

# Statistical Learning Library in DROP

**v3.15** 4 October 2017

# Section I: Probabilistic Bounds and Concentration of Measure of Sequences

**Probabilistic Bounds**

**Motivation**

1. Design of Probabilistic Bounds: The primary objective of estimating bounds is to really bound the “bad metric”, and it usually takes the form

or

Here is the “bad metric”, and is the bounds (or cut-off) on the “bad metric”.

1. Bounds Setting Techniques:
   1. Markov Inequality
   2. Jensen’s Convexity Inequality
   3. Cauchy’s Joint Bounding Inequality
   4. Tail Probability Bounds Approaches (in several chapters below)
   5. Sample Independence Inequality
   6. Local Taylor Expansion Inequality
   7. Local Extrema (mini-max) Inequality
   8. Cross Function (e.g., Hypothesis) Union Bounding Inequality
   9. Symmetrization Inequality

**Tail Probability Bounds Estimation - Survey**

1. Martingale Methods: Milman and Schechtman (1986) and McDiarmid (1989, 1998).
2. Information-Theoretic Methods: Ahlswede, Gacs, and Korner (1976), Marton (1986, 1996a, 1996b), Dembo (1997), Massart (1998), and Rio (2001).
3. Talagrand’s Induction Methods: Talagrand (1995, 1996a, 1996b), Panchenko (2001), McDiarmid (2002), Panchenko (2002, 2003), and Luczak and McDiarmid (2003).
4. Entropy Methods: This is based on logarithmic Sobolev inequalities, see Ledoux (1996, 1997), Bobkov and Ledoux (1997), Boucheron, Lugosi, and Massart (2000), Massart (2000), Rio (2001), Bousquet (2002), Boucheron, Lugosi, and Massart (2003), and Lugosi (2009).
5. Random Graph Theory: Other problem specific methods in random graph theory are contained in Janson, Luczak, and Rucinski (2000).

**Basic Probability Inequalities**

1. Jensen’s Inequality for Convex Functions:
2. Maximized Function Space Inequality:

since the former automatically filters out and retains only the measure space maxima.

1. Markov Inequality: Applies to the situation where are strictly non-decreasing. Starting from the “bad metric” bound, and taking expectations on both sides, we get

or

This upper bounds the probability of a “bad outcome”.

1. Chebyshev Inequality: Here we set

Then

This is the Chebyshev’s inequality.

1. Generalized Moment Bounds: Set

Using the treatment above, we get

Here

1. Chernoff Bound: Here

so

In practice, an additional coefficient that allows for optimization is inserted, so

for

1. Chernoff vs. Moment Bounds: While both Chernoff bounds and generalized moment bounds allow for customization and/or optimization using the parameter and the parameter respectively, the range/width of moment bounding possible often produces more tight bounds than exponential/Chernoff bounds.

**Cauchy-Schwartz Inequality**

1. Statement: The Cauchy-Schwartz inequality states that if random variables and have finite second moments –

and

then

* 1. Consider 2 sequences of the random variables of and - and . Now,

for obvious reasons. Expanding, we get

Since

and

and

the result follows.

1. Chebyshev-Cantelli Inequality: Let

Then

* 1. Proof => Re-casting

does not alter. For all

where denotes the indicator function. Thus, for

from the Cauchy-Schwartz inequality

that is

**Association Inequalities**

1. Motivation: These deal with inequalities dealing with non-increasing and non-decreasing sequences. Chebyshev inequality (Hall, Littlewood, and Polya (1952)) is a simple univariate inequality – Dubdashi and Ranjan (1998) show several instances where association properties may be used to derive concentration inequalities.
2. Chebyshev’s Association Inequality: Let and both be non-decreasing (or non-increasing) real-valued functions defined on the real line. If is a real-valued random variable, then

If is non-increasing and is non-decreasing (or vice-versa), then

* 1. Proof => Let the random variable be distributed as and be independent of it. If and are non-decreasing, then

so that

Expand this expectation to obtain the first inequality. The proof of the second inequality is similar.

1. Multi-variate Association Inequality: An important generalization of the Chebyshev’s association inequality is the following. A real-valued function defined on is said to be non-decreasing (non-increasing) in each variable while keeping all the other variables fixed at any value.
2. Harris’ Inequality: Let

be a non-decreasing functions. Let be independent real-valued random variables and define the random variable

taking values in . Then

Similarly, if is non-increasing and is non-decreasing, then

* 1. Proof Step #1 => As before, it suffices to prove the first inequality. We proceed by induction. For

the statement is just Chebyshev’s association inequality.

* 1. Proof Step #2 => Now suppose that the statement is true for

Then

because, given , both and are non-decreasing functions of the variable. Now it follows by independence that the functions

defined by

and

are non-decreasing functions, so by the induction hypothesis

**Moment, Gaussian, and Exponential Bounds**

1. Moment Bounds vs. Chernoff Bounds: The moment bounds for tail probabilities are always better Chernoff bounds for tail probabilities. More precisely, let be a non-negative random variable and set

The best moment bound for tail probability is

where the minimum is taken over all positive integers . The best Chernoff bound is

It can be shown that

1. First/Second Moments on Integer-valued Distributions: If is a non-negative integer-valued random variable then it is obvious that

Further, it can be shown that

1. Sub-Gaussian Bounds: We say that a random variable has a sub-Gaussian distribution if there exists a constant

such that for all

It can be shown that there exists a universal constant such that if is sub-Gaussian, then for every possible integer ,

1. Sub-Gaussian Bounds - Converse: Let be a random variable such that there exists a constant such that

(1.19)

for every positive integer . Then is a sub-Gaussian – more precisely, for every ,

(1.20)

1. Sub-Exponential Bounds: We say that a random variable has a sub-exponential distribution if there exists a constant such that for all ,

(1.21)

If is sub-exponential, then for every positive integer ,

(1.22)

1. Sub-Exponential Bounds - Converse: Let be a random variable such that there exists a constant

such that

for every positive integer . Then is sub exponential – more precisely, for any

**Bounding Sum of Independent Random Variables**

1. Motivation – Sum of Independent Random Variables: Chebyshev’s inequality and independence immediately imply

Writing

we get

This is referred to as the ***Weak Law of Large Numbers***.

1. Inadequacy of the Weak Law: Recall that, under some additional regularity conditions, the central limit theorem states that

from which, within a certain range of parameters, we expect something like

1. General Bounding Technique:
   1. Independence of individual variables => The first step is to exploit the independence of the random variables to extract a bound on the sum. Here the exponential individual bound is convenient, since that reduces to a joint products.
   2. Bounding individual Errors => Use Markov bounds in conjunction with moment/exponential bounds to bound the individual empirical errors.
   3. Upper Bounding Individual Variable => Find additional methods for upper-bounding each individual random variable. Transforming them using a monotone convex functions make this convenient (Hoeffding does this).
2. Literature: Hoeffding’s inequality appears in Hoeffding (1963). Ledoux and Talagrand (1992) contain the proof of the contraction principle. Proof of an equivalent inequality for binomial random variables was provided by Chernoff (1952) and Okamoto (1958).

**Non-Moment Based Bounding - Hoeffding Bound**

1. Principle of Hoeffding’s Bounding:
   1. Hoeffding’s inequality bounds the individual variable exponential probability in terms of another.
   2. This bound, along with “exponential Chernoff scaling”, results in a convex optimization problem that may be optimized for an optimal “free parameter”.
2. Hoeffding’s Lemma: Suppose is a real variable with zero mean such that

Then

Note that if

and

then

and

However, if

or

it is easy to see that the inequality follows.

1. Proof of Hoeffding’s Lemma:

* Exploit Convexity =>

Applying expectation to both sides

Using

the LHS becomes

where

* Substitution => Let

and define

and

for

Thus

* Invoke Taylor’s Expansion Inequality => By Taylor’s theorem, for every real there exists a between and such that

Note that

and

Recognize that

* Bring it together =>

1. Hoeffding’s Inequality Statement: Let be independent random variables such that

Set

Then

* 1. Proof Step #1: For

the Chernoff inequality and the independence of implies

* 1. Proof Step #2: To get the best possible upper bound, we find the minimum of the RHS as a function of . Defining

achieves its minimum at

Substituting this value of establishes the inequality.

1. Boundedness of Hoeffding:
   1. Recall that the Hoeffding inequality may only be applied for bounded (albeit negative), zero-mean random variables.
   2. The predictor ordinate boundedness of Hoeffding’s inequality enables is applicable to piece-wise spline state estimators (esp. in the context of real-valued functions).
2. Hoeffding Inequality vs. Central Limit Theorem Inequality: As can be seen above, Hoeffding inequality has the same form as that of the Central Limit Theorem above, except for the fact that the average variance is replaced by

Thus, Hoeffding’s inequality ignores the information on the variance of , thus an alternate approach that accommodates this becomes necessary.

* 1. Inequality of Absolute Value Deviation => For ERM, we are also interested in absolute value deviations. The absolute value deviation around mean is looser (and symmetric)
  2. In general, optimized moment bounds and its variants (including optimized factorial moment bounds for integer valued random variables and other types we see later) offer tighter, asymmetric bounds in a number of cases.

**Moment Based Bounds**

1. Moment-Bound Series: Without loss of generality, assume that for all

We look for the bounds of . Setting

and

and using

and

we have

1. Bounding Individual : Assume that each is bounded such that

Then for each

Thus

Thus, we have obtained

1. Bounding Moments: The above expression for may be bounded differently, in other ways. Possible options are:
   1. Applying the moment bounds considered earlier along with potential sub-Gaussian-ness.
   2. Applying the moment bounds considered earlier along with potential sub-exponential-ness.
   3. Applying extraneous “noise variance” criterion such as the Massart’s or the Tsybakov’s noise conditions (to be considered later).
2. Sum of Independent Variables: Using the Chernoff bound for bounding the sums of independent random variables we get

Optimizing the choice of , we find that the upper bound is minimized by

1. Bennett’s Inequality (Bennett (1962)): Let be independent real-valued random variables with zero mean, and assume

with probability one. Let

Then, for any

where

* 1. Bound for Real-Valued Regression => The bound

is quiet general – it may be applied for specific empirical situations in a customized manner (e.g., real-valued regression in a piece-wise constant sense where we expect the basis spline functions to be bounded).

1. Bernstein Inequality Motivation: The meaning behind Bennett’s inequality is best seen if we do additional bounding. Applying the basis inequality

(which may be seen by comparing the derivatives of both sides), we obtain the classical (but more loose) inequality of Bernstein (Bernstein (1946)).

* 1. Statement => Under the same conditions as Bennett’s inequality, for any

1. Bernstein’s Inequality vs. Central Limit Theorem Gaussian Bounds: Except for the term in the denominator of the exponent, we can see that the Bernstein’s inequality is qualitatively right/tight when we compare it with the Central Limit Theorem.
2. Bernstein’s Inequality vs. Poisson Bounds: If

then the upper bound given by Bernstein’s inequality behaves as instead of the guaranteed by Central Limit Theorem/Hoeffding. This may be intuitively explained by recalling that a distribution can be approximated, for large , by a distribution whose tail decreases as .

**Binomial Tails**

1. Setup: We consider functions that are binary-valued, appropriate to the binary classifier setting. Thus, for a fixed function , the distribution of is just a binomial law of parameters and , since we are assuming i.i.d. random variables which can be either or , and are equal to with probability
2. Formulation: Denoting

we can have an exact expression for the deviations of from as

Since this is not easy to manipulate, the treatment seen earlier has used an upper bound as provided by Hoeffding’s inequality. However there exist other sharper bounds.

1. Alternate Upper Bounds: The following quantities are an upper bound on :
   1. Hoeffding:
   2. Bernstein:
   3. Bennett:
   4. Exponential:
2. Extremes of Binomial Distribution: From the above bounds, using inversion, we can say that, roughly speaking, small deviations of have a Gaussian behavior of the form (i.e., Gaussian with variance ) while large deviations have a Poisson behavior of the form .
3. Reduction to Hoeffding’s Inequality: Thus, typical binomial tails are heavier then Gaussian. Hoeffding’s inequality lies in upper bounding the tails with a Gaussian of maximum variance (i.e., ), hence the term .
4. Empirical Bernstein Bound: Each function has a different variance, i.e.,

For each , using Bernstein’s inequality, with a probability of at least ,

1. Intuition Behind Common Concentration of Measure Inequalities: The Gaussian part dominates when is not too small, or if is large enough, and it depends on . Thus, several concentration of measure inequality formulations strive to combine Bernstein’s inequality with the union bound and symmetrization.

**Custom Bounds for Special i.i.d. Sequences**

1. Bounded Random Variables: Let be independent random variables taking values in . Denoting

for any

Chernoff bounding can be used to show that

* 1. Corollary => From the above, it can be seen that

and for all

where is the function defined in Bennett’s inequality. Finally,

(Karp (1988), Hagerup and Rub (1990)).

1. Poisson Random Variable: Comparing

with the best Chernoff bound for the tail of a Poisson random variable

reveals that

1. Sampling with Replacement: Let be a finite set with elements, and let be a random sample without replacement from and a random sample with replacement from . For any convex real-valued function , it may be shown that

In particular, by taking

we see that all inequalities derived from sums of independent random variables using Chernoff bounding remain true for the sum of ’s (See Hoeffding (1963)).

**References**

* Ahlswede, R., P. Gacs, and J. Korner (1976): Bounds on Conditional Probabilities with Applications in multi-user Communication *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **34** 157-177 (with corrections in **39** 353-354 (1977)).
* Bennett, G. (1962): Probability Inequalities for the Sum of Independent Random Variables *Journal of the American Statistical Association* **57** 33-45.
* Bernstein, S. N. (1946): *The Theory of Probabilities* **Gastehizdat Publishing House** Moscow.
* Bobkov, S., and M. Ledoux (1997): Poincare’s Inequalities and Talagrand’s Concentration Phenomenon for the Exponential Distribution *Probability Theory and Related Fields* **107** 383-400.
* Boucheron, S., G. Lugosi, and P. Massart (2000): A Sharp Concentration Inequality with Applications *Random Structures and Algorithms* **16** 277-292.
* Boucheron, S., G. Lugosi, and P. Massart (2003): Concentration Inequalities using the Entropy Method *Annals of Probability* **31** 1583-1614.
* Bousquet, O. (2002): A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes *C. R. Acad. Sci. Paris* **334** 495-500.
* Chernoff, H. (1952): A Measure of Asymptotic Efficiency of Tests of a Hypothesis based on a Sum of Observations *Annals of Mathematical Sciences* **23** 493-507.
* Dembo, A. (1997): Information Inequalities and Concentration of Measure *Annals of Probability* **25** 927-939.
* Dubdashi, D., and D. Ranjan (1998): Balls and Bins: A Study in Negative Dependence *Random Structures and Algorithms* 99-124.
* Hagerup, T., and C. Rub (1990): A Guided Tour of Chernoff Bounds *Information Processing Letters* **33** 305-308.
* Hall, G. H., J. E. Littlewood, and G. Polya (1952): *Inequalities* **Cambridge University Press** London.
* Hoeffding, W. (1963): Probability Inequalities for Sums of Bounded Random Variables *Journal of the American Statistical Association* **58** 13-30.
* Janson, S., T. Luczak, and A. Rucinski (2000): *Random Graphs* **John Wiley** New York.
* Karp, R. M. (1988): *Probabilistic Analysis of Algorithms* **University of California, Berkeley**.
* Ledoux, M., and M. Talagrand (1991): *Probability in Banach Space* **Springer-Verlag** New York.
* Ledoux, M. (1996): On Talagrand’s Deviation Inequalities for Product Measures *ESAIM: Probability and Statistics* **1** 63-87.
* Ledoux, M. (1997): Isoperimetry and Gaussian Analysis *Lectures on Probability Theory and Statistics (editor: P. Bernard)* **Ecole d’Ete de Probabilites de St-Flour XXIV-1994** 165-294.
* Luczak, M. J., and C. McDiarmid (2003): Concentration for Locally Acting Permutations *Discrete Mathematics* **265** 159-171.
* Lugosi, G. (2009): [*Concentration of Measure Inequalities*](http://www.econ.upf.edu/~lugosi/anu.pdf).
* Marton, K. (1986): A Simple Proof of the Blowing-up Lemma *IEEE Transactions on Information Theory* **32** 445-446.
* Marton, K. (1996a): Bounding -distance by Informational Divergence: A Way to prove Measure Concentration *Annals of Probability* **24** 857-866.
* Marton, K. (1996b): A Measure Concentration Inequality for contracting Markov Chains *Geometric and Functional Analysis* **6** 556-571 (Erratum: **7** 609-613 (1997)).
* Massart, P. (1998): Optimal Constants for Hoeffding Type Inequalities *Technical Report* ***98.86****, Mathematiques* **Universite de Paris-Sud**.
* Massart, P. (2000): About the Constants in the Talagrand’s Concentration Inequalities for Empirical Processes *Annals of Probability* **28** 863-884.
* McDiarmid, C. (1989): On the Method of Bounded Differences, in: *Surveys in Combinatorics 1989* 148-188 **Cambridge University Press** Cambridge.
* McDiarmid, C. (1998): Concentration, in: *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed) 195-248 **Springer** New York.
* McDiarmid, C. (2002): Concentration for Independent Permutations *Combinatorics, Probability, and Computing* **2** 163-178.
* Milman, V., and G. Schechtman (1986): *Asymptotic Theory of Finite-dimensional Normed Spaces* **Springer-Verlag** New York.
* Okamoto, M. (1958): Some Inequalities relating to the partial Sum of Binomial Probabilities *Annals of the Institute of Statistical Mathematics* **10** 29-35.
* Panchenko, D. (2001): A Note on Talagrand’s Concentration Inequality *Electronic Communications in Probability* **6**.
* Panchenko, D. (2002): Some Extensions of an Inequality of Vapnik and Chervonenkis *Electronic Communications in Probability* **7**.
* Panchenko, D. (2003): Symmetrization Approach to Concentration Inequalities for Empirical Processes *Annals of Probability* **31 (4)** 2068-2081.
* Rio, E. (2001): Inegalities de concentration pour les processus empiriques de classes de parties *Probability Theory and Related Fields* **119** 163-175.
* Talagrand, M. (1995): Concentration of Measures and Isoperimetric Inequalities in Product Spaces *Publications Mathematiques de l’I.H.E.S.* **81** 73-205.
* Talagrand, M. (1996a): A New Look at Independence *Annals of Probability* **24** 1-34 (Special Invited Paper).
* Talagrand, M. (1996b): New Concentration Inequalities in Product Spaces *Inventiones Mathematicae* **126** 505-563.

**Efron-Stein Bounds**

**Introduction**

1. Motivation: The main purpose of this section is to demonstrate how the many tail inequalities for the sums of independent variables seen earlier can be extended to general functions of independent random variables. The simplest, yet the most powerful inequality of this kind is called the *Efron-Stein* inequality, which attempts to bound the variance of a general function. To obtain tail inequalities, one may simply use the variance to extend the bound using the Chebyshev’s inequality.
2. Setup: Let be some set, and let

be a measurable function of variables. We derive inequalities for the random variable

and its expected value where are arbitrary independent (but not necessarily identically distributed) random variables in .

1. Notation: The main Efron-Stein inequality follows from the next result – the *Martingale Differences Sum Inequality*. To simplify the notation, we use for the expected value with respect to the variable , that is,

**Martingale Differences Sum Inequality**

1. Statement:

i.e., is less than or equal to the sum of the conditional variances of along each random variable dimension.

1. Proof:
   1. The proof is based on the elementary properties of conditional expectations. Recall that if and are arbitrary bounded random variables, then
   2. Introduce

where the expectation is across all , and define

Under this notation, note that

and

where refers to a NULL variable set.

* 1. Clearly

since the individual telescope out. Further, note that and are martingales in and , respectively. Thus, is written as a sum of martingale differences.

* 1. Now

since, for any

as is a martingale.

* 1. To bound

note that by independence of

and therefore

the penultimate step resulting from Jensen’s inequality. Taking expected values on both sides, we obtain the statement inequality.

**Efron-Stein Inequality**

1. Statement: The Efron-Stein inequality (Efron and Stein (1981), Steele (1986)) follows from the above. To state the theorem, let form an independent copy of , and set

Then

1. Proof of the Inequality: The statement follows from the Martingale Differences Sum Inequality by using the elementary fact that if and are independent and identically distributed random variables, then

and therefore

1. Relation to the Case of Sum of Independent Random Variables: Observe that in the case when

is a sum of independent random variables of finite variance, the Efron-Stein inequality becomes an equality. Thus, in this sense, the bound produced by the Efron-Stein inequality cannot be improved. This also shows that, among all the functions of independent random variables, sums are, in some sense, the least concentrated.

1. Extended Efron-Stein Inequality: The Efron-Stein theorem may be extended to arbitrary measurable functions. A very useful corollary is obtained by recalling that, for any random variable ,

for any constant . Using this fact, we have for every

where

for arbitrary measurable functions

of variables. Taking expectations, we get an extension of the Efron-Stein inequality:

**Bounded Differences Inequality**

1. Functions with Bounded Differences: We say that function

has the *Bounded Differences Property* if for some non-negative constants

In other words, if we change the variable of keeping all the others fixed, the value of the function cannot change by more than .

1. Bounded Differences Inequality – Statement: If has the bounded differences property with constants , then

This bound comes in particularly handy when the direct estimation of the variance becomes involved, but the bound is obtained effortlessly.

* 1. Bounded Differences with Probability Bounds => If the random variables are independent and binary -valued with

and that has the bounded differences property with constants then it is easy to show that

1. Empirical Estimation of Multivariate Martingale Differences Inequality: For a sequence of size with multivariates, the empirical estimation goes as . Thus, even for a modest sample of size

martingale differences empirical estimates can be , i.e., the computational demands blow up very, very fast.

1. Empirical Estimation of Multivariate Symmetrized Differences: Both conceptually as well as empirically/implementation-wise, symmetrized differences inequality (a.k.a. Efron-Stein-Steele Multivariate Variance Upper Bound Inequality) is easier to handle, as drawing a parallel i.i.d. ghost variable set is easier than computing conditional multivariate sequence metrics (e.g., expectation/variance). Of course, the multivariate bounded differences inequality is the easiest to work with, but has the loosest bound.

