April 1992

Convergence of Stochastic Processes

Robert Mandelbaum
University of Pennsylvania

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-92-30.

Convergence of Stochastic Processes

Abstract

Often the best way to adumbrate a dark and dense assemblage of material is to describe the background in contrast to which the edges of the nebulosity may be clearly discerned. Hence, perhaps the most appropriate way to introduce this paper is to describe what it is *not*. It is *not* a comprehensive study of stochastic processes, nor an in-depth treatment of convergence. In fact, on the surface, the material covered in this paper is nothing more than a compendium of seemingly loosely-connected and barely-miscible theorems, methods and conclusions from the three main papers surveyed ([VC71], [Pol89] and [DL91]).

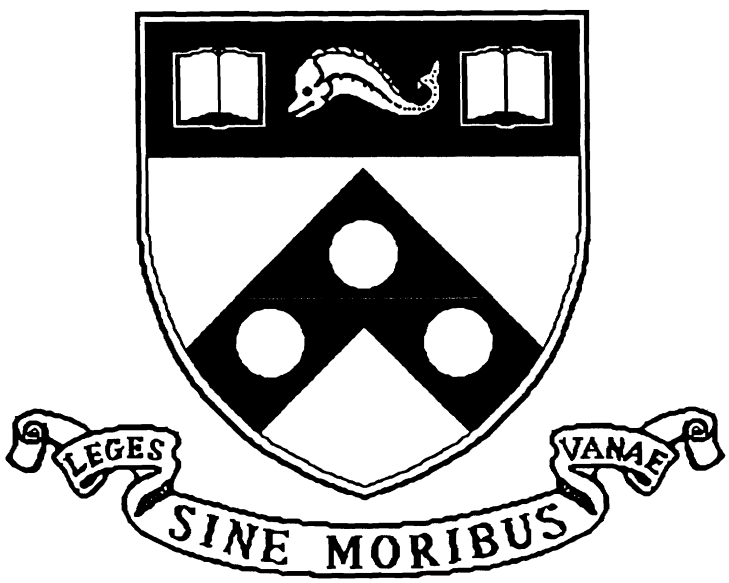
Comments

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-92-30.

Convergence of Stochastic Processes

MS-CIS-92-30
GRASP LAB 311

Robert Mandelbaum



University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

April 1992

Convergence of Stochastic Processes

Robert Mandelbaum
Department of Computer and Information Science
University of Pennsylvania

February 1992

Special Area Examination

Advisor: Max Mintz

Contents

Notation

Following the example of [Pol89] and [Pol84], linear function notation is used whenever it can cause no ambiguity. Hence, instead of $\int g(x)\,Q(dx)$ or $\int g\,dQ$ for the integral with respect to a measure Q, we write $Q(g)$ or simply Qg.

The Cartesian product symbol is $\otimes$. Maximum and minimum are represented by $\vee$ and $\wedge$ respectively. The integral symbol $\int$ appearing without limits refers to integration over the entire space. The inner product of two vectors x and ϕ is denoted $(x,\, \phi)$. Random variables are denoted X, Y and Z. CDF, PDF and PMF stand for Cumulative Distributive Function, Probability Density Function and Probability Mass Function respectively. The notation $Z \sim P$ indicates that random variable Z has distribution P, while $\mathbb{E}_P(Z)$ denotes the expectation of Z according to the distribution P. In keeping with the notation used in [?], expectation of Z with respect to the underlying probability measure will also be denoted by $\mathbb{P}\,Z$, as against $\mathbb{P}\,\{a \in A\}$ which denotes the measure of set A.

$N(m,\, \sigma^2)$ usually symbolises the Normal distribution of mean m and variance σ^2. Boldface type is reserved for sets and vectors. Calligraphic symbols are generally used as follows: $\mathcal{A}$ and $\mathcal{B}$ refer to classes of sets; $\mathcal{D}$, $\mathcal{G}$ and $\mathcal{H}$ refer to classes of functions; $\mathcal{F}$, $\mathcal{P}$ and $\mathcal{Q}$ refer to a classes of distributions; $\mathcal{L}^1$ and $\mathcal{L}^2$ are defined as in Section 2.6, while $\mathcal{P}(A)$ refers to the power set of a set A. $\Re$ represents the set of real numbers, while $\Re^+$ denotes the nonnegative reals.

$g = O(f)$ signifies that function g grows no faster than f. Similarly, $g = \Theta(f)$ signifies that g is order f, while $g = o(f)$ indicates that g has asymptotic growth strictly smaller than f. Finally, $g = O_P(f)$, $g = \Theta_P(f)$ and $g = o_P(f)$ indicate that the respective growth rates of g and f converge in probability to their prescribed asymptotic relationship.

All other symbols are defined on site.

For science, G-d is simply the stream of tendency by which all things seek to fulfil the law of their being.

LITERATURE AND DOGMA
William Arnold

1 Introduction

Often the best way to adumbrate a dark and dense assemblage of material is to describe the background in contrast to which the edges of the nebulosity may be clearly discerned. Hence, perhaps the most appropriate way to introduce this paper is to describe what it is *not*. It is *not* a comprehensive study of stochastic processes, nor an in-depth treatment of convergence. In fact, on the surface, the material covered in this paper is nothing more than a compendium of seemingly loosely-connected and barely-miscible theorems, methods and conclusions from the three main papers surveyed ([VC71], [Pol89] and [DL91]).

And yet, closer inspection reveals a common thread running steadily through the papers and delicately weaving them into a coherent and tightly-knit tapestry. It is the ambition of this paper both to describe the content and significance of each of the papers individually as well as to expose this elegant intertwining and interdependence.

The classical Bernoulli theorem states that in a sequence of n independent trials, the relative frequency of an event A converges (in probability) to the probability of that event as $n \to \infty$ [VC71]. The need often arises to ensure that this convergence is uniform over an entire class of events $\mathcal{A}$. In other words, representing the relative frequency of a set $A \in \mathcal{A}$ after n trials by $\nu_A^{(n)}$ and the probability of A by P_A, we require that for arbitrarily small ϵ,

$$\mathbb{P}\{\sup_{A \in \mathcal{A}} |\nu_A^{(n)} - P_A| > \epsilon\} \to 0 \text{ as } n \to \infty. \tag{1}$$

For instance, for a distribution function P over the real line $\Re$, and a class $\mathcal{A} = \{(-\infty, t] : t \in \Re\}$, the strong law of large numbers guarantees that the proportion

of points in an interval $(-\infty, t]$ converges almost surely to $P(t)$, while the classical Glivenko-Cantelli theorem strengthens the result by adding uniformity of convergence over all t [Pol84]. However, it turns out that even in the simplest of examples, this type of uniform convergence does **not** necessarily hold. The first of the three papers to be discussed here, [VC71], supplies criteria on the basis of which one may judge whether a given combination of distribution P and class $\mathcal{A}$ boasts such uniform convergence (see Section 3). In particular, the paper demonstrates that for an *arbitrary* P, any so-called Vapnik-Chervonenkis (VC) class of sets will exhibit uniform convergence.

While the main thread of the second paper, [Pol89], is claimed by the author to be merely a glimpse into the theory of empirical processes, one may also view the material there as a direct extension of the ideas presented in [VC71] and the generalization of the concept of uniform convergence to classes of *functions*. Instead of the relative frequency of a set $A \in \mathcal{A}$ after n trials, $\nu_A^{(n)}$, we now speak of the expectation of a function $g \in \mathcal{G}$ with respect to the *empirical measure* P_n (which puts mass n^{-1} at each of the sample points — see Section 4); instead of the probability of A, P_A, we now speak of the expectation of g with respect to the underlying distribution P. The uniformity result we are now after is that for arbitrarily small ϵ,

$$\mathbb{P}\{\sup_{g \in \mathcal{G}} (n^{1/2} |P_n g - P g|) > \epsilon\} \to 0 \text{ as } n \to \infty. \tag{2}$$

This extension is more closely entwined with the ideas in [VC71] than may at first be apparent. Indeed, if one considers the class of indicator functions $\mathcal{G} = \{I_A : A \in \mathcal{A}\}$ corresponding to the class of sets $\mathcal{A}$, then, besides a rescaling factor of $n^{1/2}$, the two uniformity goals (1) and (2) are seen to be identical: Since the empirical measure P_n puts mass n^{-1} at each of the sample points, $P_n I_A$ is easily identifiable as the relative frequency $\nu_A^{(n)}$, while in a similar fashion, $P I_A = P(A)$.

Moreover, just as [VC71] shows that a VC class of sets will satisfy requirement (1), [Pol89] shows that goal (2) is achieved for what Pollard terms *manageable* classes of functions. But the plexus does not end there: It turns out that if a class of functions $\mathcal{G} = \{g : \Psi \to \Re\}$ with bounded envelope G is such that $\{subgraph(g) : g \in \mathcal{G}\}$ is a *VC class* of subsets of $\Psi \otimes \Re$, then $\mathcal{G}$ is, in fact, a *manageable class* of functions [Pol89]. See Section 4 for the definition of $subgraph(g)$ as well for an exposé of the intricate relationship between VC classes of subsets and manageable classes of functions.

Any discourse on asymptotics must go hand in hand with a discussion of the *rates* at which convergence takes place. Indeed, concepts of 'rates of convergence' form the

very seam binding the delicate filigree of the asymptotic with the rather coarser and more ragged burlap of the finite.

The third paper surveyed, [DL91], considers a bound on the rate of convergence of an estimate $T_n(\boldsymbol{X}_n)$ (where $\boldsymbol{X}_n$ is the vector of n i.i.d. F sample points) to the value of a functional $T(F)$ of an unknown *distribution* $F \in \mathcal{F}$ uniformly over a class of distributions $\mathcal{F}$. The bound involves the *modulus of continuity* $b(\epsilon)$ [DL91] of the functional T over $\mathcal{F}$, and is shown to be attainable, *to within constants*, whenever T is linear and $\mathcal{F}$ is convex. See Section 5 for a general discussion of [DL91] as well as Subsection 5.5 where the implications of the caveat "to within constants" are analyzed from the perspective of Estimation Theory.

Once again, close scrutiny reveals a fine enmeshment of the ideas of [DL91] with those of [VC71] and [Pol89] which a cursory consideration may dispute. Indeed, given a class of functions $\mathcal{G}$, we can consider each function $g \in \mathcal{G}$ as a random variable with respect to the probability space $\langle \Re, \mathcal{B}, P \rangle$, where $\mathcal{B}$ denotes the Borel field on the real line $\Re$ and P is some probability measure of finite variance. Let $\mathcal{F}$ be the class of marginal distributions of the resultant stochastic process, and choose the (linear) functional $T(F)$, $F \in \mathcal{F}$ to be the expected value of F, i.e. $T(F) = P\,g$ where $F \in \mathcal{F}$ is the distribution of the random variable $g \in \mathcal{G}$. Convexifying $\mathcal{F}$ yields a form to which the results of Donoho&Liu are applicable, so that a bound on the rate of convergence to $T(F)$ of *any* estimate $T_n(\boldsymbol{X}_n)$, *including the empirical expectation* $P_n\,g$, uniformly over $\mathcal{F}$ may be deduced via the modulus of continuity $b(\epsilon)$. This is the approach taken in Section 6 where the methods of [DL91] are implemented to establish bounds on rates of uniform convergence for a VC class of subsets.

Of course, the results of [DL91] extend far beyond these rather constrained and contrived cases to incorporate *any* convex class of distributions $\mathcal{F}$ and *any* linear functional T (not just the expected value with respect to the probability space $\langle \Re, \mathcal{B}, P \rangle$). In many situations, even the conditions of convexity and linearity are not necessary; in fact, the power and generality of the results of [DL91] are such that they may very well assume a pivotal role in future research within this field.

This survey has the following structure: In Section 2 we review various concepts fundamental to the subsequent discussion. Sections 3, 4 and 5 comprise synopses of each of [VC71], [Pol89] and [DL91] in turn. Interconnections and interdependencies are elaborated upon where appropriate. Finally, Section 6 demonstrates how the results of [DL91] may be applied to classes of sets delineated in [VC71], while Section 6.3 gives a brief outline of how to extend the application to classes of functions described in [Pol89].

2 Revision of Basic Concepts

2.1 Linearity, Convexity, The Hölder Condition, Jensen's Inequality

2.1.1 Linearity

A nonempty set L is said to be a **linear space** ([KF70], page 118) if the following three axioms are satisfied:

1. L forms an Abelian group with respect to an operation '+'.[1]
2. Any field element α and any element $x \in L$ uniquely determine an element $\alpha x \in L$, called the *product* of α and x, such that $\alpha(\beta x) = (\alpha\beta)x$ and $1x = x$.
3. The operations of addition and multiplication defined above obey two distributivity laws: For all x, y, $\in L$
   a) $(\alpha + \beta)x = \alpha x + \beta x$;
   b) $\alpha(x + y) = \alpha x + \alpha y$.

A functional f defined on a linear topological space L is said to be **linear** on L if, for all x, $y \in L$ and arbitrary numbers α, β,

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y).$$

[1] In other words, any two elements x, $y \in L$ uniquely determine a third element $x + y \in L$, called the *sum* of x and y, such that

a) $x + y = y + x$ *(commutativity)*;
b) $\forall z \in L$, $(x + y) + z = x + (y + z)$ *(associativity)*;
c) There exists an *identity element* $0 \in L$ such that $\forall x \in L$, $x + 0 = x$;
d) For every $x \in L$, there exists an *inverse element* $-x$ such that $x + (-x) = 0$.

2.1.2 Convexity

Given a *real* linear space L, let x and y be any two points of L. Then the *segment* in L joining x and y refers to the set of all points in L of the form $\epsilon x + (1 - \epsilon)x$ for $0 \leq \epsilon \leq 1$.

A *set* $M \subset L$ is said to be **convex** if, whenever it contains two points x and y, it also contains the segment joining x and y.

A *functional* p defined on L is said to be **convex** if ([KF70], page 130)

1. $\forall x \in L,\ p(x) \geq 0$ *(non-negativity)*;
2. $\forall x \in L,\ \forall \alpha \geq 0,\ p(\alpha x) = \alpha p(x)$;
3. $\forall x, y \in L,\ p(x + y) \leq p(x) + p(y)$.

2.1.3 The Hölder Condition

A real-valued function f defined on an interval $X \in \Re$ is said to satisfy a **Hölder condition of exponent α** if

$$\forall x, y \in X,\ |f(x) - f(y)| \leq c|x - y|^{\alpha}$$

for some constant c ([Fal90], page 8). This property is also referred to as a **Lipschitz condition of exponent** α in many texts.

2.1.4 Jensen's Inequality

Let $Z : \Psi \rightarrow \Re^n$ be a random variable defined on the probability space $\langle \Psi, \mathcal{B}, P \rangle$, and let $g : \Re^n \rightarrow \Re$ denote a convex function. Under the assumption that both $I\!E(|Z|)$ and $I\!E(|g(Z)|)$ exist, **Jensen's Inequality** states that

$$g\left(I\!E(Z)\right) \leq I\!E\left(g(Z)\right).$$

2.2 Some Convergence Concepts

Let $\{Z_n : n = 1, 2, \ldots\}$ denote a sequence of real random variables $Z_n : \Psi \to \Re$ defined on the probability space $\langle \Psi, \mathcal{B}, P \rangle$. Denote the CDF of Z_n by F_{Z_n}.

$\{Z_n\}$ **converges in probability** to the random variable $Y : \Psi \to \Re$ if (this is denoted $Z_n \xrightarrow{P} Y$)

$$\forall \epsilon > 0, \forall \delta > 0, \exists N(\epsilon, \delta), \forall n > N(\epsilon, \delta), \; I\!P\{\psi \in \Psi : |Z_n(\psi) - Y(\psi)| < \epsilon\} > 1 - \delta.$$

$\{Z_n\}$ **converges almost surely** to the random variable $Y : \Psi \to \Re$ if

$$\forall \epsilon > 0, \forall \delta > 0, \exists N(\epsilon, \delta), \; I\!P\left[\bigcap_{k > N(\epsilon, \delta)} \{\psi \in \Psi : |Z_k(\psi) - Y(\psi)| < \epsilon\}\right] > 1 - \delta.$$

Almost sure convergence implies a joint occurrence of an infinite number of events having probability greater than $1 - \delta$. It is also known as *convergence with probability one (wp1)* ([CB90], page 214).

It is clear that almost sure convergence implies convergence in probability.

$\{Z_n\}$ **converges in distribution** to the random variable Y if

$$\lim_{n \to \infty} F_{Z_n}(x) = F_Y(x)$$

at all points $x \in \Re$ where $F_Y(x)$ is continuous ([CB90], page 216).

2.3 A Central Limit Theorem

Let $\{Z_n\}$ be a sequence of independent identically distributed (i.i.d.) random variables with finite mean m and variance σ^2. Then

$$\frac{\sum_{k=0}^{n-1} (Z_i - m)}{\sqrt{n}}$$

converges in distribution to a Gaussian random variable with mean 0 and variance σ^2 ([GD86], page 281).

2.4 The First Borel-Cantelli Lemma

Let $\{A_k \ : \ k = 1, 2, \ldots\}$ denote a sequence of events on the probability space $\langle \Psi, \mathcal{B}, P \rangle$. If $\sum_{k=0}^{\infty} P(A_k) < \infty$ then

$$P\{\psi \in \Psi \ : \ \psi \in A_k \text{ for infinitely many } k\} = 0.$$

Consult [Bil79], page 46 for proof and elaboration.

2.5 Stochastic Processes, Sample paths and Separability

A real-valued **stochastic process** is a collection $\{Z_t \ : \ t \in T\}$ of real random variables, all defined on a common probability space $\langle \Psi, \mathcal{B}, P \rangle$. The random variable Z_t depends on both t and the point $\psi \in \Psi$ at which it is evaluated. To emphasize its role as a function of two variables, write it as $Z(\psi, t)$. For fixed t, the function $Z(\cdot, t)$ is a measureable map from Ψ into $\Re$. For fixed ψ, the function $Z(\psi, \cdot)$ is called a **sample path** of the stochastic process. Consult [Pol84], page 1.

Let $\mathcal{A}$ denote a collection of Borel sets on the real line. A real stochastic process $\{Z_t \ : \ t \in T\}$ with a linear index set T is said to be **separable relative to $\mathcal{A}$** if there is a sequence $\{t_j\}$ of parameter values and a subset $\Lambda \subset \Psi$ of probability zero such that for any $A \in \mathcal{A}$ and any open interval I, the sets

$$\begin{aligned} S_1 &= \bigcap_{t \in I \cup T} \{\psi \ : \ Z_t(\psi) \in A\} \\ S_2 &= \bigcap_{t_j \in I \cup T} \{\psi \ : \ Z_{t_j}(\psi) \in A\} \end{aligned}$$

differ by at most a subset of Λ. Of particular importance in this paper is that if the class $\mathcal{A}$ is taken to be the class of closed sets, then for a separable process, the supremum and infimum over arbitrary intervals are measureable. This is because they agree almost everywhere with the supremum and infimum over countable parameter sets.

2.6 Metrics, Pseudometrics and $\mathcal{L}^1$ and $\mathcal{L}^2$ Norms.

2.6.1 Metrics and Pseudometrics

A **metric** for a nonempty set L is defined as a single-valued, nonnegative, real function $\rho : L \otimes L \rightarrow \Re^+$ which has the following three properties: For all x, y, $z \in L$,

1. $\rho(x, y) = 0$ if and only if $x = y$;
2. $\rho(x, y) = \rho(y, x)$ *(symmetry)*;
3. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ *(triangle inequality)*.

A **pseudometric** is defined similarly except with respect to property (1); for a pseudometric, $\rho(x, y)$ could be zero for some distinct pair x, y.

2.6.2 $\mathcal{L}^1$ and $\mathcal{L}^2$ Norms.

A functional p defined on a linear space L is said to be a **norm** in L if it has the following three properties:

1. p is finite and convex;
2. $p(x) = 0$ only if $x = 0$[2];
3. $p(\alpha x) = |\alpha| p(x)$ for all $x \in L$ and all α.

Recalling the definition of a convex functional, we deduce that a **norm** in L is a finite functional on L such that for all x, $y \in L$,

1. $p(x) > 0$, where $p(x) = 0$ if and only if $x = 0$;
2. $p(\alpha x) = |\alpha| p(x)$ for all α;
3. $p(x + y) \leq p(x) + p(y)$.

[2]For p an $\mathcal{L}^\alpha$ **norm** (see later), $x = 0$ almost everywhere.

Let L be a linear space equipped with a measure μ. Then $\mathcal{L}^1$ refers to the normed linear space of all real measureable functions g such that

$$\| g \|_1 \triangleq \int |g(x)| \, d\mu < \infty.$$

$\| g \|_1$ denotes the **$\mathcal{L}^1$-norm.**

$\mathcal{L}^2$ denotes the normed linear space of all real measureable functions such that $\int g^2(x) \, d\mu < \infty$. The **$\mathcal{L}^2$-norm** is defined as

$$\| g \|_2 \triangleq \sqrt{\int g^2(x) \, d\mu}$$

Consult [KF70], page 381 for details.

