$ be continuous at w_0. Then, provided it converges absolutely for large enough b,*

$$\int_{\mathcal{W}} g(w) e^{-b^2/2\sigma^2}\, dw \simeq \sqrt{2\pi}^{d} g(w_0) \left(\frac{\sigma_0}{b}\right)^{d} \left|\frac{-H/2}{\sigma_0^2}\right|^{-1/2} e^{-b^2/2\sigma_0^2}$$

where $H = \nabla\nabla\sigma^2(w)|_{w_0}$.

B Bounds by the Second-Moment Method

Here we show another connection between the exceedance probability estimate based on bundle size and rigorous bounds on that probability. We begin with the simple

B.1 Proposition *Any random variable $D \geq 0$ satisfies*

$$P(D > 0) \geq \frac{(ED)^2}{E(D^2)} \quad .$$

Proof. Apply the Cauchy-Schwarz inequality to $D1_{(0,\infty)}(D)$. □

As pointed out in [6, sec. A15], the relevance of this inequality to the PCH is as follows.

B.2 Corollary *If* $\mathrm{vol}(\mathcal{W}) < \infty$ *and* $Z(w)$ *is continuous then*

$$P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \geq \frac{\bar{\Phi}(b)\,\mathrm{vol}(\mathcal{W})^2}{\int_{\mathcal{W}} E[D_b \,|\, Z(w)/\sigma(w) > b]\, dw} \quad .$$

Proof. Take θ as Lebesgue measure and D as D_b in proposition B.1. The variance $\sigma^2(w)$ must be continuous, so $Z(w)/\sigma(w)$ is again continuous, and the preimage $Z^{-1}((b,\infty)) \subseteq \mathcal{W}$ is open, so the preimage is either empty or of positive Lebesgue measure. This means that

$$\begin{aligned} P(D_b > 0) &= P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \\ &\geq \frac{\big(\bar{\Phi}(b)\,\mathrm{vol}(\mathcal{W})\big)^2}{\iint P\left(\frac{Z(w')}{\sigma(w')} > b, \frac{Z(w)}{\sigma(w)} > b\right)\, dw'\, dw} \qquad \text{(B.1)} \\ &= \frac{\bar{\Phi}(b)\,\mathrm{vol}(\mathcal{W})^2}{\iint P\left(\frac{Z(w')}{\sigma(w')} > b \,\middle|\, \frac{Z(w)}{\sigma(w)} > b\right)\, dw'\, dw} \quad . \end{aligned}$$

For the normalized process Z/σ, the expression for $E[D_b \,|\, Z(w) > b]$ in (5.5) becomes the denominator, yielding the corollary. □

The lower bound developed in §5.1 for exceedance probability in the PCH context is

$$\begin{aligned} P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) &\geq \int_{\mathcal{W}} \frac{\bar{\Phi}(b)}{E[D_b \,|\, Z(w)/\sigma(w) > b]}\, dw \\ &\geq \frac{\bar{\Phi}(b)\,\mathrm{vol}(\mathcal{W})^2}{\int_{\mathcal{W}} E[D_b \,|\, Z(w)/\sigma(w) > b]\, dw} \end{aligned} \qquad \text{(B.2)}$$

where the second line follows via the harmonic mean inequality $Ef \geq (Ef^{-1})^{-1}$ (for $f > 0$). The second-moment method thus yields a lower bound on exceedance probability that is also obtainable as a further lower bound to the harmonic mean method result of corollary 5.6.

The same correspondence does not appear to obtain for the unnormalized process because the trick of converting the joint probability into a conditional one, used in (B.1), does not work.

C Calculations

§C.1 Vapnik Estimate of Sample Size

Here we derive the sample size estimate (2.2) in §2.1. We seek the sample size sufficient for

$$P(\| \, |\nu_{\mathcal{T}}(w) - \mathcal{E}(w)| \, \|_{\mathcal{W}} > \epsilon) \leq \tau \quad , \tag{C.1}$$

say. This will occur if the Vapnik upper bound (2.6) falls below the threshold τ:

$$\log 6 + v \log \frac{2e}{v} + v \log n - \epsilon^2 \frac{n}{4} \leq \log \tau \quad . \tag{C.2}$$

Deal with the $\log n$ term by approximating it linearly at a point $\alpha > 0$ to be determined:

$$\log n/\alpha \leq n/\alpha - 1 \implies \log n \leq \log \alpha + n/\alpha - 1 \quad , \tag{C.3}$$

which is an equality at $\alpha = n$. Using this bound, a sufficient condition for (C.2) is that

$$n \left(\frac{\epsilon^2}{4} - \frac{v}{\alpha} \right) \geq v \log \frac{2\alpha}{v} + \log \frac{6}{\tau} \quad .$$

Rewriting $\alpha = 4v\beta/\epsilon^2$ and rearranging yields

$$n \geq \frac{4v}{\epsilon^2(1-\beta^{-1})} \left[\log \frac{8\beta}{\epsilon^2} + v^{-1} \log \frac{6}{\tau} \right]$$

which is valid for any $\beta > 1$. Finding the minimizing β is difficult, but making a near-optimal choice is easy since in this regime the bracketed term is insensitive to β while the initial factor calls for $\beta \gg 1$. Choosing $\beta = 8$ (close to optimal for $\epsilon = 0.1$) yields

$$n \geq \frac{4.6v}{\epsilon^2} \left[\log \frac{64}{\epsilon^2} + v^{-1} \log \frac{6}{\tau} \right] \tag{C.4}$$

which is sufficient for (C.1). No choice of β could give an initial factor better than $4v/\epsilon^2$, so little has been lost beyond the imprecision already in (C.2).