**Bounded Differences Inequality - Applications**

1. Application #1 - Bin Packing:

* One of the basic problems in operations research is as follows: Given numbers

the question is the following: what is the minimal number of bins into which these numbers can be packed such that the sum of the numbers in each bind doesn’t exceed one?

* Let be this number. The behavior of

when are independent has been extensively studied (Rhee and Talagrand (1987), Rhee (1993), Talagrand (1995)). By changing one of the ’s, the value of cannot change by more than one, so we have

However, sharper bounds may be established using Talagrand’s convex distance inequality discussed later.

1. Longest Common Sub-sequence: This problem has been treated in depth in Chvatal and Sankoff (1975), Deken (1979), Steele (1982), Dancik and Peterson (1994), and Steele (1996). The simplest version is the following: Let and be 2 sequences of coin flips. Define as the length of longest common sub-sequence that appears in both sequences, i.e.,
   1. Behavior of the Expectation => The behavior of has been investigated in many papers. It is known that converges to some number whose value is unknown, but conjectured to be , and is known to fall between and .
   2. Variance of => Changing one bit cannot change the length of the longest common sub-sequence by more than one, so satisfies the bounded differences inequality with

Thus

(Steele (1986)). Thus, by Chebyshev’s inequality, with large probability, is within a constant times of its expected value. In other words, it is strongly concentrated around the mean, which means that the results on really tell us about the behavior of the longest common sub-sequence of two random strings.

1. Uniform Deviations of Indicator Functions: Let be i.i.d. random variables taking their values in some set , and let be a collection of sub-sets of . Let be a distribution of , i.e.,

and let denote the empirical distribution

The quantity of interest is

* 1. Variance Bound of => If

for every distribution of , then is called a *Uniform Glivenko-Cantelli Class*, and Vapnik and Chervonenkis (1971) provide a combinatorial characterization of such classes. But regardless of , by changing a single , can change by at most , so regardless of the behavior of , we always have

(in other words, need not be uniform). More information on the behavior of and its role in learning theory may be found in Devroye, Gyorfi, and Lugosi (1996), van der Waart and Wellner (1996), Vapnik (1998), and Dudley (1999).

* 1. Efron-Stein Bound Motivation => Next we show how a closer look at the Efron-Stein inequality implies a significantly better bound for the variance of . We do this in a slightly more general framework of the empirical process.
  2. Efron-Stein Bound Formulation => Let be a class of real-valued functions (no boundedness is assumed), and define

Observe that by symmetry, the Efron-Stein bound may be written as

Let

denote the function that achieves the supremum in the definition of , that is,

Then, clearly,

and therefore

* 1. Efron-Stein Bound vs. Simple Bounded Differences => For functions

taking values in , from just the bounded differences property we derived

The new bound may be a significant improvement whenever the maximum of of the functions in is small. More importantly, in deriving the new bounds we have not assumed any boundedness of the functions in .

* 1. Exponential Tail Inequality => The exponential tail inequality due to Talagrand (1996a) extends this variance inequality, and is one of the most important recent results in the theory of empirical processes, see also Ledoux (1997), Massart (2000a), Rio (2001), and Bousquet (2002).

1. First Passage Time in Oriented Percolation: Consider a directed graph such that the weight is associated with the edge such that the are non-negative independent random variables with a second moment

Let and be fixed vertices of the graph. We are interests in the total weight of the path from to with minimum weight. Set

where the minimum is taken over all the paths from to .

* 1. Denote the optimal path by . By replacing with , we can see that the total minimum weight can only increase if the edge is on , and therefore

Thus,

where is the length of the longest path between and .

1. Minimum of the Empirical Loss: Concentration inequalities have been used as a key tool in recent developments of model selection methods in statistical learning theory. For the background, we refer to the recent work of Massart (2000b), Koltchinskii and Panchenko (2002), Bartlett, Boucheron, and Lugosi (2002), Bousquet (2003), and Lugosi and Wegkamp (2004).
   1. Efron-Stein Formulation => Let denote a class of -valued functions on some space . For simplicity of exposition we assume that is finite, although the results below remain true as long as the measurability issues are taken care of. Given an i.i.d. sample

of pairs of random variables taking values in , for each we define the empirical loss

where the loss functional is defined on by

* 1. Efron-Stein Bound – Steps => In non-parametric classification and learning theory, it is common to select an element

by minimizing the empirical loss. The quantity of interest in this section is the minimal empirical loss

The bounded Efron-Stein bound implies that

However, a more careful application of the Efron-Stein inequality reveals that may be much more concentrated than predicted by this simple inequality. Getting tight results for the fluctuations of provides better insight into the calibration of penalties in certain model selection methods.

* 1. Tighter Bounding of the Variance => Let

and let be defined as in the Efron-Stein inequality, that is,

where is independent of and has the same distribution as . The convenient form of the Efron-Stein inequality to use is the following:

* 1. Tighter Efron-Stein Bound => Let denote a possibly non-unique minimizer of the empirical risk so that

The key observation is that

Thus,

where denotes expectation with respect to the variables and , and for each

is the true expected of . Therefore the Efron-Stein inequality implies that

* 1. Improvement over the Naïve Efron-Stein Bound => The result above is a significant improvement over the bound whenever is much smaller than . This is very often the case. For example, we have

so that we obtain

* 1. Bounding => In most cases of interest, may be bounded by a constant term depending on times (Lugosi (2002)), and the second term on the RHS is of the order of . Boucheron, Lugosi, and Massart (2003) detail exponential concentration inequalities for .

1. Kernel Density Estimation: Let be i.i.d. samples drawn according to some unknown density function on the real line. The density is estimated by the kernel estimate

where

is a smoothing parameter, and is a non-negative function with

The performance of the estimate is measured by the error

* 1. Bounding the Kernel Density Estimate => It is easy to see that

so without further work, we get

* 1. Bounds – Bounding the Error => It is known that for every ,

(Devroye and Gyorfi (1985)) which implies, by Chebyshev’s inequality, that for every

as

That is,

in probability, or in other words, is *relatively stable*. This means that the random -error behaves like its expected value (Devroye (1988, 1991)). For more on the behavior of the -error of the kernel density estimate, we refer to Devroye and Gyorfi (1985) and Devroye and Lugosi (2000).

**Self-Bounding Functions**

1. Introduction: Another simple property that is satisfied in many important functions is the *Self-Bounding Property*. We say that a non-negative function

has the self-bounding property if there exists

such that for all

and for all

and also

* 1. Concentration Properties => Concentration properties for such functions have been studied by Boucheron, Lugosi, and Massart (2000), Rio (2001), Bousquet (2002). For self-bounding functions we clearly have
  2. Variance Bound for Self-Bounding Functions => Using the extended Efron-Stein inequality

and setting

and

on applying the previous observation, and finally, using the squared form above, we get

It turns out that in many cases, the obtained bound form

is a significant improvement over what may be obtained by the bounded differences inequality.

1. Relative Stability: Bounding the variance of in many cases by its expected value implies the relative stability of . A sequence of non-negative random variables is said to be relatively stable if

in probability. This property guarantees that the random fluctuations of around its expectation are of negligible size when compared to the expectation, and therefore the most important information about the size of is given by . As seen above, if has the self-bounding property, then, by Chebyshev’s inequality, for all

Thus, for relative stability, it suffices to have

1. Empirical Processes: A typical example of self-bounding functions is the supremum of non-negative empirical processes. Let be a class of functions taking values in , and consider

A special case of the above is mentioned in the example of uniform deviations.

* 1. Formulation => Define

with

so that

Let

be a function for which

one obviously has

and therefore

* 1. Variance Bounding => Here we have assumed that the supremum is always achieved. The modification of the argument for the general case is straightforward. Thus by the variance bound, we get

Note that the bounded differences inequality implies

In some applications, may be significantly smaller than , and the improvement is essential.

1. Rademacher Averages: A less trivial example for the self-bounding functions is the Rademacher complexity. Let be a class of functions taking values in . If denote independent symmetric -valued random variables independent of the ’s (the so-called Rademacher variables), we define the *Conditional Rademacher Average* as

Thus, the expected values are taken with respect to the Rademacher variables , and is a function therefore of ’s - this makes data dependent.

* 1. Learning Class Complexity => Quantities like have been known to measure effectively the complexity of model classes in statistical learning theory (Koltchinskii (2001), Bartlett, Boucheron, and Lugosi (2002), Bartlett and Mendelson (2002), and Bartlett, Bousquet, and Mendelson (2002)).
  2. Variance Bounding => It is immediately apparent that has the bounded differences property, and therefore the bounded differences inequality implies that

However, this bound may be improved by observing that also has the self-bounding property, and therefore

Indeed, defining

it is easy to see that

and

* 1. Improving the Variance Bound => The above improvement is essential since it is well-known in the empirical process theory and the statistical learning theory that in many cases (where is a relatively small class of functions), may be bounded by something like where the constant depends on the class (van der Waart and Wellner (1996), Vapnik (1998), and Dudley (1999)).

**Configuration Functions**

1. Introduction: An important class of functions satisfying the self-bounding property consists of the so-called *Configuration Functions* defined in Talagrand (1995). The definitions presented here are a slight modification of Talagrand’s (Boucheron, Lugosi, and Massart (2000)).
   1. Definition and Setup => Assume that we have a property defined over a union of finite products of a set , that is, a sequence of sets

We say that

satisfies the property if

We assume that is hereditary in the sense that if satisfies , so does any sub-sequence of . The function that maps any tuple to the size of the largest sub-sequence satisfying is the *Configuration Function* associated with .

* 1. Bound => Let be a configuration function, and let

where are independent random variables. Then

* 1. Bound Derivation => In order to apply the self-bounding function bound, it suffices to show that any configuration function is self-bounding. Let

The condition

is trivially satisfied. On the other hand, assume that

and let

be a sub-sequence of cardinality such that

Clearly if the index is such that

Then

and therefore

is also satisfied, which concludes the proof.

1. Number of Distinct Values in a Discrete Sample: Let be independent, identically distributed random variables taking values on a set of positive integers such that

and let denote the number of distinct values taken by these random variables. Then we may write

and thus the expected value of is computed easily as

* 1. Basic Bounds => It is easy to see that

as

In fact the variance of may also be computed explicitly. Clearly satisfies the bounded differences property with

so the bounded differences inequality implies

so, by Chebyshev’s inequality

in probability.

* 1. Configuration Function Bounds => On the other hand, it is obvious that is a configuration function associated with the property of “distinctness”, and by the Bounding of the Configuration Function we have

which is a significant improvement since

as

1. VC Dimension: One of the central quantities in statistical learning theory is the *Vapnik-Chervonenkis Dimension* (Vapnik and Chervonenkis (1971, 1974), Blumer, Ehrenfeucht, Haussler, and Warmuth (1989), Devroye, Gyorfi, and Lugosi (1996), Vapnik (1998), Anthony and Bartlett (1999)).
   1. Setup => Let be an arbitrary collection of subsets of , and let

be a vector points of . Define the *Trace* of by

The *Shatter Coefficient* (or *Vapnik-Chervonenkis Growth Function*) of in is

the size of the trace. is the number of different sub-sets of the -point set generated by intersecting it with the elements of .

* 1. Formal Definition => A subset of is said to be *shattered* if

The VC Dimension of (with respect to ) is the cardinality of the largest subset of .

* 1. Configuration Function Bounding => From the definition, it is obvious that

is a configuration function associated with the property of “shatteredness”, and therefore if are independent random variables, then

1. Increasing Sub-sequences: Consider a vector

of different numbers in . The positive integers

form an *increasing sub-sequence* if

where

and

* 1. Configuration Function Bound => Let denote the length of the longest increasing sub-sequence.

is clearly a configuration function associated with the “increasing sequence” property, and therefore if are independent random variables such that they are all different with probability one (it suffices if every has an absolutely continuous distribution), then

* 1. Estimate => If the ’s are uniformly distributed in , it is known that

(Logan and Shepp (1977), Groeneboom (2002)). A tighter, but more difficult result, obtained implies that

(Baik, Deift, and Johansson (2000)). Frieze (1991), Bollobas and Brightwell (1992), and Talagrand (1995) contain reviews on some of the early results on the concentration of .

**References**

* Anthony, M., and P. L. Bartlett (1999): *Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge.
* Baik, L., P. Deift, and K. Johansson (2000): On the Distribution of the Length of the Second Row of a Young Diagram under Plancherel Measure *Geometric and Functional Analysis* **10** 702-731.
* Bartlett, P., S. Boucheron, and G. Lugosi (2002): Model Selection and Error Estimation *Machine Learning* **48** 85-113.
* Bartlett, P., O. Bousquet, and S. Mendelson (2002): Localized Rademacher Complexities *Proceedings of the 15th Annual Conference on Machine Learning* 44-48.
* Bartlett, P., and S. Mendelson (2002): Rademacher and Gaussian Complexities: Risk Bounds and Structural Results *Journal of Machine Learning Research* **3** 463-482.
* Blumer, A., A. Ehrenfeucht, D. Haussler, and M. K. Warmuth (1989): Learnability and the Vapnik-Chervonenkis Dimension *Journal of the ACM* **36** 929-965.
* Bollobas, B., and G. Brightwell (1992): The Height of Random Partial Order: Concentration of Measure *Annals of Applied Probability* **2** 1009-1018.
* Boucheron, S., G. Lugosi, and P. Massart (2000): A Sharp Concentration Inequality with Applications *Random Structures and Algorithms* **16** 277-292.
* Boucheron, S., G. Lugosi, and P. Massart (2003): Concentration Inequalities using the Entropy Method *Annals of Probability* **31** 1583-1614.
* Bousquet, O. (2002): A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes *C. R. Acad. Sci. Paris* **334** 495-500.
* Bousquet, O. (2003): New Approaches to Statistical Learning Theory *Annals of the Institute of Statistical Mathematics*.
* Chvatal, V., and D. Sankoff (1975): Longest common sub-sequences of two random sequences *Journal of applied Probability* **12** 306-315.
* Dancik, V. and M. Paterson (1994): Upper Bound for the Expected, in: *Proceedings of STACS ’94: Lecture Notes in Computer Science* **775** 669-678 **Springer** New York.
* Deken, J. P. (1979): Some Limit Results for Longest Common Sub-sequences *Discrete Mathematics* **26** 17-31.
* Devroye, L., and L. Gyorfi (1985): *Non-parametric Density Estimation: The -View* **John Wiley** New York.
* Devroye, L. (1988): The Kernel Estimate is Relatively Stable *Probability Theory and Related Fields* **77** 521-536.
* Devroye, L. (1991): Exponential Inequalities in Non-parametric Estimation, in: *Non-parametric Functional Estimation and Related Topics* (editor: G. Roussas) 31-44 NATO ASI Series **Kluwer Academic Publishers** Dordrecht.
* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer-Verlag** New York.
* Devroye, L., and G. Lugosi (2000): *Combinatorial Methods in Density Estimation* **Springer-Verlag** New York.
* Dudley, R. M. (1999): *Uniform Central Limit Theorems* **Cambridge University Press** Cambridge.
* Efron, B., and C. Stein (1981): The Jack-knife Estimate of variance *Annals of Statistics* **9** 586-596.
* Frieze, A. M. (1991): On the Length of the Longest Monotone Sub-sequence in a Random Permutation *Annals of Applied Probability* **1** 301-305.
* Groeneboom, P. (2002): Hydro-dynamical Methods for Analyzing Longest Increasing Sub-sequences; Probabilistic Methods in Combinatorics and Combinatorial Optimization *Journal of Computational and Applied Mathematics* **142** 83-105.
* Koltchinskii, V. (2001): Rademacher Penalties and Structural Risk Minimization *IEEE Transactions on Information Theory* **47** 1902-1914.
* Koltchinskii, V., and D. Panchenko (2002): Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers *Annals of Statistics* **30**.
* Ledoux, M. (1997): On Talagrand’s Deviation Inequalities for Product Measures *ESAIM: Probability and Statistics* **1** 63-87.
* Logan, B. F., and L. A. Shepp (1977): A Variational Problem for the Young Tableaux **26** 206-222.
* Lugosi, G. (2002): Pattern Classification and Learning Theory, in: *Principles of Non-Parametric Learning* (editor: L. Gyorfi) **Springer** Wien.
* Lugosi, G., and M. Wegkamp (2004): Complexity Regularization via Localized Random Penalties *Annals of Statistics* **32** 1679-1697.
* Massart, P. (2000a): About the Constants in Talagrand’s Concentration Inequalities for Empirical Processes *Annals of Probability* **28** 863-884.
* Massart, P. (2000b): Some Applications of Concentration Inequalities to Statistics *Annales de la Faculte des Sciences de Toulouse* **IX** 245-303.
* Rhee, W., and M. Talagrand (1987): Martingales, Inequalities, and NP-Complete Problems *Mathematics of Operations Research* **12** 177-181.
* Rhee, W. (1993): A Matching Problem and Sub-additive Euclidean Functionals *Annals of Applied Probability* **3** 794-801.
* Rio, E. (2001): Inegalities de concentration pour les processus empiriques de classes de parties *Probability Theory and Related Fields* **119** 163-175.
* Steele, J. M. (1982): Long Common Sub-sequences and the Proximity of Two Random Strings *SIAM Journal of Applied Mathematics* **42** 731-737.
* Steele, J. M. (1986): An Efron-Stein Inequality for Non-symmetric Statistics *Annals of Statistics* **14** 753-758.
* Steele, J. M. (1996): *Probability Theory and Combinatorial Optimization* **SIAM CBMS-NSF Regional Conference Series in Applied Mathematics** 69, 3600 University City Science Center, Philadelphia, PA 19104.
* Talagrand, M. (1995): Concentration of Measure and Isoperimetric Inequalities in Product Spaces *Publications Mathematiques de l’I.H.E.S* **81** 73-205.
* Talagrand, M. (1996a): New Concentration Inequalities in Product Spaces *Inventiones Mathematicae* **126** 505-563.
* Van der Waart, A. W., and J. A. Wellner (1996): *Weak Convergence and the Empirical Processes* **Springer-Verlag** New York.
* Vapnik, V. N., and A. Y. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and Applications* **16** 264-280.
* Vapnik, V. N., and A. Y. Chervonenkis (1974): *Theory of Pattern Recognition* (in Russian) **Nauka** Moscow. German Translation: *Theorie der Zeichenerkennung* **Academie Verlag** Berlin (1979).
* Vapnik, V. N. (1998): *Statistical Learning Theory* **John Wiley** New York.

**Entropy Methods**

**Introduction**

1. Motivation: The Efron-Stein based inequalities produce variance-based linear/polynomial tail bounds. However we can do better – so we seek exponential tail bounds, which leads us to the entropy methods.
2. Past Approaches: Originally, Martingale methods dominated the research (Rhee and Talagrand (1987), Shamir and Spencer (1987), McDiarmid (1989, 1998)), but independently information-theoretic methods were also used with success (Ahlswede, Gacs, and Korner (1977), Marton (1986, 1996a, 1996b), Massart (1998), Samson (2000), Rio (2001)).
3. Talagrand’s Induction Method: Talagrand’s induction method (Talagrand (1995, 1996a, 1996b)) caused an important breakthrough both in the theory and in the applications of exponential concentration inequalities.
4. The Entropy Method: Here we focus on the so-called “entropy method” based on logarithmic Sobolev inequalities developed by Ledoux (1996, 1997) – see also Bobkov and Ledoux (1997), Massart (2000a), Boucheron, Lugosi, and Massart (2000), Rio (2001), Bousquet (2002), and Boucheron, Lugosi, and Massart (2003). This method makes it possible to derive exponential analogues of the Efron-Stein inequality in possibly the simplest way.

**Information Theory - Basics**

1. Introduction: Here we summarize some of the basic properties of the entropy of a discrete-valued random variable. For a good introductory book on information theory, we refer to Cover and Thomas (1991).
2. Types of Entropy: Let be a random variable taking values in a countable set with distribution

The *entropy* of is defined by

where denotes the natural logarithm and

If is a pair of discrete random variables taking values in , the *Joint Entropy* of and is defined as the entropy of the pair . The *Conditional Entropy* is defined as

1. Conditional Entropy Inequality:

Thus

1. Chain Rule for Entropy: It is also easy to see that the definition of the conditional entropy remains true conditionally, i.e., for any 3 discrete random variables ,

This may be easily seen by adding to both sides and by using the definition of conditional entropy. A repeated application of this yields the *Chain Rule for Entropy*: for arbitrary random variables

1. Kullback-Leibler Divergence: Let and be two probability distributions over a countable set with a probability mass functions and . Then the *Kullback-Leibler Divergence* or the *Relative Entropy* of and is
2. The Relative Entropy Inequality: Since

thus

and equals zero if and only if

* 1. Consequence of the relative entropy inequality => If is a finite set with elements in it, and is a random variable with distribution , and we take to be the uniform distribution over , then

since

and therefore the entropy of never exceeds the logarithm of the cardinality of its range.

1. Conditioning the Relative Entropy: Consider a pair of random variables with joint distribution and marginal distributions and . Then

Observing that

* 1. Consequence of Conditioning the Relative Entropy => Thus

and the non-negativity of the relative implies that

that is, conditioning reduces entropy. It is similarly easy to see that this fact remains true for conditional entropies as well, that

1. Han’s Inequality (Han (1978)): Let be discrete random variables. Then
   1. Proof => For any

by the definition of conditional entropy and the fact that conditioning reduces entropy

Summing both sides, we get

Re-arranging we get

1. Han’s Inequality for Relative Entropy:
   1. Setup => Here we follow the layout of Massart (2000b). Let be a countable set, and let and be probability distributions on such that

is a product measure. We denote the elements of by

and denote

for the vector obtained by leaving out the component of .

* 1. Marginal Distributions of and => Denote by and the marginal distributions of according to and , that is

and

Just as we used to represent a uniform measure earlier, here we use to represent an *independence* product measure, and will be used later.

* 1. Statement =>

or, equivalently

* 1. Proof:
     1. Re-casting Han’s Inequality => Han’s inequality may be directly re-cast as stating
     2. Re-casting the Relative Entropy => Since

and

it is sufficient to show that

This is easily seen from the product property of - we have

for all , as well as

therefore

* + 1. Re-arrangement => Re-arranging the above, we obtain

where we have used the defining property of .