3 On the Uniform Convergence of Relative Frequencies of Events to their Probabilities

A synopsis of [VC71] by Vapnik and Chervonenkis

As discussed in the Introduction, the classical Bernoulli theorem states that in a sequence of l i.i.d. trials, the relative frequency of an event A converges (in probability) to the probability of that event as $l \to \infty$ [VC71]. The need often arises to ensure that this convergence is uniform over an entire class of events $\mathcal{A}$. In other words, representing the relative frequency of a set $A \in \mathcal{A}$ after l trials by $v_A^{(l)}$ and the probability of A by P_A, we require that for arbitrarily small ϵ,

$$\begin{aligned} I\!P\{\pi^{(l)} > \epsilon\} &\rightarrow 0 \text{ as } l \rightarrow \infty, \text{ where} \\ \pi^{(l)} &= \sup_{A \in \mathcal{A}} |v_A^{(l)} - P_A| \end{aligned}$$

The main thread of [VC71] comprises two strands:

(1) Sufficient conditions on $\mathcal{A}$ for uniform convergence are derived. These conditions do *not* depend on the probability distribution P, and are discussed in section 3.2. Classes of sets which satisfy these conditions have been dubbed *'Vapnik Chervonenkis (VC)'* classes, [Pol89] or classes *of polynomial discrimination* [Pol84].

(2) Sufficient *and* necessary conditions for uniform convergence are deduced. These conditions *do* depend on the probability distribution P and are elaborated upon in Section 3.3.

Before describing these results and their elegant derivations, we need a few supporting definitions and concepts.

3.1 Cake-Cutting, Growth functions and the Shattering of Classes of Sets

3.1.1 The Cake-Cutting Conundrum

Any enthusiast for conundra and puzzles is no doubt familiar with the problem of determining, as a function of r, the maximum number of pieces into which a cake E may be partitioned using at most r slices. Let us extend the problem to include n-dimensional cakes being partitioned by means of r slices ($(n-1)$-dimensional hyperplanes). Denote by $\Phi(n, r)$ the maximum number of pieces obtainable.

In order to obtain a recurrence relation for $\Phi(n, r)$, consider the case where the first $r-1$ hyperplanes have already been placed so as to maximize the number of compartments into which the 'cake' E^n has been partitioned. All that remains to be done is to place the final rth hyperplane.

Now, for $n \geq 2$, any two non-parallel $(n-1)$-dimensional hyperplanes intersect along an $(n-2)$-dimensional hyperplane. Hence, when the rth hyperplane is inserted, it will be traversed by at most $r-1$ hyperplanes, each of which is $(n-2)$-dimensional. Further, since the rth hyperplane will form one of the boundaries of any new compartments added, the maximum number of new compartments will equal the maximum number of $(n-1)$-dimensional segments into which these $(n-2)$-dimensional hyperplanes partition the rth hyperplane itself [Wen62], [Sch50]. See Figure 1.

Hence, $\Phi(n, r)$ is seen to obey the recurrence relation

$$\Phi(n, r) = \Phi(n, r-1) + \Phi(n-1, r-1), \text{ where } \Phi(0, r) = 1 \text{ and } \Phi(n, 0) = 1$$

It is not difficult to show by induction that

$$\Phi(n, r) = \begin{cases} \sum_{k=0}^{n} \binom{r}{k} & \text{if } r > n \\ 2^r & \text{if } r \leq n \end{cases}$$

and, hence, that for $n > 0$ and $r \geq 0$,

$$\Phi(n, r) \leq r^n + 1 \tag{3}$$

In what follows, essential use is made both of $\Phi(n, r)$ and of inequality (3).

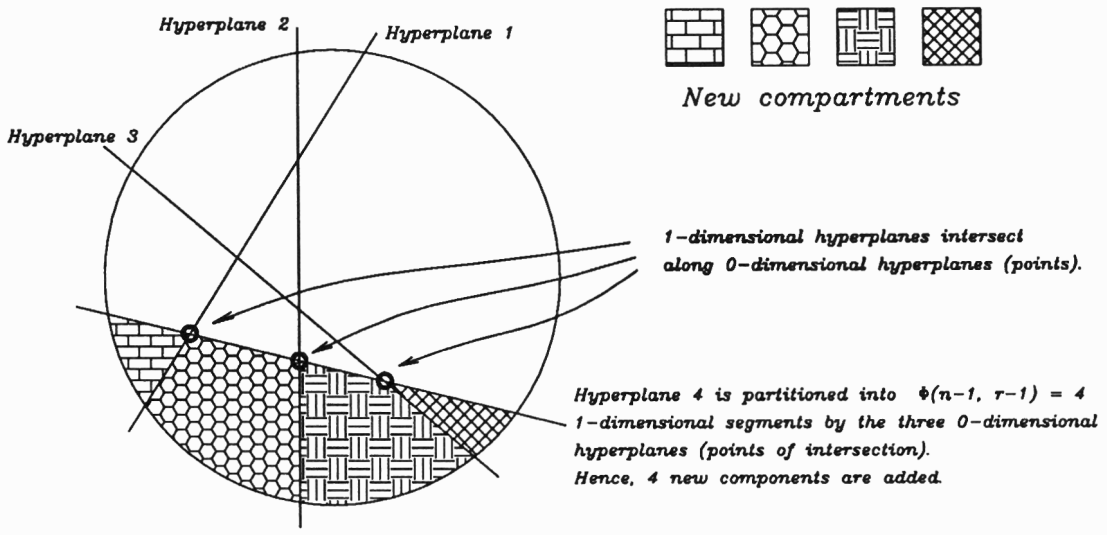


Figure 1: **2-Dimensional Cake being partitioned using 4 one-dimensional hyperplanes.** The number of new compartments added by the fourth slice is seen to be $\Phi(n-1, r-1) = 4$.

3.1.2 Fruit Cakes and Growth Functions

Let us now consider a slightly different cake-cutting problem. Let there be a set $\boldsymbol{X}_r$ of r different fruit chunks scattered throughout the cake E^n. Denote the positions of the fruit chunks within the cake by $x_1, x_2, \ldots, x_r$.

Instead of a knife with which to trace out hyperplanes, we have a host of implements with which it is possible to extract any one of a class $\mathcal{A}$ of cake pieces. Note that $\mathcal{A}$ does *not* necessarily delineate a partition of the cake since the potential pieces of cakes may intersect one another.

Now, each piece $A \in \mathcal{A}$ picks out or *induces* the subsample $\boldsymbol{X}_r^A$ of fruit chunks. The problem is to calculate the number of *different groupings* of fruit chunks which may be extracted by the class $\mathcal{A}$. We term this number the *index of* $\mathcal{A}$ *with respect to* $\boldsymbol{X}_r$ and denote it by $\Delta^{\mathcal{A}}(x_1, x_2, \ldots, x_r)$. Obviously, $\Delta^{\mathcal{A}}(x_1, x_2, \ldots, x_r)$ is always at most 2^r, the cardinality of $\mathcal{P}(\boldsymbol{X}_r)$. The maximum of $\Delta^{\mathcal{A}}(x_1, x_2, \ldots, x_r)$ over all possible positionings of the fruit chunks is called the *growth function* and is denoted by $m^{\mathcal{A}}(r)$.

In what follows, we generalize from cakes to any set $\boldsymbol{X}$. For a more formal definition of the growth function $m^{\mathcal{A}}(r)$, see [VC71], subsidiary definition 1.1.

Example 1: If $\boldsymbol{X}$ is the real line $\Re$ and $\mathcal{A}$ is the set of semi-infinite intervals of the form $(-\infty, a]$, $a \in \Re$, then $m^{\mathcal{A}}(r) = r + 1$. [VC71]

Example 2: If $\boldsymbol{X}$ is Euclidean 2-space, E^2, and $\mathcal{A}$ is the set of quadrants of the form $(-\infty, t]$, $t \in \Re \otimes \Re$, then $m^{\mathcal{A}}(r) \leq (r+1)^2$ since there are at most $r + 1$ places to set each of the horizontal and vertical boundaries [Pol84]. More precisely, $m^{\mathcal{A}}(r) = 1 + \frac{r}{2} + \frac{r^2}{2}$. ([Pol84], problem II.8).

Example 3: If $\boldsymbol{X}$ is the segment [0, 1] and $\mathcal{A}$ is the class of all open sets, then $m^{\mathcal{A}}(r) = 2^r$. [VC71]

Example 4: Let $\boldsymbol{X}$ be Euclidean n-space E^n and $\mathcal{A}$ be the class of all half-spaces of the form $(x, \phi) \geq 1$, $x \in \boldsymbol{X}$, for all fixed n-vectors ϕ. Let E^n be the space of vectors x and $\overline{E^n}$ be the space of vectors ϕ.

As shown in Figure 2, to each vector x_k there corresponds a hyperplane in $\overline{E^n}$ dividing $\overline{E^n}$ into the two half-planes

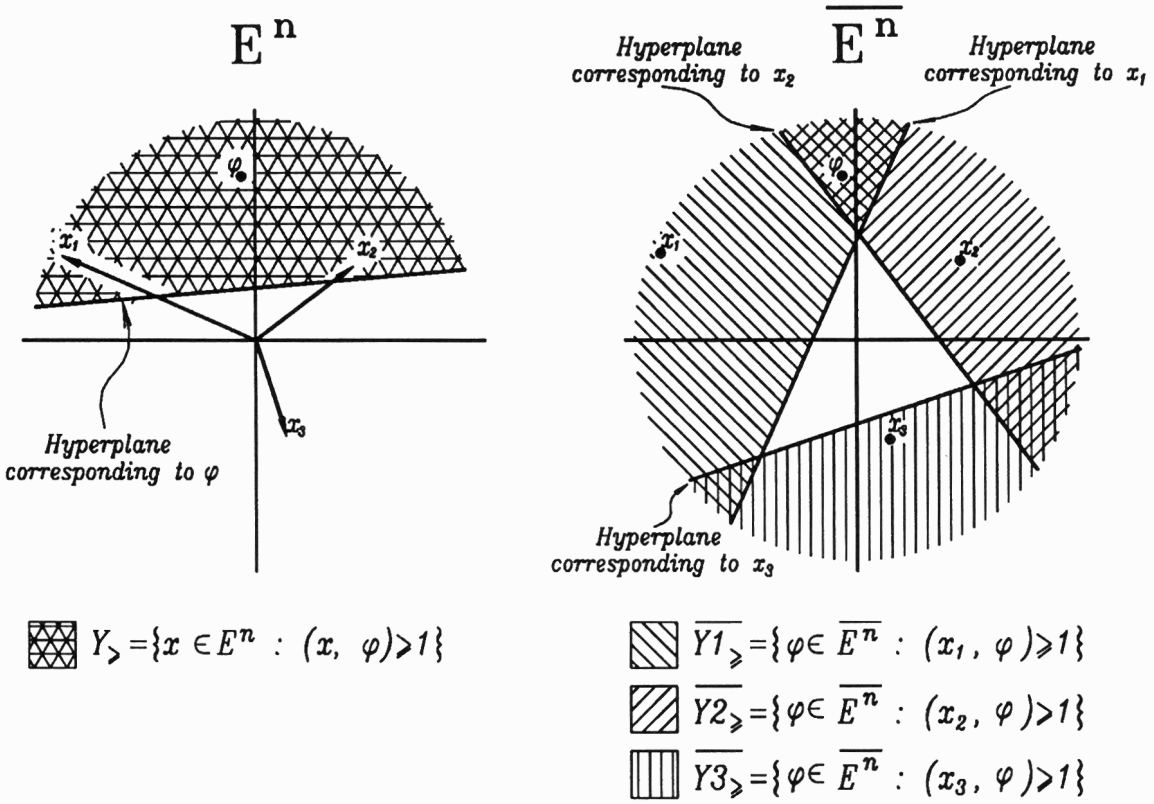
$$\begin{aligned} \overline{Y_{\geq}} &= \{\phi \in \overline{E^n} \,:\, (x_k, \phi) \geq 1\} \text{ and} \\ \overline{Y_{<}} &= \{\phi \in \overline{E^n} \,:\, (x_k, \phi) < 1\} \end{aligned}$$

Making the return journey to E^n, we find that each ϕ_k partitions E^n similarly into

$$\begin{aligned} Y_{\geq} &= \{x \in E^n \,:\, (x, \phi_k) \geq 1\} \text{ and} \\ Y_{<} &= \{x \in E^n \,:\, (x, \phi_k) < 1\} \end{aligned}$$

The critical observation is that for a fixed vector $x_k \in E^n$, if ϕ_k is *any* vector in $\overline{Y_{\geq}}$, then x_k is in $Y_{\geq}$. Similarly, $\phi_k \in \overline{Y_{<}} \Rightarrow x_k \in Y_{<}$.

Hence, any set of r points in E^n, $\boldsymbol{X}_r$=$x_1, x_2, \ldots, x_r$ induces a set of r hyperplanes in $\overline{E^n}$ which partition $\overline{E^n}$ into a number of compartments such that the vectors ϕ



Compartment of $\overline{E^n}$:							
Subset of $\{x_1, x_2, x_3\}$ induced:	$\{x_1\}$	$\{x_2\}$	$\{x_3\}$	$\{\}$	$\{x_1, x_2\}$	$\{x_2, x_3\}$	$\{x_1, x_3\}$

Figure 2: **Correspondence between subsamples of the set $\boldsymbol{X}_r = x_1, x_2, \ldots, x_r$ and compartments of $\overline{E^n}$.** Each point x_k is seen to induce a hyperplane r_k in $\overline{E^n}$, while *any* vector ϕ in a certain compartment of $\overline{E^n}$ induces the same subsample of $\boldsymbol{X}_r$.

from any single compartment induce half-planes $Y_{\geq}$ and $Y_{<}$ in E^n which, though different for each ϕ, *all pick out the same subsample* $\boldsymbol{X}_r^Y$ of $\boldsymbol{X}_r$. Thus, finding the growth function for this example is equivalent to finding the maximum number of compartments into which $\overline{E^n}$ may be partitioned, a problem already addressed in Section 3.1.1. i.e. for this example, $m^{\mathcal{A}}(r) = \Phi(n, r)$.

3.1.3 Shattering Classes of Sets

A class $\mathcal{A}$ of subsets of a universe $\boldsymbol{X}$ is said to *shatter* a set of points $\boldsymbol{X}_r \subset \boldsymbol{X}$ if it can pick out every possible subset of $\boldsymbol{X}_r$ [Pol84]. In other words, $\mathcal{A}$ shatters $\boldsymbol{X}_r$ if $\Delta^{\mathcal{A}}(x_1, x_2, \ldots, x_r) = 2^r$. As pointed out in [Pol84], the choice of the term 'shatter' is perhaps inappropriate, implying violent fragmentation of $\boldsymbol{X}_r$ rather than meticulous extraction of each individual subset, "... but at least it is vivid" [Pol84].

Example 5: Consider the class $\mathcal{A}$ of closed disks in E^2: $\mathcal{A}$ can shatter any set of three non-collinear points, but cannot shatter *any* set of four points [Pol84].

All of the above definitions and concepts are elegantly united in **Theorem 1** of [VC71] which states that for any class of sets $\mathcal{A}$, $m^{\mathcal{A}}(r)$ is either identically equal to 2^r or else is majorized by $\Phi(n, r)$, where n is the smallest sample size which $\mathcal{A}$ cannot shatter, no matter what the sample configuration (e.g. in **Example 5** above, $n = 4$). In turn, we have shown in Equation (3) that for $r > 0$, $\Phi(n, r) < r^n + 1$.

Hence, $m^{\mathcal{A}}(r)$ is either equal to 2^r or is polynomial in nature, with the order of the polynomial being the value n as defined above (For a proof of this theorem, see [VC71]). Classes $\mathcal{A}$ for which the latter condition hold are said to be *of polynomial discrimination* since they pick out at most a polynomial number of subsamples of $\boldsymbol{X}_r$; they have also been dubbed *Vapnik Chervonenkis* classes in the literature.

3.2 Distribution-Independent Sufficient Conditions

We now return to the problem of finding conditions under which we can be assured of uniform convergence of relative frequency to probability over a class $\mathcal{A} \subset \mathcal{B}$ of events with respect to the probability space $\langle \boldsymbol{X}, \mathcal{B}, P \rangle$.

To relieve the reader of the torments of suspense, we state the main result of the first part of [VC71] here: It turns out (See [VC71], **Corollary** to **Theorem 2**) that a sufficient condition for this uniform convergence to occur is merely that the class of events $\mathcal{A}$ be of polynomial discrimination with respect to the whole space $\boldsymbol{X}$.

This simple result is a consequence of some rather involved yet elegant applications of concepts from combinatorics and probability theory. We give here a brief outline of the general argument and refer the reader to [VC71] for the details.

Step 1: Symmetrization. Instead of working with $\pi^{(l)} = \sup_{A \in \mathcal{A}} |\nu_A^{(l)} - P_A|$ directly, define a class of new random variables $\rho_A^{(l)} = |\nu'_A - \nu''_A|$, $A \in \mathcal{A}$, where ν'_A and ν''_A are the relative frequencies of a set $A \in \mathcal{A}$ for two independent samples of size l. Define further the maximum difference between ν'_A and ν''_A over the entire class $\mathcal{A}$, $\rho^{(l)} = \sup_{A \in \mathcal{A}}(\rho_A^{(l)})$. We assume throughout that both stochastic processes $\{\rho_A^{(l)} : A \in \mathcal{A}\}$ and $\{|\nu_A^{(l)} - P_A| : A \in \mathcal{A}\}$ are separable, or at least that $\rho^{(l)}$ and $\pi^{(l)}$ are measureable.[3]

[3]As an example of a $\pi^{(1)}$ which would **not** be measureable, consider a universe $\boldsymbol{X} = [0, 1]$ and an index set $T = [0, 1]$. Let S be a non-measureable subset of $\boldsymbol{X}$, and let P be Lebesque measure. Define the class $\mathcal{A} = \{A_t : t \in T\}$ as follows:

$$A_t = \begin{cases} [0, 1/2] - \{t\} & \text{if } t \in S, t \leq 1/2 \\ [0, 1/2] \cup \{t\} & \text{if } t \in S, t > 1/2 \\ [0, 1] - \{t\} & \text{if } t \notin S \end{cases}$$

Hence, for $x \in \boldsymbol{X}$,

$$|\nu_{A_t}^{(1)} - P_{A_t}|(x) = \begin{cases} 1/2 & \text{for all } x \in \boldsymbol{X}, t \in S \\ 0 & \text{for } x \neq t, t \notin S \\ 1 & \text{for } x = t, t \notin S \end{cases}$$

Thus, $\pi^{(1)} = \frac{1}{2} I_S + I_{S'}$ which is non-measureable and hence **not** a random variable. (Adapted from [Mintz], page 304).

Lemma 2 of [VC71] establishes that there is a strong relationship between $\rho^{(l)}$ and $\pi^{(l)}$. More precisely,

$$I\!P\left\{\pi^{(l)} > \epsilon\right\} \leq I\!P\left\{\rho^{(l)} \geq \frac{\epsilon}{2}\right\} \tag{4}$$

In other words, if $\rho^{(l)} \to 0$ as $l \to \infty$, then $\pi^{(l)} \to 0$ (in probability) which is the result we are after. In **Step 2** below this type of convergence of $\rho^{(l)}$ is demonstrated for classes $\mathcal{A}$ of polynomial discrimination.

Step 2: Permutations. Now all that remains to be done is to place bounds on

$$I\!P\left\{\rho^{(l)} \geq \frac{\epsilon}{2}\right\} = I\!P\left\{(\boldsymbol{X}'_l, \boldsymbol{X}''_l) \in \boldsymbol{X}^l \otimes \boldsymbol{X}^l \ : \ \sup_{A\in\mathcal{A}} |\upsilon_A^{(l)}(\boldsymbol{X}'_l) - \upsilon_A^{(l)}(\boldsymbol{X}''_l)| \geq \frac{\epsilon}{2}\right\}.$$

We note three simplifications which are immediately applicable:

(1) Independence allows us to concatenate the two l-samples used in calculation of υ'_A and υ''_A into a single $2l$-sample, $\boldsymbol{X}_{2l}$,

(2) For a fixed sample $\boldsymbol{X}_{2l}$, instead of taking the supremum of $|\upsilon'_A - \upsilon''_A|$ over *all* of $\mathcal{A}$, we need consider only those sets which induce essentially *different* subsamples in $\boldsymbol{X}_{2l}$. Denote the class of all such sets by $\mathcal{A}'$. By definition, $|\mathcal{A}'| = \Delta^{\mathcal{A}}(\boldsymbol{X}_{2l})$, and

(3) We can partition the class of all *ordered samples* of size $2l$ of the universe $\boldsymbol{X}$ into equivalence classes, each indexed by a *subset* $X \subset \boldsymbol{X}$ of size $2l$, where

$$[X] = \{\boldsymbol{X}_{2l} = T_i(X) \ : \ i \in \{1, 2, \ldots, (2l)!\}\}$$

and T_i is a permutation of the elements of X.