Once this exponential bound drops below unity, it decreases precipitously toward zero. (E.g. even demanding $\tau = \epsilon^v$, infinitesimal for large v, does not affect the sample size materially because the second term in the bracketed factor is negligible.) It is thus convenient to find the "critical" sample size at which the exponential bound falls below unity. Dropping the negligible term shows this to be very nearly

$$n_c = \frac{9.2v}{\epsilon^2} \log \frac{8}{\epsilon} \quad . \tag{C.5}$$

We use this idea of a critical sample size, at which the exponential bound first becomes informative and almost simultaneously drives the probability of interest infinitesimally low, repeatedly in this work.

§C.2 Hyperbola Volume

The differential volume element used in (4.10) is

$$\mathrm{vol}(\{w \in [0,1]^d \ : \ z \leq \textstyle\prod_1^d w_i \leq z + dz\}) = dz \, \frac{(\log 1/z)^{d-1}}{(d-1)!} \quad , \tag{C.6}$$

which is the (negative) rate of change of the volume

$$I_d(z) := \{w \in [0,1]^d \ : \ \textstyle\prod_1^d w_i \geq z\} \tag{C.7}$$

of a hyperbolic region. Develop a recurrence for $I_d(z)$ by noting that it is built out of a continuum of lower-dimensional hyperbolic slices:

$$I_1(z) = \int_z^1 dw = 1 - z \tag{C.8a}$$

$$I_d(z) = \int_z^1 I_{d-1}(z/w)\, dw \qquad (d > 1). \tag{C.8b}$$

Computing the first few functions shows a simple pattern.

C.1 Lemma

$$I_d(z) = I_{d-1}(z) - z\, \frac{(\log 1/z)^{d-1}}{(d-1)!} \tag{$*$}$$

Proof. The hypothesis $(*)$ is easily checked for $d = 2$. Proceeding by induction, we compute

$$\begin{aligned} I_{d+1}(z) &= \int_z^1 I_d(z/w)\, dw \\ &= \int_z^1 I_{d-1}(z/w) - (z/w) \frac{(\log w/z)^{d-1}}{(d-1)!}\, dw \\ &= I_d(z) - z\, \frac{1}{(d-1)!} \int_1^{1/z} \frac{(\log r)^{d-1}}{r}\, dr \\ &= I_d(z) - z\, \frac{(\log 1/z)^d}{d!} \quad . \quad \square \end{aligned} \tag{C.9}$$

The lemma is then iterated to obtain

$$\begin{aligned} I_d(z) &= 1 - z \sum_{k=0}^{d-1} \frac{(\log 1/z)^k}{k!} \\ &\overset{\text{(a)}}{=} \frac{\gamma(d; \log 1/z)}{\Gamma(d)} \\ &\overset{\text{(b)}}{=} z\, \frac{(\log 1/z)^d}{d!}\, M(1, d; \log 1/z) \quad . \end{aligned} \tag{C.10}$$

Relation (a) follows from [1, eq. 6.5.13], where $\gamma(m,x)$ is the incomplete gamma function, and (b) follows from [1, eq. 6.5.12], with M the Kummer function. Differentiating the first expression yields a telescoping series which simplifies to (C.6).

§C.3 Rectangle Constant

We solve the recurrence (4.15):

$$I_1(z) := \int_{1/2z}^{1} dy \tag{C.11a}$$

$$I_d(z) := \int_{1/2z}^{1} \frac{1-y}{y} I_{d-1}(zy)\,dy \qquad (d > 1); \tag{C.11b}$$

Note that $I_d(\cdot)$ is only evaluated on $[1/2, 1]$.

C.2 Lemma

$$I_d(z) = \frac{1}{2z}\frac{(\log 2z)^{2d-1}}{(2d-1)!}\,M(d, 2d; \log 2z)$$

where M is the Kummer function [1, ch. 13].

Proof. For $d = 1$, note $M(1, 2; \log 2z) = (2z-1)/\log(2z)$ and the formula in question gives $I_1(z) = 1 - 1/2z$ as desired. For the inductive step, if the hypothesized formula holds everywhere on $[1/2, 1]$ for d then

$$\begin{aligned} I_{d+1}(z) &= \int_{1/2z}^{1} \frac{1-y}{y}\cdot\frac{1}{2zy}\frac{(\log 2zy)^{2d-1}}{(2d-1)!}\sum_{k=0}^{\infty}\frac{d^{\bar{k}}}{(2d)^{\bar{k}}}\frac{(\log 2zy)^k}{k!}\,dy \\ &= \frac{1}{2z}\sum_{k=0}^{\infty}\frac{\binom{d+k-1}{k}}{(2d+k-1)!}\int_{1/2z}^{1}\frac{1-y}{y^2}(\log 2zy)^{2d+k-1}\,dy \end{aligned}$$