**Tensorization of the Entropy**

1. Introduction: We now use the results above to prove our main exponential concentration inequality. As before, we let be independent random variables, and investigate the concentration properties of
2. Re-casting the Martingale Differences Inequality: We extend the martingale differences inequality in a powerful way. Noting that it may be re-cast as

we generalize the martingale differences inequality to a large class of convex functions (by exploiting Jensen’s inequality for convex functions) as

Using

recovers the original martingale differences inequality.

1. Applicability of the Extended Martingale Differences Inequality: It turns out that this inequality remains valid for a large class of convex functions (Beckner (1989), Ledoux (1997), Latala and Oleszkiewicz (2000), Chafai (2002)).
2. Usage for the Relative Entropy Function: The function of interest in our case is

- the relative entropy function. For this function, as can be seen below, the LHS of the inequality may be written as the relative entropy induced by on and the distribution of . Hence the name “Tensorization of the entropy inequality” (Ledoux (1997)).

1. Tensorization of Entropy Theorem - Statement: Let

for

Let be independent random variables taking values in and let be a positive-valued function on . Letting

we have

* 1. Proof Step #1 – Outline => We only prove the statement for discrete random variables . The extension to the general case is technical but straightforward. The inequality is a direct consequence of Han’s inequality for relative entropies. First, note that if the inequality is true for a random variable , then it is also true for where is a positive constant. Hence, without loss of generality, we assume that
  2. Proof Step #2 – The Tensorial Measure => We set the probability measure on to be

where denotes the distribution of

Then clearly

Using Han’s inequality for relative entropy, cannot exceed . However, straightforward calculation shows that

and the original statement of inequality follows.

1. Ledoux’s Entropy Method: The main idea behind Ledoux’s entropy method for proving concentration bounds is the application of the tensorization inequality to the positive random variable

Denoting the moment generating function of by

the LHS of the tensorization of entropy inequality becomes

* 1. Approach Behind the Ledoux’s Entropy Method => The strategy then is to derive upper bounds for , and then derive the tail bounds using Chernoff bounding. To do this in a convenient way, however, requires additional bounds for the RHS of the tensorization of the entropy inequality.

**Logarithmic Sobolev Inequalities**

1. Introduction: As before, we set

where is some function over . Here we further develop the RHS of the tensorization of the entropy inequality to obtain further inequalities that serve as basis for deriving the exponential concentration inequalities. These inequalities are closely related to the so-called *Logarithmic Sobolev Inequalities Of Analysis* (Ledoux (1997, 1999), Massart (2000a), Ledoux (2001)).

1. Relative Entropy Inequality Lemma: Let be a positive random variable. Then, for any
   1. Proof => Since for any

we have

hence

upon re-arranging which, you retrieve the original statement.

1. Logarithmic Sobolev Inequality: Denote

Then

* 1. Proof => We bound each term on the RHS of the tensorial entropy inequality. Note that the relative entropy inequality implies that if is a positive function of , then

Applying the above inequality to the variables

and

one gets

1. Symmetrized Logarithmic Sobolev Inequality (Massart (2000a)): If is defined as in the logarithmic Sobolev inequality, then

Moreover, denote

Then, for all

and

* 1. Proof => The first inequality is proved exactly as in the logarithmic Sobolev inequality, by nothing that, just like , is also independent of . To prove the 2nd and the 3rd inequalities, write

By symmetry, the conditional expectation of the 2nd term may be written as

Summarizing, we have

The 2nd inequality of the theorem follows simply by noting that

The last inequality follows similarly.

**Logarithmic Sobolev Inequalities - Applications**

1. Introduction: Here we show how the logarithmic Sobolev inequalities developed in the previous section maybe used to obtain powerful exponential concentration inequalities. The first result is fairly easy to obtain, yet it turns out to be very useful. Also, its proof is proto-typical in the sense that, it shows in a transparent way, the main ideas.
2. Difference Square Bound: Assume that there exists a positive constant such that, almost surely

Then, for all

* 1. Proof Step #1 – Application of the Symmetrized Logarithmic Sobolev Inequality => Observe that for

and therefore, for any

the symmetrized logarithmic inequality implies

where we have used the assumptions of the theorem.

* 1. Proof Step #2 – Bounding of the Expectation => Denoting the moment generating function of by

the above inequality may be re-written as

After diving both sides by , we observe that the LHS is just a derivative of

that is, we obtain the inequality

By l’Hospital’s rule, we note that

so by integrating the above inequality, we get

or, in other words

* 1. Proof Step #3 – Apply Markov’s Inequality and Chernoff Bounds => By Markov’s inequality

The choice of

makes the upper bound . Replace by to obtain the same upper bound for .

1. Two-sided Bounded Differences Inequality: It is clear from the proof above that under the condition

one has a two-sided inequality

An immediate corollary of this is that functions following the bounded-square differences inequality also satisfy sub-Gaussian tail probability.

1. McDiarmid Bounded Differences Inequality: Under the conditions of bounded square differences, McDiarmid (1989) showed that

This may be seen by writing as a sum of Martingale differences as in the Martingale Differences Theorem, using Chernoff bounding, and proceeding as in the proof for Hoeffding’s inequality, since that argument also works for sums of Martingale differences.

1. Bounded Square Differences Inequality - Statement: Assume that the function satisfies the bounded square differences assumption with constants , then

where

1. Usage of the Bounded Square Differences Inequality: We remark here that the constant appearing in the exponent above may be improved (e.g., the Martingale Differences Technique used in McDiarmid (1989) shows an approach). Thus, we have been able to extend the bounded differences (in addition to the bounded square differences condition) to an exponential concentration inequality. Note that by combining the variance bound of the bounded differences inequality with Chebyshev’s inequality, we only obtained

therefore the exponential improvement obtained here is substantial. All the sample applications that use the bounded differences inequality are now improved in an essential way without any further work.

1. Relaxing the Bounded Differences Square Bound: The differences square bound maybe relaxed in the following way to produce expected data-dependent tail bounds. If

then for all

(This can be seen by using the Association Inequality for the monotonic multi-variate functions and ). Further, if

we can see that

1. Example - Hoeffding’s Inequality in Hilbert Space: As a simple illustration of the power of bounded differences inequality, we derive a Hoeffding-type inequality for the sums of random variables taking values in a Hilbert space. In particular, we show that if are independent zero-mean random variables taking values in a separable Hilbert space such that

with probability one, then for all

where

* 1. Proof => The above follows simply by observing that, by triangle inequality

satisfies the bounded differences inequality with constants , and therefore

The proof is completed by observing that, by independence

Application of the Markov/Chebyshev inequality produces the final result.

1. Bounded Square Differences vs. Bounded Differences: Note that the bounded square differences criterion is much stronger than the bounded differences criterion. This is because the former does not require that be bounded – all that is required is that

an obviously milder requirement. Thus, as seen in the next example, in situations where the bounded differences technique may not work, the bounded square differences inequality may provide a sharp bound.

1. Example - Largest Eigen-value of a Random Symmetric Matrix: Using the bounded square differences inequality, we derive a result of Alon, Krivelevich, and Vu (2002). Let be a symmetric real matrix whose entries

are independent random variables with absolute value bounded by . We seek to show that if

is the largest eigenvalue of , then

The property of the largest eigenvalue we need is that if

is an eigenvector corresponding to the largest eigenvalue with

then

* 1. Using the Bounded Square Differences Theorem => To use the bounded square differences inequality, consider the symmetric matrix obtained by replacing in by the independent copy while keeping all other variables fixed. Let denote the largest eigenvalue of the obtained matrix. Then, using the above-mentioned property of the largest eigenvalue

Therefore

The result now follows from the bounded differences inequality.

* 1. Comparison with Efron-Stein => Note that by the Efron-Stein inequality, we also have

A similar exponential inequality, though with a somewhat worse constant in the exponent, can also be derived for the lower tail. In particular using a theorem from below, it can be shown that for

* 1. Alternate Eigenvalues => Notice that the same proof works for the smallest eigenvalue as well. Alon, Krivelevich, and Vu (2002) show that, with a simple extension of the argument, if is the largest eigenvalue (or the smallest eigenvalue), then the upper bound becomes , although it is not clear whether the factor in the exponent is necessary.

**Exponential Inequalities for Self-Bounding Functions**

1. Introduction: Recall that the bounded differences inequality implies that for self-bounding functions

Based on the logarithmic Sobolev inequality, we may now obtain exponential concentration inequality bounds (Boucheron, Lugosi, and Massart (2000), Massart (2000a)).

1. Bennett and Logarithmic Sobolev Functions: Recall the definition of the following two functions that we have already seen above in the Bennett’s inequality and in the logarithmic Sobolev inequalities:
2. Logarithmic Sobolev Inequality for Self-Bounding Functions: Assume that satisfies the self-bounding property. Then for every

Moreover for every

and for every

* 1. Corollary from the re-cast => By recalling that

for

(we have already used this in the proof for Bernstein’s inequality), and observing that

for

we obtain the following two corollaries:

* + 1. For every
    2. For every

1. Proof of Logarithmic Sobolev Inequality for Self-Bounding Functions:
   1. Step #1 – Applying the Logarithmic Sobolev Inequality => Since the function is convex with

for any and any

Thus, since

we have that, for every

Therefore the combination of the logarithmic inequality and the condition

implies that

* 1. Step #2 – Variable Transformation => Introduce

and define for any

The inequality above then becomes

which, on writing

implies

* 1. Step #3 – Inequality for => Observe that the function

is a solution to the ODE

Here we intend to show that

In fact, if

then

Thus, defining

we have

This clearly indicates that is non-positive, and therefore is non-increasing.

* 1. Step #4 – Proof that => Now, since is centered

Using the fact that

as

we conclude that

as

This shows that is non-positive over and non-negative over , hence is non-positive everywhere, therefore

This proves the first inequality.

* 1. Step #5 – Tail Probabilities Calculation => The proof of inequalities for the tail probabilities maybe completed by Chernoff Bounding:
  2. Step #6 – Reduced Form Bounds => The final step of the proof is done using the following easy-to-check and well-known relations:

for

for

**Combinatorial Entropy**

1. VC Entropy: The VC Entropy is closely related to the VC dimension discussed earlier. Let be an arbitrary collection of subsets of , and let

be a vector of points of . Recall that the *shatter coefficient* is defined as the size of the trace of on , that is

The VC entropy is defined as the log of the shatter coefficient, that is

1. VC Entropy Self-Bounding Property: The VC Entropy has the self-bounding property.
   1. Proof Step #1 – Setup => We need to show that there exists a function of variables such that for all

writing

the following true self-bounding conditions are valid:

and

* 1. Proof Step #2 – Outline => We define in the natural way, that is, as the entropy based on the points in its arguments. Then, clearly for any

and the difference cannot be more than one. The non-trivial part of the proof is to demonstrate the validity of the second condition. We do this using Han’s inequality for joint entropy.

* 1. Proof Step #3 – Invoking Uniform Outcome Distribution => Consider the uniform distribution over the outcome set . This defines a random vector

Then clearly

where is the joint entropy of . Since the uniform distribution maximizes the entropy (in other words, the uniform distribution in the outcome space – the uniform trace – maximizes the outcome entropy), we also have for all

* 1. Proof Step #4 – Invoking Han’s Inequality => Since by Han’s inequality for joint entropy

we have

as desired.

1. VC Entropy Probabilistic Bounds: Let be a set of independent random variables taking values in and let

denote the random VC entropy. Then the following are true:

For every

For every

* 1. Moreover, for the random shatter coefficient , we have

1. VC Entropy Probabilistic Bounds - Proof: The LHS of the VC Entropy Bounds statement follows from the Jensen’s inequality, while the RHS arises by taking

in the first inequality of the exponential inequality for self-bounding functions. This last statement shows that the expected VC entropy and the annealed VC Entropy are tightly connected regardless of the class of sets and the distribution of ’s.

1. Non-trivial ERM Consistency (Vapnik (1995)): The probabilistic VC entropy bounds above answers in a positive way an open question raised by Vapnik (1995): the EMR procedure is *non-trivially consistent* and *rapidly convergent* if and only if the annealed entropy rate converges to zero.
2. Bounded Supremum of the Self Bounding Exponential Inequality: Let and denote two positive real numbers and denote

It is easy to show that the local supremum is given from

and that the supremum is attained at

1. Unbounded Supremum of the Self-Bounding Exponential Inequality: Further

if

and the supremum is attained at

1. Generalization of the VC Entropy: The proof of concentration if the VC entropy may be generalized in a straightforward way to a class of functions called *combinatorial entropies* defined below.
2. Combinatorial Entropy - Setup: Let

be an -vector of elements

to which we associate a set

of -vectors whose component are elements of a possibly different set . We assume that for each

and

the set

is a projection of along the co-ordinate, that is

The associated combinatorial entropy is

where is an arbitrary positive number .

1. Self-Bounding Property of Combinatorial Entropy: Just as in the case of VC entropy, the combinatorial entropy may also be shown to possess the self-bounding property, i.e., assuming

is a combinatorial entropy such that for all

and

then has the self-bounding property.

1. Combinatorial Entropy - Exponential Bounds Generalization: Assume that

is a combinatorial entropy such that for all

and

If

is a vector of -independent random variables taking values in , then the random combinatorial entropy

satisfies

for

and

for

Moreover

1. Combinatorial Entropy Example - Increasing Sub-sequences: Recall the setup of the example of the increasing sub-sequences earlier, and let denote the number of different increasing sub-sequences of . Observe that is a combinatorial entropy. This is easy to see by considering

and by assigning, to each increasing sub-sequence

of a binary -vector

such that

if and only if

for some

(i.e., the indices appearing in the increasing sub-sequence are marked by ).

* 1. Exponential Bounds => Now that the conditions for combinatorial entropy are met

satisfies all the inequalities associated with the combinatorial entropy exponential inequality. This result improves a concentration inequality obtained by Frieze (1991) for .

**Variations on the Theme of Self-Bounding Functions**

1. Introduction: Here we show how the techniques for the entropy method for proving concentration inequalities may be used in various situations not considered so far. The versions differ in the assumptions on how is controlled by different functions of . For various other versions with applications, we refer to Boucheron, Lugosi, and Massart (2003).
2. Generalized Self-Bounding Functions - Probabilistic Bounds: In all cases the upper bound is roughly of the form where is the corresponding Efron-Stein upper bound of . The first inequality may now be regarded as a generalization of the exponential bounds for self-bounding functions.
3. Exponential Bounds for Generalized Self-Bounding Functions: Assume that there exist positive constants and such that

Then, for

and for all

* 1. Reduction to Expectations Bound => Let

Just as in the first step of the proof for the Bounded Square Differences Exponential Bounds, we use the fact that for

and therefore by the Symmetrized Logarithmic Inequality we have

where at the final step we have used the assumptions of our postulate.

* 1. Inequality ODE for => Denoting, again

the above inequality becomes

After scaling both sides by , we notice that the LHS is just the derivative of

so we obtain

* 1. Bounding the Expectation and Extracting Exponential Inequality => Using the fact that

and

and integrating the inequality, we obtain

or, if

thereby proving the first inequality. The inequality for the upper tail follows from the Markov inequality along with the bounded/unbounded suprema of the self-bounding exponential inequality.

1. Distinction Between the Upper and the Lower Bounds: Bounds for the lower tail maybe more easily derived by exploiting the association inequalities seen before, under much more general conditions on . Notice the difference between this and the quantity appearing in the upper bound above.
2. Generalized Self-Bounding Lower Tail: Assume that for some non-decreasing function

Then for all

* 1. Proof Step #1 - Applying the Symmetrized Logarithmic Sobolev Inequality => To prove the lower tail inequalities, we obtain the upper bounds for

with

By the third inequality of the symmetrized logarithmic Sobolev inequality

By using

and

for

the RHS above becomes

* 1. Prof Step #2 – Apply Chebyshev’s Association Inequality => Since is a non-decreasing function, and is a decreasing function of , Chebyshev’s association inequality implies that
  2. Proof Step #3 – Extraction of the Lower Tail => Scaling both sides of the above inequality by and writing

we obtain

Integrating in the interval , we obtain

Finally, applying Markov’s inequality and carrying out the optimization results in the statement of the inequality.

1. Data-Dependent Lower Tail: Assume that

where are independent real-valued random variables, and that is a non-decreasing function of each variable. Suppose there exists another non-decreasing function

such that

Then, for all

it can be shown, following above, that

1. Alternate Approach to the Lower Tail Bound: The next result is useful when one is interested in lower tail-bounds, but is difficult to handle. In certain situations the right symmetrization pivot, i.e., , is easier to bound. In such a situation, however, we need the additional guarantee that remains bounded. Without loss of generality, we assume that this bound is .
2. Lower Tail Bound Using the Right Symmetrization Pivot: Assume that there exists a non-decreasing function such that

and for any value of and

Then the following are true:

for all

and

for all

and

1. Proof Step #1 - Using the Right Symmetrization Pivot: The key observation is that the function

is increasing if

Choose

Thus, for

the right symmetrization pivot of the symmetrized logarithmic Sobolev inequality implies that

where at the last step, we have used the assumption of our postulate.

1. Proof Step #2 - Use of the Association Inequality: Just as in the proof for generalized self-bounding functions lower tail inequality, we bound by . For the rest of the proof, we proceed as in the proof for self-bounding functions lower tail inequality. Here we have taken

**References**

* Ahlswede, R., P. Gacs, and J. Korner (1976): Bounds on Conditional Probabilities with Applications in multi-user Communication *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **34** 157-177 (with corrections in **39** 353-354 (1977)).
* Alon, N., Krivelevich, M., and V. H. Vu (2002): On the Concentration of Eigen-values of Random Symmetric Matrices *Israeli Mathematical Journal* **131** 259-267.
* Beckner, W. (1989): A Generalized Poincare’s Inequality for Gaussian Measures *Proceedings of the American Mathematical Society* **105** 397-400.
* Bobkov, S., and M. Ledoux (1997): Poincare’s Inequalities and Talagrand’s Concentration Phenomenon for the Exponential Distribution *Probability Theory and Related Fields* **107** 383-400.
* Boucheron, S., G. Lugosi, and P. Massart (2000): A Sharp Concentration Inequality with Applications *Random Structures and Algorithms* **16** 277-292.
* Boucheron, S., G. Lugosi, and P. Massart (2003): Concentration Inequalities using the Entropy Method *Annals of Probability* **31** 1583-1614.
* Bousquet, O. (2002): A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes *C. R. Acad. Sci. Paris* **334** 495-500.
* Chafai, D. (2002): [*On -entropies and -Sobolev Inequalities*](http://arxiv.org/pdf/math/0211103v1.pdf) **arXiv**.
* Cover, T. M., and J. A. Thomas (1991): *Elements of Information Theory* **John Wiley** New York.
* Dembo, A. (1997): Information Inequalities and Concentration of Measure *Annals of Probability* **25** 927-939.
* Frieze, A. M. (1991): On the Length of the Longest Monotone Sub-sequence in a Random Permutation *Annals of Applied Probability* **1** 301-305.
* Han, T. S. (1978): Non-negative Entropy Measures of Multi-variate Symmetric Correlations *Information and Control* **36**.
* Latala, R., and C. Oleszkiewicz (2000): Between Sobolev and Poincare, in: *Geometric Aspects of Functional Analysis, Israeli Seminar (GAFA), 1996-2000* 147-168 **Springer** Lecture Notes in Mathematics **1745**.
* Ledoux, M. (1996): Isoperimetry and Gaussian Analysis *Lectures on Probability Theory and Statistics (editor: P. Bernard)* **Ecole d’Ete de Probabilites de St-Flour XXIV-1994** 165-294.
* Ledoux, M. (1997): On Talagrand’s Deviation Inequalities for Product Measures *ESAIM: Probability and Statistics* **1** 63-87.
* Ledoux, M. (1999): Concentration of Measure and Logarithmic Sobolev Inequalities, in: *Seminaire de Probabilites XXXIII, Lecture Notes in Mathematics* **1709** 120-216 **Springer**.
* Ledoux, M. (2001): *The Concentration of Measure Phenomenon* **American Mathematical Society** Providence, RI.
* Marton, K. (1986): A Simple Proof of the Blowing-up Lemma *IEEE Transactions on Information Theory* **32** 445-446.
* Marton, K. (1996a): Bounding -distance by Informational Divergence: A Way to prove Measure Concentration *Annals of Probability* **24** 857-866.
* Marton, K. (1996b): A Measure Concentration Inequality for contracting Markov Chains *Geometric and Functional Analysis* **6** 556-571 (Erratum: **7** 609-613 (1997)).
* Massart, P. (1998): Optimal Constants for Hoeffding Type Inequalities *Technical Report* ***98.86****, Mathematiques* **Universite de Paris-Sud**.
* Massart, P. (2000a): About the Constants in Talagrand’s Concentration Inequalities for Empirical Processes *Annals of Probability* **28** 863-884.
* Massart, P. (2000b): Some Applications of Concentration Inequalities to Statistics *Annales de la Faculte des Sciences de Toulouse* **IX** 245-303.
* McDiarmid, C. (1989): On the Method of Bounded Differences, in: *Surveys in Combinatorics 1989* 148-188 **Cambridge University Press** Cambridge.
* McDiarmid, C. (1998): Concentration, in: *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed) 195-248 **Springer** New York.
* Rhee, W., and M. Talagrand (1987): Martingales, Inequalities, and NP-Complete Problems *Mathematics of Operations Research* **12** 177-181.
* Rio, E. (2001): Inegalities de concentration pour les processus empiriques de classes de parties *Probability Theory and Related Fields* **119** 163-175.
* Samson, P. M. (2000): Concentration of Measure Inequalities for Markov Chains and -mixing Processes *Annals of Probability* **28** 416-461.
* Shamir, E., and J. Spencer (1987): Sharp Concentration of Chromatic Number on Random Graphs *Combinatorica* **7** 374-387.
* Talagrand, M. (1995): Concentration of Measure and Isoperimetric Inequalities in Product Spaces *Publications Mathematiques de l’I.H.E.S* **81** 73-205.
* Talagrand, M. (1996a): New Concentration Inequalities in Product Spaces *Inventiones Mathematicae* **126** 505-563.
* Talagrand, M. (1996b): A New Look at Independence *Annals of Probability* **24** 1-34 (Special Invited Paper).
* Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory* **Springer-Verlag** New York.