Hence,

$$I\!P\left\{\boldsymbol{X}_{2l} \in \boldsymbol{X}^{2l} \ : \ \sup_{A\in\mathcal{A}} \rho_A^{(l)}(\boldsymbol{X}_{2l}) > \frac{\epsilon}{2}\right\} = \int \theta\left[\sup_{A\in\mathcal{A}'}\left(\rho_A^{(l)}(\boldsymbol{X}_{2l}) - \frac{\epsilon}{2}\right)\right] dP$$

where $\mathcal{A}'$ depends on the choice of $\boldsymbol{X}_{2l}$ pursuant to simplification (2) above, and $\theta : \Re \to \{0, 1\}$ is the indicator function for the subset $[0, \infty) \subset \Re$.

In turn, since the supremum over a class of non-negative functions cannot be greater than the superposition of the functions,

$$\int \theta \left[\sup_{A \in \mathcal{A}'} \left(\rho_A^{(l)}(\boldsymbol{X}_{2l}) - \frac{\epsilon}{2} \right) \right] dP \leq \int \sum_{A \in \mathcal{A}'} \theta \left(\rho_A^{(l)}(\boldsymbol{X}_{2l}) - \frac{\epsilon}{2} \right) dP$$

And now for the step which makes use of permutations of the sample $\boldsymbol{X}_{2l}$. Since all samples $\boldsymbol{X}_{2l}$ in an equivalence class $[X]$ generate the same class of sets $\mathcal{A}'$, we have

$$\int \sum_{A \in \mathcal{A}'} \theta \left(\rho_A^{(l)}(\boldsymbol{X}_{2l}) - \frac{\epsilon}{2} \right) dP = \int \sum_{A \in \mathcal{A}'} \left[\frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta \left(\rho_A^{(l)}(T_i \boldsymbol{X}_{2l}) - \frac{\epsilon}{2} \right) \right] dP$$

The crucial observation is that the innermost summation represents the total number of arrangements of a fixed sample $\boldsymbol{X}_{2l}$ for which $\rho_A^{(l)} \geq \epsilon/2$. But if A picks out m elements in $\boldsymbol{X}_{2l}$, then $\rho_A^{(l)} \geq \epsilon/2$ for any arrangement in which k of these m elements fall in one l-sample and $|v'_A - v''_A| = |\frac{k}{l} - \frac{m-k}{l}| \geq \epsilon/2$. Hence, the expression in brackets, call it Γ, may be rewritten as

$$\Gamma = \sum_{\{k \,:\, |2k/l - m/l| \geq \epsilon/2\}} \frac{\binom{m}{k}\binom{2l-m}{l-k}}{\binom{2l}{l}}$$

Now, since $|\mathcal{A}'| \leq m^{\mathcal{A}}(2l)$ for all samples $\boldsymbol{X}_{2l}$ and Γ satisfies $\Gamma \leq 2e^{-\epsilon^2 l/8}$, we can combine all the relations back to Equation (4) to yield the succint inequality

$$I\!P\left\{\pi^{(l)} > \epsilon\right\} \;\leq\; 4m^{\mathcal{A}}(2l)\, e^{-\epsilon^2 l/8} \tag{5}$$

Finally, for any Vapnik-Chervonenkis class $\mathcal{A}$, $m^{\mathcal{A}}(2l) \leq (2l)^n$ so that Inequality (5) implies uniform convergence:

$$\lim_{l \to \infty} I\!P\{\pi^{(l)} > \epsilon\} \leq 4 \lim_{l \to \infty} (2l)^n\, e^{-\epsilon^2 l/8} = 0$$

Actually, an even stronger result follows from Inequality (5): A simple application of the first Borel-Cantelli Lemma (See Section 2.4) guarantees almost sure convergence. For details, consult [VC71].

Note that nowhere in this derivation did we have to impose criteria on the properties of the distribution P. This is a testament to the power of the result.

3.3 Distribution-Dependent Necessary and Sufficient Conditions

The second major strand of the [VC71] paper completes the finely woven arras by providing a sufficient *and* necessary condition for relative frequencies to converge (in probability) to probabilities uniformly over a class of events $\mathcal{A}$.

Since the mathematical justification of the validity of this condition is relatively complex and does not lend itself readily to simplification, nor does it contribute to the conceptual clarity of the ideas, we omit most of it here and refer the reader to [VC71] itself. Instead we merely state the results and discuss their importance.

Once again, we need first a definition.

Entropy. In section 3.1.2 we defined the **index** of a class $\mathcal{A}$ with respect to a sample $\boldsymbol{X}_l$ as the number of different subsamples of $\boldsymbol{X}_l$ which $\mathcal{A}$ can pick out. We denoted this index by $\Delta^{\mathcal{A}}(\boldsymbol{X}_l)$. We also defined the **growth function** $m^{\mathcal{A}}(l)$ as the *maximum* value of $\Delta^{\mathcal{A}}(\boldsymbol{X}_l)$ over all possible samples of size l. We now turn our attention to a function which reflects the *expected* value of $\Delta^{\mathcal{A}}(\boldsymbol{X}_l)$ with respect to the underlying distribution P. Define

$$H^{\mathcal{A}}(l) = \mathbb{E}_P \log_2 \Delta^{\mathcal{A}}(\boldsymbol{X}_l)$$

$H^{\mathcal{A}}(l)$ is dubbed the **entropy** of the system of events $\mathcal{A}$ in samples of size l [VC71]. The concept correlates well with the thermodynamic idea of entropy; indeed the greater the entropy of $\mathcal{A}$ within samples of size $\boldsymbol{X}_l$, the greater is $\mathcal{A}$'s discriminatory power, and the less the l elements of $\boldsymbol{X}_l$ are permitted to 'cluster' together.

Our main interest is in the ratio of entropy to sample size, $H^{\mathcal{A}}(l)/l$, as $l \to \infty$. In fact, the key result is that convergence of $H^{\mathcal{A}}(l)/l \to 0$ as $l \to \infty$ is both a sufficient and necessary condition for the desired uniform convergence of relative frequencies to probabilities.

We give here an outline of the argument validating this claim. First of all, we define the random variable $\xi^{(l)} = [\log_2 \Delta^{\mathcal{A}}(\boldsymbol{X}_l)]/l$, so that $H^{\mathcal{A}}(l)/l = \mathbb{E}_P \xi^{(l)}$. Now, **Lemma 3** of [VC71] states that $H^{\mathcal{A}}(l)$ has a limit c, $0 \leq c \leq 1$, as $l \to \infty$. **Lemma 4** augments this by showing that for large l, the distribution of $\xi^{(l)}$ is concentrated near c. Indeed, for any $\epsilon > 0$, $\lim_{l\to\infty} \mathbb{P}\{|\xi^{(l)} - c| > \epsilon\} = 0$, showing convergence in probability of $\xi^{(l)}$ to c. Observe that the requirement that $H^{\mathcal{A}}(l) \to 0$ as $l \to \infty$ is equivalent to the requirement that $c = 0$.

3.3.1 Proof of Sufficiency

To prove *sufficiency* of this requirement, let us now partition the space of l-samples $\boldsymbol{X}^l$ into two regions: $\boldsymbol{X}^l_1 = \{\log_2 \Delta^{\mathcal{A}}(\boldsymbol{X}_l) \leq \epsilon^2 l/8\}$ for some $\epsilon > 0$, and $\boldsymbol{X}^l_2 = \boldsymbol{X}^l - \boldsymbol{X}^l_1$. Since these sets are disjoint and exhaustive, invoking elementary set theory[4],

$$\begin{aligned} I\!P\left\{\pi^{(l)} > \epsilon\right\} &= I\!P\left(\left\{\pi^{(l)} > \epsilon\right\} \cap \boldsymbol{X}^l_1\right) + I\!P\left(\left\{\pi^{(l)} > \epsilon\right\} \cap \boldsymbol{X}^l_2\right) \\ &\leq I\!P\left(\left\{\pi^{(l)} > \epsilon\right\} \cap \boldsymbol{X}^l_1\right) + I\!P\left(\boldsymbol{X}^l_2\right) \end{aligned}$$

Now, by definition, within $\boldsymbol{X}^l_1$, $\Delta^{\mathcal{A}}(\boldsymbol{X}_l) \leq 2^{\epsilon^2 l/8}$ Further, with $c = 0$, $I\!P\left(\boldsymbol{X}^l_2\right) = I\!P\left\{\xi^{(l)} > \epsilon^2/8\right\} \to 0$ as $l \to \infty$ by **Lemma 4** of [VC71]. Hence, invoking Equation 5 above, we see that

$$I\!P\left\{\pi^{(l)} > \epsilon\right\} \leq 4 \cdot 2^{\epsilon^2 l/8} e^{-\epsilon^2 l/8} + I\!P\left(\boldsymbol{X}^l_2\right) = 4(2/e)^{\epsilon^2 l/8} + I\!P\left(\boldsymbol{X}^l_2\right)$$

The right hand expression converges to zero as l goes to infinity. Hence,

$$I\!P\left\{\pi^{(l)} > \epsilon\right\} \to 0 \text{ as } l \to \infty$$

3.3.2 Proof of Necessity

To establish *necessity* of the condition $\lim_{l\to\infty} H^{\mathcal{A}}(l)/l = 0$, we resort to an argument by contradiction, showing that the supposition $\lim_{l\to\infty} H^{\mathcal{A}}(l)/l = c > 0$ implies the existence of a positive ϵ such that $\lim_{l\to\infty} I\!P\{\sup_{A\in\mathcal{A}} |\upsilon'_A - \upsilon''_A| > 2\epsilon\} = 1$. A bound similar to Inequality 4, namely

$$I\!P\left\{\sup_{A\in\mathcal{A}} |\upsilon^{(l)}_A - P_A| > \epsilon\right\} \geq \frac{1}{2} I\!P\left\{\sup_{A\in\mathcal{A}} |\upsilon'_A - \upsilon''_A| > 2\epsilon\right\}$$

then abrogates uniform convergence of relative frequencies to probabilities. (See [VC71] for details).

Intuitively, this condition imposed on $H^{\mathcal{A}}(l)$ amounts to ensuring that the expected value of the index of $\mathcal{A}$ increase at a rate strictly smaller than the rate of proliferation of subsets of the sample $\boldsymbol{X}_l$ with l. In other words, even if the growth function $m^{\mathcal{A}}(l)$

[4] We assume here that $\boldsymbol{X}^l_1$, $\boldsymbol{X}^l_2 \in \mathcal{B}$ where $\mathcal{B}$ is the set of events in the probability space $\langle \boldsymbol{X}^l, \mathcal{B}, P\rangle$.

increases exponentially, uniform convergence is assured as long as the *expected* value of $\Delta^{\mathcal{A}}(\boldsymbol{X}_l)$ is a member of the $o(2^l)$ class[5].

As a final note, we observe that though this is a fine result, it is attained at the expense of both independence from distribution properties as well as almost surety of convergence. It is shown in [Pol84] that both of these desireable properties may be reinstated with a slight alteration of the condition $H^{\mathcal{A}}(l)/l \to 0$ as $l \to \infty$. Indeed, **Theorem 21** of [Pol84] states that a necessary and sufficient condition for *almost sure* convergence of relative frequencies of events in a class $\mathcal{A}$ to their probabilities is $(n_l/l) \xrightarrow{P} 0$ where $n_l = n_l(\boldsymbol{X}_l)$ is the smallest integer such that $\mathcal{A}$ shatters no collection of n_l points from $\boldsymbol{X}_l$. We refer the reader to [Pol84], Section II.4 and problems II.11 and II.12 for a proof of sufficiency and necessity.

[5]Actually, the conditions are less stringent even than this: Thanks to the concavity of the logarithmic function,

$$H^{\mathcal{A}}(l) = \mathbb{E}_P \log_2 \Delta^{\mathcal{A}}(\boldsymbol{X}_l) \leq \log_2 \mathbb{E}_P \Delta^{\mathcal{A}}(\boldsymbol{X}_l)$$

so that $H^{\mathcal{A}}(l)$ could still satisfy the criterion even if $\mathbb{E}_P \Delta^{\mathcal{A}}(\boldsymbol{X}_l)$ exhibited exponential growth.

4 Asymptotics via Empirical Processes

A synopsis of [Pol89] by David Pollard

In Section 3, we discussed conditions under which the relative frequencies of events in a class $\mathcal{A}$ serve as asymptotically good estimates of the probabilities of the events *uniformly* over $\mathcal{A}$. Consider now the extension of these ideas to a class of functions $\mathcal{G} = \{g : \Psi \to \Re\}$, where, for each $g \in \mathcal{G}$, we are interested in Pg, the expected value of g with respect to some probability measure P in the probability space $\langle \Psi, \mathcal{B}, P \rangle$.

Define the **empirical measure** P_n as that measure which places mass n^{-1} at each of n sample points, $x_1, \ldots, x_n \in \Psi$. An intuitive estimate for Pg is then the expected value of g with respect to this empirical measure. In other words, we estimate the mean of g by the average of the n evaluations of g at the sample points $x_1, \ldots x_n$.

Our quest is then criteria under which $P_n g$ provides an asymptotically good estimate of Pg *uniformly* over $\mathcal{G}$. This is the subject of this section.

Note that seen in this light, the material covered in [VC71] and reviewed in Section 3 emerges as a special case of the more general case involving function classes. Indeed, with $\mathcal{G}$ as the class of indicator functions $\mathcal{G} = \{I_A : A \in \mathcal{A}\}$, the determination of probabilistic bounds on the worst case difference between the true mean of a function and its expectation with respect to the empirical measure reduces to the determination of probabilistic bounds on the worst case difference between the relative frequency and the probability of a set.

The main topic of [Pol89] is an exposition of a very powerful technique for the analysis of the entire family of problems involving averages of functions of independent observations, of which the problem scrutinized here — that of finding criteria under which these averages converge uniformly to the expected values of the functions — is a member.

Let us now cast the problem into notation consistent with that used in [Pol89]. Define the **empirical process** $\nu_n = \{n^{1/2}(P_n - P)g : g \in \mathcal{G}\}$ for a class of functions $\mathcal{G}$. ν_n may be thought of as an operator acting on g to produce a properly standardized sample average [Pol89]. As stated in the Introduction, the uniformity result we are now after is that for arbitrarily small ϵ,

$$I\!P\{\sup_{g \in \mathcal{G}} (n^{1/2}|P_n g - P g|) > \epsilon\} = I\!P\{\sup_{g \in \mathcal{G}} |\nu_n g| > \epsilon\} \to 0 \text{ as } n \to \infty.$$

The empirical process method for establishing criteria under which the above result holds comprises four main steps [Pol89]:

1. Beginning with a family of averages, symmetrize via the introduction of a new source of randomness. Instead of analyzing the difference between an empirical expectation and the true mean, we are now looking at the difference between two independent empirical expectations.

2. Transform the symmetrized process of averages into a conditionally Gaussian stochastic process.

3. Apply a recursive method known as *chaining* to exploit the rapid decay of Gaussian tails and bound the process probabilistically by an integral involving a *capacity* function.

4. Derive conditions on the function class $\mathcal{G}$ subject to which the necessary uniform bound on the capacity function is attained. This bound then percolates through the integral derived in STEP 3 above, and manifests itself as the required bound on the original empirical process.

Figure 3 presents schematically the thread of our mini-tour through the labyrinth of empirical processes. In order to present the material in a modular fashion, we will discuss Gaussian Processes and the Chaining method first and then return to the four-step method outlined above. Though this ordering may seem haphazard, familiarity with Gaussian Processes and the Chaining method *in principle* will later obviate the need to break the continuity of the argument with a meandering excursion into clarification of the supporting definitions.

4.1 Maximal Inequalities for Gaussian Processes

As stated in Section 2.5, a stochastic process is any collection of random variables $\{Y_t : t \in T\}$. A process is said to be **Gaussian** if every finite subcollection of these random variables has a *joint normal distribution* [Pol89]. Let us now consider the problem of finding a bound on the expectation of $\sup_{t \in T} |Y_t|$ where $\{Y_t : t \in T\}$ is a Gaussian process.

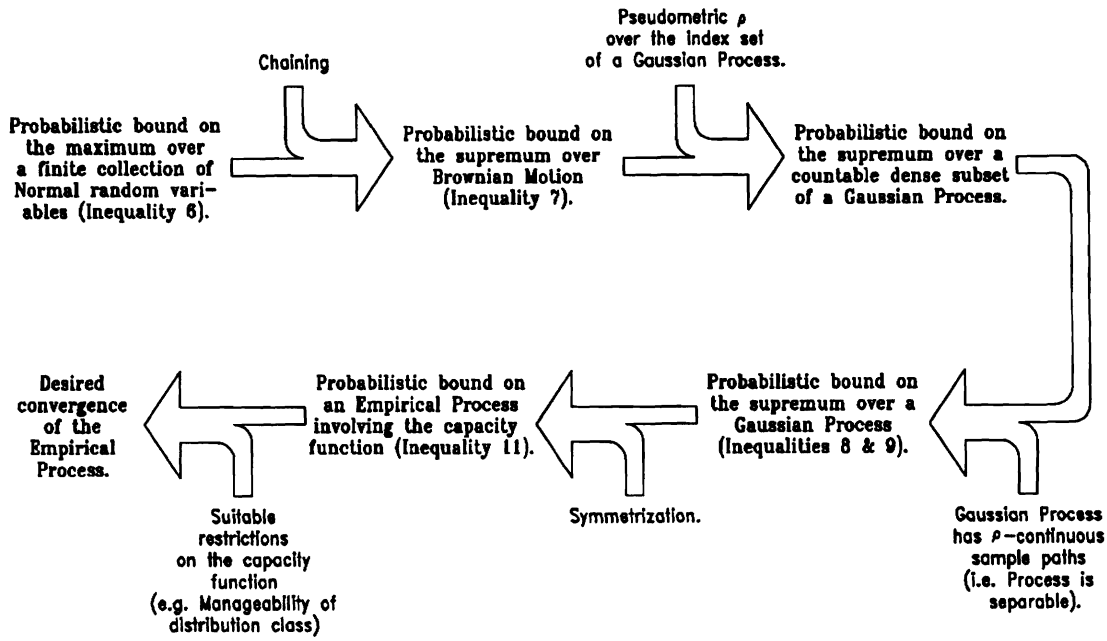


Figure 3: **Schematic of the main thread through the labyrinth of empirical processes.**

4.1.1 Finite Collections of Normal Random Variables

Consider first the related problem of estimating the maximum of a finite collection of normal random variables $\{Z_i \sim N(0, \sigma_i^2) \ : \ i = 1, \ldots, n\}$ where nothing is known about the joint distributions. Define $\sigma = \max(\sigma_1, \ldots, \sigma_n)$. A crude bound on $\max_i |Z_i|$ is $\sum_{i=1}^n |Z_i|$. Since

$$\mathbb{P}\,|Z_i| = \mathbb{P}\,|N(0, \sigma_i)| = \sqrt{\frac{2}{\pi}}\,\sigma_i \leq \sqrt{\frac{2}{\pi}}\,\sigma,$$

we conclude that

$$\mathbb{P} \max_i |Z_i| \leq \mathbb{P} \sum_{i=1}^n |Z_i| = \sum_{i=1}^n \mathbb{P}\,|Z_i| \leq \sqrt{\frac{2}{\pi}}\,\sigma n.$$

The problem with this bound is that we have placed identical emphasis on the contribution of each $|Z_i|$ towards $\sum_{i=1}^n |Z_i|$. In order to improve on this bound, we need somehow to stress the contribution of whichever $|Z_i|$ is the *true* maximum, while simultaneously suppressing the contributions from the other $|Z_i|$'s as much as possible.

We do this by transforming the $|Z_i|$'s via a nonnegative, convex, increasing function $M(\cdot)$ on the positive half-line:

From Jensen's Inequality (see Section 2.1) followed by the crude bound,

$$\begin{aligned} I\!P \max_i |Z_i| = M^{-1}\left[M\left(I\!P \max_i |Z_i|\right)\right] &\leq M^{-1}\left[I\!P \max_i M(|Z_i|)\right] \\ &\leq M^{-1}\left[\sum_{i=1}^n I\!P\, M(|Z_i|)\right] \end{aligned}$$

In order to exploit convexity as much as possible, we make H increase as fast as the tails of $|Z_i|$ can bear without allowing the sum of expectations $\sum_{i=1}^n I\!P\, M(|Z_i|)$ to exceed a multiple of n [Pol89]. It is straightforward to show that for normal tails, the function $M(x) = e^{x^2/4\sigma^2}$ suffices to ensure that $I\!P\, M(|Z_i|) \leq \sqrt{2}$ for all i. Thus,

$$M^{-1}\left[\sum_{i=1}^n I\!P\, M(|Z_i|)\right] \leq M^{-1}\left[\sqrt{2}\, n\right] < M^{-1}\left[n^2\right] \leq 2\sqrt{2}\,\sigma\,(\log n)^{1/2} \text{ for } n \geq 2,$$

$$\text{whence } I\!P \max_i |Z_i| < 3 \max_i \sigma_i\,(\log n)^{1/2} \text{ for } n \geq 2. \tag{6}$$

The *chaining method* of Section 4.1.2 makes use of repeated applications of Inequality (6) to obtain a surprisingly good bound on the supremum of a Gaussian process.