Express the integrand as terms like $(\log 1/y)^m/y$ via the power series

$$\frac{1}{y^2} - \frac{1}{y} = \sum_{m=1}^{\infty}\frac{1}{m!}\frac{(\log 1/y)^m}{y}$$

and the binomial theorem, leaving us to integrate

$$\begin{aligned} I_{d+1}(z) &= \frac{1}{2z}\sum_{k=0}^{\infty}\frac{\binom{d+k-1}{k}}{(2d+k-1)!}\sum_{m=1}^{\infty}\frac{1}{m!}\sum_{n=0}^{2d+k-1}\binom{2d+k-1}{n} \\ &\qquad (\log 2z)^{2d+k-n-1}(-1)^n\int_{1/2z}^{1}\frac{(\log 1/y)^{m+n}}{y}\,dy \qquad (**) \\ &= \frac{1}{2z}\sum_{k=0}^{\infty}\frac{\binom{d+k-1}{k}}{(2d+k-1)!}\sum_{m=1}^{\infty}\frac{1}{m!}(\log 2z)^{2d+k+m}\times \\ &\qquad \sum_{n=0}^{2d+k-1}\binom{2d+k-1}{n}\frac{(-1)^n}{m+n+1} \quad . \end{aligned}$$

Because $2z \geq 1$, the only change in order of summation involving some negative quantities occurs with the *finite* sum on n in $(**)$. Since

$\sum_{r=0}^{N}\binom{N}{r}(-1)^r/(x+r) = \binom{x+N}{N}^{-1}/x$ (see (5.4.1) of Knuth's helpful book [27]) the final sum simplifies, and after rearrangement we have

$$\begin{aligned}
I_{d+1}(z) &= \frac{1}{2z}\sum_{k=0}^{\infty}\binom{d+k-1}{k}\sum_{m=1}^{\infty}\frac{(\log 2z)^{2d+k+m}}{(2d+k+m)!}\\
&= \frac{1}{2z}(\log 2z)^{2d+1}\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\binom{d+k-1}{k}\frac{(\log 2z)^{k+m}}{(2d+k+m+1)!}\\
&= \frac{1}{2z}(\log 2z)^{2d+1}\sum_{k=0}^{\infty}\sum_{r=k}^{\infty}\binom{d+k-1}{k}\frac{(\log 2z)^{r}}{(2d+1+r)!}\\
&= \frac{1}{2z}(\log 2z)^{2d+1}\sum_{r=0}^{\infty}\frac{(\log 2z)^{r}}{(2d+1+r)!}\sum_{k=0}^{r}\binom{d+k-1}{k}\\
&= \frac{1}{2z}(\log 2z)^{2d+1}\sum_{r=0}^{\infty}\frac{(\log 2z)^{r}}{(2d+1+r)!}\binom{d+r}{r}\\
&= \frac{1}{2z}\frac{(\log 2z)^{2d+1}}{(2d+1)!}M(d+1,2(d+1);\log 2z) \quad . \quad \square
\end{aligned}$$

We note in passing that the leading constant in learning orthants led to a similar recurrence whose solution also involved a Kummer function.

Originally we were interested only in $I_d = I_d(1)$. There is an equivalent formula for this which involves smaller powers of $\log 2$:

$$2I_{d+1} = (-1)^d\sum_{k=0}^{d}(2-(-1)^k)\frac{1}{k!}\binom{2d-k}{d}(-\log 2)^k \quad . \tag{C.12}$$

The equivalence can be checked by a lengthy manipulation of the generating functions.

Finally, there is the simple bound

C.3 Lemma

$$\sqrt{2} \le M(d,2d;\log 2) \le 2$$

Proof. Since $2^{-k} \le d^{\bar{k}}/(2d)^{\bar{k}} \le 1$,

$$e^{z/2} = \sum_{k=0}^{\infty}\frac{z^k}{k!}2^{-k} \le M(d,2d;z) \le \sum_{k=0}^{\infty}\frac{z^k}{k!} = e^z \quad . \qquad \square$$

§C.4 Perceptron Sample Size

In §4.4, we find

$$P(\|Z(w)\|_{\mathcal{W}} > b) \simeq \frac{\pi}{4}\,\frac{8^d}{\pi^{d/2}\Gamma(d/2)}\,K_{d,1}b^{2d-2}e^{-2b^2} \quad ; \tag{C.13}$$

the latter expression is defined as Q. As shown in §A.1, the crude form of Stirling's formula is adequate for our purposes here:

$$\frac{1}{d}\log Q \simeq \frac{1}{d}\log K_{d,1} + \frac{1}{2}\log\frac{128e}{\pi} + \log\frac{b^2}{\sqrt{d}} - 2\frac{b^2}{d} \tag{C.14}$$

The isotropic constant $K_{d,1}$ is bounded in result 4.12 as $d^{-d/2} \leq K_{d,1} \leq 1$. The lower approximation to exceedance probability becomes

$$\frac{1}{d} \log Q \geq \frac{1}{2} \log \frac{128e}{\pi} + \log \frac{b^2}{d} - 2\frac{b^2}{d} \tag{C.15}$$

which falls below zero at $b^2/d = 1.3$.

For the upper bound, with $\gamma = b^2/d$, (C.14) becomes

$$\frac{1}{d} \log Q \leq \log 10.5 + \log \gamma\sqrt{d} - 2\gamma \leq \log \frac{10.5}{\alpha e} - \gamma(2 - \alpha\sqrt{d})$$

where we have used

$$\log \gamma\sqrt{d} \leq -\log \alpha + \alpha\gamma\sqrt{d} - 1$$

as in (C.3). This means

$$\gamma = \frac{\log(10.5/\alpha e)}{2 - \alpha\sqrt{d}}$$

is enough to force $Q \leq 1$. Choosing the nearly optimal $\alpha = 1/(2\sqrt{d})$ shows

$$\gamma = b^2/d = 1.4 + (1/3) \log d \tag{C.16}$$

is sufficient.