**Concentration of Measure**

**Introduction**

1. Overview: In this section we address the *isoperimetric* approach to concentration inequalities, developed in large parts by Talagrand (1995, 1996a, 1996b).
2. Equivalent Bounded Differences Inequality: First, we provide an equivalent formulation of the bounded differences inequality, which shows that any not-too-small set in a product probability space has the property that the probability of those points whose Hamming distance from the set is much larger than is exponentially small.
3. Convex Distance Inequality: Using the full power of the bounded square differences inequality, we provide a significant improvement on this concentration-of-measure result, also called Talagrand’ *Convex Distance Inequality*.

**Equivalent Bounded Differences Inequality**

1. Setup: Consider independent random variables taking their values in a measurable set , and denote a vector of these variables by

taking its values in .

1. Hamming Distance: Let

be an arbitrary measurable set, and write

The *Hamming Distance* between the vectors

is defined as the number of co-ordinates in which and differ. Introduce

as the Hamming distance between set and the point .

1. Hamming Distance Inequality: For any
   1. Proof Step #1 – Bounded Differences Property => Observe that the function

cannot change more than by altering one component of , that is, it follows the bounded differences property with constants

Thus, by the bounded square differences inequality, with the optimal constants provided by the McDiarmid’s inequality

* 1. Proof Step #2 - Inequality for => By taking

the LHS becomes

so the above inequality implies

By using the bounded differences inequality again, we obtain

as desired.

1. *t-Blowup* of the Set : Observe that on the RHS we have a measure of the complement of the *t-Blowup* of the set , that is, the measure of the set of points whose Hamming distance from is at least . If we consider a set, say with

we see something very surprising; the measure of the set of points whose Hamming distance to is more than is smaller than ! In other words, product measures are concentrated on extremely small *sets* – hence the name concentration of measure.

1. Recovering the Bounded Differences Inequality: Observe that the bounded differences inequality may also be derived from the above theorem. Indeed, if we consider a function on having the bounded differences property with constants

(for simplicity), then we may let

where denotes the median of the random variable

Then clearly

so the above inequality implies

Note that the Hamming distance in this case of

can be represented by .

1. Reduction to the Bounded Differences Form: This has the same form as the bounded differences inequality, except that the expected value of is replaced by its mean. This difference is usually negligible, since

so whenever the deviation of from its mean is small, its expected value must also be close to the mean.

* 1. Mean-Median Closeness Statement => Let be a random variable with median such that there exist positive constants and such that for all

Then

**Convex Distance Inequality**

1. Motivation: Talagrand (1995, 1996a, 1996b) developed an induction method to prove powerful concentration results in many cases where the bounded differences inequality fails. The most widely used among these is the so-called *convex distance inequality* – Steele (1996) and McDiarmid (1998) contain surveys with several interesting applications.
2. Bounded Square Differences Convex Differences Inequality: Here we use the bounded square differences inequality to derive a version of the convex distance inequality. Talagrand (1995, 1996a, 1996b) contains extensions and several variations.
3. Weighted Hamming Distance: The following simple argument first presented in McDiarmid (1998) helps understand the rationale behind Talagrand’s inequality. First, observe that the *Equivalent Bounded Square Differences Theorem* may be easily generalized by allowing the distance of point from the set to be measured by a *Weighted Hamming Distance*

where

is a vector of non-negative numbers.

1. Applying the Equivalent Bounded Square Differences to the Weighted Hamming Distance: Repeating the argument used in the proof of the *Equivalent Bounded Square Differences Theorem* for the Weighted Hamming Distance, we obtain, for all

where

denotes the Euclidean norm of - recall that serves the role of

in the bounded square differences inequality.

1. Reduction to Euclidean Norm: For all vectors with unit norm

Thus, denoting

for any

we get

1. Reduction to the Talagrand Form: On the one hand, if

On the other hand, since

for

for any

the inequality implies

Thus, for all

we have

1. Talagrand Supremum Form: The main impact of the Talagrand’s inequality is that the above inequality remains that even if the supremum over with unit norm is taken with probability, i.e.
2. Relation of to Chernoff/Hoeffding Type Optimization Parameters: To make above statement more precise, we introduce, for any

the *convex distance* of from the set by

Using this convex distance metric, we are now ready to derive the prototypical result from Talagrand (1995) – for an even stronger concentration-of-measure result, refer to Talagrand (1996a).

**Convex Distance Inequality - Proof**

1. Statement: For any subset

with

and

1. Step #1 - Saddle Point Setup: Define the random variable

First we observe that can be expressed as a saddle-point. Let denote the set of probability measure on . Then

where is distributed according to

thereby the saddle point is achieved.

1. Step #2 - Saddle Point Establishment: We apply Sion’s theorem (Sion (1958)), which may be viewed at the association inequality applied to functional space saddle points. Sion’s minimax theorem states that if denotes a function

that is convex and lower semi-continuous with respect to , concave and upper semi-continuous with respect to , where is convex and compact, then

Additional details relating to checking the conditions of Sion’s theorem are available in Boucheron, Lugosi, and Massart (2003).

1. Step #3 - Applying the Saddle Point: Let now be a saddle point for . We have

where

if

and

Let denote the distribution on that achieves the infimum in the latter expression. Now we have

Hence we get

1. Step #4 - Use of the Bounded Square Differences Inequality: From above, we can see that

Thus, by the bounded square differences inequality (or, more precisely, using its generalized expectations), for any

1. Step #5 - Application of the Equivalent Bounded Square Differences Inequality for the Upper Tail: Applying the above to the lower tail, we get

which by taking

implies

Thus, this means

1. Step #6 - Reduction to the Talagrand Form: Now, if

then

On the other hand, if

then

where the second inequality follows from the upper-tail inequality above.

1. Step #7 - Bringing it all together: In conclusion, for all

which concludes the proof of the convex distance inequality. The constant in the exponent may be improved by other means – say using McDiarmid’s inequality.

**Application of the Convex Distance Inequality – Bin Packing**

1. Overview: Here we apply the convex distance inequality to the bin packing problem (Talagrand (1995)). Let denote the minimum number of bins of size into which the numbers

can be packed. We consider the random variable

where are independent, and take values in .

1. Convex Distance Inequality for Bin Packing: For each
2. Proof Step #1 - Symmetrization of Bin Packing: We first establish (and this is the only specific property of that we use in the proof) that for any

To see this, it suffices to show that those for which

can be packed into at most bins. For this, it is enough to find a packing such that at most one bin is less than half full. But such a packing must exist, since we can always pack the contents of two half-empty bins into one.

1. Proof Step #2 - Recasting this into the Hamming Distance Metric: Denoting by

the unit vector , we clearly have

1. Proof Step #3 - Symmetrization Bounding Using Convex Distance: Let be a positive number, and define the set

Then, by the argument above and using the definition of convex distance, for each

there exists a

such that

from which we set that for each

1. Proof Step #4 - Applying the Hamming Distance Metric: From the above, we have for any
2. Proof Step #5 - Application of the Bernstein Inequality: Given that are independent random variables taking values in , it may be shown that, by applying the Bernstein inequality to the sum

Using this along with the Hamming metric in Step #7, we get

1. Proof Step #6 - Acquiring the Bounds: To obtain the desired inequality, we incorporate two difference choices of on the inequality above. To derive the bound for the upper tail, we take

Then

and the convex distance inequality yields

We obtain a similar equality in the same way for by taking

**References**

* Boucheron, S., G. Lugosi, and P. Massart (2003): Concentration Inequalities using the Entropy Method *Annals of Probability* **31** 1583-1614.
* McDiarmid, C. (1998): Concentration, in: *Probabilistic Methods for Algorithmic Discrete Mathematics* (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed) 195-248 **Springer** New York.
* Sion, M. (1958): On General Minimax Theorems *Pacific Journal of Mathematics* **8** 171-176.
* Steele, J. M. (1996): Probability Theory and Combinatorial Optimization *SIAM, CBMS-NSF Regional Conference Series in Applied Mathematics* **Philadelphia**.
* Talagrand, M. (1995): Concentration of Measure and Isoperimetric Inequalities in Product Spaces *Publications Mathematiques de l’I.H.E.S* **81** 73-205.
* Talagrand, M. (1996a): New Concentration Inequalities in Product Spaces *Inventiones Mathematicae* **126** 505-563.
* Talagrand, M. (1996b): A New Look at Independence *Annals of Probability* **24** 1-34 (Special Invited Paper).

**Section II: Statistical Learning Theory – Foundation and Framework**

**Standard SLT Framework**

**Statistical Learning Theory Monographs**

1. Mathematical and Technical Surveys: Fukunaga (1972), Duda and Hart (1973), Vapnik and Chervonenkis (1974), Vapnik (1982), Breiman, Friedman, Olshen, and Stone (1984), Natarajan (1991), McLachlan (1992), Kearns and Vazirani (1994), Vapnik (1995), Devroye, Gyorfi, and Lugosi (1996), Kulkarni, Lugosi, and Venkatesh (1998), Vapnik (1998), Anthony and Bartlett (1999), Herbrich and Williamson (2002), Lugosi (2002), Scholkopf and Smola (2002), Bosquet, Boucheron, and Lugosi (2003), Mandelson (2003), Boucheron, Bousquet, and Lugosi (2005).

**Computational Learning Theory**

1. Purpose: Computational learning theory is a mathematical field related to the analysis of machine learning algorithms. It seeks to address issues of:
   1. Learnability/Learn Feasibility/Learn Complexity
   2. Learning Speed/Performance
   3. Learning Formulation Criterion
   4. Learning Basis/Representation Choice
   5. Probabilistic Learning Framework and Viability
2. Learn Complexity: In addition to performance bounds computational learning theory studies the time complexity and flexibility of learning. In computational learning theory, a computation is considered feasible if it can be performed in polynomial time (Computational Learning Theory (Wiki)). There are two kinds of time complexity results:
   1. Positive Results => Showing that a certain class of functions is learnable in polynomial time.
   2. Negative Results => Showing that a certain class of functions cannot be learned in polynomial time.
3. Non Polynomial Time Results: Negative results often rely on commonly believed, but as yet unproven assumptions such as:
   1. Computational Complexity –
   2. Cryptographic – One-way functions exist
4. Types of Approaches: There are several different approaches to computational learning theory. These differences are based on making assumptions about inference principles used to generalize from limited data. These include different conceptions of probability (frequentist, Bayesian) and different assumptions on the generation of samples.
5. Computational Learning Theory Approaches:
   1. Probably Approximately Correct Learning (PAC) proposed by Valiant (1984)
   2. VC Theory proposed by Vapnik and Chervonenkis (1971)
   3. Bayesian Inference
   4. Algorithmic learning theory, from the work of Gold (1967)
   5. Online machine learning algorithms, from the work of Nick Littlestone
6. Practical Algorithms from Computational Learning Theory: PAC theory inspired boosting, VC theory led to support vector machines, and Bayesian inference led to belief networks (by Judea Pearl).

**Probably Approximately Correct (PAC) Learning**

1. PAC Learning Motivation: PAC learning is a framework for mathematical analysis of machine learning (Probably Approximately Correct Learning (Wiki), Valiant (1984)). In this framework, the learner selects a generalization hypothesis from a certain class of possible functions. The goal is that, with high probability (the “probably” part) the selected hypothesis function will have a low generalization error (The “approximately” part). The learner must be able to learn the concept given any arbitrary approximation ratio, probability of success, or distribution of samples. The original model has been expanded later to treat noise (misclassified samples).
   1. Probably/Approximately => This model uses , which is the “probably” part, and within that chooses the low error, the “approximately” part. Since is explicit, the treatment in regards to empirical risk is different.
2. The PAC Approach and Goals: An important innovation of the PAC framework is the introduction of the computational complexity theory concepts to machine learning (for e.g., VC deals primarily with “hypothesis complexity”). In particular, the learner is expected to find efficient functions (i.e., the time and space requirements are bounded to a polynomial on the sample size), and the learner itself must implement an efficient procedure (requiring an example count to be bounded to a polynomial of the concept size, modified by the approximation and the likelihood bounds).

**PAC Definitions and Terminology**

1. Sample Background: To elaborate on the PAC learnability, we use the 2 samples employed in Natarajan (1991) and Kearns and Vazirani (1994). The first is the problem of character recognition given an array of bits encoding a binary image. The other is the problem of locating the interval that correctly classifies points inside as positive and outside as negative.
2. The PAC Instance Space: Let be a set called the instance space or encoding of all the samples, and each instance has a length assigned. In the character recognition problem the instance space is . In the interval problem, this would be

where denotes the set of all real numbers.

1. PAC Concept and the Concept Class: A concept is the subset

One concept is the set of all patterns in bits

that encode the picture of the letter P. An example concept from the second example is the set of all numbers between and . A concept class is a set of concepts over . For instance, this could be the set of all the subsets of array of bits that are skeletonized 4-connected (the width of the font being ).

1. The PAC Procedure: Let be a procedure that draws an example using a probability distribution and gives the correct label
2. PAC Learnability and the Learning Algorithm: Say that there is an algorithm that, given access to alongwith inputs and that, with a probability of at least , outputs a hypothesis that has error less than or equal to with examples drawn from using the distribution . If there is an algorithm for every concept

for every distribution over , and for all

and

then is PAC learnable (or distribution-free PAC learnable). Further, we can also say that is a PAC learning algorithm for .

1. Efficient PAC Algorithm: An algorithm runs in time if it draws at most samples and requires at most time steps. A concept class is efficiently PAC learnable if it is PAC learnable by an algorithm that runs in time polynomial in , , and the instance length.

**SLT Introduction**

1. Motivation: Statistical Learning Theory provides the framework for machine learning by drawing from the fields of statistics and functional analysis (Statistical Learning Theory (Wiki), Mohri, Rostamizadeh, and Talwalkar (2012)), with a wide variety of applications (e.g., Sidhu and Caffo (2014))
2. Binary Classification in SLT: The reason that the binary classifier is well-studied is that the classifier prediction error is easily reducible to a) estimation error/precision, and b) approximation error/accuracy. For multinomial and/or linear/non-linear classifiers, other factors will need to be set.
3. Error Factors in Multinomial Classification and Regression: In multinomial classification, the error factor contributions come from each of the component error contributions per label/class above. For regression, in addition to the nodal “best-fit” error contributions (due to the components above), penalization may be applied to limit curvature and improve smoothness.
4. SLT Goals: These are spelt out in von Luxburg and Scholkopf (2008):
   1. Which learning tasks can be performed by computers in general (producing positive/negative results)?
   2. What kind of assumptions do we have to make to ensure that such machine learning tasks are successful?
   3. What key properties does a learning algorithm need to satisfy in order to be successful?
   4. What performance guarantees can we provide on certain learning algorithms?

**The Setup**

1. Empirical Error Determinants: In general, the empirical error depends upon:
   1. The Joint Probability Distribution
   2. The Sample Size
   3. The Function Space/Sub-space Complexity
2. Sample vs. Population Mean: Likewise, the probability that the sample mean approaches the population mean for a finite sample of size to within a tolerance δ depends on:
   1. The tolerance δ
   2. The sample size
   3. The Joint Probability Distribution
   4. The Function Complexity – the complexity of the function proxying
3. SLT Objectives:
   1. No assumptions on
   2. Labels can be non-deterministic so as to be able to accommodate label noise and overlapping label classes
   3. Sampling i.i.d.
   4. is unknown at the time of learning
4. Use of in SLT: In light of the assumptions above, SLT seeks to make statements on the empirical error and sample mean departure probability without making references to specific forms for , i.e., only using sample size, function space complexity, and tolerance.
5. Loss and Risk: The loss function is the cost of misclassifying one observation point, and is written as

Risk of a function is the average of the loss over the data points generated according to the underlying distribution , so the distribution DOES factor in

Empirical risk, however, is simply a miscalculation counter.

1. Bayes’ Classifier:

Almost certainly, Bayes’ classifier does not possess zero empirical risk. Further, it has explicit dependence on (or ).

1. SLT Problem Statement: Given a set of training points that have been drawn from an unknown distribution , and given a loss function , how can we construct a function

that has its risk as close as possible to that of the Bayes’ classifier?

**Algorithms for Reducing Over-fitting**

1. Statistical Regularization: Overfitting is symptomatic of unstable solution, i.e., small training perturbations can cause large shifts in the learned functions. It can be shown that if the stability for the solution can be guaranteed, generalization and consistency are guaranteed as well (Vapnik, Chervonenkis (1971), Mukherjee, Niyogi, Poggio, and Rifkin (2006)). Regularization can address over-fitting and provide problem stability.
2. The Strategy: Essentially 2 ways. The first way is to restrict the function class space (say, via the classical SLT and ERM). The second is to modify the criterion to be minimized using a penalizer for “complicated” functions. Structural Risk Minimization, Standardized Regularization, and Normalization all combine the first and the second.
3. Structure Risk Minimization: We discuss this in detail later on. The idea here is to choose an infinite sequence

of models of increasing size and to minimize the empirical risk in each model with an added penalty term for the model size, i.e.

The penalty gives preference to those models whose estimation error is small, and thus measures the size/capacity of the model.

1. Standardized Regularization: This easier to implement approach consists in choosing of a large model (possibly dense in continuous functions, for example), and to define on a regularizer, usually a norm . One then minimizes the regularized empirical risk

Most existing and successful methods can be thought of as a variant of the regularization method.

1. Regularization by Hypothesis Restriction: Regularization can also occur by restricting the hypothesis space. A common example is the restriction of to linear functions – thereby reducing the problem to that of linear regression. may also be reduced to polynomials of degree , exponentials, or bounded functions on . Restrictions of the hypothesis space avoids over-fitting because the form of the possible functions is limited, and therefore may avoid choosing the function that gives the empirical risk to be arbitrarily close to zero.
2. Tikhonov Regularization: This consists of minimizing

where is a fixed positive parameter referred to as the regularization parameter. Tikhonov regularization ensures the existence, uniqueness, and stability of the solution (Poggio, Rosasco, Ciliberto, Frogner, Evangelopoulus (2012)).

1. Normalized Regularization: In addition to the above regularization approaches, there are other possible approaches where the regularizer can, in some sense, be “normalized”, i.e., when it corresponds to some probability distribution over .

**Bayesian Normalized Regularizer Setup**

1. Regularization from Probability: Given a probability distribution (referred to as a prior) defined on , one can use as the regularizer. In case is countably/uncountably infinite/continuous, we use the density associated with instead.
2. “Prior” from the Regularizer: Reciprocally, from the regularizer of the form , if there exists a measure on such that

for some

then one may construct a prior corresponding to this regularizer.

* 1. As an example, if is a set of hyper-planes in going through the origin, can be identified with , and taking as the Lebesgue measure, it is possible to go from Euclidean norm regularizer to spherical Gaussian measure on as a prior.

1. RKHS Generalization: Generalization to the infinite dimensional Hilbert spaces can also be done, but requires more care. One can, for example, establish a correspondence between the norm of a Reproducing Kernel Hilbert Space and a Gaussian process prior whose covariance function is the kernel of this space.
2. Posterior: From this type of normalized regularizer (or its posterior), we can construct another probability distribution on (typically called the posterior) as

where

is a free parameter and is a normalization factor.

1. Uses of the Posterior:
   1. MAP as the Regularization Process => If we maximize , we recover the original regularization framework as

where the regularizer is (note that maximizing is equivalent to minimizing ).

* 1. Randomization of Predictions => can also be used to randomize the predictions. In this case, before computing the predicted labels for the input , one samples a function according to and computes . This procedure is called Gibbs classification.
  2. Bayesian Averaging => Another way in which the constructed above can be used is by taking the expected prediction of the functions in ;

This is called Bayesian averaging.

**References**

* Anthony, M., and P. L. Bartlett (1999): *Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge.
* Boucheron, S., O. Bousquet, and G. Lugosi (2005): Theory of Classification *ESAIM: Probability and Statistics* **9** 323-375.
* Bousquet, O., S. Boucheron, and G. Lugosi (2003): Introduction to Statistical Learning Theory 169-207 *Advances in Machine Learning* (editors: Bousquet, O., U. von Luxburg, and G. Ratsch) **Springer** Berlin.
* Breiman, L., Friedman, J., Olshen, R., and C. Stone (1984): *Classification and Regression Trees* **Wadsworth International** Belmont, CA.
* Computational Learning Theory (Wiki): [*Wikipedia Entry for Computational Learning Theory*](http://en.wikipedia.org/wiki/Computational_learning_theory).
* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Duda, R., and P. Hart (1973): *Pattern Classification and Scene Analysis* **John Wiley** New York.
* Fukunaga, K. (1972): *Introduction to Statistical Pattern Recognition* **Academic Press** New York.
* Gold, E. (1967): Language Identification in the Limit, *Information and Control* **10 (5)**: 447-474.
* Herbrich, R., and R.C. Williamson (2002): Learning and Generalization: Theoretical Bounds, in *Handbook of Brain Theory and Neural Networks* (editor: M Arbib).
* Kearns, M., and U. Vazirani (1994): *An Introduction to Computational Learning Theory* **MIT Press** Cambridge, MA.
* Kulkarni, S., G. Lugosi, and S. Venkatesh (1998): Learning Pattern Classification – A Survey *IEEE Transaction on Information Theory* **44** 2178-2206 *Information Theory: 1948-1998. Commemorative Special Issue*.
* Lugosi, G. (2002): Pattern Classification and Learning Theory, in: *Principles of Nonparametric Learning* (editor: L Gyorfi) 5-62 **Springer** Vienna.
* Mandelson, S. (2003): A few Notes on Statistical Learning Theory *Advanced Lectures on Machine Learning* **LNCS** 1-40 **Springer**.
* McLachlan, G. (1992): *Discriminant Analysis and Statistical Pattern Recognition* **John Wiley** New York.
* Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012): *Foundations of Machine Learning* **The MIT Press**.
* Mukherjee, S., P. Niyogi, T. Poggio, and R. Rifkin (2006): Learning Theory: Stability is sufficient for Generalization, and necessary and sufficient for Consistency of Empirical Risk Minimization *Advances in Computational Mathematics* **25** 161-193.
* Natarajan, B. K. (1991): *Machine Learning – A Theoretical Approach* **Morgan Kaufmann Publishers**.
* Poggio, T., L. Rosasco, C. Ciliberto, C. Frogner, and G. Evangelopoulus (2012): [*Statistical Learning Theory and Applications*](http://www.mit.edu/~9.520/)
* Probably Approximately Correct Learning (Wiki): [*Wikipedia Entry for Probably Approximately Correct Learning*](http://en.wikipedia.org/wiki/Probably_approximately_correct_learning).
* Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
* Sidhu, G., and B. Caffo (2014): Exploiting Pitcher Decision-making using Reinforcement Learning *Annals of Applied Statistics* **8 (2)** 926-955.
* Statistical Learning Theory (Wiki): [*Wikipedia Entry for Statistical Learning Theory*](http://en.wikipedia.org/wiki/Statistical_learning_theory).
* Valiant, L. (1984): A Theory of the Learnable *Communications of the ACM* **27 (11)** 1134-1142.
* Vapnik, V. and A. Chervonenkis (1971): On the Uniform Convergence of relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16 (2)** 264-280.
* Vapnik, V., and A. Chervonenkis (1974): *Theory of Pattern Recognition (in Russian)* **Nauka** Moscow.
* Vapnik, V. (1982): *Estimation of Dependencies based on Empirical Data* **Springer Verlag** New York.
* Vapnik, V. (1995): *The Nature of Statistical Learning Theory* **Springer Verlag** New York.
* Vapnik, V. (1998): *Statistical Learning Theory* **Springer** New York.
* von Luxburg, U., and B. Scholkopf (2005): Statistical Learning Theory: Models, Concepts, and Results, in *Handbook of the History of Logic* (editors: D. Gabbay, S. Hartmann, and J. Woods).