4.1.2 Brownian Motion and Chaining

Before addressing Gaussian processes in their full generality, consider next the special case of Brownian Motion on the bounded index set $[0, \delta]$.

Brownian Motion or the **Wiener Process** on $[0, \delta]$ is defined to be a Gaussian process $\{B(t) : 0 \leq t \leq \delta\}$ with the following properties ([Bil79], page 442):

1. With probability 1, $B(0) = 0$ (Process begins at the origin).
2. For $0 \leq t_1 \leq t_2 \leq \ldots \leq t_{2m} \leq \delta$, the nonoverlapping increments $B(t_2) - B(t_1), \ldots, B(t_{2m}) - B(t_{2m-1})$ are independent.
3. For any $t, s \in [0, \delta]$, the increment $B(t) - B(s)$ is distributed $N(0, |t - s|)$.

Once again, we are interested in a bound on the expectation of $\sup_{t\in[0,\delta]} |B(t)|$. The main idea of the **chaining** method is to approximate this supremum by the maximum

taken over a succession of finite subsets of $[0, \delta]$ each more finely spaced than the last. For $k = 0, 1, \ldots$, define $\delta_k = \delta/2^k$ and let $T(k)$ denote the set of 2^k equally spaced points $\{\delta_k, 2\delta_k, \ldots, 2^k\delta_k\}$. Owing to sample path continuity, the maximum of $B(t)$ taken over $T(k)$ increases monotonically to $\sup_{t\in[0,\delta]} |B(t)|$, whence

$$I\!P \max_{t\in t(k)} |B(t)| \to I\!P \sup_{t\in[0,\delta]} |B(t)| \text{ as } k \to \infty.$$

Figure 4 represents a systematic way of relating the maxima over successive sets $T(k)$. In a way, the hunt for the supremum of $B(t)$ is akin to a parallel binary tree-search over $[0, \delta]$. The crucial observation is that for any $k \geq 1$, to each t in $T(k)$ there corresponds a t' in $T(k-1)$ at most a distance of δ_{k-1} away. Thus, for each t, t' pair, by the triangle inequality,

$$|B(t)| \leq |B(t')| + |B(t) - B(t')|.$$

So, when attempting to find the maximum of $|B(t)|$ over a set $T(k)$, $k \geq 1$, one need only find the maximum over the set $T(k-1)$ and then add to this value the maximum discrepancy $\max_{t\in T(k), t'\in T(k-1)} |B(t) - B(t')|$:

$$\max_{t\in T(k)} |B(t)| \leq \max_{t'\in T(k-1)} |B(t')| + \max_{t\in T(k)} |B(t) - B(t')|.$$

Now, for Brownian Motion, each increment $B(t) - B(t')$ is distributed $N(0, \delta_k)$, so that, by Inequality (6), $I\!P \max_{t\in T(k)} |B(t) - B(t')| \leq 3\delta_{k-1}^{1/2}(\log 2^k)^{1/2}$. Hence, taking expected values of both sides of the above inequality yields the recurrence relation

$$\begin{aligned} I\!P \max_{t\in T(k)} |B(t)| &\leq I\!P \max_{t\in T(k-1)} |B(t)| + 3\sqrt{\delta_{k-1} \log 2^k}, \\ I\!P \max_{t\in T(0)} |B(t)| &= I\!P\, |B(\delta)| \end{aligned}$$

whose solution is

$$I\!P \max_{t\in T(k)} |B(t)| \leq I\!P\, |B(\delta)| + \sum_{i=1}^{k} 3\sqrt{\delta_{i-1} \log 2^i}.$$

Hence, making use of the identity $\delta_i = \delta_{i-1}/2$, $i \geq 1$ and the fact $B(\delta) \sim N(0, \delta)$, and letting $k \to \infty$,

$$\begin{aligned} I\!P \sup_{t\in[0,\delta]} |B(t)| &\leq \sqrt{\delta}\, I\!P|N(0, 1)| + \sqrt{\delta} \sum_{i=1}^{\infty} 3\sqrt{i\,(\tfrac{1}{2})^{i-1} \log 2} \qquad (7) \\ &\leq K\sqrt{\delta} \text{ since the infinite sum converges.} \end{aligned}$$

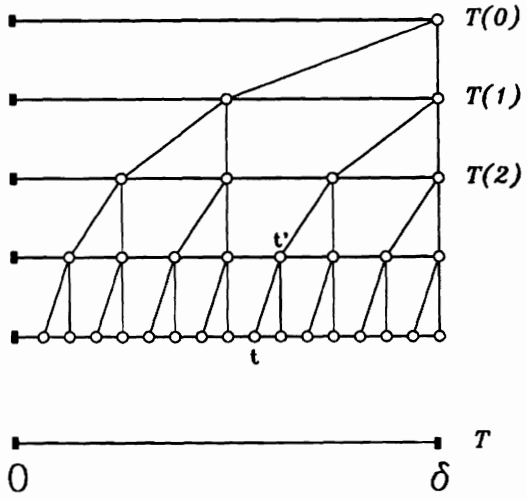


Figure 4: **Chaining.**

4.1.3 Generalization to Gaussian Processes

The above bound on the supremum over Brownian Motion on $[0, \delta]$ can be carried over to a Gaussian Process $\{Y_t : t \in T\}$ where a pseudometric[6] ρ defined over T controls the increments of the process[7]:

$$\mathbb{P}\,|Y(s) - Y(t)|^2 \leq \rho(s, t)^2 \text{ for all } s, t \text{ in } T.$$

The following adjustments complete the generalization:

- With $T(0)$ the singleton $\{t_0\}$, $\delta = \sup_{t \in T} \rho(t, t_0)$,
- The subsets $T(k) \subset T$, $k = 0, 1, \ldots$ are now maximal sets of points greater than $\delta_k = \delta/2^k$ apart, so that for all $t \in T(k)$, there exists a $t' \in T(k-1)$ such that $\rho(t, t') \leq \delta_{k-1}$, and
- The size of $T(k)$ is measured by the function $D(\epsilon) = D(\epsilon, T, \rho)$, defined as the largest n for which there are points $t_1, \ldots, t_n$ in T with $\rho(t_i, t_j) > \epsilon$ for $i \neq j$ [8].

[6] see Section 2.6 for definition.

[7] For Brownian Motion the usual Euclidean metric is replaced by $\rho(s, t) = \sqrt{|s - t|}$

[8] $\log D(\epsilon)$ is called the *ϵ-capacity* of T. Also, $\lim_{\epsilon \to 0} \frac{\log D(\epsilon)}{-\log \epsilon}$ may be shown to be equivalent to the *box dimension* of T. See [Fal90] for details.

A chaining argument similar to that for Brownian Motion leads to the recurrence relation

$$\mathbb{P} \max_{t \in T(k)} |Y(t)| \leq \mathbb{P} \max_{t \in T(k-1)} |Y(t)| + 3\delta_{k-1}\sqrt{\log D(\delta_k)},$$

whose solution in the limit is, for some t_0,

$$\mathbb{P} \sup |Y(t)| \leq \mathbb{P} |Y(t_0)| + 3 \sum_{i=1}^{\infty} \delta_{i-1}\sqrt{\log D(\delta_i)},$$

where the supremum is over a countable dense subset of T. If Y has ρ-continuous sample paths[9] then the supremum over a dense subset of T is equal to the supremum over *all* of T. Further, if the sum is treated as a lower step function approximation to an integral, then for all $t_0 \in T$ the solution simplifies to

$$\begin{aligned} \mathbb{P} \sup_{t \in T} |Y(t)| &\leq \mathbb{P} |Y(t_0)| + K \int_0^{\delta} \sqrt{\log D(x, T, \rho)}\, dx \qquad (8) \\ \text{where } \delta &= \sup_{t \in T} \rho(t, t_0) \end{aligned}$$

Of course, this inequality has meaning only when the integral is finite. It can be shown that in this case, there exists a version of the process having continuous sample paths, so that the assumption of sample path continuity becomes superfluous.

As a final note, we observe that a result similar to Inequality (8) would be obtained if the expectations in the recurrence relation were replaced by any $\mathcal{L}^{\alpha}(\mathbb{P})$ norm, $\alpha \in [1, 2]$. For example, with $\mathcal{L}^2(\mathbb{P})$ norms, for all $t_0 \in T$,

$$\begin{aligned} \left(\mathbb{P} \sup_{t \in T} |Y(t)|^2 \right)^{1/2} &\leq \left(\mathbb{P} |Y(t_0)|^2 \right)^{1/2} + K \int_0^{\delta} \sqrt{\log D(x, T, \rho)}\, dx \qquad (9) \\ \text{where } \delta &= \sup_{t \in T} \rho(t, t_0) \end{aligned}$$

This result is strictly stronger than Inequality (8) in the sense that an upper bound on an $\mathcal{L}^2(\mathbb{P})$ norm implies a corresponding upper bound on the $\mathcal{L}^1(\mathbb{P})$ norm, but the reverse does not necessarily hold. In what follows, we focus our attention on the $\mathcal{L}^2(\mathbb{P})$ norm, bearing in mind that for a class of functions $\mathcal{G}$, convergence of $\mathbb{P} \sup_{g \in \mathcal{G}} |\nu_n g|^2 \to 0$ as $n \to \infty$ guarantees the uniformity result that we are after, $\mathbb{P}\{\sup_{g \in \mathcal{G}} |\nu_n g| > \epsilon\} \to 0$ as $n \to \infty$.

[9] This would make Y a *separable* process — see Section 2.5.

4.2 Symmetrization

Now that we have found a bound on the expectation of the supremum of a Gaussian process in the form of Inequalities (8) and (9), all that remains to be done is to transform the original empirical process $\nu_n = \{n^{1/2}(P_n - P)g \ : \ g \in \mathcal{G}\}$ into a form to which the results for Gaussian processes become applicable. Although ν_n is, by the Central Limit Theorem[10], approximately asymptotically a Gaussian process, a surprising amount of manoeuvring is needed to obtain a strict inequality. To avoid tedium, only the general approach and main results are stated here; the reader is referred to [Pol89] for the details.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\} \in \Psi^n$ and $\boldsymbol{x}' = \{x'_1, \ldots, x'_n\} \in \Psi^n$ be two independent sequences of observations, with each observation sampled according to the distribution P in the probability space $\langle \Psi, \mathcal{B}, P \rangle$. Further, let $\{\sigma_i\}$ be a sequence of sign variables for which $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$.

By an approach strongly reminiscent of the SYMMETRIZATION STEP of Section 3.2, we avoid dealing with $\nu_n = \{n^{1/2}(P_n - P)g \ : \ g \in \mathcal{G}\}$, and work rather with the rescaled difference between the two empirical measures, $\{n^{1/2}|(P_n - P'_n)g| \ : \ g \in \mathcal{G}\}$. Then, exploiting the symmetry between $\boldsymbol{x}$ and $\boldsymbol{x}'$, we may introduce the sign variables without affecting expected values (see [Pol89]). Further symmetrization arguments replace terms involving x'_i by their x_i counterparts, yielding ultimately the inequality

$$\mathbb{P} \sup_{g \in \mathcal{G}} |\nu_n g|^2 \ \leq \ 4\mathbb{P} \sup_{g \in \mathcal{G}} n^{-1} \left| \sum_1^n \sigma_i g(x_i) \right|^2 . \tag{10}$$

Now consider the construction of the sign variables from a sequence $\{q_i\}$ of independent $N(0, 1)$ variables, $\sigma_i = q_i/|q_i|$, and define the process

$$\{Y_n(g, \boldsymbol{x}) = n^{-1/2} \sum_1^n q_i g(x_i) \ : \ g \in \mathcal{G}\}$$

which is Gaussian conditionally on $\boldsymbol{x}$ and has increments controlled by the $\mathcal{L}^2(P_n)$ norm. Some arithmetic (see [Pol89]) then reduces Inequality (10) to

$$\mathbb{P} \sup_{g \in \mathcal{G}} |\nu_n g|^2 \leq 2\pi \mathbb{P} \sup_{g \in \mathcal{G}} n^{-1} |Y_n(g, \boldsymbol{x})|^2.$$

[10] See Section 2.3.

Now, the right hand side may be bounded by use of Inequality (9), with the corresponding ϵ-capacity of $\mathcal{G}$ denoted by $\log D_2(\epsilon, \mathcal{G}, P_n)$ [11]: For fixed $g_0 \in \mathcal{G}$,

$$\left(\mathbb{P} \sup_{g \in \mathcal{G}} |\nu_n g|^2 \right)^{1/2} \leq \sqrt{2\pi (P_n g_0^2)} + K \int_0^{\delta(\boldsymbol{x})} \sqrt{\log D_2(y, \mathcal{G}, P_n)}\, dy \qquad (11)$$

$$\text{where } \delta(\boldsymbol{x}) = \sup_{g \in \mathcal{G}} (P_n |g - g_0|^2)^{1/2}.$$

4.3 Manageable Classes

Although we have found a bound for $\mathbb{P} \sup_{g \in \mathcal{G}} |\nu_n g|^2$ which depends on $\mathcal{G}$ only through its capacity, it remains for us to find criteria for $\mathcal{G}$ subject to which this bound will converge to zero, hence validating $P_n g$ as an asymptotically good estimate of $P g$ *uniformly* over $\mathcal{G}$.

It turns out that function classes which exhibit a property known as *manageability* (to be defined in Section 4.3.1) are prime candidates for this uniform convergence. In fact, the climax of our four-step tour of empirical processes is encapsulated in **Theorem 4.4** of [Pol89]:

> Let $\mathcal{G}$ be a manageable class for an envelope G with $P G^2 < \infty$. Let $\mathcal{G}(n)$, $n = 1, 2, \ldots,$ be subclasses for which
>
> 1. $0 \in \mathcal{G}(n)$ for all n, and
> 2. $\sup_{g \in \mathcal{G}(n)} P|g| \to 0$ as $n \to 0$,
>
> Then
>
> $$\mathbb{P} \sup_{g \in \mathcal{G}(n)} |\nu_n g|^2 \to 0 \text{ as } n \to \infty.$$

The proof of this theorem as well as the intricate details of how *manageability* leads to a simplification of Inequality (11) are omitted here. Instead, the remainder of this section is devoted to the definition of manageable classes of functions and a discourse on their intimate relationship with VC classes of sets.

[11] In other words, $D_2(\epsilon, \mathcal{G}, P_n)$ equals the largest N for which there are functions $g_1, \ldots, g_N$ in $\mathcal{G}$ with

$$(P_n |g_i - g_j|^2)^{1/2} > \epsilon \text{ for } i \neq j.$$

4.3.1 Definition of Manageability

Let $\mathcal{G}$ be a class of functions with envelope G[12]. $\mathcal{G}$ is said to be **manageable** for G if there exists a decreasing function $D(\cdot)$ for which

1. $\int_0^1 \sqrt{\log\, D(x)}\, dx < \infty$, and
2. for every measure Q with finite support,

$$D_2\left(\epsilon\sqrt{QG^2}, \mathcal{G}, Q\right) \leq D(\epsilon) \text{ for } 0 < \epsilon \leq 1.$$

It is seldom possible to calculate directly the uniform bound on capacities required by this definition [Pol89]. How, then, are we to establish manageability of a function class and hence exploit the results of the previous section? The answer to this question involves VC classes of sets and is perhaps as remarkable as it is elegant.

4.3.2 VC Classes of Sets and Manageable Classes of Functions

Define the **subgraph** of a function $g : \Psi \rightarrow \Re$ as a subset of the product space $\Psi \otimes \Re$:

$$subgraph(g) = \{(\psi, x) \in \Psi \otimes \Re : 0 < x < g(\psi) \text{ or } g(\psi) < x < 0\}.$$

Define also a **Euclidean** function class as a class $\mathcal{G}$ with envelope G for which, for measures Q of finite support,

$$\sup_Q D_1(\epsilon QG, \mathcal{G}, Q) = O(\epsilon^{-V}) \text{ for some } V,$$

where D_1 is the $\mathcal{L}^1(Q)$ capacity of $\mathcal{G}$.

The crucial connection between VC classes, subgraphs and Euclidean function classes appears as **Lemma II.25** of [Pol84]:

[12]That is, $G \geq |g|$ for every $g \in \mathcal{G}$

Let $\mathcal{G}$ be a class of functions on Ψ with envelope G, and let Q be a measure on Ψ with finite support. If the class of subgraphs of functions in $\mathcal{G}$ form a VC class of subsets of $\Psi \otimes \Re$, then $\mathcal{G}$ is Euclidean.

From here, the final leap is easy: Elementary inequalities involving the $\mathcal{L}^1(P)$ and $\mathcal{L}^2(Q)$ seminorms, where P has density G with respect to Q, show that Euclidean classes of functions have analogous bounds on their $\mathcal{L}^2$ capacities ([Pol84], **Lemma 36**). In particular,

Every Euclidean class is manageable.

Hence, in short, a function class $\mathcal{G}$ with envelope G whose subgraphs form a VC class is, in fact, a manageable class. Conceptually, the above arguments may be summarized thus:

Subgraphs of $\mathcal{G}$ form a VC Class	$\Rightarrow$	$\mathcal{G}$ is a Euclidean class	$\Rightarrow$	$\mathcal{G}$ is a manageable class	$\Rightarrow$	Desired convergence of $\mathbb{P}\sup_{g\in\mathcal{G}} \lvert\nu_n g\rvert^2$

For completion, we mention the existence of another connection between the VC property and manageability: A class of functions $\mathcal{G}$ is called a **VC major class** if there exists a VC class of sets $\mathcal{C}$ such that $\{\phi : g(\phi) \geq \alpha\}$ is a member of $\mathcal{C}$ for all $g \in \mathcal{G}$ and for all $\alpha \in \Re$. Dudley (1987) has shown that

Every uniformly bounded VC major class is manageable for a constant envelope.

Example: As an example application, consider the first of the two asymptotic problems dealt with in [Pol89]. Glossing over the reduction to empirical process notation, we pick up the analysis at the stage where we need to show uniform convergence in shrinking neighborhoods of a point t_0,

$$\mathbb{P} \sup_{t\in[\delta_n - t_0,\, \delta_n + t_0]} \lvert\nu_n\left(\lvert\cdot - t\rvert - \lvert\cdot - t_0\rvert\right)\rvert \to 0 \text{ as } n \to 0$$

for every sequence of positive numbers $\{\delta_n\}$ converging to zero.

Consider a member $g(\cdot, t)$ of the function class $\mathcal{G} = \{g(\cdot, t) = (|\cdot - t| - |\cdot - t_0|) : |t_o - t| \leq \delta_n\}$. It has a constant value of $t - t_0$ on the interval $(-\infty, \min\{t, t_0\}]$, a constant value of $t_0 - t$ on $(\max\{t, t_0\}, \infty,]$, and inbetween follows the straight line joining $(t - t_0)$ to $(t_0 - t)$.

Hence, for any $\alpha \in \Re$ and $t \in [\delta_n - t_0, \delta_n + t_0]$, the inverse image $C = \{x : g(x, t) \geq \alpha\}$ is a semi-infinite interval on the real line. Now, the class $\mathcal{C}$ of all such intervals has been shown in Section 3.1.2 to be a VC class. Thus, $\mathcal{G}$ is a uniformly bounded VC major class, and is, therefore, manageable for constant envelope δ_1.

Further, $g(t_0, t_0) = 0 \in \mathcal{G}$ and for all $g \in \mathcal{G}$, $Pg \leq |t - t_0| \leq \delta_n \to 0$ as $n \to 0$, whence all the hypotheses of **Theorem 4.4** of [Pol89] (see Section 4.3) are satisfied. It follows that uniform convergence is, indeed, attained.

4.3.3 Properties of Manageable Classes

We conclude this chapter with a few subsidiary remarks about the nature of manageability, and the construction of more complicated manageable classes once the basic classes have been established by way of VC classes of subgraphs or VC major classes of functions.

The first three of the following properties of manageable classes are deduced from elementary $\mathcal{L}^2$ inequalities; the reader is referred to [Pol89] for a sample derivation. The last property is taken from Dudley (1987), **Theorem 5.3**.

If $\mathcal{G}$ is manageable for envelope G and $\mathcal{H}$ is manageable for envelope H, then

1. $\mathcal{G} \diamond \mathcal{H} = \{g \diamond h : g \in \mathcal{G}, h \in \mathcal{H}\}$ is manageable for envelope $G + H$, where the symbolic operator $\diamond$ represents addition $(+)$, maximum $(\vee)$, or minimum $(\wedge)$.
2. $\mathcal{G} \star \mathcal{H} = \{gh : g \in \mathcal{G}, h \in \mathcal{H}\}$ of products is manageable for envelope GH.
3. The closure of $\mathcal{G}$ under convergence is manageable for envelope G.
4. The symmetric convex hull of $\mathcal{G}$, denoted by $sco(\mathcal{G})$ and consisting of all finite linear combinations $\sum \alpha_j g_j$ of functions $g_j \in \mathcal{G}$ for which $\sum |\alpha_j| \leq 1$, is manageable for envelope G.