§C.5 Smooth Network Sample Size

In corollary result 4.18 of §4.5, we seek the value of b for which

$$K^d (2\pi)^{d/2} b^d \, \bar{\Phi}(b) \leq 1 \tag{C.17}$$

Taking logs, dividing by d, and letting $\gamma = b^2/d$ gives

$$\log(2\pi K^2) + \log d + \log \gamma - \gamma \leq 0 \quad .$$

Using the standard upper bound to the logarithm (C.3) yields the sufficient condition

$$\gamma(1 - \alpha) \geq \log 2\pi K^2 + \log d - 1$$

or

$$\gamma \geq \frac{1}{1 - \alpha} \log \frac{2\pi K^2 d}{e\alpha} \quad .$$

For the tightest bound, select α so that $\alpha\gamma \approx 1$; somewhat arbitrarily we use $\alpha = 1/6$ resulting in a critical value of

$$\gamma = 3.2 + 1.2 \log(K^2 d) \quad . \tag{C.18}$$

§C.6 Finding Bundle Sizes

In §5.3 we introduced an upper bound and approximation for the mean clump size; here we compute this approximation for three covariance models. The two 'rough' covariances (cusped and isotropic) are treated simultaneously by defining the vector p-norm

$$|w|_p := \Big(\sum_{j=1}^{d} |w_j|^p\Big)^{1/p} \tag{C.19}$$

for $1 \le p \le \infty$. The two covariances can be written locally about w as

$$R(w, w+v) = \sigma^2 - |\Gamma v|_p \quad . \tag{C.20}$$

The cusped case is $p = 1$ and Γ diagonal; the isotropic case is $p = 2$ and Γ a multiple of the identity.

If the variance is constant, or if w is a local maximum of a smooth variance function, then variations of $R(w, w + v)$ dominate those of $\sigma(w + v)$. As far as ρ and ζ are concerned, the variance is constant so

$$\begin{aligned}
\rho(w, w+v) &\simeq 1 - \left|\tfrac{1}{\sigma^2}\,\Gamma v\right|_p \\
\zeta(w, w+v) &\simeq \left(\tfrac{1-\rho}{1+\rho}\right)^{1/2} \simeq \left(\tfrac{1}{2}\left|\tfrac{1}{\sigma^2}\,\Gamma v\right|_p\right)^{1/2} \quad .
\end{aligned}$$

The estimate of clump size is

$$\begin{aligned}
E[D_b \,|\, Z(w) > b] &\simeq \int \bar{\Phi}\big((b/\sigma)\zeta(w, w+v)\big)\, dv \\
&\simeq \int \bar{\Phi}\left(\left|\tfrac{b^2}{2\sigma^2}\,\tfrac{1}{\sigma^2}\,\Gamma v\right|_p^{1/2}\right) dv \\
&= 2^d \left|\Gamma/\sigma^2\right|^{-1} \left(\frac{\sigma}{b}\right)^{2d} \int \bar{\Phi}\left(|u|_p^{1/2}\right) du \\
&= 2^d \kappa_{p,d} \left|\Gamma/\sigma^2\right|^{-1} \left(\frac{\sigma}{b}\right)^{2d} d \int_0^\infty z^{d-1}\, \bar{\Phi}\big(\sqrt{z}\big)\, dz \quad .
\end{aligned} \tag{C.21}$$

In the final line we let $z = |u|_p$, and the volume element is the derivative of the volume of a generalized ball in R^d:

$$\begin{aligned}
\kappa_{p,d}\, z^d &:= \mathrm{vol}(\{u \in R^d \,:\, |u|_p \le z\}) = \frac{2^d \Gamma(1/p)^d}{p^{d-1} d\, \Gamma(d/p)}\, z^d \\
\mathrm{vol}(\{u \in R^d \,:\, |u|_p \in (z, z+dz)\}) &= \frac{d}{dz} \kappa_{p,d}\, z^d = d\, \kappa_{p,d}\, z^{d-1}\, dz \quad .
\end{aligned}$$

After a substitution, the remaining integral is done by parts.

$$\begin{aligned}
d \int_0^\infty z^{d-1}\, \bar{\Phi}(\sqrt{z})\, dz &= 2d \int_0^\infty r^{2d-1}\, \bar{\Phi}(r)\, dr \\
&= r^{2d}\, \bar{\Phi}(r)\Big|_0^\infty + \int_0^\infty r^{2d} \phi(r)\, dr \\
&= \frac{1}{2}\, \frac{(2d)!}{2^d d!} = \frac{1}{2\sqrt{\pi}}\, 2^d\, \Gamma(d + (1/2)) \quad ;
\end{aligned}$$

the last line is half the $2d$-th moment of $N(0, 1)$.

The mean bundle size becomes

$$E[D_b \,|\, Z(w) > b] \simeq \frac{1}{2}\,\frac{(2d)!}{d!}\,\frac{2^d\Gamma(1/p)^d}{p^{d-1}d\,\Gamma(d/p)}\,\left|\Gamma/\sigma^2\right|^{-1}\left(\frac{\sigma}{b}\right)^{2d} \tag{C.22}$$