**Generalization and Consistency**

**Concept Motivation**

1. Good Generalizer: A classifier generalizes well if the difference between the overall risk and the empirical risk is small. This does NOT mean that the classifier has a small overall error ; it simply means that the empirical error is a good estimate of the true error (Scholkopf and Smola (2002)).
2. The Concept of Consistency: The notion of consistency in SLT is the same as that in statistics; it aims to make a statement about what happens to the learning with more and more sample points – the expectation for an appropriate algorithm would be to converge to the “optimal” solution (in hopefully one direction, asymptotically). As opposed to generalization, SLT consistency is not a property of , it is a property of the hypothesis space ℑ (von Luxburg and Scholkopf (2008)).

**Types of Consistency**

1. The Setup: Let be an infinite sequence of training points that have been drawn independently from some probability distribution . Let be a loss function. For each

let be a classifier constructed by some learning algorithm on the basis of the first training points.

1. Consistency with respect to ℑ and : The learning algorithm is called consistent with respect to ℑ and if the risk converges in probability to the risk of the best classifier in , that is, for all

as

1. Bayes’ Consistent: The learning algorithm is called Bayes’ consistent with respect to if the risk converges to of the Bayes’ classifier, that is, for all

as

1. Universal Consistency: The learning algorithm is called universally consistent with respect to ℑ (respectively universally Bayes’ consistent) if it is consistent with respect to ℑ (respectively Bayes’ consistent) for all probability distributions .
2. Convergence vs Weak Consistency: The consistency as stated above is called *weak consistency* in probability, as it is a statement about the consistency in probability. The analogous statement for convergence *almost surely* (i.e., in the sample size limit) would be called *strong consistency* (Devroye, Gyorfi, and Lugosi (1996)).

**Bias-Variance or Estimation-Approximation Trade-off**

1. Definition: Bias Variance Dilemma (or trade off) refers to the problem of simultaneously minimizing the bias (i.e., how accurate the model is across different training sets) and the variance of the model error (how sensitive is the model to small changes in the training set) (Bias-Variance Dilemma (Wiki)).
2. Applicability Suite: This trade off applies to all forms of supervised learning – classification, function fitting (Geman, Bienenstock, and Doursat (1992)), and structured output learning.
3. Motivation: Ideally one wants to choose a model that captures the irregularities in the training data, which at the same time generalizes well to unseen data.
4. High Bias Models: High-bias models are intuitively simple models, and imposed restrictions on the kinds of irregularities that can be learned (examples include linear classifiers). Problem is that they under-fit, i.e., they do not learn the relationship between the predicted (i.e., target) variables and the features.
5. High Variance Models: These can learn many kinds of complex irregularities, which unfortunately includes the noise in the training data as well (i.e., over-fitting).
6. Origin of the Tradeoff: A common model selection criterion is that decrease of bias with an increase in model complexity results in increase of variance. Other considerations such as error losses and complexity costs lead to alternate tradeoff criteria. The choice of model may also introduce biases that correspond to useful previous information, e.g., the output may need to be range bound.

**Bias Variance Decomposition**

1. Error Decomposition:

is called the Variance or the Estimation Error. Traditionally this has occupied much of the focus of SLT. is called the Bias or the Approximation Error. This does not depend on the sample size, but it does depend on the underlying probability distribution .

1. Setup: We employ the terminology of Vijayakumar (2007). Let the data specified

be derived from the true model

where is a random error with

Using this we train a model to approximate . We also set

and

1. Mean Squared Error:

The quantity of interest is the expectation of the across the different realizations of the data:

1. MSE Expansion: Re-writing as , we recognize that is the noise term above, and is simply the proxying error. Thus

Given that is deterministic

Likewise

if we assume that

Thus the last term also vanishes. Therefore

1. MSE Reduction:

Here

Thus

1. MSE Final Form:

Here is the irreducible sample error (assuming infinite sample), is the square of the bias term, and is the model variance term. Cast in another manner

**Bias Variance Optimization**

1. Impact of Feature Addition: Feature selection/filtering and dimensionality reduction can decrease variance by simplifying the models. Adding features (predictors) tends to decrease bias at the expense of introducing extra variance.
2. Approaches for Bias Variance Tuning: Learning algorithms typically have some tunable parameters that control bias and variance. For example:
   1. Generalized linear models can be regularized to increase their bias.
   2. In neural nets, deeper models with more layers will have stronger variance, this regularization (as in GLM) is applied.
   3. In k-nearest neighbor models, a high value of k leads to low variance.
   4. In Instance based learning, regularization can be achieved by varying the mixture of prototypes and examplars (Gagliardi (2011)).
   5. In decision trees, the depth of the tree determines the variance. Decision trees are commonly pruned to control the variance (James, Witten, Hastie, and Tibshirani (2013)).
3. Mixture Models and Ensemble Learning: These provide another way to address the bias-variance tradeoff (Ting, Vijayakumar, and Schaal (2011), Fortmann-Roe (2012)). For example, as seen earlier, boosting combines many “weak” (high bias) models in an ensemble that has greater variance than the individual models, while bagging combines strong learners in a way that reduces their variance.

**Generalization and Consistency for kNN**

1. Setup: Assume that there exists a distance metric function

that assigns a distance value to each pair of training points . Then , the nearest neighbor to among all the training points is

(Stone (1977), Scholkopf and Smola (2002)). The corresponding classifier is

where

1. kNN Bayes’ Risk vs. Risk: It is easy to construct instances to see where the Bayes’ risk for 1NN is quiet different from . However, by expanding the nearest neighbor out to kNN, we can see the difference between and decreases, thereby improving the consistency.
2. Stone (1977) kNN Consistency Theorem: Let be the kNN classifier constructed on sample points. If

and

such that

then

for all probability distributions . Therefore, kNN classification thus constructed can be made universally consistent.

* 1. Interpretation => Essentially, one has to allow the sample size of the neighborhood under consideration to grow with the sample size in a controlled way. For e.g., if the parameter grows slowly with , say

then kNN becomes universally Bayes’ consistent.

1. Intuition behind the kNN Function Complexity: Treating as the hypothesis variant dimension in the space of functions , the extremes to observe are: a) 1NN, where potentially each node can switch its label depending upon its neighbors, and b) kNN with

the sample size, which has only one label. In this situation, for a location independent

(Devroye, Gyorfi, and Lugosi (1996)).

**References**

* Bias-variance dilemma (Wiki): [*Wikipedia Entry for Bias-variance dilemma*](http://en.wikipedia.org/wiki/Bias-variance_dilemma).
* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Fortmann-Roe, S. (2012): [*Understanding the Bias-Variance Tradeoff*](http://scott.fortmann-roe.com/docs/BiasVariance.html)
* Gagliardi, F. (2011): Instance-based Classifiers applied to Medical Databases *Artificial Intelligence in Medicine* **52 (3)** 123-139.
* Geman, S., E. Bienenstock, and R. Doursat (1992): Neural Networks and the Bias/Variance Dilemma *Neural Computation* **4** 1-58.
* James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): [*An Introduction to Statistical Learning*](http://www-bcf.usc.edu/~gareth/ISL/)
* Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
* Stone, C. J. (1977): Consistent non-parametric regression (with discussion) *Annals of Statistics* **5** 595-645.
* Ting, J. A., S. Vijayakumar, and S. Schaal (2011): Locally Weighted Regression for Control in *Encyclopedia of Machine Learning* (eds. C. Sammut and G. I. Webb) **Springer** 615.
* Vijayakumar, S. (2007): [*The Bias-Variance Tradeoff*](http://www.inf.ed.ac.uk/teaching/courses/mlsc/Notes/Lecture4/BiasVariance.pdf).
* von Luxburg, U., and B. Scholkopf (2005): Statistical Learning Theory: Models, Concepts, and Results, in *Handbook of the History of Logic* (editors: D. Gabbay, S. Hartmann, and J. Woods).

**Section III: Empirical Risk Minimization – Principles and Techniques**

**Empirical Risk Minimization**

**ERM Literature and Introduction**

1. Empirical Process Theory - Literature: Uniform deviations of averages form their expectations is one of the central problems of empirical process theory, and comprehensive coverages include, Gine (1996), van der Waart and Wellner (1996), Vapnik (1998), and Dudley (1999).
2. Empirical Processes in Classification: The use of empirical processes in classification was pioneered by Vapnik and Chervonenkis (1971, 1974) and re-discovered 20 years later by Blumer, Ehrenfeucht, Haussler, Warmuth (1989), and Ehrenfeucht, Haussler, Kearns, and Valiant (1989). For surveys, see Natarajan (1991), Kearns and Vazirani (1994), Vapnik (1995), Devroye, Gyorfi, and Lugosi (1996), Vapnik (1998), and Anthony and Bartlett (1999).
3. ERM Induction Principle: The proxying of the risk corresponding to the full population to the empirical risk of the sample, choosing that particular

that minimizes the empirical risk of the given finite sample is called the ERM induction principle

1. Principle Caveats: With every sample set change, will vary. Ideally, should be part of , so that as

Otherwise

but

even as

this produces sizeable estimation error as well as approximation error .

**Overview**

1. Definition: ERM is a principle in statistical learning which defines a family of learning algorithms constructed using a specific risk principle, and is used to provide performance bounds on these algorithms.
2. Background/Setting: Consider the following situation typical of supervised learning problems. We have 2 spaces of objects and and would like to learn a function (called a hypothesis)

which outputs an object

given

We have a set of training examples where

is the input and

is the corresponding response, and we wish to get the response predictor hypothesis .

1. Training Set Generation: Formally we assume that there is a joint probability distribution over and , and the training set consists of instances drawn i.i.d. from . Note that the assumptions of joint probability allows us to model uncertainties in the predictors (e.g., noise from data) because is not a deterministic function of , but rather a random variable with a conditional distribution for a fixed .
2. ERM Loss Function: We also assume that we are given a non-negative real-valued loss-function that quantifies how different the prediction of the hypothesis is from the true outcome . The risk associated with the hypothesis is then defined as the expectation of the loss function:

A commonly used loss function for classification is the loss function

where is the indicator notation.

* 1. Expectation Maximization Insight => The least squares loss function corresponds to maximizing the joint normal probability . If is not Gaussian, and/or if there is a non-uniform (i.e., Bayesian, not frequentist) prior, then the corresponding for the loss function may also be derived (it will not be Gaussian in general).

1. The Goal of the Computational Learning Algorithm: To find the hypothesis among the fixed class of hypotheses that minimizes the risk :

**The Loss Function and the Empirical Risk Minimization Principles**

1. Choice of Loss Function: The choice of loss function is critical in computing the hypothesis – this choice significantly impacts the algorithm convergence rate. Further it is also important for the loss function to be convex (Rosasco, Vito, Caponnetto, Piana, and Verri (2004)).
2. ERM Definition: In general cannot be computed since is unknown to the learning algorithm (this situation is referred to as agnostic learning). Empirical risk is the approximation that results from averaging the loss function evenly across the training set (this simply corresponds to a uniform ):

(See Empirical Risk Minimization (Wiki)).

1. Problem Statement: Empirical Risk Minimization Principle states that the learning algorithm should choose a hypothesis that minimizes the empirical risk:

Depending upon the nature of the learner, this base formulation may end up as a combinatorial optimization problem for which globally optimal solutions are infeasible, thereby forcing a resort to meta-heuristic techniques.

**Application of the Central Limit Theorem (CLT) and Law of Large Numbers (LLN)**

1. CLT vs. LLN over Function Spaces: CLT and LLN are alternate views at the same fact; CLT looks at

expected mean of the sample, whereas LLN looks at the probability of the sample exceeding the population mean. If CLT is computed over the hypotheses spaces, it is called -Donsker; LLN over hypothesis space is, likewise, called Glivenko-Cantelli. Thus, one implies the other.

1. Glivenko-Cantelli Literature: The question of how

is also called the Glivenko-Cantelli, and is covered in Vapnik and Chervonenkis (1971), Dudley (1978), Vapnik and Chervonenkis (1981), Dudley (1984, 1987), Talagrand (1987, 1994), and Alon, Ben-David, Cesa-Bianchi, and Haussler (1997).

1. ERM LLN: LLN simply states that

as

where and are drawn from the same distribution. LLN is valid under very mild conditions, and is easily extensible to ERM as follows:

as

Again, is drawn from the same distribution that generates

1. Probability Bounds for LLN: A famous inequality by Chernoff (1952) expanded later to the i.i.d. observation set by Hoeffding (1963) characterizes how well the sample empirical mean approaches the population mean as a function of the sample size (the result is independent of , and is quiet conservative). Applied to ERM, it is:
2. Other Probability Bounds: Both Chernoff and Hoeffding bounds are derived from the Markov inequality. As such, they are conservative and provide only loose bounds, since they do not use at all. Even among these agnostic bounds, moment bounds (and factorial moment bounds) can be much tighter than the Hoeffding bound (which does not use moments at all). Examples include the Bennett and the Bernstein bounds (naïve usage of limited number of moments, like the Chebyshev bounds which uses the 2nd moment based bounds, provide even looser convergence than Hoeffding).

**Inconsistency of Empirical Risk Minimizers**

1. Memorizing Classifiers: While memorizing classifiers can produce zero empirical risk, thereby being the “perfect” empirical risk minimizer, they would often need to “wiggle” enormously, thereby possessing a high “complexity” quotient i.e.,

This makes them terrible learners.

1. Small Wavelength Trigonometric Functions: Is a small λ sine/cosine (or any periodic function) as example of a perfect memorizer? It appears so.

**Uniform Convergence**

1. Motivation: As it turns out, the conditions required to render ERM consistent involve restricting the set of admissible functions (thereby limiting complexity). The main insight of VC theory is that the consistency of ERM is determined by the worst case behavior over all

that the learning machine could choose.

1. The ERM Approach: For each function

LLN tells us that as

However, this does not necessarily mean that the ER minimizer function produces a risk that approaches that of as

The latter is what results in statistical/functional class consistency. For this to be true, we explicitly require that

as

for all

(Scholkopf and Smola (2002)).

1. Sample Size producing Uniform Convergence: Of course, from LLN, there will exist some large enough such that

for all

- we write this as

Ideally we hope that this

is not too large.

1. ERM Consistency Workout: For the hypothesis class estimation error consistency, we require that be bounded. Now

**ERM Complexity**

1. ERM Complexity Analysis: ERM for a classification problem with a loss function is known to be NP-hard even for relatively simple class of functions such as linear classifiers (owing to the supra-exponential growth of the combinatorial optimization search space – Feldman, Guruswamy, Raghavendra, and Wu (2010)). However, when the minimal empirical risk is zero AND when the data is linearly separable, it can be solved efficiently.
2. ERM Complexity Handling Approaches: In practice, machine learning algorithms cope with NP-hard situations as above either by using a convex approximation to loss function (e.g., hinge loss for SVM – this alters the combinatorial search space onto a convex real search space) that is easier to work with/optimize, or by posing simplifying assumptions on (thereby not being agnostic anymore).
   1. Loss Functions for Classification => The indicator loss function can be implemented via the Heaviside step function. To convert it to a convex function, a hinge loss function is often used:

**References**

* Alon, N, S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997): Scale-sensitive Dimensions, Uniform Convergence, and Learnability *Journal of the ACM* **44** 615-631.
* Anthony, M., and P. L. Bartlett (1999): *Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge.
* Blumer, A., A. Ehrenfeucht, D. Haussler, and M. Warmuth (1989): Learnability and the Vapnik Chervonenkis Dimension *Journal of the ACM* **36** 929-965.
* Ehrenfeucht, A., D. Haussler, M. Kearns, and L. Valiant (1989): A General Lower Bound on the Number of Sample needed for Learning *Information and Computation* **82** 247-261.
* Empirical Risk Minimization (Wiki): [*Wikipedia Entry for Empirical Risk Minimization*](http://en.wikipedia.org/wiki/Empirical_risk_minimization).
* Chernoff, H. (1952): A Measure of Asymptotic Efficiency of Tests of a Hypothesis based on the Sum of Observations *Annals of Mathematical Statistics* **23** 493-507.
* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Dudley, R. (1978): Central Limit Theorems for Empirical Measures *Annals of Probability* **6** 899-929.
* Dudley, R. (1984): Empirical Processes, in: *Ecole de Probabilite de St. Fleur 1982, Lecture Notes in Mathematics #1097* **Springer-Verlag** New York.
* Dudley, R. (1987): Universal Donsker Classes and Metric Entropy *Annals of Probability* **15** 1306-1326.
* Dudley, R. (1999): *Uniform Central Limit Theorems* **Cambridge University Press** Cambridge.
* Feldman, V., V. Guruswamy, P. Raghavendra, and Y. Wu (2010): [*Agnostic Learning of Monomials by Halfspaces is Hard*](http://arxiv.org/pdf/1012.0729v1.pdf) **arXiV**.
* Gine, E. (1996): Empirical Processes and Applications: An Overview *Bernoulli* **2** 1-28.
* Hoeffding, W. (1963): Probability Inequalities for Sums of Bounded Random Variables *Journal of American Statistical Association* **58** 13-30.
* Kearns, M., and U. Vazirani (1994): *An Introduction to Computational Learning Theory* **MIT Press** Cambridge, MA.
* Natarajan, B. (1991): *Machine Learning: A Theoretical Approach* **Morgan Kauffman** San Mateo, CA.
* Rosasco, L., E. Vito, A. Caponnetto, M. Piana, and A. Verri (2004): Are All Loss Functions of Same? *Neural Computation* **5 (16)** 1063-1076.
* Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
* Talagrand, M. (1987): The Glivenko-Cantelli Problem *Annals of Probability* **15** 837-870.
* Talagrand, M. (1994): Sharper Bounds for Gaussian and Empirical Processes *Annals of Probability* **22** 28-76.
* Van der Waart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes* **Springer Verlag** New York.
* Vapnik, V., and A. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16** 264-280.
* Vapnik, V., and A. Chervonenkis (1974): *Theory of Pattern Recognition (in Russian)* **Nauka** Moscow.
* Vapnik, V., and A. Chervonenkis (1981): The Necessary and Sufficient Conditions for the Uniform Convergence of Averages to their Expected Values *Teoriya Veroyatnostei i Ee Primeneniya* **26 (3)** 543-564.
* Vapnik, V. (1995): *The Nature of Statistical Learning Theory* **Springer Verlag** New York.
* Vapnik, V. (1998): *Statistical Learning Theory* **Wiley** New York.

**Symmetrization**

**Introduction**

1. Motivation: The purpose of the Symmetrization step is to replace in

(since is not an observable) by , which is an observable. To this end, we introduce a new copy of our original sample, and refer to this as our ghost sample. Thus all expectations for now occur over .

1. Symmetrization is Very Generic: Since Symmetrization converts all the expectation calculations to happen over the ghost sample, it may be very effectively applied across all operators that perform expectations, e.g., moment bounds, Chernoff/Chebyshev bounds etc.
2. Vapnik Chervonenkis Symmetrization Lemma: For

we have

The on the LHS refers to the distribution of an i.i.d. sample, while the second one refers to the distribution of 2 samples of size each – the original and the ghost samples – thereby an i.i.d. distribution of size .

1. Transformation onto the Outcome Space: Using the lemma, we can now convert the probability bounds estimation in the infinite sample case, i.e., , to samples over the dual sequence of and , i.e., sample outcomes. Remember that the set can only assume a finite number of outcomes, i.e., in the case of binary classifiers, at most

outcomes.

1. The Continuous Case: Since we are evaluating functions only on their outcomes, we treat all functions that produce the same outcome identically. Thus, even in the infinite function space case, as long as the space only results in a fixed output tuple, that’s all we use to evaluate. For a binary classifier this reduces to cases.

**Proof of the Symmetrization Lemma**

1. Literature: Treatment in detail of the Symmetrization lemma may be found in Vapnik and Chervonenkis (1971, 1974) and Gine and Zinn (1984).
2. Symmetrization Lemma - Statement: For any

such that

where is the corresponding empirical measure taken over an independent ghost sample .

1. #1 - Applying Indicator Functions over the Split Subsamples: Let be the function that achieves the supremum (note that it depends on ). Using to represent the conjunction of the 2 events we get
2. #2 - Expectation over the Ghost Sample: Taking expectations over the ghost samples gives
3. #3 - Applying Chebyshev Inequality to the Ghost Sample: By Chebyshev’s inequality

since the random variable in the range has a variance . Thus, switching from to

1. #4 Expectation over the Original Sample: Taking expectation with respect to the original sample (), and observing that

we recover the statement we set out to prove, i.e., the Symmetrization lemma.

1. Applying Hoeffding’s Union Bound:

Inversion of this proves the VC theorem for risk bound using the growth function; For

and, with probability at least

**References**

* Gine, E., and J. Zinn (1984): Some Limit Theorems for Empirical Processes *Annals of Probability* **12** 929-989.
* Vapnik, V., and A. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16** 264-280.
* Vapnik, V., and A. Chervonenkis (1974): *Theory of Pattern Recognition (in Russian)* **Nauka** Moscow.