5 Geometrizing Rates of Convergence

A synopsis of [DL91] by Donoho and Liu

As mentioned in the introduction, any discussion on asymptotics must go hand in hand with a treatment of *rates* of convergence. The theme of [DL91] revolves around a bound on the rate of convergence of an estimate $T_n(\boldsymbol{X}_n)$[13] to the value of a functional $T(F)$ of an unknown distribution $F \in \mathcal{F}$ uniformly over a class of distributions $\mathcal{F}$. The main result is two-fold: First, it turns out that for estimating a *linear* functional over a *convex* distribution class $\mathcal{F}$, the geometry of the problem, expressed in terms of an index known as the *modulus of continuity* $b(\epsilon)$, determines the optimal rate of convergence. Second, and perhaps more startling, is that this optimal rate is, in fact, *attainable*, provided only that $b(\epsilon)$ is Hölderian[14].

The result is established by way of another bound on the rate on convergence, denoted by Δ_A and involving the difficulty of testing the composite hypothesis[15] $H_0 : T(F) \leq t$ against the composite hypothesis $H_1 : T(F) > t + \Delta$. *Under certain asymptotic conditions*, Δ_A is *always attainable*, to within constants, regardless of the characteristics of the functional T or the class of distributions $\mathcal{F}$. Linearity of T and convexity of $\mathcal{F}$ then guarantee that the modulus bound agrees with Δ_A, to within constants. From here, verification that $b(\epsilon)$ is a Hölder function is all that is necessary to ensure that the supporting asymptotic conditions are met.

Yet that is not all. Donoho&Liu show that for the modulus bound to agree with Δ_A, the prerequisites of linearity and convexity may be discarded, provided that the essence of the geometry is preserved: A new criterion is that the hardest two-point subproblem of testing $T(F) \leq t$ versus $T(F) \geq t + \Delta$ should be roughly as difficult, from a minimax risk point of view, as the full composite hypothesis-testing problem. Moreover, in one example, Donoho&Liu show that even this last condition may be dropped. On the other hand, in *all* cases satisfaction of a Hölder condition by the modulus of continuity is necessary in order to preserve the *attainability* of the optimal rates. For clarity, the relationships among these concepts are shown graphically in Figure 5.

In this section, we review the definitions and properties of concepts vital to later developments. We then identify $\Delta_A(n, \alpha)$ as a *lower bound*, to within constants, on the rate of convergence of $T_n(\boldsymbol{X}_n)$ to $T(F)$ (**Theorem 2.1** of [DL91]). Next, we

[13] where $\boldsymbol{X}_n$ is a vector of n i.i.d. F sample points

[14] See Section 2.1 for definition.

[15] See [CB90], Section 8.3, for a very lucid treatment of Hypothesis testing.

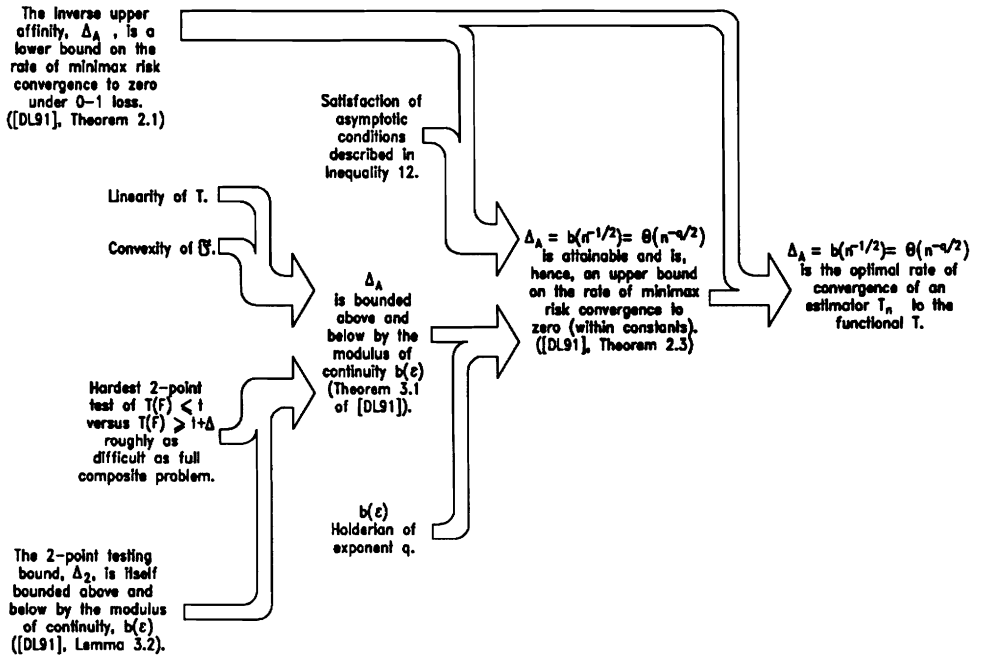


Figure 5: **Graphical interpretation of the relationships among Δ_A, the modulus bound, linearity, convexity and the Hölder property of $b(\epsilon)$.**

present an estimator which, again to within constants, *attains* the rate $\Delta_A(n, \alpha)$, provided that the tails of $\Delta_A(n, \alpha)$ behave appropriately. Finally, we show how linearity of T, convexity of $\mathcal{F}$ and the satisfaction of a Hölder condition by $b(\epsilon)$ ensure such behaviour.

The generalized case (where linearity of T and convexity of $\mathcal{F}$ need not be assumed) is dealt with in Section 5.4.2, while Section 5.5 discusses the implications of the caveat "to within constants" from a Decision-Theoretic point of view.

The concepts discussed in [VC71] and [Pol89] [16] can be placed within the current framework of estimators and functionals on classes of distributions. In this way, *rates* may be deduced for the convergences involved. Though often relatively straightforward, detailed derivations of this nature can be lengthy and the reader is referred to Section 6 for an example exposition.

[16] Respectively those of uniform convergence of relative frequency to probability over a class of events and uniform convergence of sample mean to true mean over a class of functions.

5.1 Definitions

The following definitions are concerned with the distinguishability of distributions within a class and the difficulty of estimating a functional T over such a class.

5.1.1 Hellinger Affinity

Hellinger Affinity $\rho(P, Q)$ is a measure of the 'closeness' of two measures P and Q and is defined as

$$\rho(P, Q) = \int \sqrt{p}\sqrt{q}\, d\mu,$$

where $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ for any measure μ which dominates P and Q [LY90].

5.1.2 Hellinger Distance

The **Hellinger Distance** $H(P, Q)$ between two probability measures P and Q is defined as

$$H^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\mu$$

where, as before, $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ for any measure μ which dominates P and Q [LY90].

It can easily be shown that

$$H^2(P, Q) = 2\,(1 - \rho(P, Q))$$

5.1.3 Modulus of Continuity

The **modulus of continuity** of a functional T over a class of distributions $\mathcal{F}$, with respect to Hellinger distance, is defined as

$$b(\epsilon) = \sup \{|T(F_1) - T(F_0)| \ : \ H(F_1, F_0) \leq \epsilon, \, F_1, \, F_0 \in \mathcal{F}\}.$$

In words, the modulus of continuity measures, as a function of ϵ, the greatest variation of the functional over any Hellinger ball of radius ϵ. In a way, it gives an indication of

the extent of correlation between the "shape" of a distribution and the value of the functional of that distribution: For functionals which exhibit very little dependence on how the probability mass is distributed[17] and can change wildly for very similar (in a Hellinger sense) distributions, the modulus will be large, even for small ϵ. On the other hand, functionals which are highly dependent on mass distribution[18] will tend to have small moduli, perhaps linear or polynomial in ϵ for $\epsilon \to 0$.

Throughout the remainder of this paper, we will be interested in the asymptotic behaviour of $b(\epsilon)$ for $\epsilon \to 0$.

In order to build up some intuition regarding the nature of the modulus of continuity, we look at a few examples of functionals and classes of distributions and derive their moduli.

Example 1: Location Parameters. Consider the class of shifted Gaussian distributions $\mathcal{F} = \{N(a, 1) : a \in \Re\}$ and a functional T which returns some location parameter such as the mean. Let F_0 and F_1 be distributions whose locations differ by $\Delta \geq 0$. Then the Hellinger affinity between F_0 and F_1 is seen to be

$$\rho(F_0, F_1) = \int_{-\infty}^{\infty} \frac{e^{-(x-a)^2/4} e^{-(x-a-\Delta)^2/4}}{\sqrt{2\pi}} dx = \frac{1}{e^{\Delta^2/8}}.$$

Thus the Hellinger distance between the two distributions is

$$H(F_0, F_1) = \sqrt{2\,(1 - \rho(F_0, F_1))} = \sqrt{2}\sqrt{1 - e^{-\Delta^2/8}}.$$

Since this is a monotonically increasing function of Δ, we see that in any Hellinger ball of radius ϵ, the distributions whose locations are furthest apart lie on the *skin* of the ball. Hence, the modulus of continuity is simply the inverse function of the Hellinger distance:

$$b(\epsilon) = \Delta = \sqrt{-8 \log \left(1 - \frac{\epsilon^2}{2}\right)}$$

A Taylor Series expansion of the above yields $b(\epsilon) = 2\epsilon + \frac{\epsilon^3}{4} + O(\epsilon^5)$, whence it is clear that $b(\epsilon)$ is dominated by the linear term for $\epsilon \to 0$.

[17] Consider, for instance, a functional which counts the number of local maxima in the probability density.

[18] Such as mean or median.

In a similar fashion we may show that the modulus is linear for location parameters over classes of Laplacian or Cauchy distributions:

In the case of Laplacians, $\mathcal{F} = \{F(\cdot) = \int_{-\infty}^{\cdot} e^{-|t-a|/2}\,dt \;:\; a \in \Re\}$ so that some algebra leads to $\rho(F_0, F_1) = \frac{2+\Delta}{2e^{\Delta/2}}$, where, as before, the locations of F_0 and F_1 differ by $\Delta \geq 0$. Hence, $H(F_0, F_1) = \sqrt{2 - \frac{2+\Delta}{e^{\Delta/2}}}$ which, once again, is monotonically increasing. Some numerical analysis then confirms that $b(\epsilon)$, the function inverse of H, is dominated by a linear term for small ϵ.

The case of Cauchy distributions yields similar calculations.

Example 2: A Nonlinear Modulus. Consider the problem of estimating the value of a density at a point. [Far72] deals with optimal rates of convergence in a very general setting. In this example we limit our analysis to a very specific class of distributions, and show that in this case, the modulus of continuity is quadratic for small ϵ.

The main idea is to choose a class of distributions for which minor differences in the value of the functional are amplified in the profiles of the densities concerned. Hence, distributions which are confined to small Hellinger balls must have very similar profiles, and even closer functional values.

We select the class of densities indexed by $\alpha \in (0, 1]$ where an arbitrary member f_α is defined by[19]:

$$f_\alpha(x) = \begin{cases} \frac{\alpha}{\sqrt{x+1}} & \text{for } 0 \leq x \leq a \text{ where } \int_0^a \frac{\alpha}{\sqrt{x+1}}\,dx = 1 \Rightarrow a = \left(1 + \frac{1}{2\alpha}\right)^2 - 1 \\ 0 & \text{elsewhere} \end{cases}$$

Let the functional be $T(F_\alpha) = F_\alpha(0) = \alpha$. Let us now calculate the Hellinger distance between two arbitrary members F_α and F_β of $\mathcal{F} = \{F_\alpha(\cdot) = \int_{-\infty}^{\cdot} f_\alpha(x)\,dx \;:\; \alpha \in (0, 1]\}$. Without loss of generality, assume $\alpha > \beta$. Hence,

$$H^2(F_\alpha, F_\beta) \;=\; \tfrac{1}{2}\int \left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2 dx$$

[19]Perhaps unexpectedly, the rate of decay of the tails of these types of densities does not seem to influence the modulus of continuity. For instance, identical results are obtained if we choose $f_\alpha(x) = \frac{\alpha}{(x+1)^{1/n}}$, or even $f_\alpha(x) = \frac{1}{\alpha}e^{-\alpha x}$, where $n \in (1, \infty)$ and $0 \leq x \leq a$ for a suitably normalizing a.

$$= \frac{1}{2}\left[\int_0^{\left(1+\frac{1}{2\alpha}\right)^2-1}\left(\frac{\sqrt{\alpha}}{(x+1)^{1/4}}-\frac{\sqrt{\beta}}{(x+1)^{1/4}}\right)^2 dx + \int_{\left(1+\frac{1}{2\alpha}\right)^2-1}^{\left(1+\frac{1}{2\beta}\right)^2-1}\frac{\alpha}{(x+1)^{1/2}}\,dx\right]$$

Substituting $y = x + 1$,

$$H^2(F_\alpha, F_\beta) = \frac{1}{2}\left[\int_1^{\left(1+\frac{1}{2\alpha}\right)^2}\frac{1}{\sqrt{y}}\left(\alpha - 2\sqrt{\alpha\beta}+\beta\right)dy + \int_{\left(1+\frac{1}{2\alpha}\right)^2}^{\left(1+\frac{1}{2\beta}\right)^2}\frac{\alpha}{\sqrt{y}}\,dy\right]$$

$$= \frac{1}{2}\left[(\alpha+\beta-2\sqrt{\alpha\beta})(\frac{1}{\alpha}) + \beta(\frac{1}{\beta}-\frac{1}{\alpha})\right], \text{ whence}$$

$$H(F_\alpha, F_\beta) = \sqrt{1-\sqrt{\beta/\alpha}}.$$

Plotting $H(F_\alpha, F_\beta)$ as a function of α and β yields the surface in Figure 6. Contour lines represent the skins of Hellinger balls, so that from Figure 6 we see that for any Hellinger ball of radius ϵ, the difference between α and β is maximized on the skin of the ball at $\alpha = 1 \geq \beta$. Hence we can derive the modulus of continuity: For any $0 \leq \epsilon \leq 1$,

$$\epsilon = H(F_1, F_{1-b(\epsilon)}) = \sqrt{1-\sqrt{1-b(\epsilon)}}$$

$$\text{so that } b(\epsilon) = 1-\left(1-\epsilon^2\right)^2$$

$$= 2e^2 - e^4.$$

For small ϵ, $b(\epsilon)$ is seen to be dominated by the quadratic term.

5.1.4 Testing Affinity

Let P and Q be probability distributions on a common space Ψ. Let $F \in \{P, Q\}$ be an unknown distribution and consider deciding the hypothesis $H_0 : F = P$ versus $H_1 : F = Q$ based on an observation $\psi \in \Psi$. Let $\phi : \Psi \to [0, 1]$ be an arbitrary member of the class Φ of measureable randomized decision rules such that $\phi(\psi)$ represents the probability of rejection. Then the **testing affinity** [LY90], [LeC86] is defined as

$$\pi(P, Q) = \inf_{\phi\in\Phi} E_P\,\phi + E_Q\,(1-\phi)$$

and is seen to be the sum of errors of the best test between P and Q. Indeed, the testing affinity may be shown to be equal to $\| P \wedge Q \| = \int (p \wedge q)\,d\mu$ where $p = \frac{dP}{d\mu}$,

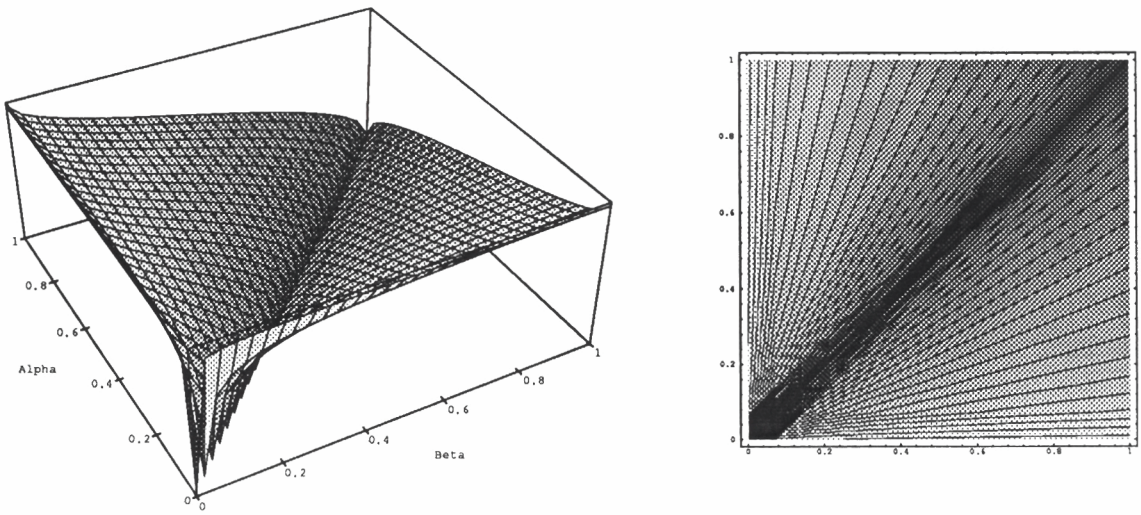


Figure 6: **3-Dimensional Plot and Contour Plot of $H(F_\alpha, F_\beta)$ as a function of α and β.**

$q = \frac{dQ}{d\mu}$ for any measure μ which dominates P and Q. Hence, the testing affinity gives a very intuitive indication of the distinguishability of P and Q.

The concept of testing affinity may be generalized to composite hypotheses: If $\mathcal{P}$ and $\mathcal{Q}$ are sets of measures, denote the hardest two-point testing problem by

$$\pi(\mathcal{P}, \mathcal{Q}) = \sup_{P \in \mathcal{P}, Q \in \mathcal{Q}} \pi(P, Q).$$

An observation crucial to future developments is that the minimax risk[20] in separating $\mathcal{P}$ and $\mathcal{Q}$ is $\pi(conv(\mathcal{P}), conv(\mathcal{Q}))$ where $conv(\mathcal{F})$ is the class of measures which are convex combinations of members of $\mathcal{F}$ [LY90] and [LeC86].

5.1.5 Upper Affinity and Inverse Upper Affinity

As before, let T be the functional of interest, acting over the class of distributions $\mathcal{F}$. As in [DL91], let $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{\geq t+\Delta}$ denote the subsets of $\mathcal{F}$ where T takes values

[20] Assuming a zero-one loss function

$\leq t$ and $\geq t + \Delta$ respectively. Now, let $\mathcal{F}^{(n)}_{\leq t}$ denote the set of product measures $\{F^{(n)} : F \in \mathcal{F}_{\leq t}\}$, and similarly for $\mathcal{F}^{(n)}_{\geq t+\Delta}$. Then the **upper affinity** $\alpha_A(n, \Delta)$ of the estimation problem is defined as

$$\alpha_A(n, \Delta) = \sup_t \pi \left(conv(\mathcal{F}^{(n)}_{\leq t}), \, conv(\mathcal{F}^{(n)}_{\geq t+\Delta}) \right)$$

Assuming a zero-one loss function, this is the minimax risk of the hardest problem of distinguishing $H_0 : \mathcal{F}_{\leq t}$ and $H_1 : \mathcal{F}_{\geq t+\Delta}$ with a sample of size n.

The **inverse upper affinity** $\Delta_A(n, \alpha)$ is defined as

$$\Delta_A(n, \alpha) = \sup \{\Delta : \alpha_A(n, \Delta) \geq \alpha\}.$$

Δ_A is the largest Δ at which, for a sample of size n, one cannot test hypotheses $H_0 : \mathcal{F}_{\leq t}$ and $H_1 : \mathcal{F}_{\geq t+\Delta}$ with sum of errors less than α [DL91]. In other words, Δ_A is the largest amount by which the subclasses $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{\geq t+\Delta}$ can be separated while still guaranteeing a minimum threshold error level of α. In effect, $\Delta_A(n, \alpha)$ places a bound on how well the functional T can be estimated: Any estimator T_n of T gives rise to the test where we accept H_0 whenever $T_n \leq t + \Delta/2$ and accept H_1 when $T_n \geq t + \Delta/2$. For this reason, $\Delta_A(n, \alpha)$ will be vitally important in the discussion which follows.

5.2 The Lower Bound

As mentioned in the previous paragraph, Δ_A places limits on how well T can be estimated. Indeed, we show here that for some $\alpha \in (0, 1)$ and for any symmetric increasing loss function $L(\cdot)$, the minimax risk is bounded from below by $\frac{\alpha}{2} L\left(\frac{\Delta_A}{2}\right)$. The result is a simple corollary of **Theorem 2.1** of [DL91] which states that

$$\inf_{T_n} \sup_{F \in \mathcal{F}} P_F\{|T_n - T(F)| \geq \Delta_A(n, \alpha)/2\} \geq \alpha/2.$$

The proof of this theorem appears in [DL91] and will not be presented here. Instead, as we have been wont to do in previous chapters, we give a brief overview of the main argument:

Basically, in testing the hypotheses $H_0 : \mathcal{F}_{\leq t}$ versus $H_1 : \mathcal{F}_{\geq t+\Delta}$, an inverse upper

affinity of $\Delta_A(n, \alpha)$ implies that Type I and Type II errors[21] together sum to a minimum risk level of α[22], so that at least one of these error Types must incur a risk of $\frac{\alpha}{2}$. Now, with a test which decides H_0 if $T_n \leq t + \frac{\Delta}{2}$ and H_1 if $T_n > t + \frac{\Delta}{2}$, a Type I or Type II error can occur only if the estimate T_n is on the 'wrong' side of the point $t + \frac{\Delta}{2}$; in other words, only if $|T_n - T(F)| \geq \frac{\Delta}{2}$.