The cusped covariance is the case $p = 1$ and the mean bundle size reduces to (5.14). With $p = 2$, using the alternate expression for the normal integral,

$$E[D_b \,|\, Z(w) > b] \simeq \frac{1}{d\sqrt{\pi}}\,(4\sqrt{\pi})^d\,\frac{\Gamma(d + (1/2))}{\Gamma(d/2)}\,\left(\frac{\sigma^2}{\gamma}\right)^d\left(\frac{\sigma}{b}\right)^{2d} \tag{C.23}$$

which has the same dependence on b and γ as $EC_b(w)$. The ratio of this to $EC_b(w)$ depends on the isotropic constant $K_{d,1}$ (see result 4.12) for which only bounds are known. Retaining only the factorials and exponential factors and using Stirling's formula, the ratio is

$$\begin{aligned}\frac{E[D_b \,|\, Z(w) > b]}{EC_b(w)} &\approx (4\sqrt{\pi})^d\,\frac{\Gamma(d + (1/2))}{\Gamma(d/2)}\,K_{d,1}\\ &\le (4\sqrt{\pi})^d\,\frac{\Gamma(d + (1/2))}{\Gamma(d/2)}\\ &\approx (8\sqrt{\pi})^d (d/2)! \quad ;\end{aligned} \tag{C.24}$$

$$\begin{aligned}\frac{E[D_b \,|\, Z(w) > b]}{EC_b(w)} &\ge (4\sqrt{\pi})^d\,\frac{\Gamma(d + (1/2))}{\Gamma(d/2)}\,d^{-d/2}\\ &\approx \left(8\sqrt{\pi/(2e)}\right)^d \quad .\end{aligned} \tag{C.25}$$

In the case of the smooth process, nonstationarity can play a role because changes in the variance are not dominated by other terms, thus affecting ρ and ζ. The local expansion (4.28) we have used is

$$R(w + v, w + v') \simeq \sigma_0^2 - \frac{1}{2}\begin{bmatrix} v^\mathsf{T} & v'^\mathsf{T}\end{bmatrix}\begin{bmatrix}\Lambda_{02} & -\Lambda_{11}\\ -\Lambda_{11} & \Lambda_{02}\end{bmatrix}\begin{bmatrix} v\\ v'\end{bmatrix}$$

and in these terms, after some simple algebra,

$$\rho(w, w + v) \simeq 1 - \tfrac{1}{2} v^\mathsf{T}\tfrac{\Lambda_{11}}{\sigma^2}\,v \tag{C.26a}$$

$$\zeta(w, w + v) \simeq \tfrac{1}{2}\left(v^\mathsf{T}\tfrac{\Lambda_{02}}{\sigma^2}\,v\right)^{1/2}\left(\frac{v^\mathsf{T}\Lambda_{02}\,v}{v^\mathsf{T}\Lambda_{11}\,v}\right)^{1/2} \quad . \tag{C.26b}$$

While integration of $\bar{\Phi}((b/\sigma)\zeta)$ is difficult, it is easy to obtain an upper bound that is tight enough to illustrate our point. As remarked previously, at a local maximum of variance, $-\nabla\nabla\sigma^2(w) = 2(\Lambda_{02} - \Lambda_{11})$ so the latter must be non-negative definite, implying that the Rayleigh quotient involved in ζ is at least unity. (In the constant-variance case the Hessian is zero so the quotient equals unity.) Then

$$\begin{aligned} E[D_b \,|\, Z(w)\!>\!b] &\simeq \int \bar{\Phi}((b/\sigma)\,\zeta(w,w+v))\,dv \\ &\le \int \bar{\Phi}\Big(\tfrac{b}{2\sigma}\,\big(v^{\mathsf{T}}\tfrac{\Lambda_{02}}{\sigma^2}\,v\big)^{1/2}\Big)\,dv \\ &= \int \bar{\Phi}\Big(\tfrac{b}{2\sigma}\,\big(v^{\mathsf{T}}\tfrac{\Lambda_{02}}{\sigma^2}\,v\big)^{1/2}\Big)\,dv \\ &= \big|\tfrac{\Lambda_{02}}{\sigma^2}\big|^{-1/2}\,\big(\tfrac{2\sigma}{b}\big)^d \int \bar{\Phi}\Big(\big(u^{\mathsf{T}}u\big)^{1/2}\Big)\,du \end{aligned} \tag{C.27}$$

and the remaining integral is done just as above:

$$\begin{aligned} \int \bar{\Phi}\Big(\big(u^{\mathsf{T}}u\big)^{1/2}\Big)\,du &= d\,\kappa_d \int_0^\infty \bar{\Phi}(r)r^{d-1}\,dr \\ &= \kappa_d \int_0^\infty \phi(r)r^d\,dr \\ &= \kappa_d\,\frac{1}{2\sqrt{\pi}}\,2^{d/2}\,\Gamma(\tfrac{1}{2}(d+1)) \\ &\approx (2\pi)^{d/2} \quad . \end{aligned} \tag{C.28}$$

In the last line we have used Stirling's formula and discarded $\sqrt{d}$ and small constant factors. For our purposes,

$$\begin{aligned} E[D_b \,|\, Z(w)\!>\!b] &\le 2^d\,(2\pi)^{d/2}\,\big|\Lambda_{02}/\sigma^2\big|^{-1/2}\left(\frac{\sigma}{b}\right)^d \\ &= 2^d\,EC_b(w) \end{aligned} \tag{C.29}$$

by comparison with result 4.14.