**Generalization Bounds**

**Union Bound**

1. Motivation: Thus far, all we have are bounds for a single function (across the sample point space), “in the limit”. We seek bounds that work across the entire function space for finite sample sizes, i.e., we seek
2. Setup: Given a finite sample error probability for a single function, union bound seeks to find the probability that any one function in the function space exceeds the specified error bound. This is necessary to compute a conservative lower bound for the empirical case.
3. Formulation: We work out the discrete hypothesis set case:

where is the number of (discrete) functions in the function space , and is the number of data points. Note that we seek to optimize the modulus .

1. Hoeffding for the Function Space: We start from

Note that

and

Further

and

Therefore

Thus

1. Discrete Union Bound - Convergence: Obviously, as

This ensures statistical sample consistency. Extension to the continuous function space case requires a few more techniques, including replacing with a more general capacity measure.

1. Refined Union Bound and the Countable Case #1: For each

for each

(possibly depending on , which we write as ), Hoeffding’s inequality says that

Hence, if we have a countable set , the union bound immediately yields

1. Countably Infinite Case: Choosing

this makes the RHS above equal to , and we get the following result; with probability at least

We notice that if is finite in , taking a uniform results in as before.

1. ERM Consistency Bounds: Applying the bounds separately to each counterpart above, we see that

Expressing this in terms of probabilities using LLN produces the following:

1. Necessary and Sufficient Condition for ERM Convergence: Vapnik and Chervonenkis (1971), Devroye, Gyorfi, and Lugosi (1996), and Mendelson (2003) contain the details.

as

for all

is a necessary and sufficient condition for the consistency of ERM with respect to ℑ.

**Shattering Coefficient**

1. Motivation: The infinite function space case may now be reduced to

(over the original and the ghost) subset. Of course, in many classifiers, the full spectrum will never be produced, so is used as the corresponding outcome “capacity measure”, i.e., the outcome spectrum of the hypothesis set.

1. Definition: Let

be a given sample set. Denote by the cardinality of when restricted to , i.e., the number of in that can be distinguished by their values on . The shattering coefficient is defined as

1. Intuition: Shattering coefficient is simply a count of the number of ways the function space can classify the input pattern set. When

this means that there a sample of size on which all possible separations can be achieved. Then is said to shatter into points.

1. Continuous Supremum Bound: It is easy to see that

Convergence therefore depends on the growth of with .

1. Polynomial Growth of the Shattering Coefficient: If

then

Thus

as

1. Shattering Coefficient growth as : In this case

thus this does not support convergence. However, since the supremum limits we’ve worked on so far are very conservative, this does not indicate non-conformance either.

1. Necessary and Sufficient Conditions for ERM Convergence: As shown in Vapnik and Chervonenkis (1971, 1981), Devroye, Gyorfi, and Lugosi (1996), and Mendelson (2003), the necessary and sufficient condition for EMR convergence is

**Empirical Risk Generalization Bound**

1. Probabilistic Bounds: Expressing the probability bound the other way, with a probability of at least

satisfies

For polynomial growth functions

as

therefore it converges. However, for

therefore that DOES NOT converge.

1. Union Bound Contribution - Inversion: To recap, for a specific , with probability at least

For all

with probability of at least

The extra term contribution from the union bound may also be interpreted as the number of additional bits needed to encode the hypothesis space . It turns out that this kind of coding interpretation of the generalization bounds is often possible, and can be used to obtain error estimates (von Luxburg, Bousquet, and Scholkopf (2004)).

1. Alternate Generalization Bounds: Several other types of bounds for the probability and ERM exist, and they differ in the constants (in front of the exponential and as its argument), as exponent of , and in the way they measure capacity (Devroye, Gyorfi, and Lugosi (1996), Vapnik (1998)).
2. Improvements over Classical Bounds: There are several things that can be improved:
   1. Hoeffding’s inequality only uses the boundedness of the functions, not their variance.
   2. The union bound is as bad as if all the functions in the class were independent (i.e., if and were independent).
   3. The supremum over of is not necessarily what the algorithm would choose, so that upper bounding by the supremum might be loose.

**Large Margin Bounds**

1. Motivation: Large Margin Bounds are a specialized capacity measure of the function classes, where the sole purpose is to classify points in a 2D data space into separate classes using a straight line. A generalization would be linear classifiers in .
2. Setup: Given a set of training points and a classifier that would perfectly separate them, we define the *Margin of the Classifier* as the smallest distance of any training point to the separating line . A similar margin can be defined in for an arbitrary dimension
3. VC Dimension: The VC dimension of a class of linear classifiers with all having a margin of at least can be essentially bounded by the ratio of the radius of the smallest sphere enclosing the data points with the margin , that is

(Vapnik (1995)).

1. Capacity: Thus, larger the , smaller is the VC dimension. Thus, one can use the margin of the classifiers as a capacity concept. SVM builds on this concept, and is one of the best known classifiers (Scholkopf and Smola (2002)).
2. Definition: Assume that the data lies in a ball of radius in . Consider the set of linear classifiers with a margin of at least . Assume that we are given training examples. Use to denote the fraction of the training examples with margin smaller than , or which are wrongly classified by a classifier

Then, with probability of at least , the true error of any

can be bounded by

where is a universal constant (Scholkopf and Smola (2002)).

**References**

* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Mandelson, S. (2003): A few Notes on Statistical Learning Theory *Advanced Lectures on Machine Learning* **LNCS** 1-40 **Springer**.
* Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
* Vapnik, V., and A. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16** 264-280.
* Vapnik, V., and A. Chervonenkis (1981): The Necessary and Sufficient Conditions for the Uniform Convergence of Averages to their Expected Values *Teoriya Veroyatnostei i Ee Primeneniya* **26 (3)** 543-564.
* Vapnik, V. (1995): *The Nature of Statistical Learning Theory* **Springer Verlag** New York.
* Vapnik, V. (1998): *Statistical Learning Theory* **Wiley** New York.
* Von Luxburg, U., O. Bousquet, and B. Scholkopf (2004): A Compression Approach to support Vector Model Selection *The Journal of Machine Learning Research* **5** 293-323.

**Rademacher Complexity**

**Setup and Definition**

1. Motivation: Rademacher complexity is an alternate metric of the function space capacity. Compared to shattering coefficient and the VC dimension, using the Rademacher complexity in conjunction with the underlying distribution can produce much tighter bounds. The mathematical treatment can also become simpler (Bousquet, Boucheron, and Lugosi (2003), Mendelson (2003), Boucheron, Bousquet, and Lugosi (2005)).
2. Problem Space: Given a space and a fixed distribution , let be a set of samples drawn from . Let be a class of functions .
3. Rademacher Complexity - Definition: The *Empirical Rademacher Complexity* of ℑ is defined to be where are independent random variables chosen from . We refer to these as Rademacher variables. The *Rademacher Complexity* is defined as .
4. Empirical Rademacher Complexity - Intuition: The intuitively measures, for a given set and the Rademacher vector , the maximum correlation between and over all . The expectation over , i.e., the empirical Rademacher complexity of measures the ability of functions from (when applied to the fixed set ) to fit random noise.
5. Rademacher Complexity – Intuition: The Rademacher complexity of measures the expected noise-fitting-ability of over all the data sets that could be drawn according to the distribution .
6. Sample Space Generalization: The Rademacher complexity can be defined even more generally on sets by making the over (in place of ), and replacing each with . Of course, setting recovers the earlier definitions.

**Rademacher-based Uniform Convergence**

1. The Concept: The intention is to bound each function that is part of any class if functions by their empirical averages, the Rademacher complexity of the class, and an error term that depends on the probabilistic confidence interval and the sample size.
2. Rademacher Bounding - Steps: The Rademacher bounding typically involves extracting sequential bounds, and it typically employs the following steps:
   1. Use *concentration* to relate

to its expectation.

* 1. Use *symmetrization* to relate the expectation to the Rademacher average.
  2. Use *concentration* again to relate the unconditional Rademacher average to its conditional one.

1. Symbology and Steps: Denote the empirical average over a sample as

Thus for a function , the definition of leads to

We set

and try to bound it by using the McDiarmid inequality.

* 1. McDiarmid Bounded Differences Inequality: Let be independent random variables taking on values in set , and let be positive real constants. If

satisfies

then for

1. Supremum Bounds: as defined above satisfies

where is the number of data points.

* 1. Proof Setup => Let

and

We then define

* 1. Optimizer Function => Letting be the maximizing function for the supremum in
  2. Optimizer Function => By definition of the supremum, can also, at best, maximize , so
  3. Value Realization => Using the above 2 inequalities, and noting the negative sign ahead of
  4. Bounds Estimator => Since and differ in only element , the above becomes

where the last step follows from the fact that

so

1. McDiarmid Bounds on the Inequality : Using the bound

we apply McDiarmid’s inequality to get

Setting this probability to be less than and solving for , we get

1. First ERM Bounds: Thus, with a probability of at least ,

To obtain further refinement/tightness to these bounds, we introduce the concept of ghost variables, and then estimate the bounds in terms of Rademacher Complexity.

* 1. Impact of McDiarmid Bounds => Effectively , applying the McDiarmid bounds on each of the sample point transforms the agnostic supremum over to an expectation of the supremum over (plus a bound that depends on the probability), i.e.

1. Estimation of the Sample Bound for the Supremum: To estimate

we do the following steps:

* 1. Draw a “ghost sample”

that is i.i.d. and independent of

* 1. Get as an expectation over these ghost samples.
  2. Observe that

and that

1. Sample Expectation of the Supremum:
2. Convexity of the Supremum Function: Since it seeks out the supremum, by definition *sup* is a convex function. Thus, we can apply Jensen’s inequality to push the *sup* inside the expectation n as:

Remember that

since are independent.

1. Advantage of the Rademacher Formulation: There are 2 main advantages to introducing Rademacher variables:
   1. Multiplying each term in the summation by a Rademacher variable independent of does not alter the expectation;
   2. Negating a Rademacher variable uniformly does not alter its distribution.
2. Reduction using Rademacher Variables:
3. Rademacher Uniform Convergence Bound:
4. Empirical Rademacher Bounds Supremum Estimation: For

it is bounded precisely the same way as before using the McDiarmid bounded difference inequality as

This is easy to show, since all the arguments for the earlier bounding of the supremum apply in the presence of Rademacher variable coefficients as well.

1. Rademacher Average Uniform Convergence Bound #2: Thus, the second application of the McDiarmid’ inequality (now using confidence limits of for each) bound in terms of its expectation gives
2. Bounding the Rademacher Complexity: To set up the use of Hoeffding’s inequality, we start by taking exponential of the Empirical Rademacher complexity, multiply by some positive constant , apply Jensen’s inequality, and optimize over . In other words, to compute

start with

1. Bounding Rademacher Complexity – Steps:
2. Hoeffding’s Application Criterion: The last step above exploits the fact that the ’s are independent. Now we can apply Hoeffding’s inequality since

and

where

1. Application of Hoeffding’s Inequality:

where is the cardinality of the space , and

1. Optimization around : Taking log of both sides above we get

which has a maximum at

1. Rademacher Complexity Bounds: Substituting this value for back into the previous bound, and dividing both sides by gives
2. Application to VC Theory Finite Concept Class: For any finite concept class

and data points

we can take

Thus

and

on setting each

Thus, in this case

1. Application to the VC Theory - The General Case: In general, irrespective of whether is finite or not, we can take

the set of distinct labels of points in using concepts in . Then

the shatter coefficient of on points, and

By Sauer’s lemma

where is the VC dimension of . Thus the above may be further simplified into

1. Rademacher Average on the Loss Class vs. Hypothesis Class:

Using this we can relate the empirical loss bound to the empirical hypothesis bound.

**VC Entropy**

1. Distribution Dependent Bounds: The treatment so far has been distribution independent. We modify the treatment above to get a distribution dependent bound. We use the notation
2. VC Entropy Definition: The annealed VC entropy is defined as

Using this definition, for any , with probability at least , we now show that, by extending the VC theory

1. Distribution Dependent Bounds Estimator: As before, we begin with the Symmetrization lemma, so we have to upper bound the quantity
2. Distribution Dependent Rademacher Bounds: Using Rademacher variables , we note that

and

have the same distribution, since changing one corresponds to exchanging and .

1. Distribution Dependent Bounds - Union Bounding: Applying Rademacher variables as shown above, we get

and union bounding followed by the application of the distribution-dependent shattering leads to

1. Distribution Dependent Bounds - Final Step: Noting that

and applying Hoeffding’s inequality finally gives

Finally, observe that just as

before we now have

Applying this last step, we get the bounds above.

**Chaining Technique**

1. Introduction: Although the Rademacher bounding produces bounds that are comparable to the traditional VC theory, i.e.

the dependence on can be improved by using the *Chaining* technique. The idea is to use the covering numbers at all scales to capture the geometry of the class in a better way than the VC entropy class.

1. Dudley-Haussler Entropy Bound: Using the chaining technique, one has the following so-called Dudley’s entropy bound:

As a consequence, along with the Haussler’s upper bound, we get the following result:

We can, with this approach, thus remove the unnecessary factor of the VC bound.

**Literature**

1. Bounded Differences Inequality - Martingale Methods: The bounded differences inequality was first formulated explicitly by McDiarmid (1989), who proved it using the Martingale method (McDiarmid (1989, 1998)).
2. Bounded Differences Inequality – Information Theoretic Methods: Concentration results closely related to martingale methods have been obtained using information theoretic methods (Ahlswede, Gacs, and Korner (1976), Marton (1986, 1996a, 1996b), Dembo (1997), Massart (1998), and Rio (2001)).
3. Bounded Differences Inequality – Talagrand’s Induction Method: See Talagrand (1995, 1996a, 1996b), Panchenko (2001, 2002), McDiarmid (2002), Luczak and McDiarmid (2003), Panchenko (2003).
4. Bounded Differences Inequality – Entropy Methods: The *Entropy Method*, based on logarithmic Sobolev inequalities, was developed by Ledoux (1996, 1997). See also Bobkov and Ledoux (1997), Massart (2000), Boucheron, Lugosi, and Massart (2000), Rio (2001), Bousquet (2002), Boucheron, Lugosi, and Massart (2003), and Boucheron, Bousquet, Lugosi, and Massart (2004).
5. Rademacher Averages: The use of Rademacher averages in classification was first promoted by Koltchinskii (2001), and Bartlett, Boucheron, and Lugosi (2001). Details are available in additional surveys by Koltchinskii and Panchenko (2000, 2002), Bartlett and Mendelson (2002), Bartlett, Bousquet, and Mendelson (2002), Bousquet, Koltchinskii, and Panchenko (2002), and Antos, Kegl, Linder, and Lugosi (2002).

**References**

* Ahlswede, R., P. Gacs, and J. Korner (1976): Bounds on Conditional Probabilities with Applications in multi-user Communication *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **34** 157-177 (with corrections in **39** 353-354 (1977)).
* Antos, A., B. Kegl, T. Linder, and G. Lugosi (2002): Data-dependent Margin-based Generalization Bounds for Classification *Journal of Machine Learning Research* **3** 73-98.
* Bartlett, P., S. Boucheron, and G. Lugosi (2001): Model Selection and Error Estimation *Machine Learning* **48** 85-113.
* Bartlett, P., O. Bousquet, and S. Mendelson (2002): Localized Rademacher Complexities, in: *Proceedings of the 15th Annual Conference on Computational Learning Theory* 44-48.
* Bartlett, P., and S. Mendelson (2002): Rademacher and Gaussian Complexities: Risk Bounds and Structural Results *Journal of Machine Learning Research* **3** 463-482.
* Bobkov, S., and M. Ledoux (1997): Poincare’s Inequalities and Talagrand’s Concentration Phenomenon for the Exponential Distribution *Probability Theory and Related Fields* **107** 383-400.
* Boucheron, S., G. Lugosi, and P. Massart (2000): A Sharp Concentration Inequality with Applications *Random Structures and Algorithms* **16** 277-292.
* Boucheron, S., G. Lugosi, and P. Massart (2003): Concentration Inequalities using the Entropy Method *Annals of Probability* **31** 1583-1614.
* Boucheron, S., O. Bousquet, G. Lugosi, and P. Massart (2004): Moment Inequalities for Functions of Independent Random Variables *Annals of Probability* **33 (2)** 514-560.
* Boucheron, S., O. Bousquet, and G. Lugosi (2005): Theory of Classification *ESAIM: Probability and Statistics* **9** 323-375.
* Bousquet, O. (2002): A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes *C. R. Acad. Sci. Paris* **334** 495-500.
* Bousquet, O., V. Koltchinskii, and D. Panchenko (2002): Some Local Measures of Complexities of Convex Hulls and Generalization Bounds, in: *Proceedings of the 15th Annual Conference on Computational Learning Theory* **Springer** 59-73.
* Bousquet, O., S. Boucheron, and G. Lugosi (2003): Introduction to Statistical Learning Theory 169-207 *Advances in Machine Learning* (editors: Bousquet, O., U. von Luxburg, and G. Ratsch) **Springer** Berlin.
* Dembo, A. (1997): Information Inequalities and Concentration of Measure *Annals of Probability* **25** 927-939.
* Koltchinskii, V., and D. Panchenko (2000): Rademacher Processes and Bounding the Risk of Function Learning, in: *High Dimensional Probability II (editors: E. Gine, D. Mason, and J. Wellner)* 443-459.
* Koltchinskii, V. (2001): Rademacher Penalties and Structural Risk Minimization *IEEE Transactions on Information Theory* **47** 1902-1914.
* Koltchinskii, V. and D. Panchenko (2002): Empirical Margin Distribution and Bounding the Generalization Error of Combined Classifiers *Annals of Statistics* **30**.
* Ledoux, M. (1996): On Talagrand’s Deviation Inequalities for Product Measures *ESAIM: Probability and Statistics* **1** 63-87.
* Ledoux, M. (1997): Isoperimetry and Gaussian Analysis *Lectures on Probability Theory and Statistics (editor: P. Bernard)* **Ecole d’Ete de Probabilites de St-Flour XXIV-1994** 165-294.
* Luczak, M. J., and C. McDiarmid (2003): Concentration for Locally Acting Permutations *Discrete Mathematics* **265** 159-171.
* Mandelson, S. (2003): A few Notes on Statistical Learning Theory *Advanced Lectures on Machine Learning* **LNCS** 1-40 **Springer**.
* Marton, K. (1986): A Simple Proof of the Blowing-up Lemma *IEEE Transactions on Information Theory* **32** 445-446.
* Marton, K. (1996a): Bounding -distance by Informational Divergence: A Way to prove Measure Concentration *Annals of Probability* **24** 857-866.
* Marton, K. (1996b): A Measure Concentration Inequality for contracting Markov Chains *Geometric and Functional Analysis* **6** 556-571 (Erratum: **7** 609-613 (1997)).
* Massart, P. (1998): Optimal Constants for Hoeffding Type Inequalities *Technical Report* ***98.86****, Mathematiques* **Universite de Paris-Sud**.
* Massart, P. (2000): About the Constants in the Talagrand’s Concentration Inequalities for Empirical Processes *Annals of Probability* **28** 863-884.
* McDiarmid, C. (1989): On the Method of Bounded Differences, in: *Surveys in Combinatorics* 148-188 **Cambridge University Press** Cambridge.
* McDiarmid, C. (1989): Concentration, in: *Probabilistic Methods for Algorithmic Discrete Mathematics (M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed: editors)* 195-248 **Springer** New York.
* McDiarmid, C. (2002): Concentration for Independent Permutations *Combinatorics, Probability, and Computing* **2** 163-178.
* Panchenko, D. (2001): A Note on Talagrand’s Concentration Inequality *Electronic Communications in Probability* **6**.
* Panchenko, D. (2002): Some Extensions of an Inequality of Vapnik and Chervonenkis *Electronic Communications in Probability* **7**.
* Panchenko, D. (2003): Symmetrization Approach to Concentration Inequalities for Empirical Processes *Annals of Probability* **31 (4)** 2068-2081.
* Rio, E. (2001): Inegalities de concentration pour les processus empiriques de classes de parties *Probability Theory and Related Fields* **119** 163-175.
* Talagrand, M. (1995): Concentration of Measures and Isoperimetric Inequalities in Product Spaces *Publications Mathematiques de l’I.H.E.S.* **81** 73-205.
* Talagrand, M. (1996a): A New Look at Independence *Annals of Probability* **24** 1-34 (Special Invited Paper).
* Talagrand, M. (1996b): New Concentration Inequalities in Product Spaces *Inventiones Mathematicae* **126** 505-563.

**Local Rademacher Averages**

**Introduction**

1. Definition: Local Rademacher averages refers to Rademacher averages of subsets of function class determined by a condition on the variance of the function. Formally, the local Rademacher average at a radius

for the class is defined as

1. Importance of the Local Variance Based Families: The rationale for the above definition/construction is that, as just seen, the crucial ingredient for better rates of convergence is to use the variance of the function. Localizing the Rademacher average allows us to focus on that part of the function class where the fast rate phenomenon occurs, i.e., the functions with small variance.

**Star-Hull and Sub-root Functions**

1. Sub-root Function: A function

is a sub-root if:

* 1. is non-decreasing
  2. is non-negative, and
  3. is non-increasing.

These characteristics ensure that the sub-root function is continuous, and has a unique (non-zero) fixed-point satisfying

1. Star-Hull Definition: Let be a set of functions. Its star-Hull is defined as
2. Motivation Behind the Star-Hull Construction: By taking the star-Hull of a class of functions, we are guaranteed that the local Rademacher average behaves as a sub-root function, and thus has a unique fixed-point. This fixed point is a key quantity in relative error bounds. Formally, we start that, for any class of functions , is a sub-root.
3. Star-Hull Hypothesis Space Expansion: One way to estimate the enlargement of the size of the function class space is to compare the metric entropy (i.e., the log of the covering numbers) of and . It is possible to see that the metric entropy increases only by a logarithmic factor, which is essentially negligible.