Combining these two observations, the probability[23] that T_n differs from $T(F)$ by at least $\frac{\Delta}{2}$ is lower bounded by $\frac{\alpha}{2}$ for the worst case F.

5.3 Attaining the Lower Bound: The Binary Search Estimator

In the previous subsection, we established $\Delta_A(n, \alpha)/2$ as a lower bound on the rate of risk convergence to zero for some fixed $\alpha \in (0, 1)$. The proof involved showing that *even with the best of all possible estimators* T_n, hypothesis testing techniques would always yield a worst-case risk proportional to $L\left(\frac{\Delta_A(n,\alpha)}{2}\right)$.

In this section, we describe an *actual* estimator T_n which is optimal to within constants, in the sense that, under certain conditions, it too converges to $T(F)$ at a rate which is a constant multiple of $\Delta_A(n, \alpha)$. In this case, though, the actual worst-case risk is proportional to $L(K\Delta_A(n, \alpha))$, where K may be substantially larger than $\frac{1}{2}$.[24]

Nevertheless, the fact that Δ_A forms a lower bound together with its (near) attainability establishes it as an *optimal* rate of convergence.

The estimator proposed in [DL91] assumes a compact parameter space $T(F) \in \Omega = [-d, d]$. Consider an estimator constructed from repeated hypothesis tests where each successive test enables us to shrink the search space and to home in on $T(F)$ in a manner akin to a Binary Search. During each phase we perform a minimax hypothesis test between the lower third and upper third of the current search space, with the middle third adopting the role of 'Δ' in our previous discussions on hypothesis testing. The new search space is formed by deleting whichever third — upper or lower — is rejected by the test. Hence, each phase of the search yields a search space $\frac{2}{3}$ the size of the previous one; after a prescribed N phases we are left with an interval a

[21] False rejection and false acceptance respectively

[22] Consult definition of inverse upper affinity above

[23] measured according to the distribution F whose parameter we are attempting to estimate

[24] See Section 5.5 for details.

fraction $(\frac{2}{3})^N$ of the length of the initial space, whereupon we select the midpoint as our estimate.

Let us refine this idea with a few details: First of all, we endow our 'Binary Search Estimator' T_{Bin} with a 'tuning constant' Δ_n which varies with sample size but is fixed for any given n. Δ_n is the length that we wish the search space finally to have after all N tests; if no error occurs during any of the hypothesis tests, then T_{Bin} will differ from $T(F)$ by no more than $\frac{\Delta_n}{2}$. The number of successive tests we need perform is then $N = N(d, \Delta_n)$, the smallest integer such that $d' = (\frac{3}{2})^N \frac{\Delta_n}{2} \geq d$, while the starting search space is $[-d', d'] \supset \Omega$ and the initial hypothesis test compares $H_0 : \mathcal{F}_{\leq -d'/3}$ against $H_0 : \mathcal{F}_{\geq d'/3}$.

Naturally, the minimax risk associated with T_{Bin} is the accumulated risk from all N hypothesis tests. More precisely, using T_{Bin} as an estimator[25],

$$\sup_{F \in \mathcal{F}} P_F\{|T_{Bin} - T(F)| > \frac{\Delta_n}{2}\} \leq \sum_{k=0}^{N-1} \alpha_A \left(n, (\frac{3}{2})^k \frac{\Delta_n}{2}\right),$$

where $\alpha_A \left(n, \frac{\Delta_n}{2}(\frac{3}{2})^k\right)$, is the upper affinity of the $(N-k)$th hypothesis test. Though this last sum may look unwieldy, if Δ_n is made a constant multiple[26] of the inverse upper affinity $\Delta_A(n, \alpha)$, **Theorem 2.3** of [DL91] shows that the sum of upper affinities can be made arbitrarily small provided only that Δ_A exhibits appropriate tail behaviour. Hence, under this condition, T_{Bin} is seen to attain the desired rate of convergence, a constant multiple of Δ_A.

5.4 Ensuring Appropriate Tail Behaviour of Δ_A

The reader would be justified in surmising that it may prove difficult to obtain $\Delta_A(n, \alpha)$ in closed form, let alone derive properties concerning its tail behaviour. In this section we side-step the former problem, and instead focus on the latter, showing that asymptotic behaviour of Δ_A may be derived indirectly by way of the modulus of continuity.

The required tail behaviour we would like Δ_A to exhibit is:

[25] It should be noted that T_{Bin} incorporates N 'sub-estimators' $T_{n,1}, \ldots, T_{n,N}$ (one for each successive hypothesis test), no two of which need be the same.

[26] Once again, the reader is referred to Section 5.5 for a discussion of the magnitude of this constant.

For fixed $\alpha \in (0, 1)$, there should exist $q > 0$ and $0 < A_0 \leq A_1 < \infty$ such that

$$A_0 \left[\frac{|\log \alpha|}{n}\right]^{q/2} \leq \Delta_A(n, \alpha) \leq A_1 \left[\frac{|\log \alpha|}{n}\right]^{q/2} \tag{12}$$

for suitably large n.

If this is indeed the case, then the supporting conditions of **Theorem 2.3** of [DL91] are met, and, as discussed in the previous section, T_{Bin} is seen to achieve the desired rate of convergence. This rate is proportional to Δ_A; from Inequality (12) the rate is, hence, $O\left(n^{-q/2}\right)$.

Now, in general, the validity of Inequality (12) may be difficult to show. However, it is possible to show that if the geometry of T and $\mathcal{F}$ conform to certain criteria, then

$$b\left(c\sqrt{\frac{|\log \alpha|}{n}}\right) \leq \Delta_A(n, \alpha) \leq b\left(C\sqrt{\frac{|\log \alpha|}{n}}\right) \tag{13}$$

where $b(\epsilon)$ is the modulus of continuity described in Section 5.1.3. The geometric criteria necessary as well as the derivation of the above inequality are discussed in Sections 5.4.1 and 5.4.2. In the meanwhile we note that if Inequality (13) can indeed be established, then the problem is simplified dramatically: The establishment of $b(\epsilon)$ as a Hölder function is all that is necessary to transform Inequality (13) into a form which satisfies Inequality (12). The rate of convergence is, thus, $b(n^{-1/2})$. See Figure 5.

It would seem at first glance that we have simply replaced one set of criteria with another. This is indeed the case; however, as we will see in Sections 5.4.1 and 5.4.2, the replacement criteria are far easier to deal with.

5.4.1 The Case of Linear T and Convex $\mathcal{F}$

In this section, we show that sufficent conditions for Inequality (13) are linearity of the functional T and convexity of the distribution class $\mathcal{F}$. It can be shown that the lower bound of Inequality (13) can always be established; it is the upper bound which needs some work.

We begin by extending the notions of Hellinger Affinity and Hellinger Distance[27] to *classes* of distributions. For two classes of distributions $\mathcal{P}$ and $\mathcal{Q}$, define

$$\begin{aligned} \rho(\mathcal{P}, \mathcal{Q}) &= \sup\{\rho(P, Q) : P \in \mathcal{P}, Q \in \mathcal{Q}\}, \text{ and} \\ H(\mathcal{P}, \mathcal{Q}) &= \inf\{H(P, Q) : P \in \mathcal{P}, Q \in \mathcal{Q}\}. \end{aligned}$$

Now, for $\hat{F} \in \mathcal{F}$, consider the Hellinger ball $B_H(\epsilon, \hat{F}) = \{F \in \mathcal{F} : H(F, \hat{F}) \leq \epsilon\}$. By the definition of the modulus of continuity[28], $\{T(F) : F \in B_H(\epsilon, \hat{F})\} \subset [T(\hat{F}) - b(\epsilon), T(\hat{F}) + b(\epsilon)]$, so that for any t in the parameter space Ω, $H(\mathcal{F}_{\leq t}, \mathcal{F}_{\geq t+b(\epsilon)}) \geq \epsilon$. Recalling the identity $H^2(P, Q) = 2(1 - \rho(P, Q))$, this leads to

$$\rho(\mathcal{F}_{\leq t}, \mathcal{F}_{\geq t+b(\epsilon)}) \leq 1 - \frac{\epsilon^2}{2}$$

Now the crucial observation is that if T is a linear functional and $\mathcal{F}$ is a convex class, then $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{\geq t+\Delta}$ are *both convex*, for all t and all Δ. Hence, $\mathcal{F}_{\leq t} = conv(\mathcal{F}_{\leq t})$ and $\mathcal{F}_{\geq t+b(\epsilon)} = conv(\mathcal{F}_{\geq t+b(\epsilon)})$, so that

$$\rho\left(conv(\mathcal{F}_{\leq t}), conv(\mathcal{F}_{\geq t+b(\epsilon)})\right) \leq 1 - \frac{\epsilon^2}{2} \tag{14}$$

Now, Le Cam has established ([LeC86], page 477) that

$$\rho\left(conv(\mathcal{P}^{(n)}), conv(\mathcal{Q}^{(n)})\right) \leq \rho\left(conv(\mathcal{P}), conv(\mathcal{Q})\right)^n \tag{15}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are distribution classes and $\mathcal{P}^{(n)}$ and $\mathcal{Q}^{(n)}$ are the classes of corresponding product measures. Le Cam has also shown that $\rho(P, Q) \geq \pi(P, Q)$ where $\pi(P, Q)$ is the testing affinity[29] between distributions P and Q. If we put $\mathcal{P} = \mathcal{F}_{\leq t}$ and $\mathcal{Q} = \mathcal{F}_{\geq t+b(\epsilon)}$ and take suprema over t, Inequality (15) then yields

$$\begin{aligned} \sup_{t\in\Omega} \rho\left(conv(\mathcal{F}_{\leq t}), conv(\mathcal{F}_{\geq t+b(\epsilon)})\right)^n &\geq \sup_{t\in\Omega} \rho\left(conv(\mathcal{F}^{(n)}_{\leq t}), conv(\mathcal{F}^{(n)}_{\geq t+b(\epsilon)})\right) \\ &= \alpha_A(n, b(\epsilon)). \end{aligned}$$

Finally we substitute Inequality (14) into this last, to yield

$$\begin{aligned} \alpha_A(n, b(\epsilon)) &\leq \left(1 - \frac{\epsilon^2}{2}\right)^n, \text{ whence} \\ \Delta_A(n, \alpha) &\leq b\left(\sqrt{2(1 - \alpha^{1/n})}\right). \end{aligned}$$

[27] See Sections 5.1.1 and 5.1.2 for definitions.
[28] See Section 5.1.3 for definition.
[29] See Section 5.1.4 for definition.

A simple application of the inequality $\left(1-\alpha^{1/n}\right) \leq \frac{|\log \alpha|}{n}$, which appears as **Lemma 3.3** in [DL91], produces the upper bound of Inequality (13). As mentioned previously, once Inequality (13) has been validated, establishment of the modulus as a Hölder function suffices to guarantee that T_{Bin} attains the desired convergence rate of $b\left(n^{-1/2}\right)$.

Section 6 contain an elaborate example of a case where Δ_A is linked to the modulus via linearity of T and convexity of $\mathcal{F}$.

5.4.2 The General Case

In the previous section, we employed linearity of T and convexity of $\mathcal{F}$ to establish Inequality (12). Though these are *sufficient* conditions, we demonstrate here that they are not *necessary*. Indeed, the link between Δ_A and the modulus follows from an underlying geometric property, of which linearity of T and convexity of $\mathcal{F}$ constitute just one of many possible manifestations.

The essential geometric property is that the hardest *two-point* subproblem of testing $T(F) \leq t$ versus $T(F) \geq t+\Delta$ should be roughly as difficult, from a minimax risk point of view, as the full composite hypothesis-testing problem (i.e. testing $H_0 : conv(\mathcal{F}_{\leq t}^{(n)})$ versus $H_1 : conv(\mathcal{F}_{\geq t+\Delta}^{(n)})$ for sample size n).

Define the **two-point upper affinity** $\alpha_2(n, \Delta)$ and the **two-point testing bound** $\Delta_2(n, \alpha)$ as

$$\begin{aligned} \alpha_2(n, \Delta) &= \sup_{t \in \Omega} \pi\left(\mathcal{F}_{\leq t}^{(n)}, \mathcal{F}_{\geq t+\Delta}^{(n)}\right) \\ \Delta_2(n, \alpha) &= \sup\{\Delta : \alpha_2(n, \Delta) \geq \alpha\} \end{aligned}$$

Note the omission of convexification in comparison with the definitions of α_A and Δ_A. Now, some identities concerning Hellinger Affinities and Hellinger Distances combined with a little algebraic manipulation lead directly to an inequality identical to that of (12), but involving Δ_2 in place of Δ_A.[30] The details may be found in [DL91]. It now becomes apparent that this new inequality leads immediately to Inequality (12) provided only that Δ_A is roughly equal to Δ_2. Consult [DL91] for more precise criteria.

[30] Indeed this new inequality, combined with the fact that $\Delta_2(n, \alpha) \leq \Delta_A(n, \alpha)$ since $\alpha_2 \leq \alpha_A$, accounts for the lower bound of Inequality (12).

Yet, even this last criterion concerning the hardest two-point subproblem may be discarded, as long as the connection between Δ_A and the modulus can still be made. Indeed, In **Example 5.2** of [DL91], Donoho&Liu show that a suitable relationship between Δ_2 and Δ_A is elusive and we are forced to resort to other methods to find a link between Δ_A and the modulus of continuity.

Nevertheless, no matter how the connection with the modulus is made, $b(\epsilon)$ still needs to be a Hölder function in order to establish Inequality (12) and hence ensure that T_{Bin} achieves the desired convergence rate of $b\left(n^{-1/2}\right)$.

5.5 Link with Estimation Theory

In Section 5.2 we identified $\Delta_A(n, \alpha)$ as a *lower bound*, to within constants, on the rate of convergence of $T_n(\boldsymbol{X}_n)$ to $T(F)$. More specifally, we reiterated **Theorem 2.1** of [DL91], which states that

$$\inf_{T_n} \sup_{F \in \mathcal{F}} P_F\{|T_n - T(F)| \geq \Delta_A(n, \alpha)/2\} \geq \alpha/2.$$

We then went on to show that *for some constant* K, $K\Delta_A(n, \alpha)$ is an *attainable* rate of convergence, whence $\Delta_A(n, \alpha)$ emerges as the optimal rate of convergence *to within constants.*

In this section, we analyze the caveat "to within constants" from a Decision-Theoretic standpoint. We show that this description, though accurate, may be misleading in the sense that the constants involved will *not*, as is often the case, be swallowed up during our passage to the infinite. Moreover, the constants, though finite, are unbounded: classes of distributions can be found for which the constants are arbitrarily large and, correspondingly, the rate of uniform convergence arbitrarily reduced.

Consider the following problem of minimax location parameter estimation [ZM84]: Let F be some distribution with an even, unimodal density function. Define the distribution class $\mathcal{F} = \{F(\cdot - \theta) \ : \ \theta \in \Omega = [-d, d]\}$. The functional we are attempting to estimate is $T(F_\theta \in \mathcal{F}) = \theta$. Let Λ denote the action space of the statistician[31], and let $L(T_n(\boldsymbol{X}_n), \theta)$ denote the zero-one loss function defined on $\Lambda \otimes \Omega$:

$$L(T_n(\boldsymbol{X}_n), \theta) = \begin{cases} 0 & \text{if } |T_n(\boldsymbol{X}_n) - \theta| \leq e \\ \\ 1 & \text{if } |T_n(\boldsymbol{X}_n) - \theta| > e \end{cases}$$

[31] In this case, $\Lambda = [-d, d]$.

where $e > 0$ is given.

In [ZM84] it is shown that for the case $d = 2e$, the minimax risk is $\alpha = F(-e)$, while at the other extreme, for $d \to \infty e$, $\alpha \to 2F(-e)$. In other words,

$$\inf_{T_n} \sup_{F_\theta \in \mathcal{F}} P_{F_\theta}\{|T_n - T(F)| \geq e\} = \begin{cases} F(-e) & \text{if } d = 2e \\ 2F(-e) & \text{if } d = \infty e \end{cases} \tag{16}$$

If we fix the risk level at α and compare the relative sizes of the confidence intervals e_2 and e_∞ in the two cases, we see that, thanks to the evenness of F,

$$\begin{array}{rclcrcl} \alpha & = & F(-e_2) & = & 1 - F(e_2) & \Rightarrow & e_2 = F^{-1}(1-\alpha) \\ \alpha & = & 2F(-e_\infty) & = & 2(1 - F(e_\infty)) & \Rightarrow & e_\infty = F^{-1}\left(1 - \frac{\alpha}{2}\right) \end{array}$$

so that

$$\frac{e_\infty}{e_2} = \frac{F^{-1}\left(1 - \frac{\alpha}{2}\right)}{F^{-1}(1-\alpha)} = \frac{F^{-1}\left(\frac{\alpha}{2}\right)}{F^{-1}(\alpha)}.$$

We now make the connection with $\Delta_A(n, \alpha)$. Note that for any $t \in \Omega$, $conv(\mathcal{F}_{\leq t}) = \mathcal{F}_{\leq t}$ and $conv(\mathcal{F}_{\geq t+\Delta}) = \mathcal{F}_{\geq t+\Delta}$. Hence, owing to the unimodality and evenness of F, the upper affinity of the estimation problem is, for $n = 1$,

$$\begin{aligned} \alpha_A(1, \Delta) &= \sup_{t \in \Omega} \pi\left(\mathcal{F}_{\leq t}, \mathcal{F}_{\geq t+\Delta}\right) \\ &= \| F_{-\Delta/2} \wedge F_{\Delta/2} \| \\ &= 2F\left(-\frac{\Delta}{2}\right). \end{aligned}$$

Thus, the inverse upper affinity for $n = 1$ may be derived:

$$\Delta_A(1, \alpha) = 2F^{-1}\left(\frac{\alpha}{2}\right).$$

Now, in both the fixed-size confidence procedure and the hypothesis-testing settings, loss can be incurred only if the estimate differs from the true value of the location parameter by at least e or $\Delta_A(1, \alpha)/2$. Hence, $\Delta_A(1, \alpha)/2$ can be identified with e in Equation (16) above. But *which* e should we use, e_2 or e_∞?

If we set $\Delta_A(1, \alpha)/2 = e_2$ for some α, the lower bound in **Theorem 2.1** of [DL91] is seen to correspond exactly with Equation (16) above in the $d = 2e$ case, with

minimax risk of $\frac{\alpha}{2}$. However, for the $d = \infty e$ case, the minimax risk then becomes *twice* the lower bound of **Theorem 2.1.**

In order to preserve the risk level at $\frac{\alpha}{2}$ and hence accommodate all scenarios, we resort to setting $\Delta_A(1, \alpha)/2 = e_\infty$. Thus we see that *the optimal rate of convergence of estimator to parameter, $\Delta_A(1, \alpha)$, must be slowed by a factor of $e_\infty/e_2 = F^{-1}\left(\frac{\alpha}{2}\right)/F^{-1}(\alpha)$.*

This reduction in rate of convergence would not be as noteworthy if it were not for the fact that relatively simple classes of distributions may be found which satisfy the unimodality and evenness criteria and for which the factor $F^{-1}\left(\frac{\alpha}{2}\right)/F^{-1}(\alpha)$ can be made arbitrarily large.

Example: The Contaminated Normal distribution Consider a Contaminated Normal distribution, $F_{\epsilon,v} = (1 - \epsilon)N(0, 1) + \epsilon N(0, v)$ for some $0 \leq \epsilon \leq 1$ and $v \in [1, \infty)$.

Figure 7 shows a plot of $F_{\epsilon,v}^{-1}\left(\frac{\alpha}{2}\right)/F_{\epsilon,v}^{-1}(\alpha)$ for $v = 100$, $\epsilon = \frac{1}{10}$ and $0 \leq \alpha \leq 0.25$. For comparison, the curves for Normal, Cauchy, and Laplacian F are also shown.

Let us now focus our attention on the Contaminated Normal distribution: Figure 8 illustrates $F_{\epsilon,v}^{-1}\left(\frac{\alpha}{2}\right)/F_{\epsilon,v}^{-1}(\alpha)$ for various values of variance v of the contaminant, with ϵ held constant at $\frac{1}{10}$. The shape of the surface bears out what many evaluations of $F_{\epsilon,v}^{-1}\left(\frac{\alpha}{2}\right)/F_{\epsilon,v}^{-1}(\alpha)$ for $\epsilon = \frac{1}{10}$, $\alpha = 0.05$ and v ranging from 1 through 10^9 seem to suggest: that the maximum value of the curve increases as $C\sqrt{v}$ where $C \approx 0.2355$. In other words, the factor by which the rate of convergence is slowed seems to be proportional to the standard deviation of the contaminant, and is therefore unbounded.