§C.7 Finding Correlation Volumes

In §5.5 we wish to compute the correlation volume for three covariance models. As in §C.6, both 'rough' covariances can be treated by the model (C.20); for such covariances

$$\begin{aligned} V_\tau(w) &= \mathrm{vol}\big(\{v \in R^d \,:\, \zeta \le \tau\}\big) \\ &= \mathrm{vol}\big(\{v \in R^d \,:\, \big|\tfrac{1}{\sigma^2}\,\Gamma v\big|_p \le 2\tau^2\}\big) \\ &= 2^d\,\big|\Gamma/\sigma^2\big|^{-1}\,\tau^{2d}\,\mathrm{vol}\big(\{u \in R^d \,:\, |v|_p \le 1\}\big) \\ &= 2^d\,\kappa_{p,d}\,\big|\Gamma/\sigma^2\big|^{-1}\,\tau^{2d} \end{aligned} \tag{C.30}$$

The bound to the mean bundle size is simply

$$E[D_b \,|\, Z(w)\!>\!b] \ge 2^d\,\kappa_{p,d}\,\big|\Gamma/\sigma^2\big|^{-1}\,\tau^{2d}\,\bar{\Phi}((b/\sigma)\,\tau) \quad .$$

Application of the asymptotic expansion for $\bar{\Phi}$ and differentiation shows the best threshold is

$$\tau^2(b/\sigma)^2 = 2d-1 \simeq 2d \quad ;$$

this yields

$$E[D_b \,|\, Z(w)\!>\!b] \ge (4d)^d\,\bar{\Phi}(\sqrt{2d}\,)\,\kappa_{p,d}\,\big|\Gamma/\sigma^2\big|^{-1}\,(\sigma/b)^{2d} \quad .$$

This estimate is exactly (C.22) with a different constant. The ratio between the exact integral and its lower bound is just

$$\frac{(2d)!/(2\,d!)}{(4d)^d\,\bar{\Phi}\,\sqrt{2d}} \simeq \sqrt{2\pi d} \quad , \tag{C.31}$$

as usual via Stirling's formula and the expansion for $\bar{\Phi}$, for any process parameters and level b.

As for the smooth process, its ζ function is above as (C.26) so

$$\mathcal{V}_\tau(w) = \{v \in R^d \,:\, \left(v^\mathsf{T} \tfrac{\Lambda_{02}}{\sigma^2}\, v\right) \left[\frac{v^\mathsf{T}\Lambda_{02}\, v}{v^\mathsf{T}\Lambda_{11}\, v}\right] \le 4\tau^2\} \quad .$$

The region is not an ellipse;[1] however, it can be bounded from within and without by ellipses. As discussed in §C.6, at a variance maximum the bracketed factor is at least one, so replacing it by unity yields a larger (elliptical) set. Since we wish to compare to $E[D_b \,|\, Z(w)\!>\!b]$, let us assume that the variance is constant, so $\Lambda_{02} = \Lambda_{11}$ and the bracketed term is exactly unity. This way both the integral for $E[D_b \,|\, Z(w) > b]$ and the correlation volume are exact results. The volume of the ellipse is trivially

$$V_\tau(w) = 2^d\,\kappa_d\,\left|\Lambda_{02}/\sigma^2\right|^{-1/2}\tau^d \quad . \tag{C.32}$$

The bundle size is at least as large as the correlation volume multiplied by $\bar{\Phi}((b/\sigma)\,\tau)$. Using the asymptotic expansion as before to find the tightest estimate yields

$$\tau^2(b/\sigma)^2 = d - 1 \simeq d \quad ;$$

the resulting estimate of $E[D_b \,|\, Z(w)\!>\!b]$ is

$$\begin{aligned} E[D_b \,|\, Z(w)\!>\!b] &\le 2^d d^{d/2} \kappa_d\,\bar{\Phi}(\sqrt{d})\,\left|\Lambda_{02}/\sigma^2\right|^{-1/2}\left(\frac{\sigma}{b}\right)^d && \text{(C.33)} \\ &\approx 2^d\,(2\pi)^{d/2}\,\left|\Lambda_{02}/\sigma^2\right|^{-1/2}\left(\frac{\sigma}{b}\right)^d && \text{(C.34)} \end{aligned}$$

where in the last line we have used Stirling's formula and dropped small factors. To compare to $E[D_b \,|\, Z(w) > b]$, take the ratio of the exact expression for $E[D_b \,|\, Z(w)\!>\!b]$ ((C.27), (C.28)) and (C.33):

$$\frac{2^{d/2}\,\Gamma\big(\tfrac{1}{2}(d+1)\big)/2\sqrt{\pi}}{d^{d/2}\,\bar{\Phi}(\sqrt{d})} \quad .$$

After some algebra, we see that the ratio is asymptotically $\sqrt{\pi d}$.

§C.8 Correlation Volume: Orthants with Relative Distance

We can find the correlation volume analytically for the situation of learning orthants in R^d with a normalized error criterion. Following §4.2, the covariance of the original process is

$$E\,Z(w)Z(w') = |w \wedge w'| - |w||w'|$$

[1] Some rewriting shows it is actually the 'polar inverse' of an elliptical region. That is, if $x^\mathsf{T}\mathbf{M}x \ge 1$, then $(x/\left|x\right|_2)/\left|x\right|_2$ is in the polar inverse.

where we again use the notation $|w| = \prod_{j=1}^{d} w_j$. To find the correlation volume, $\zeta(w, w+\delta)$ is needed; this is in general not the same as the ζ found previously for w a variance maximum.