**Local Rademacher Averages and Fixed Point**

1. Basic Theorem: Let be a class of bounded functions, i.e.

and be the fixed point of . There exists a constant

such that, with probability at least

If, in addition, the functions in satisfy

then one obtains, with probability at least

1. Proof Step #1 - Talagrand’s Inequality for Empirical Processes: The starting point is the Talagrand’s inequality for empirical processes, a generalization of the McDiarmid’s inequality of the Bernstein’s type (i.e., it includes the variance). This inequality tells us, that with high probability

for some .

1. Proof Step #2 - Split into Variance sub-groups: The second step consists in “peeling” the class, that is, splitting the classes into sub-classes according to the variance of the functions

Note that although depends on both the suprema across the expectation as well as the variance, here we classify the function space only on the variance.

1. Proof Step #3 - Apply Talagrand’s Inequality to each sub-group: We then apply the Talagrand inequality to each sub-group separately to obtain, with high probability
2. Proof Step #4 - Rademacher Bounding to the Variance Sub-group: Now use the symmetrization lemma to introduce local Rademacher averages. We then get, with high probability
3. Proof Step #5 - Express Local Rademacher Average in terms of the Fixed Point: We then “solve” the above inequality for . Things are simple if

behaves like a square root function, since we can upper bound the local Rademacher average by its fixed point value. With high probability, we then obtain

1. Proof Step #6 - Relation Between the Variance and the Expectation: Finally we obtain the relationship between the variance and the expectation

and “solve” the inequality in to get the original result.

**Local Rademacher Average – Consequences**

1. Fast Rate: An important example in this case is where the class is of finite VC dimension . In that case, one has

so that

As a consequence, under the Tsybakov noise condition, we obtain a rate of convergence of to as

It is important to note that, in this case, the rate of convergence of to is

So we obtain a fast rate by looking at the relative error. These fast rates can be obtained provided

but it is not necessary that

This requirement can be removed if one uses structural risk minimization or regularization.

1. Conditional Data Dependent Rademacher Averages: Another related result is that, as in the global case, one can obtain a bound with data-dependent (i.e., conditional) local Rademacher averages for

The result is the same as before (but with different constants) under the same conditions. With probability at least

where is the fixed-point of a sub-root upper bound of .

1. Conditional Local Rademacher Averages - Asymptotic Noise Behavior: Hence, we get improved rates when the noise is well-behaved, and these rates interpolate between and . However, in general it is not possible to estimate the parameters and that enter in the noise condition.
2. Local Rademacher Averages - Local Capacity Measure: Although the capacity measure used seems “local”, it does depend on all the functions in the class (through the variance), but each of them is appropriately re-scaled. Indeed, in , each function

with

is considered at a scale .

**Normalized ERM**

**Background**

1. The Concept: The main idea is to consider the ratio

where

and

1. Motivation: The motivation for considering the above normalized variant is that the fluctuations of the variants from are more “uniform”. Hence, the supremum in

is not necessarily attained at those functions whose variance is large, as in the previous case. Moreover, since we know that our goal is to find functions with small error (hence small variance), the normalized supremum takes this into account.

**Computing the Normalized Empirical Risk Bounds**

1. Statement: See Vapnik and Chervonenkis (1971) and Bousquet, Boucheron, and Lugosi (2003). For

with probability at least

and also with probability at least

1. Proof Step #1 - Using Symmetrization: The first step uses a variant of the symmetrization lemma:
2. Proof Step #2 - Use of Rademacher Variables: In this step we use randomization using Rademacher variables, i.e.
3. Proof Step #3 - Bernstein + Union Bounding: Finally, one applies the union bound in conjunction with a tail bound of the Bernstein type.

**Denormalized Bounds**

1. Denormalized Bound Estimation: From the fact that for positive, , , and ,

we get

1. Ideal, Noiseless Minimizer: In such a situation, there is no noise, i.e.

almost surely, and

Denoting by the empirical minimizer, we have

as well as

in this case. Thus, in the situation where is a function space with VC dimension we get

This clearly demonstrates that we de facto interpolate between the best rate of convergence

and the worst rate of convergence

(the factor cannot be removed in this case).

1. Corresponding ERM Bound: Casting

in the approach above, we get

thus when

and the convergence is , while if

the rate contains contribution from the term. Therefore it is not possible to obtain a rate with a power of between and .

1. Challenges with Denormalized Bounding: The main challenge with the Denormalized ERM approach is that the factor of the square root term (which arises out of the denormalization term ) is not a convenient entity to use here, since it does not vary with . However, if we use as the corresponding entity under the square root, this converges to zero with increasing (see later on the relative error classes). However, the denormalized approach cannot be used for , so we need another approach.

**References**

* Bousquet, O., S. Boucheron, and G. Lugosi (2003): Introduction to Statistical Learning Theory 169-207 *Advances in Machine Learning* (editors: Bousquet, O., U. von Luxburg, and G. Ratsch) **Springer** Berlin.
* Vapnik, V., and A. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16** 264-280.

**Noise Conditions**

**SLT Analysis Metrics**

1. Regression Function: Note that

We define the regression function as

and the target function (i.e., the Bayes’ classifier) as

1. Noise Level: We define the noise level as

In practice, a more useful definition (one that does not involve ) is

In the deterministic case

almost surely. Thus, the Bayes’ risk is

1. Noise Condition Motivation: To improve the treatment/results above, we seek a refinement that requires us to make assumptions about the noise function - in the ideal case

everywhere, thus

and

So we use the above-mentioned metrics to quantify how well-behaved the noise function is.

1. Range of the Noise and the Regression Function: From a classification point-of-view, the favorable classifier situation occurs when is not too close to zero. Indeed

means that the noise is a maximum at that and the corresponding

Here the label is completely undetermined, as any prediction results in an error with a probability of .

**Types of Noise Conditions**

1. Type 1 - Massart Noise Condition: For some , assume

almost surely ( may be viewed as a measure of the “odds against”). This condition implies that there is no region where the decision is completely random, i.e., that the noise is bounded away from .

1. Type 2 - Tsybakov Noise Condition: Let

and assume that one of the following equivalent conditions is satisfied:

This condition is the easiest to interpret, as it indicates that is close to the critical value with low probability.

1. Equivalence of the Tsybakov Conditions:
   1. Equivalence between a) and c) => It is easy to check that

For each function , there exists a set such that

* 1. Equivalence between b) and c) => Let

where

* 1. Equivalence between Tsybakov Conditions a) and c) => We write

Setting

finally gives

and setting

recovers a).

1. Restriction on the Range of : The parameter has to be in the range . Otherwise we end up with the opposite inequality

which is incompatible with condition a) if

(i.e., this is a special case).

1. Equivalence of the Tsybakov and the Massart Cases: When

the Tsybakov condition becomes invalid, and when

it is equivalent to the Massart’s condition.

**Relative Loss Class**

1. Motivation: The conditions above that we impose on the noise yield a crucial relationship between the variance and the expectation of functions within the so-called relative loss class (i.e., the difference between the sample-specific error and the empirical Bayes’ error) defined as

This relationship will allow exploiting the Bernstein-type inequalities applied to this latter class.

1. Massart and Tsybakov Noise Relative Error: Under Massart’s condition, one has

or equivalently, for

Under Tsybakov’s condition, this becomes, for

and for

1. Bernstein Inequality on Tsybakov Condition: In the finite case, with

one can apply the Bernstein inequality to and get (in conjunction with the finite union bound), with a probability of at least , for all

1. Bernstein-Tsybakov Bound - Case: When

is the minimizer of the empirical error and hence

one has

which is always better than for

and is valid even if

**Section IV: VC Theory and Capacity Measure Analysis**

**VC Theory and VC Dimension**

**Introduction**

1. Purpose: A form of the computational learning theory, VC theory aims to explain the learning process from a statistical point of view by applying the empirical processes independent of the probability distribution (VC Theory (Wiki)).
2. Components of the VC Theory: As detailed in Vapnik (1989, 2000), the VC theory covers 4 main parts:
   1. Theory of consistency of the learning process => What are the necessary and sufficient conditions for the consistency of a learning process based on empirical risk minimization?
   2. Non-asymptotic theory of the rate of convergence of a learning process => How fast is the rate of convergence of the learning process?
   3. Theory of control of the generalization ability of a learning process => How can one control the rate of convergence (i.e., the generalization ability) of the learning process?
   4. Theory of construction learning machines => How can one construct the algorithms that control the generalization ability?

**Empirical Processes**

1. Background: Let be the random elements defined on a measurable space . Define the empirical measure

where stands for the Dirac notation. Further, using the notation of van der Vaart and Wellner (2000), denote

for any probability measure .

1. The Empirical Measure Map: Let ℑ be a class of measurable functions

The empirical measure above induces a map from ℑ to given by

Notice the similarity between the success counting measure and the empirical loss function.

1. Empirical Process Theory: Let

where is the supremum operator on the set, representing the least upper bound of . Empirical process theory aims at identifying the classes ℑ for which the following statements hold. In both cases we assume that the underlying distribution of the data is unknown in practice.

i.e., this is the uniform law of large numbers

i.e., this is the generalized uniform central theorem limit.

1. Glivenko-Cantelli and -Donsker Classes: If the law of large numbers can be explicitly established across all elements of , the class is called Glivenko-Cantelli. If the validity of the central limit theorem under the assumption

can be established, then the class is called -Donsker. Obviously a -Donsker class is Glivenko-Cantelli in probability by the application of the Slutsky’s theorem.

1. Determination of : The main challenge here is to determine that satisfies the LLN and the CLT criteria bove (under regularity conditions) for all

Intuitively, cannot be too large, and the geometry of will play an important role.

1. Size of : One way of measuring how big is is by using covering numbers. The covering number is the minimum number of balls

needed to cover fully (obviously assuming there is an underlying norm on ). The entropy number is the logarithm of the covering number.

1. Sufficiency conditions for using Glivenko-Cantelli: A class is Glivenko-Cantelli if it is -measurable inside of an envelope ₣ such that

and satisfies

for every

1. Sufficiency Conditions for being -Donsker: This criterion is a version of the Dudley’s theorem. If belongs to the class of functions such that

then is -Donsker for every probability measure such that

**Bounding the Empirical Loss Function**

1. Symmetrization: The majority of treatments on how to bound empirical processes rely on symmetrization, application of the maximal concentration inequalities, and chaining. Symmetrization is usually the first step in these proofs, and since it is used in many machine learning proofs on bounding empirical loss functions (including the proof of VC inequality) we treat it in some detail.
2. The Empirical Process and its Symmetrized Counterpart: Consider the empirical process

Here we establish the connection between the empirical processes above and its symmetrized equivalent

where is an independent random variable that can be with a probability each. This symmetrized process, therefore, is a Rademacher process, conditional on the data . Therefore this process is a sub-Gaussian process by the Hoeffding’s inequality.

1. The Symmetrization Inequality: The Symmetrization bounds the LLN criterion of the empirical process theory. For every non-decreasing convex

and a class of measurable functions

1. Conceptual idea behind the proof by Symmetrization: The proof of the symmetrization lemma lies on introducing independent copies of the original variables (sometimes referred to as the ghost sample) and replacing the inner expectation of with these copies. After an application of Jensen’s inequality, different signs could be introduced (hence the name symmetrization) without changing the expectation.
2. Proof Steps:
   1. Introduction of the Ghost Sample => The empirical measure is replaced by – thereby providing the motivation for introducing the ghost sample variables as independent copies of . For fixed values of one has
   2. Apply Jensen’s Inequality => Apply the convex operator to both sides:

We are able to switch the locations of and , because is convex, and we have applied Jensen’s inequality to convert this to an inequality.

* 1. Take expectation with respect to :

Given that only is the variate on the RHS, the expectation simply reduces to on RHS (and on LHS).

* 1. Sign Perturbation Step => Note that adding a minus sign ahead of does not alter the RHS, because it is a symmetric function of and . Thus the RHS remains the same under sign perturbation:

for a random Rademacher sequence given as

Therefore

* 1. Triangle Inequality and Convexity of =>

Since the 2 right hand terms are equal, we get

1. Proving Empirical CLT’s Using Symmetrization: Proof for the CLT’s proceed analogously to the LNN. First use symmetrization to pass the process to , and then argue conditionally on the data using the fact that Rademacher processes are simple processes with nice properties.

**VC Dimension - Introduction**

1. Motivation: Although we have formulated the generalization bound in terms of the shattering coefficient, shattering coefficients in themselves are hard to evaluate. Another alternate capacity measure - the VC dimension – can be used to characterize the growth behavior of the shattering coefficient using a single number.
2. The VC Dimension: The VC Dimension of , , is defined as the largest number such that there exists a sample of size which is shattered by ; i.e.

If the maximum does not exist, the VC dimension is defined to be infinite. For other studies and many of the properties of the VC Dimension please refer to Cover (1965), Steele (1978), Dudley (1979), Wenocur and Dudley (1981), Assouad (1983), Khovanskii (1991), Anthony and Biggs (1992), MacIntyre and Sontag (1993), Kearns and Vazirani (1994), Goldberg and Jerrum (1995), Karpinski and MacIntyre (1997), Koiran and Sontag (1997), Anthony and Bartlett (1999), and Dudley (1999) demonstrate VC dimensions for several classes of functions.

**VC Dimension - Formal Definition**

1. Problem Space: Let be a concept class over an infinite space , i.e., a set of functions from to (both and may be infinite). For any

denotes the set of all labels or dichotomies on that are induced or realized by , i.e., if

then

therefore

For any natural number , we consider to be the number of ways to split points using concepts in , that is

1. Shattering of by : If

then is said to be *shattered* by .

1. VC Dimension of : The *Vapnik-Chervonenkis* dimension of , denoted as , is the cardinality of the largest set shattered by . If arbitrarily large finite sets can be shattered by , then

**VC Dimension Examples**

1. Computing the VC Dimension: In order to show that the VC Dimension of a class is at least , we must find some shattered set of size . In order to show that the VC Dimension is at most , we must show that no set of size is shattered.
2. VC Dimension Examples:
   1. Thresholds on a Number Line => Let be the concept class of thresholds on the number line. Clearly samples of size 1 can be shattered by this class. However, no sample of size 2 can be shattered, since it is impossible to choose a threshold such that is labeled positive and is labeled negative. Hence
   2. Intervals on a Real Line => Here a sample of size is shattered, but no sample of size is shattered, since no concept can satisfy a sample whose middle point is negative and the outer points are positive. Hence the
   3. Non-intersecting Intervals => Let be the concept class of non-intersecting intervals on the real line. A sample of size shatters (simply treating each pair of points as a separate case of the previous example), but no sample of size shatters, since if the sample points are alternated negative/positive, starting with a positive point, the positive points cannot all be covered by only intervals. Hence
   4. Linear Separators in => Here 3 points can be shattered, but not 4, so

In general, one can show that the of the class of linear separators in is .

* 1. Axis-aligned Rectangles => The class of axis-aligned rectangles in a plane has

The trick here is to note that for any collection of 5 points, at least one of them must be interior to, or on the boundary of any rectangle bounded by the other 4; hence if the bounding points are positive, the interior points cannot be made negative.

**VC Dimension vs Popper’s Dimension**

1. The Comparison: Corfield, Scholkopf, and Vapnik (2005) pointed out that the VC dimension is related to Popper’s notion of the dimension of a theory. They also highlight the differences between these approaches, including characterizing complexity using the number of parameters (e.g., the example of the class of thresholded sine waves in , Vapnik (1995)).
2. Popper’s Dimension of a Theory: As stated in Popper (1959), if there exists, for a theory , a field of singular (but not necessarily basic) statements such that, for some number , the theory cannot be falsified for any -tuple of the field, although it can be falsified by certain -tuples, then we call the characteristic number of the theory with respect to that field. All statements of the of the field whose degree of composition is less than or equal to , are then compatible with the theory, and permitted by it, irrespective of the content.

**References**

* Anthony, M., and N. Biggs (1992): *Computational Learning Theory* **Cambridge University Press**.
* Anthony, M., and P. Bartlett (1999): *Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge.
* Assouad, P. (1983): Densite et Dimension *Annales de l’Institut Fourier* **33** 233-282.
* Corfield, D., B. Scholkopf, and V. Vapnik (2005): Popper, Falsification, and VC Dimension *Technical Report TR-145* **Max Planck Institute for Biological Cybernetics**.
* Cover, T. (1965): Geometric and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition *IEEE Transactions on Electronic Computers* **14** 326-334.
* Dudley, R. (1979): Balls in do not cut all sub-sets of points *Advances in Mathematics* **31 (3)** 306-308.
* Dudley, R. (1999): *Uniform Central Limit Theorems* **Cambridge University Press** Cambridge.
* Goldberg, P., and M. Jerrum (1995): Bounding the Vapnik-Chervonenkis Dimension of Concept Classes Parametrized by Real Numbers *Machine Learning* **18** 131-148.
* Karpinski, M., and A. MacIntyre (1997): Polynomial Bounds for VC Dimension of Sigmoidal and General Pfaffian Neural Networks *Journal of Computer and System Science* **54**.
* Kearns, M., and U. Vazirani (1994): *An Introduction to Computational Learning Theory* **MIT Press** Cambridge, MA.
* Khovanskii, A. G. (1991): Fewnomials *Translations of Mathematical Monographs* **88** American Mathematical Society.
* Koiran, P., and E. Sontag (1997): Neural networks with Quadratic VC Dimension *Journal of Computer and System Science* **54**.
* MacIntyre, A., and E. Sontag (1993): Finiteness Results for Sigmoidal Neural Networks, in: *Proceedings of the 25th Annual ACM Symposium on the Theory of* Computing 325-334 **Associationof Computing Machinery** New York.
* Popper, K. (1959): *The Logic of Scientific Discovery* (Hutchinson, Translation of *Logik der Forschung (1934)*).
* Steele, J. (1978): Existence of sub-matrices with all possible Columns *Journal of Combinatorial Theory* **A28** 84-88.
* Van der Waart, A. W., and J. A. Wellner (2000): *Weak Convergence and Empirical Processes with Applications to Statistics 2nd Edition* **Springer**.
* Vapnik, V. N. (1989): *Statistical Learning Theory* **Wiley Interscience**.
* Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory* **Springer Verlag** New York.
* Vapnik, V. N. (2000): *The Nature of Statistical Learning Theory, 2nd Edition* **Springer Verlag**.
* Vapnik Chervonenkis Theory (Wiki): [*Wikipedia Entry for Vapnik Chervonenkis Theory*](http://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_theory).
* Wenocur, R., and R. Dudley (1981): Some special Vapnik-Chervonenkis Classes *Discrete Mathematics* **33** 313-318.

**Sauer Lemma and VC Classifier Framework**

**Motivation**

1. Growth Behavior of the VC Dimension: A combinatorial result proved simultaneously by Vapnik and Chervonenkis (1971), Sauer (1972), and Shelah (1972) characterizes the growth behavior of the shattering coefficient and relates it to the VC dimension – this is called the Sauer lemma. Related combinatorial results are available in Frankl (1983), Haussler (1995), Alekser (1997), Alon, Ben-David, Cesa-Bianchi, and Haussler (1997), Szarek and Talagrand (1997), and Cesa-Bianchi and Haussler (1998).
2. Sauer Lemma: Let be a function class with a finite VC dimension . Then

for all

In particular for all

we have

1. Nature of : is very similar to the concept of covering numbers (in all their variants), and Sauer’s lemma is an estimate of the bound of the shattering coefficient in terms of the sample size , and the VC dimension.
2. Implication of Sauer’s Lemma: Thus, for

the shattering coefficient behaves like a polynomial in the sample size . Once we know that the VC dimension is finite, the shattering coefficients grow polynomially, and this implies ERM consistency. The converse is true too.

1. ERM Consistency Statement: ERM is consistent with respect to if and only if is finite.

**Derivation of Sauer Lemma Bounds**

1. Statement: If

then for all

where

1. Shattering Cardinality Polynomial Limit: We estimate in the limit of . Note that

so

Thus

Thus for

we have

1. General Sauer Complexity Results: Sauer’s lemma can be used to get closed form expressions on sample complexity. For the treatment below, is an arbitrary hypothesis space, an arbitrary unknown fixed probability distribution, and an arbitrary unknown target function.
2. Shattering Cardinality Bound - Recap: For any

if we draw a sample from of size satisfying

then with probability of at least , all hypothesis in with

are inconsistent with the data, i.e.

1. Sample Size Bound: For any

if we draw a sample from of size satisfying

then with probability of at least , all hypothesis in with

are inconsistent with the data, i.e.

* 1. Step #1 => From the shatter cardinality bound we’ve just seen, we get
  2. Step #2 => Applying Sauer’s lemma, and recognizing that

we can reduce the inequality for to

* 1. Step #3 => We use Taylor expansion to estimate the bound . For

we get

so

* 1. Step #4 => Re-cast the above bounding for to get

**Sauer Lemma ERM Bounds**

1. Growth Function - Definition: The growth function is the maximum number of ways into which points can be classified by the function class :
2. Risk Bound Using Growth Function: For any , with probability at least
3. Bounding with Sauer’s Lemma: Let be a class of functions with a finite VC dimension . Then, for all

and for all

This implies, that with probability at least

Thus, this indicates that the difference between the TRUE and the EMPIRICAL risk is at most of the order of

1. Behavior of the Growth Function: The growth function, which is exponential in the sample size as long as

where is the VC Dimension, becomes polynomial in for

**VC Index**

1. Motivation and Setup: Consider a collection of subsets of the sample space . The collection of sets is said to pick out a certain subset of finite set

if

from

is said to shatter if it picks out each of its subsets. The VC-index of is the smallest number for which NO SET of size is shattered by (in fact, the VC-index is similar the for an appropriately chosen classifier set).

1. Bounds on the VC Class Subset Number: Sauer’s lemma then states that the number of subsets picked out by a VC class satisfies

This is polynomial in the number of subsets (i.e.,) rather than exponential. Intuitively, this means that finite VC index implies that has an apparent simplistic structure.

1. VC Sub-graph Bounds: A similar bound can be shown (for a different constant, same polynomial order) for the so-called VC subgraph classes. For a function

the sub-graph is a subset of χ such that

A collection of is called a VC sub-graph class if all the sub-graphs form a VC-class.