As a final note, it should be emphasized that the demonstration of the possibility of unbounded constants is intended merely to warn the unwary and to discourage the blind application of the results; it is certainly not meant to detract from an otherwise very general and very powerful result.

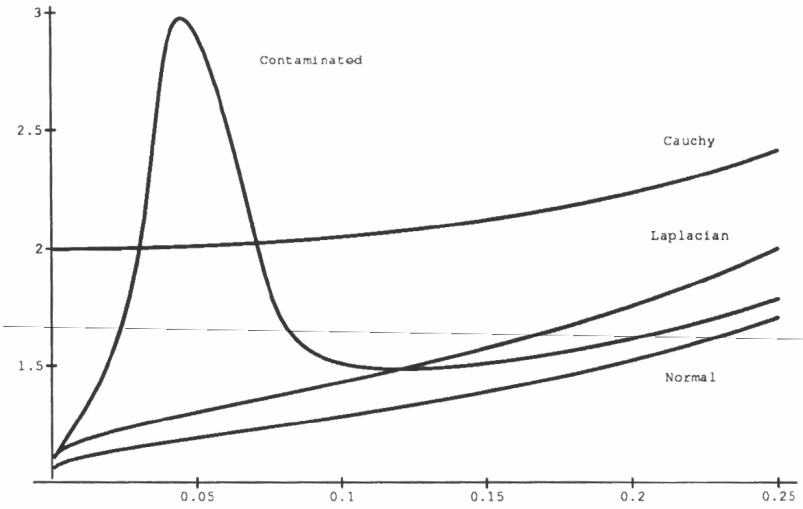


Figure 7: **Plot of $F^{-1}\left(\frac{\alpha}{2}\right)/F^{-1}(\alpha)$ as a function of $0 \leq \alpha \leq 0.25$ for F Normal, Cauchy, Laplacian and Contaminated Normal $(\frac{9}{10}N(0,1)+\frac{1}{10}N(0,100))$.**

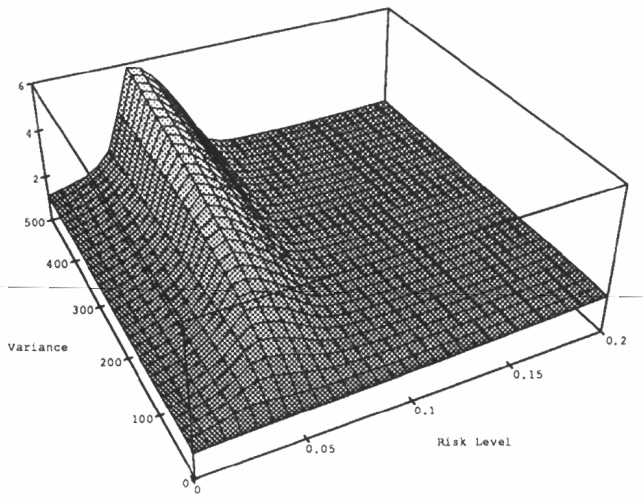


Figure 8: **Plot of the factor $e_\infty/e_2 = F_{\epsilon,v}^{-1}\left(\frac{\alpha}{2}\right)/F_{\epsilon,v}^{-1}(\alpha)$ as a function of $0 \leq \alpha \leq 0.2$ and $1 \leq v \leq 500$ for the Contaminated Normal distribution, where ϵ is held constant at $\frac{1}{10}$.**

6 Rate of Convergence over a VC Class

In order to gain insight into the manner in which the results of [DL91] may be applied, consider a relatively simple setting: Let us derive a lower bound on the rate of uniform convergence of relative frequency to probability over a class of events $\mathcal{A} \subset \mathcal{B}$ with respect to the probability space $\langle \boldsymbol{X}, \mathcal{B}, P \rangle$.

Section 3 summarized the result of [VC71] that a sufficient condition for such convergence to be *uniform* over $\mathcal{A}$ is merely that $\mathcal{A}$ be of polynomial discrimination with respect to the whole space $\boldsymbol{X}$. We therefore limit our analysis to such a class of events, and derive *rates* of convergence for estimators of $P(A)$, using methods outlined in [DL91] and described in Section 5. [32]

We begin by transforming the problem into a form to which the results of [DL91] are applicable. We need to find a suitable distribution class $\mathcal{F}$ and functional $T : \mathcal{F} \rightarrow \Re$.

Let $\theta_A : \boldsymbol{X} \rightarrow \{0, 1\}$ denote the indicator function for $A \in \mathcal{A}$.[33], so that we can define the stochastic process $\mathcal{D} = \{\theta_A \ : \ A \in \mathcal{A}\}$. Let $\mathcal{F}$ denote the corresponding class of distributions such that $F_A \in \mathcal{F}$ is the CDF corresponding to the random variable θ_A. Further, let T be the (linear) functional which returns the expected value of its argument $F \in \mathcal{F}$. Hence, for any $F_A \in \mathcal{F}$, $T(F_A) = P(A)$. By establishing a uniform rate of convergence of an estimate $T_n(\boldsymbol{X}_n)$[34] to the value of the functional $T(F_A)$ of an unknown distribution $F_A \in \mathcal{F}$ we simultaneously establish a rate on the uniform convergence of any estimate (including relative frequency) to probability over the class of events $\mathcal{A}$.[35]

Having defined $\mathcal{F}$ and T, we proceed to derive an expression for the inverse upper affinity $\Delta_A(n, \alpha)$ for some confidence level α. If $\Delta_A(n, \alpha)$ can then be shown to exhibit appropriate tail behavior, our quest will be accomplished: $\Delta_A(n, \alpha)$ will

[32] In the interests of consistency, we maintain the use of the symbol 'A' to represent an arbitrary member of $\mathcal{A}$, despite the unfortunate clash with the subscripts used in the symbolic representations of upper affinity α_A and inverse upper affinity Δ_A.

[33] We assume throughout that each event $A \in \mathcal{A}$ is measureable.

[34] where $\boldsymbol{X}_n$ is a vector of n i.i.d. F_A sample points

[35] Actually, since distinct sets of equal probability have indistinguishable images in $\mathcal{F}$, much of the structure of the class $\mathcal{A}$ may be lost in the transformation to $\mathcal{F}$. It is for this reason that we still insist that $\mathcal{A}$ be of polynomial discrimination to guarantee uniformity over $\mathcal{A}$; the uniformity result established using the results of [DL91] makes a statement only about the distribution class $\mathcal{F}$, and *cannot* be extrapolated back to the generating class of sets $\mathcal{A}$. Hence, *if* the conditions of [VC71] are satisfied by $\mathcal{A}$, then convergence will occur at best at the rate prescribed by [DL91].

represent the optimal rate of convergence, to within constants.

In Section 6.1 we derive an expression for $\Delta_A(n, \alpha)$ using various graphical and empirical techniques. For an arbitrary generating class of sets $\mathcal{A}$, we derive testing affinities and the upper affinity $\alpha_A(n, \Delta)$ en route to the inverse upper affinity $\Delta_A(n, \alpha)$ which is seen to display $n^{-1/2}$ tail behavior. Then, in Section 6.2, we show how the geometry of the problem, in the form of the modulus of continuity, may be exploited to short-cut the derivation.

6.1 Graphical and Empirical Approach

Recall the definition of the upper affinity

$$\alpha_A(n, \Delta) = \sup_t \pi \left(conv(\mathcal{F}_{\leq t}^{(n)}), conv(\mathcal{F}_{\geq t+\Delta}^{(n)}) \right)$$

where $\mathcal{F}_{\leq t}$ and $\mathcal{F}_{\geq t+\Delta}$ denote the subsets of $\mathcal{F}$ where T takes values $\leq t$ and $\geq t + \Delta$ respectively, $\mathcal{F}^{(n)}$ denotes the set of product measures $F^{(n)}$, $F \in \mathcal{F}$, and $\pi(\mathcal{P}, \mathcal{Q})$ represents the testing affinity between the classes $\mathcal{P}$ and $\mathcal{Q}$. Recall also the identity

$$\pi(\mathcal{P}, \mathcal{Q}) = \sup_{P \in \mathcal{P}, Q \in \mathcal{Q}} \| P \wedge Q \|$$

Now, for the current example, an arbitrary member $F_A : \Re \to [0, 1]$ of $\mathcal{F}$ has the profile

$$F_A(x) = \begin{cases} 0 & \text{for } x \in (-\infty, 0) \\ 1 - P(A) & \text{for } x \in [0, 1) \\ 1 & \text{for } x \in [1, \infty) \end{cases}$$

so that $T(F_A)$ also corresponds to $f_A(1)$, where f_A is the PMF of F_A. Let F_A have the dual notation F_t, where $t = T(F_A) = P(A)$. Then the product measure $F_t^{(n)}$ is seen to have a PMF $f_t^{(n)}$ defined on $\Re^n$ as:

$$f_t^{(n)}(x_1, x_2, \ldots, x_n) = \begin{cases} (t)^{(\sum_i x_i)} (1-t)^{(n - \sum_i x_i)} & \text{for } (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n \\ 0 & \text{for } (x_1, x_2, \ldots, x_n) \notin \{0, 1\}^n. \end{cases}$$

Hence, for $0 \leq \Delta \leq \frac{1}{2}$ and $0 \leq t \leq 1 - \Delta$,

$$\pi \left(F_t^{(n)}, F_{t+\Delta}^{(n)} \right) = \sum_{k=0}^{n} \binom{n}{k} \left[t^k (1-t)^{n-k} \wedge (t+\Delta)^k (1-t-\Delta)^{n-k} \right]$$

It is already clear that a purely theoretical analysis will soon become overwhelmingly complicated. We therefore resort to numerical and graphical tools for assistance.

Figure 9 shows a plot of $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ as a function of t and Δ, with $n = 6$. Similar plots for various values of n yield surfaces which all share the following algebraically verifiable properties:

1. For fixed t, $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ is monotone decreasing in Δ.
2. For fixed Δ, $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ exhibits $(n-1)$ cusps[36], as shown in Figure 10. Furthermore, $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ is maximized at the cusp nearest the hyperplane $t = \frac{1-\Delta}{2}$.
3. The surfaces are symmetrical about the hyperplane $t = \frac{1-\Delta}{2}$ over the region $0 \leq \Delta \leq 1$, $0 \leq t \leq 1 - \Delta$.

From property (1) we deduce that

$$\pi\left(\mathcal{F}_{\leq t}^{(n)}, \mathcal{F}_{\geq t+\Delta}^{(n)}\right) = \pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right).$$

In words, the most error-prone two-point testing problem between $\mathcal{F}_{\leq t}^{(n)}$ and $\mathcal{F}_{\geq t+\Delta}^{(n)}$ occurs for elements which reside on the very 'edges' of the classes, and whose functional values are, hence, as close to each other as possible. Moreover, a greater testing affinity cannot be generated even if the *convexified* classes $conv(\mathcal{F}_{\leq t}^{(n)})$ and $conv(\mathcal{F}_{\geq t+\Delta}^{(n)})$ are tested in the stead of $\mathcal{F}_{\leq t}^{(n)}$ and $\mathcal{F}_{\geq t+\Delta}^{(n)}$. This last is a consequence of the following two facts:

- The functions $t^k(1-t)^{n-k}$, $k = 1, 2, \ldots, n$ are *unimodal* over $0 \leq t \leq 1$;
- For the calculation of $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$, we select the *smaller* of $t^k(1-t)^{n-k}$ and $(t+\Delta)^k(1-t-\Delta)^{n-k}$ for all $k = 1, 2, \ldots, n$.

[36] It can be shown that for fixed Δ, the cusps occur at t satisfying

$$\left(\frac{t}{t+\Delta}\right)^k = \left(\frac{1-t-\Delta}{1-t}\right)^{n-k} \quad \text{for } k = 1, 2, \ldots, n-1$$

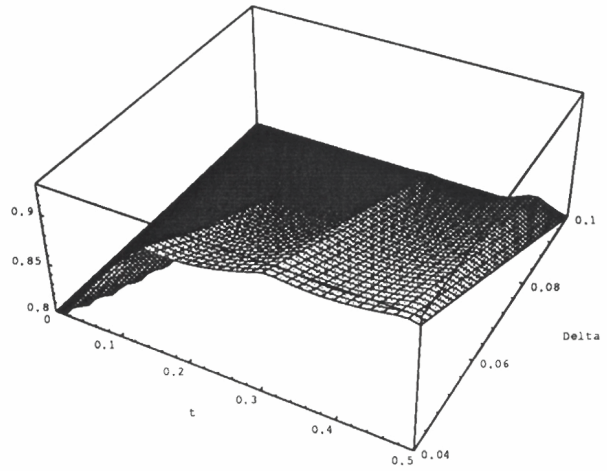


Figure 9: $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ as a function of $0 \leq t \leq 0.5$ and $0.04 \leq \Delta \leq 0.1$, with $n = 6$.

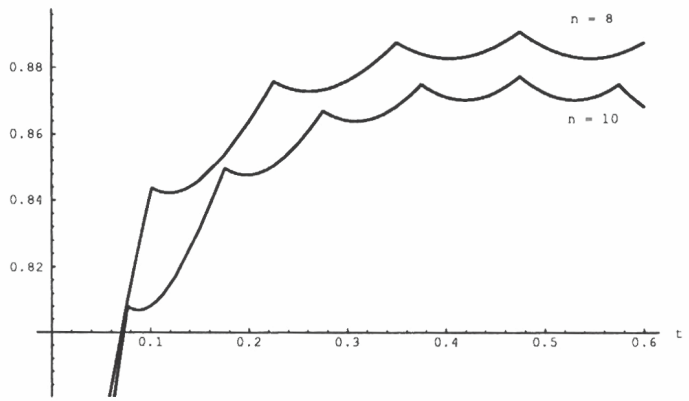


Figure 10: $\pi\left(F_t^{(n)}, F_{t+\Delta}^{(n)}\right)$ as a function of $0 \leq t \leq 0.6$ for fixed $\Delta = 0.05$ and $n = 8, 10$. Both functions are maximized at $t = \frac{1-\Delta}{2} = 0.475$.

For even n, properties (2) and (3) imply that the cusp which maximizes $\pi\left(F_t^{(n)},\, F_{t+\Delta}^{(n)}\right)$ occurs at $t = \frac{1-\Delta}{2}$.

Combining all these observations, we deduce

$$\begin{aligned}
\alpha_A(n,\,\Delta) &= \sup_t \pi\left(conv(\mathcal{F}_{\leq t}^{(n)}),\, conv(\mathcal{F}_{\geq t+\Delta}^{(n)})\right) \\
&= \sup_t \pi\left(F_t^{(n)},\, F_{t+\Delta}^{(n)}\right) \\
&= \pi\left(F_{\frac{1-\Delta}{2}}^{(n)},\, F_{\frac{1+\Delta}{2}}^{(n)}\right) \text{ for even } n \\
&= \sum_{k=0}^{n} \binom{n}{k}\left[\left(\tfrac{1-\Delta}{2}\right)^k \left(\tfrac{1+\Delta}{2}\right)^{n-k} \wedge \left(\tfrac{1+\Delta}{2}\right)^k \left(\tfrac{1-\Delta}{2}\right)^{n-k}\right]
\end{aligned}$$

Exploiting symmetry, we conclude that, for even n,

$$\alpha_A(n,\,\Delta) = \binom{n}{\frac{n}{2}}\left(\tfrac{1-\Delta^2}{4}\right)^{n/2} + 2\sum_{k=0}^{\frac{n}{2}-1} \binom{n}{k}\left[\left(\tfrac{1-\Delta}{2}\right)^k \left(\tfrac{1+\Delta}{2}\right)^{n-k}\right]$$

A plot of $\alpha(n,\,\Delta)$ is shown in Figure 11.

Recall now the definition of the inverse upper affinity

$$\Delta_A(n,\,\alpha) = \sup\{\Delta \,:\, \alpha_A(n,\,\Delta) \geq \alpha\}.$$

For a fixed confidence level α, the graph of $\Delta_A(n,\,\alpha)$ versus n corresponds to the contour line at height α on the surface of $\alpha_A(n,\,\Delta)$. Hence, the contour plot of Figure 11 actually serves also as a plot of $\Delta_A(n,\,\alpha)$ versus n for $\alpha = 0.1,\, 0.2,\, \ldots,\, 0.9$. Each curve is seen to follow $\frac{c}{\sqrt{n}}$ for some value of c, whence we conclude that $\Delta_A(n,\,\alpha)$ displays the required tail behaviour delineated in Inequality (12) of Section 5.4. The conditions of **Theorem 2.3** of [DL91] are met, and $\Delta_A(n,\,\alpha) = O\left(n^{-1/2}\right)$ emerges as the optimal rate of convergence of an estimator to $T(F_A) = P(A)$, $A \in \mathcal{A}$.

It should be emphasized that this is only an empirical result, abounding in empirical observation and heuristic deduction and lacking somewhat in rigour. It would seem that without some complicated algebra, we have reached an impasse in our quest for a precise expression for the optimal rate of convergence.

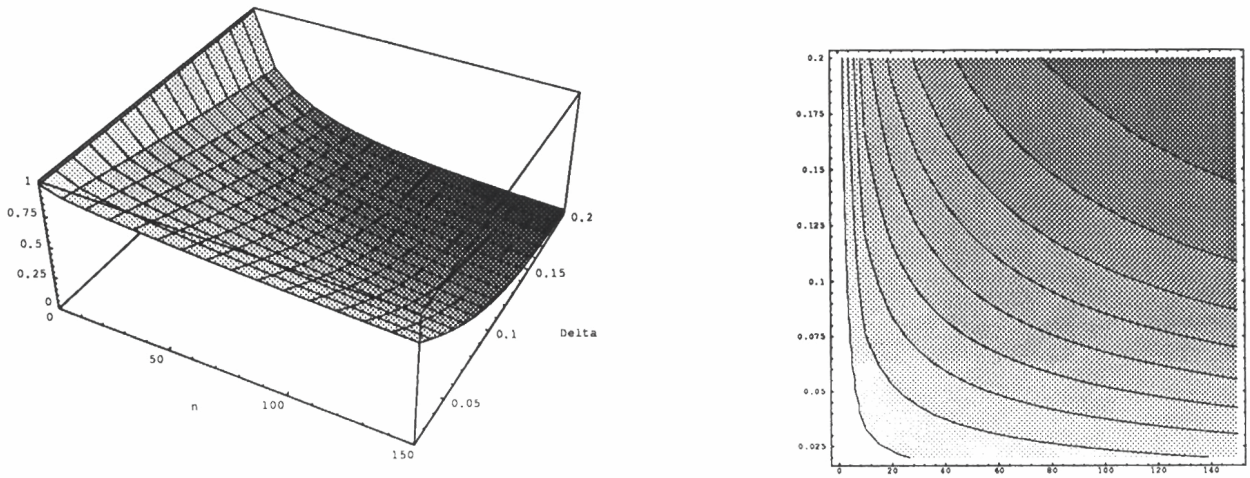


Figure 11: **3-Dimensional Plot and Contour Plot of $\alpha_A(n, \Delta)$ for $0 \leq n \leq 150$ and $0.02 \leq \Delta \leq 0.2$.**

6.2 The Modulus to the Rescue

We showed in Section 5.4.1 that the problem of deriving optimal rates of convergence is greatly simplified provided

1. The functional T is linear,
2. The class $\mathcal{F}$ is convex, and
3. The modulus of continuity of T with respect to $\mathcal{F}$, $b(\epsilon)$, is a Hölder function of exponent q.

Indeed, if the geometry of $\mathcal{F}$ and T satisfy these three criteria, we can immediately conclude that the optimal rate of convergence is $b\left(n^{-1/2}\right) = \Theta\left(n^{-q/2}\right)$.

For the current setup, the functional T returns the expected value of its argument $F \in \mathcal{F}$, and is, thus, linear. Though $\mathcal{A}$ does not necessarily generate a convex $\mathcal{F}$, we can continue the analysis using $conv(\mathcal{F})$ as the distribution class. Of course, the price to be paid is that the rates will no longer necessarily be optimal since we are

dealing with a superset of the original class[37]. Nevertheless, there are many interesting examples of VC classes $\mathcal{A}$ for which $\mathcal{F}$ is itself convex[38] and nothing is lost. We assume henceforth that $\mathcal{F}$ is convex.

The only remaining criterion concerns the modulus of continuity $b(\epsilon)$, to which we now turn our attention.

Recall the definition of the modulus of continuity of T over $\mathcal{F}$

$$b(\epsilon) = \sup \{|T(F_1) - T(F_0)| \ : \ H(F_1, F_0) \leq \epsilon, F_1, F_0 \in \mathcal{F}\}.$$

Here $H(F_1, F_0)$ is the Hellinger Distance between two probability measures F_1 and F_0 in $\mathcal{F}$ and is defined as

$$H^2(F_1, F_0) = \frac{1}{2} \int \left(\sqrt{f_1} - \sqrt{f_0}\right)^2 d\mu$$

where f_1 and f_0 are the PMF's of F_1 and F_0 respectively.