Let $\delta \in R^d$ be small and partition indices into $\mathcal{J}^+ = \{j \le d : \delta_j > 0\}$ and $\mathcal{J}^-$; write $\mathcal{J} = \{1 \cdots d\}$. Let R, σ^2, and σ'^2 be the covariance and variances of the unnormalized Z process.

$$\begin{aligned}
R(w,w') &= \prod_{\mathcal{J}^+} w_j \cdot \prod_{\mathcal{J}^-}(w_j+\delta_j) - |w| \prod_{\mathcal{J}}(w_j+\delta_j) \\
&\simeq |w| + |w| \sum_{\mathcal{J}^-}(\delta_j/w_j) - |w|^2 - |w|^2 \sum_{\mathcal{J}}(\delta_j/w_j) \\
&= \sigma^2 + |w| \sum_{\mathcal{J}^-}(\delta_j/w_j) - |w|^2 \sum_{\mathcal{J}}(\delta_j/w_j)
\end{aligned}$$

Similarly,

$$\begin{aligned}
E\,Z(w')^2 &= |w+\delta| - |w+\delta|^2 \\
&\simeq \prod_{\mathcal{J}}(w_j+\delta_j) - \prod_{\mathcal{J}}(w_j^2 + 2\delta_j w_j) \\
&\simeq |w| + |w| \sum_{\mathcal{J}}(\delta_j/w_j) - |w|^2 - 2|w|^2 \sum_{\mathcal{J}}(\delta_j/w_j) \\
&= \sigma^2 + |w|(1-2|w|) \sum_{\mathcal{J}}(\delta_j/w_j)
\end{aligned}$$

To put these together easily, note

$$\begin{aligned}
\rho(w,w') &= R(w,w')/(\sigma\sigma') \\
&= \left[R(w,w')/\sigma^2\right] / \left[\sigma'^2/\sigma^2\right]^{1/2} \\
&= (1-\Delta_1)/(1-\Delta_2)^{1/2} \\
&\simeq 1 - \Delta_1 + \Delta_2/2
\end{aligned}$$

for appropriate terms Δ_1, Δ_2 which are of order δ. Using the expressions for the covariance and variance just derived, we find

$$\begin{aligned}
\rho(w,w') &\simeq 1 + \frac{|w|}{\sigma^2} \sum_{\mathcal{J}^-}(\delta_j/w_j) - \frac{1}{2}\frac{|w|}{\sigma^2} \sum_{\mathcal{J}}(\delta_j/w_j) \\
&= 1 - \frac{1}{2(1-|w|)} \sum_{\mathcal{J}} |\delta_j|/w_j \quad .
\end{aligned} \tag{C.35}$$

The correlation volume is defined by the point where $\big((1-\rho)/(1+\rho)\big)^{1/2} \le \tau$, or approximately $1-\rho \le 2\tau^2$, or

$$\sum_{j=1}^{d} |\delta_j|/w_j \ge 4(1-|w|)\tau^2 \quad .$$

This is a 'diamond-shaped' polyhedron in R^d with extreme points at $\delta_j = \pm 4(1-|w|)w_j\tau^2$ along each axis. Its volume is

$$V_\tau(w) \simeq 8^d |w| (1-|w|)^d \tau^{2d}/d! \tag{C.36}$$

where we have used the fact that the volume of the polyhedron with extreme points at ±1 along each axis is $2^d/d!$.

§C.9 Learning Orthants Empirically

In §6.3, we found the an empirical estimate of the exceedance probability

$$P\Big(\Big\|\tfrac{Z(w)}{\sigma(w)}\Big\|_{\mathcal{W}} > b\Big) \simeq e^4\, e^{-1.3\,d}\, d^d \left(\frac{b^2}{d}\right)^d \bar{\Phi}(b) \quad ; \tag{C.37}$$

we need an estimate of sample size. Dropping the leading constant and taking logs yields the criterion

$$\log\gamma - (1/2)\gamma - 1.3 + \log d \leq 0$$

after letting $\gamma = b^2/d$. Using the approximation (C.3) to $\log\gamma$ yields for any $\alpha \geq 0$

$$\gamma((1/2) - (1/\alpha)) \geq \log\alpha - 2.3 + \log d \quad .$$

On solving for γ we find, no matter what α is used, $\gamma > 2\log(Kd)$ for some constant K. Come close by letting $\alpha = 10$ so that

$$\gamma = b^2/d \geq 2.5\log(10\, e^{-2.3}\, d) = 2.5\log d \quad . \tag{C.38}$$

References

NIPS abbreviates the proceedings volumes titled *Advances in Neural Information Processing Systems.*

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover, 1965.

[2] R. J. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes.* Inst. of Mathemat. Stat., 1990.

[3] R. J. Adler and L. D. Brown. Tail behaviour for suprema of empirical processes. *Ann. Probab.*, 14(1):1–30, 1986.

[4] R. J. Adler and G. Samordonitsky. Tail behaviour for suprema of Gaussian processes with applications to empirical processes. *Ann. Probab.*, 15(4):1339–1351, 1987.

[5] D. Aldous. The harmonic mean formula for probabilities of unions: Applications to sparse random graphs. *Discrete Mathematics*, 76:167–176, 1989.

[6] D. Aldous. *Probability Approximations via the Poisson Clumping Heuristic.* Springer, 1989.

[7] K. S. Alexander. The central limit theorem for empirical processes on Vapnik-Červonenkis classes of sets. *Ann. Probab.*, 15:178–203, 1987.

[8] S. Amari and N. Murata. Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, 5(1):140–153, 1993.

[9] M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge Univ., 1992.

[10] E. Baum and D. Haussler. What size net gives valid generalization? In D. S. Touretzky, editor, *NIPS 1*, pages 81–90. Morgan-Kauffman, 1989.

[11] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Jour. Assoc. Comp. Mach.*, 36(4):929–965, 1989.

[12] D. Cohn and G. Tesauro. How tight are the Vapnik-Chervonenkis bounds? *Neural Computation*, 4(2):249–269, 1992.

[13] J. E. Collard. A B-P ANN commodity trader. In *NIPS 3*, pages 551–556. Morgan-Kaufmann, 1991.

[14] G. W. Cottrell and J. Metcalfe. EMPATH: Face, emotion, and gender recognition using holons. In *NIPS 3*, pages 564–571. Morgan-Kaufmann, 1991.

[15] A. Delopoulos, A. Tirakis, and S. Kollias. Invariant image classification using triple-correlation-based neural networks. *IEEE Trans. Neural Networks*, 5:392–408, 1994.

[16] M. Donsker. An invariance principle for certain probability limit theorems. *Mem. Amer. Math. Soc.*, 67:1–12, 1951.

[17] J. L. Doob. Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.*, 20:393–403, 1949.

[18] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Jour. Assoc. Comput. Mach.*, 38(1):1–17, 1991.

[19] B. Efron. *The Jackknife, the Bootstrap, and Other Resampling Plans.* SIAM, 1982.

[20] H. Fredholm et al. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. In *NIPS 3*, pages 523–529. Morgan-Kaufmann, 1991.

[21] L. Atlas et al. Performance comparisons between backpropagation networks and classification trees on three real-world applications. In *NIPS 2*, pages 622–629. Morgan-Kaufmann, 1990. Vowel classification section.

[22] Y. Le Cun et al. Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, pages 41–46, 1989.

[23] Y. Le Cun et al. Handwritten digit recognition with a back-propagation network. In *NIPS 2*, pages 396–404. Morgan-Kaufmann, 1990.

[24] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, second edition, 1971.

[25] T. Ferguson. *Mathematical Statistics: A Decision-Theoretic Approach.* Academic, 1967.

[26] P. Gaenssler and W. Stute. Empirical processes: A survey of results for independent and identically distributed random variables. *Ann. Probab.*, 7(2):193–243, 1979.

[27] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Addison-Wesley, second edition, 1994.

[28] V. Gullapalli. Learning control under extreme uncertainty. In *NIPS 5*, pages 327–334. Morgan-Kaufmann, 1993.

[29] P. Hall. *Introduction to the Theory of Coverage Processes.* Wiley, 1988.

[30] D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[31] D. Haussler, M. Kearns, and R. E. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. In *Proc. Fourth Ann. Workshop on Computational Learning Theory (COLT'91)*, pages 61–74. Morgan-Kauffman, 1991. Reprinted in *Machine Learning*, **14**, 83–113, 1994.

[32] M. L. Hogan and D. Siegmund. Large deviations for the maxima of some random fields. *Adv. Appl. Math.*, 7:2–22, 1986.

[33] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT, 1994.

[34] J. Kiefer. On large deviations of the empiric D.F. of vector chance variables and a law of the iterated logarithm. *Pac. Jour. Math.*, 11:649–660, 1961.

[35] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes.* Springer, 1983.

[36] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer, 1991.

[37] L. Lovász. Geometric algorithms and algorithmic geometry. In *Proceedings of the International Congress of Mathematicians.* The Mathematical Society of Japan, 1991.

[38] Y.-D. Lyuu and I. Rivin. Tight bounds on transition to perfect generalization in perceptrons. *Neural Computation*, 4(6):854–862, 1992.

[39] M. B. Marcus and L. A. Shepp. Sample behavior of Gaussian processes. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, 2:423–442, 1971.

[40] D. A. Mighell, T. S. Wilkinson, and J. W. Goodman. Backpropagation and its application to handwritten signature verification. In *NIPS 1*, pages 340–347. Morgan-Kaufmann, 1989.

[41] R. Nekovi and Ying Sun. Back-propagation network and its configuration for blood vessel detection in angiograms. *IEEE Trans. Neural Networks*, 6:64–72, 1995.

[42] M. O. Noordewier, G. G. Towell, and J. W. Shavlik. Training knowledge-based neural networks to recognize genes in DNA sequences. In *NIPS 3*, pages 530–536. Morgan-Kaufmann, 1991.

[43] D. Pollard. *Convergence of Stochastic Processes.* Springer, 1984.

[44] D. Pollard. *Empirical Processes: Theory and Applications.* NSF-CBMS Regional Conference Series. Inst. Math. Stat., 1990.

[45] D. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *NIPS 1*, pages 305–313. Morgan-Kaufmann, 1989.

[46] N. Sauer. On the density of families of sets. *Jour. of Combinatorial Theory A*, 13:145–7, 1972.

[47] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45(8):6056–6091, 1992.

[48] J. Shawe-Taylor, M. Anthony, and N. Biggs. Bounding sample-size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 42:65–73, 1993.

[49] H. Sompolinsky, N. Tishby, and H. S. Seung. Learning from examples in large neural networks. *Phys. Rev. Lett.*, 65:1683–1686, 1990.

[50] M. Talagrand. Small tails for the supremum of a Gaussian process. *Ann. Inst. Henri Poincaré*, 24(2):307–315, 1988.

[51] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.

[52] L. G. Valiant. A theory of the learnable. *Comm. Assoc. Comput. Mach.*, 27(11):1134–1142, 1984.

[53] V. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer, 1982.

[54] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. and its Apps.*, 16(2):264–280, 1971.

[55] T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Rev. Mod. Phys.*, 65:499–556, 1993.

[56] R. S. Wenocur and R. M. Dudley. Spme special Vapnik-Chervonenkis classes. *Discrete Mathematics*, 33:313–318, 1981.

[57] R. Wong. *Asymptotic Approximations of Integrals.* Academic, 1989.