In our case, the PMF f_t is defined as

$$f_t(x) = \begin{cases} 1-t & \text{for } x = 0 \\ t & \text{for } x = 1 \\ 0 & \text{for } x \in \Re,\ x \notin \{0, 1\} \end{cases}$$

Hence, the Hellinger distance between two arbitrary members F_α and F_β of $\mathcal{F}$ is

$$H(F_\alpha, F_\beta) \ = \ \sqrt{\frac{\left(\sqrt{1-\alpha}-\sqrt{1-\beta}\right)^2 + \left(\sqrt{\alpha}-\sqrt{\beta}\right)^2}{2}} \tag{17}$$

Figure 12 shows a plot of $H(F_\alpha, F_\beta)$ as a function of α and β. Contour lines represent the skins of Hellinger balls, so that from the figure we see that for any Hellinger ball of radius ϵ, the quantity $|\alpha - \beta|$ is maximized on the skin of the ball at $\alpha = 1 - \beta$. Substituting into Equation (17) yields, for any $0 \leq \epsilon \leq 1$,

$$H(F_\alpha, F_{1-\alpha}) = \epsilon \ = \ \sqrt{\frac{\left(\sqrt{1-\alpha}-\sqrt{\alpha}\right)^2 + \left(\sqrt{\alpha}-\sqrt{1-\alpha}\right)^2}{2}}$$

[37]On the other hand, if we could find a convex *subset* of $\mathcal{F}$, the rate of convergence for the subset would form a lower bound on the optimal rate for $\mathcal{F}$. However, this lower bound might no longer be attainable.

[38]Consider, for instance, the class $\mathcal{A}$ of intervals of $\Re$.

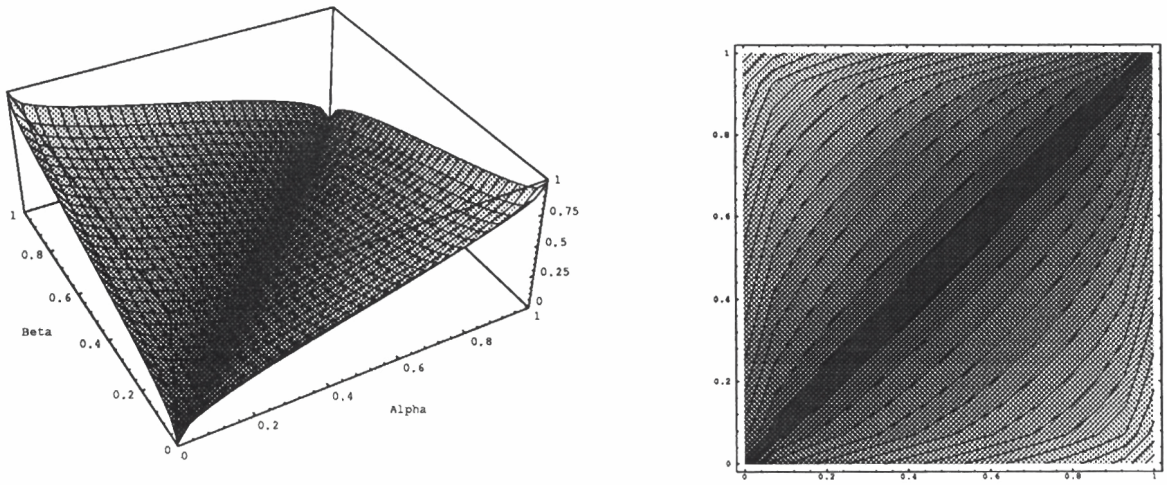


Figure 12: **3-Dimensional Plot and Contour Plot of $H(F_\alpha, F_\beta)$ as a function of α and β.**

$$\begin{aligned} &= \sqrt{\alpha} - \sqrt{1-\alpha} \\ \text{Hence, } \epsilon^2 &= 1 - 2\sqrt{\alpha(1-\alpha)} \\ \text{so that } \left(\frac{1-\epsilon^2}{2}\right)^2 &= \alpha - \alpha^2 \\ \text{whence } \alpha &= \frac{1 \mp \sqrt{1-(1-\epsilon^2)^2}}{2} \end{aligned}$$

From here we can derive the modulus of continuity: For any $0 \leq \epsilon \leq 1$,

$$\begin{aligned} b(\epsilon) &= |\alpha - \beta| \\ &= |1 - 2\alpha| \\ &= \sqrt{1-(1-\epsilon^2)^2} \\ &= \epsilon\sqrt{2-\epsilon^2}. \end{aligned}$$

A Taylor Series expansion yields $b(\epsilon) = \sqrt{2}\epsilon - \frac{1}{\sqrt{8}}\epsilon^3 - O(\epsilon^5)$, so that for small ϵ, $b(\epsilon)$ is seen to be Hölderian of exponent 1. We conclude that the rate of conver-

gence optimal for all possible estimators of $T(F_A)$, including relative frequency, is $b\left(n^{-1/2}\right) = \Theta\left(n^{-1/2}\right)$, and that this rate is, in fact, attainable.

Finally, the astute reader may be perturbed by a possible incongruity between of the above result and that of Inequality (5). Indeed, fixing $\hat{T}_n$ as the relative frequency estimator, the methods of [DL91] have lead to the lower bound

$$\sup_{F\in\mathcal{F}} P\left\{|\hat{T}_n - T(F)| \geq \Delta_A(n, \alpha)/2\right\} \geq \alpha/2.$$

On the other hand, if the generating class of events $\mathcal{A}$ is of polynomial discrimination, and we denote the order of the majorizing polynomial by d [39], then the approach of [VC71] leads to the upper bound

$$P\left\{\sup_{F\in\mathcal{F}} |\hat{T}_n - T(F)| \geq \epsilon\right\} \leq 4\,(2n+1)^d\, e^{-\epsilon^2 n/8}$$

for $n \geq d/2$.

With respect to the probability space $\langle \Psi, \mathcal{B}, P\rangle$, define, for each $F \in \mathcal{F}$ the 'bad-set' $B_F(\delta) = \{\psi \in \Psi \;:\; |\hat{T}_n - T(F)| \geq \delta\}$. The lower bound of [DL91] then limits the rate of convergence to zero of the probability of the largest[40] bad-set $\sup_{F\in\mathcal{F}} P\left(B_F(\Delta_A/2)\right)$, while the upper bound furnished by [VC71] guarantees a rate of $4\,(2n+1)^d\, e^{-\epsilon^2 n/8}$ for the convergence to zero of $P\left\{\sup_{F\in\mathcal{F}} |\hat{T}_n - T(F)| \geq \epsilon\right\} = P\left\{\bigcup_{F\in\mathcal{F}} B_F(\epsilon)\right\}$.

Hence, if we equate $\epsilon = \Delta_A(n, \alpha)/2$, elementary set theory allows us to relate the two bounds:

$$\begin{aligned} 4\,(2n+1)^d\, e^{-\epsilon^2 n/8} &\geq P\left\{\sup_{F\in\mathcal{F}} |\hat{T}_n - T(F)| \geq \epsilon\right\} = P\left\{\bigcup_{F\in\mathcal{F}} B_F(\epsilon)\right\} \\ &\geq \sup_{F\in\mathcal{F}} P\left(B_F(\epsilon)\right) = \sup_{F\in\mathcal{F}} P\left(B_F(\Delta_A/2)\right) \\ &\geq \alpha/2 \end{aligned}$$

It would seem that the exponentially decreasing upper bound may quickly diminish to less than the lower bound! The resolution of this apparent contradiction, however,

[39] d is called the *VC-dimension* of $\mathcal{A}$.

[40] with respect to measure P

lies in the fact that ϵ, now equated with $\Delta_A(n, \alpha)$, is no longer a constant, but is instead monotonically decreasing in n. Indeed, the rate of decrease of ϵ is sufficient to compensate for the apparent exponential decay of the upper bound: In the above example we showed that $\Delta_A(n, \alpha)$ is proportional to $b(n^{-1/2})$ where $b(\delta)$ is asymptotically linear in δ. Hence, putting $\epsilon = Cn^{-1/2}$ for some constant C, the upper bound is seen to degenerate into an increasing function:

$$4\,(2n+1)^d\, e^{-\epsilon^2 n/8} = 4\,(2n+1)^d\, e^{-C^2/8} \propto (2n+1)^d.$$

The integrity of the results is preserved.

6.3 Extension to Manageable Classes of Functions

As a final note, we mention that a similar application of the results of [DL91] to a manageable class of functions $\mathcal{G} = \{g_\lambda \;:\; \lambda \in \Lambda\}$ allows us to derive a lower bound on the rate of uniform convergence to zero of the empirical process[41] operating over this class with respect to the probability space $\langle \Re, \mathcal{B}, P \rangle$ [42].

As in the above example, we begin by casting the problem into a mold to which the results of [DL91] are applicable. Our first task is to define a suitable class $\mathcal{F}$ of distributions on the measureable space $\langle \Re, \mathcal{B} \rangle$ and to find a mapping $R : \mathcal{G} \to \mathcal{F}$. Furthermore, if a functional $T : \mathcal{F} \to \Re$ can be found such that for all $g \in \mathcal{G}$, $T(R(g)) = P\,g$, then establishing an optimal rate for *all* estimators $T_n : \Re^n \to \Re$ to converge to T uniformly over $\mathcal{F}$ is tantamount to establishing a lower bound on the rate of convergence to $\mathbb{E}_P$ of a *specific* estimator such as $\mathbb{E}_{P_n}$, the expectation with respect to the empirical measure[43] P_n. A scaling by a factor $n^{1/2}$ then establishes an analogous bound on the rate of convergence to zero of the empirical process.

An intuitive choice for the mapping R and the functional T is to treat g_λ as a random variable[44] on $\langle \Re, \mathcal{B}, P \rangle$, have $R(g_\lambda) = F_\lambda$ be the distribution of this random variable, and let $T\,(R(g_\lambda))$ be the expected value of this distribution F_λ. Regardless of P, the condition $T(R(g)) = P\,g$ is then satisfied for all $g \in \mathcal{G}$, provided $P\,g$ exists.

---

[41] See Section 4 for definition.

[42] Where $\mathcal{B}$ denotes the Borel field on $\Re$ and P is some probability measure of finite variance. This last condition is needed in order for the Central Limit Theorem to be applicable — see Section 2 of [Pol89].

[43] See Section 4 for definition.

[44] We assume here that g_λ is measureable for all $\lambda \in \Lambda$.

From here, an expression may be derived for the *rate* of convergence by methods inspired by [DL91]. In particular, we once again invoke one of the main results of [DL91]:

> If T is linear, $\mathcal{F}$ is convex, and the modulus of continuity of T with respect to $\mathcal{F}$, $b(\epsilon)$, is a Hölder function of exponent q, then the optimal rate of convergence is $b\left(n^{-1/2}\right) = \Theta\left(n^{-q/2}\right)$.

Since T is an expectation, it is linear. In general, however, $\mathcal{F}$ is *not* necessarily convex. The simplest remedy is to derive an optimal rate of convergence for the class $conv(\mathcal{F})$ rather than for $\mathcal{F}$. Of course, the price of the simplification is that $b(n^{-1/2})$ is demoted from its position as optimal rate to a humble status of upper bound for the minimax estimator T_{Bin} described in Section 5.3. However, provided the losses incurred by convexification do not exceed the gains furnished by the use of an *optimal* estimator instead of $\mathbb{E}_{P_n}$, the validity of the main claim — that $b(n^{-1/2})$ forms a lower bound on the empirical process rate of convergence — is not jeopardized[45].

Hence, all that remains to be done is to derive an expression for the modulus of continuity $b(\epsilon)$ of T with respect to $conv(\mathcal{F})$. Provided convexification losses are not excessive, verification that $b(\epsilon)$ is Hölderian of exponent q is sufficient to establish $b\left(n^{-1/2}\right) = \Theta\left(n^{-q/2}\right)$ as the lower bound on the rate of uniform convergence to zero of the empirical process operating over $\mathcal{G}$.

[45] Even in the case where convexification losses outstrip the gains provided by an *optimal* estimator, some significance can still be gleaned from the results: $b(n^{-1/2})$ then forms an upper bound on the *empirical* process rate of convergence, guaranteeing a certain rate. Moreover, it may yet be possible to establish $b(n^{-1/2})$ as a *lower* bound as well, using methods from [DL91] which do not rely on the convexity of $\mathcal{F}$ (See Section 5.4.2 for elaboration). Hence, even greater import is imparted to the rate $b(n^{-1/2})$: Since it forms an upper *and* lower bound on the empirical process rate, asymptotic equality of the two rates may be inferred.

7 Conclusion

It was the ambition of this survey not only to give conspectuses of the main threads of each of the three papers [VC71], [Pol89] and [DL91], but also to expose their intricate intertwinement and interdependence.

The first paper, [VC71], addresses the problem of establishing criteria subject to which one may conclude that the relative frequencies of events converge to their probabilities *uniformly* over a class $\mathcal{A} \subset \mathcal{B}$ with respect to the probability space $\langle \Psi, \mathcal{B}, P \rangle$. The results are twofold:

1. A sufficient condition for uniform convergence is that the class of events be *of polynomial discrimination.* No constraints need be imposed on the distribution P.

2. A sufficient *and* necessary condition for uniform convergence is the asymptotic approach to zero of the ratio of entropy $H^{\mathcal{A}}(l)$ to sample size l. Since the entropy of a class of events $\mathcal{A}$ is defined as the *expected* value of the index of $\mathcal{A}$, the satisfaction of this condition depends upon the distribution P.

The results of [VC71] continue to be pertinent to, and have impact upon, areas such as Neural Network Theory and Learning Theory. Though stronger results have emerged since its publication, the paper retains its pre-eminance, if not for its continuing general applicability then for the sheer elegance of its derivations.

The second paper, [Pol89], sees the extension of the ideas of [VC71] to a class of *functions* $\mathcal{G}$. The main concern is the establishment of criteria for which convergence of the empirical mean $P_n g$ to the actual mean $P g$ may be guaranteed uniformly over $\mathcal{G}$. The climax of [Pol89] is basically that the desired uniformity is attained provided $\mathcal{G}$ is a manageable class, along with a few subsidiary conditions. The link with [VC71] is strengthened by the rather remarkable result that if $\{subgraph(g) \ : \ g \in \mathcal{G}\}$ is of polynomial discrimination, then $\mathcal{G}$ is, in fact, a *manageable class* of functions.

Perhaps more than in the actual results, however, the significance of [Pol89] lies in its exposition of a very powerful technique for the analysis of the entire family of problems involving averages of functions of independent observations, of which the problem scrutinized here — that of finding criteria under which these averages converge uniformly to the expected values of the functions — is a member.

Any treatment of convergence concepts would be incomplete without a discussion of *rates* of convergence. The theme of [DL91] revolves around a bound on the rate of convergence of an estimate $T_n(\boldsymbol{X}_n)$[46] to the value of a functional $T(F)$ of an unknown distribution $F \in \mathcal{F}$ uniformly over a class of distributions $\mathcal{F}$. The main result is that for estimating a *linear* functional over a *convex* distribution class $\mathcal{F}$, the geometry of the problem, expressed in terms of the modulus of continuity $b(\epsilon)$, determines the optimal rate of convergence. Moreover, if $b(\epsilon)$ is a Hölder function of exponent q, the optimal rate is $b\left(n^{-1/2}\right) = \Theta\left(n^{-q/2}\right)$ and is, in fact, attainable. As an encore, it is further shown that the prerequisites of linearity and convexity may be discarded, provided that the essence of the geometry is preserved: A new criterion is that the hardest two-point subproblem of testing $T(F) \leq t$ versus $T(F) \geq t + \Delta$ should be roughly as difficult, from a minimax risk point of view, as the full composite hypothesis-testing problem.

As mentioned in the Introduction and demonstrated in the elaborate example of Section 6, a little reflection shows the results of [DL91] to be directly applicable to the convergence problems analyzed in [VC71] and [Pol89]. Indeed, given a class of functions $\mathcal{G}$[47], each function $g \in \mathcal{G}$ may be construed as a random variable with respect to the probability space $\langle \Re, \mathcal{B}, P \rangle$. Let $\mathcal{F}$ be the class of marginal distributions of the resultant stochastic process, and choose the (linear) functional $T(F)$, $F \in \mathcal{F}$ to be the expected value of F, i.e. $T(F) = P\,g$ where $F \in \mathcal{F}$ is the distribution of the random variable $g \in \mathcal{G}$. Establishing the modulus of continuity of T over $conv(\mathcal{F})$ as a Hölder function of exponent q places a lower bound of $b\left(n^{-1/2}\right) = \Theta\left(n^{-q/2}\right)$ on the rate of uniform convergence to $T(F)$ of *any* estimate $T_n(\boldsymbol{X}_n)$, *including the empirical expectation* $P_n\,g$.

Of course, the results of [DL91] extend far beyond these rather confined cases. Indeed, the power and generality of the results is matched only by the scope of their applicability: Nonparametric distribution classes succumb to investigation as tractably as parametric classes, and the latitude afforded in the choice of functional T is virtually unconstrained. For these reasons, the results of [DL91] may very well assume a pivotal role in future research within the field of stochastic processes and their rates of convergence.

[46] where $\boldsymbol{X}_n$ is a vector of n i.i.d. F sample points

[47] $\mathcal{G}$ could be a class of indicator functions if classes of events are involved as in [VC71].

Papers Surveyed

[VC71] **Vapnik & Chervonenkis**: (1971)
On the Uniform Convergence of Relative Frequencies of Events to their Probabilities.
Theory of Probability and its Applications, 1971, Vol. 16, No. 2, 264-280.

[Pol89] **Pollard, D.**: (1989)
Asymptotics via Empirical Processes.
Statistical Science, 1989, Vol. 4, No. 4, 341-366.

[DL91] **Donoho & Liu**: (1991)
Geometrizing Rates of Convergence, II.
The Annals of Statistics, 1991, Vol. 19, No. 2, 633-667.

References

[BD77] **Bickel & Doksum**: (1977)
Mathematical Statistics: Basic Ideas and Selected Topics, Holden-Day.

[Bil79] **Billingsley, P.**: (1979)
Probability and Measure, Wiley.

[CB90] **Casella & Berger**: (1990)
Statistical Inference, Wadsworth.

[Dud78] **Dudley, R. M.**: (1978)
Central Limit Theorems for Empirical Measures.
The Annals of Probability, 1978, No. 6, 899-929.

[Dud87] **Dudley, R. M.**: (1987)
Universal Donsker Classes and Metric Entropy.
The Annals of Probability, 1987, No. 15, 1306-1326.

[Fal90] **Falconer, K.**: (1990)
Fractal Geometry: Mathematical Foundations and Applications, Wiley.

[Far66] **Farrell, R. H.**: (1966)
On the Lack of a Uniformly Consistent Sequence of Estimators of a Density Function in Certain Cases,

[Far72] **Farrell, R. H.**: (1972)
On the Best Obtainable Asymptotic Rates of Convergence in Estimation of a Density Function at a Point.
The Annals of Mathematical Statistics, 1972, Vol. 43, No. 1, 170-180.

[GD86] **Gray & Davisson**: (1986)
Random Processes: A Mathematical Approach for Engineers, Prentice-Hall.

[IH81] **Ibragimov & Hasminskii**: (1981)
Statistical Estimation: Asymptotic Theory, Springer.

[KF70] **Kolmogorov & Fomin**: (1970)
Introductory Real Analysis, Dover.

[LY90] **Le Cam & Yang**: (1990)
Asymptotics in Statistics: Some Basic Concepts, Springer-Verlag.

[LeC86] **Le Cam, L.**: (1986)
Asymptotic Methods in Statistical Decision Theory, Springer-Verlag.

[Mal88] **Maller, R. A.**: (1988)
Asymptotic Normality of Trimmed Means in Higher Dimensions.
The Annals of Probability, 1988, Vol. 16, No. 4, 1608-1622.

[Mintz] **Mintz, M.**
A Review of some ideas in Measure Theory and Probability Theory
Lecture Notes, pp 51-500.

[Pol84] **Pollard, D.**: (1984)
Convergence of Stochastic Processes, Springer-Verlag.

[Rom88] **Romano, J. P.**: (1988)
On Weak Convergence and Optimality of Kernel Density Estimates of the Mode.
The Annals of Statistics, 1988, Vol. 16, No. 2, 629-647.

[SY81] **Sacks & Ylvisaker**: (1981)
Asymptotically Optimum Kernels for Density Estimation at a Point.
The Annals of Statistics, 1981, Vol. 9, No. 2, 334-346.

[Sch50] **Schläfli, L.**: (1950)
Gesammelte mathematische Abhandlungen, I, Basel, 1950.

[Wen62] **Wendel, J. G.**: (1962)
A Problem in Geometric Probability.
Math. Scand., 1962, Vol. 11, 109-111.

[Wol91] **Wolfram, S.**: (1991)
Mathematica, Addison-Wesley.

[Yat85] **Yatracos, Y. G.**: (1985)
Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy.
The Annals of Statistics, 1985, Vol. 13, No. 2, 768-774.

[ZM84] **Zeytinoglu & Mintz**: (1984)
Optimal Fixed Size Confidence Procedures for a Restricted Parameter Space.
The Annals of Statistics, 1984, Vol. 12, No. 3, 945-957.