1. Covering Number for the Loss Function: Consider a set of indicator functions

in for the discrete empirical measure (or any type of probability measure - discrete or not). It can be shown that for any

1. Entropy Number for the Symmetric Convex Hypotheses Hull: Next consider the symmetric convex hull of a set , which is a collection of functions of the form

with

Then

1. Convergence of the Convex Hull: The most important consequence of the above is that the power of - is strictly less than 2, which is just enough so that entropy integral is guaranteed to converge, therefore the class is going to be -Donsker.
2. VC Index of a Sub-graph Class: Any finite dimensional vector space of measurable functions

is a VC-subgraph of index smaller than or equal to .

1. Generalizations of VC Sub-graph: There are generalizations of the notion of VC subgraph class, e.g., there is the notion of pseudo-dimension (Bousquet, Boucheron, and Lugosi (2004)).

**VC Classifier Framework**

1. Introduction: Let χ be a feature space and

The function

is called the classifier. Similar to before, we define the shattering coefficient (also referred to as growth coefficient) as

1. Shattering Coefficient Relation: In terms of the previous section the shattering coefficient is precisely

with being a collection of all the sets as laid out above. Further, using Sauer’s lemma, it can be shown that is going to be a polynomial in provided that class has a finite VC dimension, or equivalently, the collection has a finite VC index.

1. VC Indicator Empirical Risk: Let

be the data set. As always, we assume that is unknown, thus we work on bounding the indicator empirical risk:

1. VC Inequality Bounds: For binary classification and loss function, we have the following generalization bounds:
2. Implication of the VC Inequality Bounds: The impact of the VC inequality relation is that, as the sample size increases, provided has a finite VC dimension, the empirical risk becomes a good proxy for the expected risk. Note that both RHS of the two inequalities will converge to 0 provided grows polynomially in .
3. Connection between the Classifier Framework and the Empirical Process Framework: With the empirical classifier framework, on deals with the modified empirical process , but the other components are the same. As before the proof of the first VC inequality relies on symmetrization, and then argue conditionally using the data with the concentration inequalities – in particular the Hoeffding’s inequality (Bousquet, Boucheron, and Lugosi (2004)).

**References**

* Alekser, S. (1997): A Remark on the Szarek-Talagrand Theorem *Combinatorics, Probability, and Computing* **6** 139-144.
* Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997): Scale-sensitive Dimension, Uniform Convergence, and Learnability *Journal of the ACM* **44** 615-631.
* Bousquet, O., S. Boucheron, and G. Lugosi (2004): Introduction to Statistical Learning Theory *Advanced Lectures in Machine Learning Lecture Notes in Artificial Intelligence - Bousquet, von Luxburg, and Ratsch (editors)* **3176** 169-207.
* Cesa-Bianchi, N., and D. Haussler (1998): A Graph-Theoretic Generalization of the Sauer-Shelah Lemma *Discrete Applied Mathematics* **86** 27-35.
* Frankl, P. (1983): On the Trace of Finite Sets *Journal of Combinatorial Theory* **A34** 41-45.
* Haussler, D. (1995): Sphere-packing Numbers for the Boolean n-cube with bounded Vapnik-Chervonenkis Dimension *Journal of Combinatorial Theory* **A69** 217-232.
* Sauer, N. (1972): On the Density of Families of Sets *Journal of Combinatorial Theory (A)* **13** 145-147.
* Shelah, S. (1972): A Combinatorial Problem: Stability and Orders for Models and Theories in Infinitary Languages *Pacific Journal of Mathematics* **41** 247-261.
* Szarek, S., and M. Talagrand (1997): On the Convexified Sauer-Shelah Theorem *Journal of Combinatorial Theory* **B69** 183-192.
* Vapnik, V., and A. Chervonenkis (1971): On the Uniform Convergence of Relative Frequencies of Events to their Probabilities *Theory of Probability and its Applications* **16** 264-280.

**Section V: Capacity/Complexity Estimation Using Covering Numbers**

**Covering and Entropy Numbers**

**Motivation**

1. Alternate Generalization Bounds: In the literature there exist many more capacity concepts, and they all assume the following form: with a probability of at least

In the simplest case, the capacity term only depends on and the confidence term on the underlying probability. Obviously, as is to be expected, these bounds are worst case bounds on *bad behavior*.

1. Shortcoming of the VC Measure: Of the three capacity measures seen so far – the VC dimension, growth function, and the VC entropy, all three are usually hard/impossible to compute. There are other measures which not only give sharper estimates, but also have properties that make their computation possible from data only.
2. Capacity/Complexity Measures: Some capacity/complexity measures are:
   1. Covering Numbers (all variants)
   2. VC Dimension
   3. Rademacher Complexity

**Nomenclature - Normed Spaces**

1. Spaces: For

denotes the -dimensional space of vectors

We define spaces as follows: as vector spaces, they are identical to , in addition, they are endowed with -norms; for

and for

Note that a different normalization of the norm is used in some papers in learning theory (e.g., Talagrand (1996)). For

1. of with respect to : Given points

we use the short-hand

Suppose is a class of functions

The norm with respect to of

is defined as

Likewise

1. Integral Norm over a -algebra Space: Given some set with a -algebra, a measure on , some

and a function

we define

if the integral exists, and

For

we let

We also let

**Covering, Entropy, and Dyadic Numbers**

1. -Covering Numbers: If is a set and is a metric on , then the -covering number of

with respect to the metric denoted is the smallest number of elements of an -cover for using the metric .

1. Entropy Number of a Set : Given a metric space

we also write the covering number as . The entropy number of a set

for

is

1. Operator Norm: Let be the set of all bounded operators between the normed spaces and , i.e., operators such that the image of the closed unit ball

is bounded. The smallest such bound is called the *Operator Norm*.

1. Entropy Numbers of an Operator: The *Entropy Numbers of an Operator*

are defined as

Note that

and that is certainly well-defined for all

if is a compact operator, i.e., for any

there exists a finite cover of with open balls over .

1. Dyadic Entropy Numbers of an Operator: The dyadic entropy numbers of an operator are defined by

Similarly, the dyadic entropy numbers of a set are defined from its entropy numbers. A very nice introduction to entropy numbers of operators is found in Carl and Stephani (1990).

1. Banach Spaces: and will always be Banach spaces - for instance, spaces with

In some cases they may be *Hilbert Spaces* , Banach spaces endowed with a dot-product giving rise to its norm via

**Background and Overview of Basic Results**

1. Setup: We start by endowing a function class ℑ with the following random metric:

This is the normalized Hamming distance of the projections on the sample. Given such a metric, we say that *covers* at a radius if

The term *cover* is used analogous to the term *space* in function space.

1. Definition: The covering number of at a radius with respect to , denoted by , is the minimum size of a cover of radius . Note that it does not matter if we apply this definition to the original class or the loss class , since
2. Growth of the Covering Function: The covering numbers characterize the size of the function class measured by the metric . The rate of growth of the logarithm of , called the metric entropy, is related to the classical concept of the vector dimension. Indeed, if is a compact set in a -dimensional Euclidean space
3. Finite Covering Numbers: When the covering numbers are finite, it is possible to approximate the class by a finite set of functions (which cover ). This again allows the use of finite union bound, provided we can relate the behavior of all functions in to that of the functions in the cover.
4. Finite Covering Number Sample Bound:
5. Covering Numbers for Real Valued Functions: Covering Numbers can also be defined for real-valued functions – see treatment below.
6. Covering Numbers and the VC Dimension: Notice that, if the functions in can only take 2 values, for all

Hence the VC entropy corresponds to the log covering number at the minimal scale, which implies

but one can have a considerably better result.

1. Haussler’s Bound: Let be a class of VC dimension . Then for all

all , and any sample

The interesting aspect of this bound is that it does not depend on the sample size .

1. Principle behind Covering Bound Improvement: The covering bound is a generalization of the VC entropy bound where the scale is adapted to the error. The result can be considerably improved by considering all scales.

**References**

* Carl, B., and I. Stephani (1990): *Entropy, Compactness, and the Approximation of Operators* **Cambridge University Press** Cambridge UK.
* Talagrand, M. (1996): The Glivenko-Cantelli Problem, 10 years later *Journal of Theoretical Probability* **9 (2)** 371-384.

**Covering Numbers for Real-Valued Function Classes**

**Introduction**

1. Literature: Covering numbers for functions have been extensively studied in a variety of literature dating way back to Kolmogorov (1956), Kolmogorov and Tihomirov (1961)). They play a central role in a number of areas of information theory and statistics, including density estimation, empirical processes, and machine learning (Pollard (1984), Birge (1987), Haussler (1992)).
2. Definition: Let be a sub-set of a metric space . For a given

the metric covering number is defined as the smallest number of sets/balls of radius whose union contains . In what follows, is omitted if the context is clear.

1. Functions of Bounded Variation under the Metric: Here our intention is to find bounds on the covering numbers of functions of bounded variations under the metric. Specifically, let be the set of all functions on taking values in with total variation of at most

It is natural to use the same for both the range and the variation, since a bound on the variation of a function implies a bound on its range.

1. Tight Bounds on : Our goal is to find tight bounds on in terms of the relevant constants. This is related to the problem of density estimation, where attention has been given to the problem of finding covering numbers for classes of densities that are unimodal or non-decreasing (Groeneboom (1986), Birge (1987)).
2. Density Estimation using Covering Numbers under Constraints: Treatment here is a super-set of the treatment in Birge (1987), since we do not impose a density constraint on the class, which accounts for the difference in function behavior as a function of the parameters. In fact, it is this density constraint that accounts for the constant in Birge (1987) rather than that we obtain here.
3. Covering Numbers for General Classes of Real-valued Functions: Here we also investigate the metric covering numbers for the general classes of real-valued functions under the family of metrics, where is a probability distribution. Upper bounds in terms of the Vapnik-Chervonenkis dimension or the pseudo-dimension of the function class were first obtained by Dudley (1978), improved in Pollard (1984), and further by Haussler (1992, 1995). Various lower bounds have also been obtained (e.g., Kulkarni, Mitter, and Tsitsiklis (1993)).
4. Scale-Sensitive Covering Number Bounds: The importance of scale-sensitive versions of several combinatorial dimensions in learning problems may be seen from Alon, Ben-David, Cesa-Bianchi, and Haussler (1993), and Bartlett and Long (1995). Using techniques due to Haussler and results from Alon, Ben-David, Cesa-Bianchi, and Haussler (1993), Lee, Bartlett, and Williamson (1995) proved an upper bound on

in terms of the scale-sensitive dimension of the function class. The treatment here improves the result and provides a lower bound. As will be shown, in general the bounds presented here cannot be significantly improved.

**Functions of Bounded Variation**

1. Upper/Lower Bounds - Statement: For all

(Bartlett, Kulkarni, and Posner (1997)).

1. Special Case of the Metric: Certainly under the metric, the classes of functions of bounded variation (or even the subset of functions under this class that are continuous) are not compact, hence

However, it has been established that the class of function of bounded variation that are also Lipschitz-smooth have covering numbers of the order of in the metric (Lorentz (1966)).

1. Function Bound vs. Class Bound: Let be all functions of variation at most that map from to for some

The theorem above can be extended to as

(Bartlett, Kulkarni, and Posner (1997)). Both upper and lower bounds are obtained by considering vertical shifts in and using the proof approach outlined below.

1. Distribution-Dependent Bounds: The upper bound (with set to ) can be extended to give the upper bound

(Bartlett, Kulkarni, and Posner (1997)), i.e., a uniform bound on the covering numbers for all weighted norms where is an arbitrary probability distribution in . The proof for this is obtained by modifying the proof presented below for an equi-probable partition of rather than a partition of equal size.

**Functions of Bounded Variation – Upper Bound Proof**

1. -Covering of the Function Class: Define to be the class of all non-decreasing functions on taking values in . Let be any -covering of . It is well-known that for any

there exist

such that

(e.g., Kolmogorov and Fomin (1970)).

1. Range of the -cover Class: More precisely, if denotes the total variation over the interval , then and can be taken to be

and

It is easy to show that both and are non-decreasing. Also, if takes values in , and has total variation bounded by , then both and take values only in the range .

1. Metric Covering Number of the -cover Set: By then definition of the cover of a set, there exist

such that

This gives

Thus, we can produce -covering of by taking all pairs from

Hence

1. Cardinality of the Equi-Partition Set: We now find an upper bound on by producing an -covering of . Partition into

sub-intervals of length , i.e.,

Let be the set of all functions that are constant on these sub-intervals, non-decreasing, and taking values only in the set

where

It is easy to see that cardinality of , denoted by , satisfies

1. Metric Covering Number - Upper Bound: Now, note that for any

there exists a

such that

To see this, consider the error to the best constant approximation to on a sub-interval, and the additional error introduced by quantizing the range. Choosing

and

gives

Hence, with this choice we get

for

By using the -covering bound of from , we get

**Functions of Bounded Variation – Lower Bound Proof**

1. Partition of : Partition into segments. Take the set of all the binary -valued functions constant over each segment. If

the bounded variation and the boundedness constraints impose

1. The Difference Function: There are functions in the set . For any two functions

define as the number of segments on which the functions have different values. It is then easy to see that

1. Bounding the Differences Function: Thus

are close if

For an arbitrary

let be the number of functions in that are -close to . Then

where denotes the largest number no bigger than . The Chernoff-Okamoto inequality (Dudley (1978)) is

for

and

Letting

we get that

The same result can also be obtained by using Hoeffding’s inequality.

1. Optimal Lower Bound: Since the cardinality of is , it is clear that we need at least functions for a -cover, i.e.

To obtain a large lower bound we want to maximize this expression subject to

Equivalently

If

we may choose

to show that this maximum is at least

Thus the lower bound on the covering number is

**General Function Classes**

1. Introduction: Here we provide lower and upper bounds on the quantity

for the general classes of -valued functions defined on the set , where the is taken over all the probability distributions on . The bounds are given in terms of the scale-sensitive dimension (Alon, Ben-David, Cesa-Bianchi, and Haussler (1993)).

1. Fat-Shattering Coefficient: Define

where a sequence

is -shattered by if there is a sequence

such that, for all

there is an

with

For example, it is easy to show that, bounded variation functions have

To see the upper bound, consider an -shattered sequence, and for a suitable sequence , find the sequence for which the corresponding function has maximum variation. An easy construction with

gives the lower bound.

1. Upper/Lower Bounds - Statement: There are constants and such that, for any permissible (i.e., benign measurability condition (Pollard (1984))) class of -valued functions defined on a set

(Bartlett, Kulkarni, and Posner (1997)).

1. Gap Between Upper and Lower Bounds: There is a gap between the lower and the upper bounds. The class of bounded variation functions shows that the lower bound in tight within a constant factor. The following example shows that some gap between the lower and the upper bounds is essential. For any positive integers , let be the class of all functions from to , and let be the uniform distribution on . Then

for

yet for

we have

1. Function Range Scaling: Clearly, the statement above can be trivially extended to classes of functions that map to an arbitrary interval by scaling by a factor .
2. Generalization of the Statement: Cesa-Bianchi and Haussler (1993) proved a version of the above statement in a considerable more general setting. Their results provides bounds on covering numbers of a class of functions that take values in an arbitrary totally bounded metric space, in terms of a scale-sensitive dimension of the class. In the special case of real-valued functions, their scale-sensitive dimension is different from that considered here, although the lemmas in Anthony and Bartlett (1995) show that the two quantities are within log factors of each other.

**General Function Class Bounds – Proof Sketch**

1. Lemma #1 of 3 - Bounds on : In the first step, we show that a bound on implies that, for any finite sequence

from , there is a small subset of that is a cover of the restriction of to , denoted

In defining the cover and the covering numbers , we use the scaled metric on defined by

We can also consider as a class of functions from to ; this is how we define .

1. Lemma #2 of 3 - Intra-Class Cover: In the second step, we use the results of the first lemma to deduce that there is a small cover for the set of absolute differences between the functions in .
2. Lemma #3 of 3 - Uniform Convergence: The third lemma implies that the above 2 indicate a uniform convergence result for this set. We use this result to show that, for some sequences of positive probability , the cover for the restriction of to the sequence induces a cover for .

**General Function Class Bounds – Lemmas**

1. Lemma #1 - Statement: Suppose

and is a set of -valued functions on . Let

where

If

there is a subset of for which is an -cover of , and

1. Lemma #1 - Proof: This lemma is implicit in Bartlett and Long (1995), which provides a slightly better bound - by a factor of - than that given in Alon, Ben-David, Cesa-Bianchi, and Haussler (1993).
2. Lemma #2 - Statement: For a class of -valued functions, let

Then, with the metric on

1. Lemma #2 - Proof: Take an -cover for . Then for all

pick

within of and , respectively. We have

It follow that

is an -cover for .

1. Lemma #3 – Statement: For a permissible class of -valued functions on a set , a probability distribution on , and an
2. Lemma #3 - Consequence: This lemma, which provides the uniform convergence property, is due to Pollard (1984). Haussler (1992) provides a related result with different constants. The class that we will consider is .

**General Function Class – Upper Bounds**

1. Reduction of the Probability Bound: We start by establishing the following chain of inequalities:
   1. Use Lemma #3 to get
   2. Use Lemma #2 to get
   3. Finally use Lemma #1 for

and

to get

1. Sample Size Bounds: It is easy to show that the probability is less than for

where is a constant. In fact, a more detailed estimation will show that

will be sufficient.

1. Cover for the Restriction of : For the above sample size, it follows that there is a

such that any

for which is an –cover for is also an -cover for . To see this, notice that for all

there is a

with

and if is chosen in the complement of the set that generates above, then

so that

In fact, for sufficiently large , a proper cover for the restriction of to almost any -sequence induces a cover of .

1. The Upper Bound - Final Form: Together with Lemma #1, the above statement shows that some

satisfies

from which the result follows.

**General Function Class – Lower Bounds**

1. The Lower Bound: Suppose

Then consider the uniform distribution on a -shattered set of size , and consider the restriction of the class to this set. The same argument as the proof for the lower bound for the functions of bounded variation gives

**References**

* Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1993): Scale-sensitive Dimensions, Uniform Convergence, and Learnability *Proceedings of the ACM Symposium on Foundations of Computer Science*.
* Anthony, M., and P. L. Bartlett (1995): Function Learning from Interpolation *Computational Learning Theory: Proceedings of the 2nd European Conference, EuroCOLT ’95* 211-221 **Springer**.
* Anthony, M. (1997): Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC Model and its Variants *Neural Computational Survey* **1** 1-47.
* Bartlett, P. L., and P. Long (1995): More Theorems about Scale-sensitive Dimensions and Learnability *Proceedings of the 8th Annual Conference on Computational Learning Theory* 392-401 **ACM Press**.
* Bartlett, P. L., S. R. Kulkarni, and S. E. Posner (1997): Covering Numbers for Real-Valued Function Classes *IEEE Transactions on Information Theory* **43 (5)** 1721-1724.
* Birge, L. (1987): Estimating a Density under Order Restrictions: Non-asymptotic Minimax Risk *Annals of Statistics* **15** 995-1012.
* Cesa-Bianchi, N., and D. Haussler (1996): A Graph-theoretic Generalization of the Sauer-Shelah Lemma *Internal Report 170 96* **DSI Universita di Milano**.
* Dudley, R. M. (1978): Central Limit Theorems for Empirical Measures *Annals of Probability* **6 (6)** 899-929.
* Groeneboom, P. (1986): *Some Current Developments in Density Estimation* **CWI Monographs** North-Holland.
* Haussler, D. (1992): Decision-theoretic Generalizations of the PAC Model for Neural Net and other Learning Applications *Information and Computation* **100** 78-150.
* Haussler, D. (1995): Sphere Packing Numbers for the Subsets of the Boolean n-cube with Bounded Vapnik-Chervonenkis Dimension *Journal of Combinatorial Theory A* **69 (2)** 217.
* Kolmogorov, A. N. (1956): Asymptotic Characteristics of some Completely Bounded Metric Spaces *Dokl. Acad. Nauk. SSSR* **108** 585-589.
* Kolmogorov, A. N., and V. M. Tihomirov (1961): -entropy and -capacity of Sets in Function Spaces *Amer. Math. Soc. Transl. (2)* **17** 277-364.
* Kolmogorov, A. N., and S. V. Fomin (1970): *Introductory Real Analysis* **Dover** New York.
* Kulkarni, S. R., S. K. Mitter, and J. N. Tsitsiklis (1993): Active Learning using Binary-valued Queries *Machine Learning* **11** 23-35.
* Lee, W. S., P. L. Bartlett, and R. C. Williamson (1995): On Efficient Agnostic Learning of Linear Combinations of Basis Functions *Proceedings of the 8th Annual Conference on Computational Learning Theory* 369-376 **ACM Press**.
* Lorentz, G. G. (1966): Metric Entropy and Approximation *Bull. Amer. Math. Soc.* **72** 903-937.
* Pollard, D. (1984): *Convergence of Stochastic Processes* **Springer** New York.

**Operator Theory Methods for Entropy Numbers**

**Introduction and Setup**

1. Classical Statistical Theory Viewpoint: Under the traditional viewpoint of the statistical learning theory, one is given a class of functions , and the generalization performance attainable using is determined via the covering numbers of .
2. Uniform Covering Numbers of : More precisely, for some set , and

for

define *Uniform Covering Numbers* of the function class on by

where is the -covering number of with respect to . Recall

Many generalization bounds can be expressed in terms of .

1. Combinatorial Dimension Bounding of Entropy Numbers: Traditionally, lemmas such as the Sauer lemma have been used in function learning to extract dimensions such as VC-dimension of real-valued functions, a variation due to Pollard called pseudo-dimension, or a scale-sensitive dimension thereof.
2. Bounding Covering Numbers: These dimensions reduce the computation of to that of a single “dimension” like quantity independent of . An overview of these various dimensions, details of their history, and examples of their computation can be found in Anthony (1997), and Anthony and Bartlett (1999).
3. Information Theory Viewpoint:
   1. The Kernel Operator => An alternate view is that of a function class being induced by an operator that depends on some kernel function - thus is the image of a base class under .
   2. Covering Number in terms of Properties => Thus, the determination of can be done in terms of the properties of the operator , as the latter plays a constructive role in controlling the complexity of , and hence the difficulty of the learning task (Smola, Williamson, Mika, and Scholkopf (1999), Williamson, Shawe-Taylor, Scholkopf, and Smola (1999), Smola, Elisseff, Scholkopf, and Williamson (2000), Williamson, Smola, and Scholkopf (2000, 2001). Often the determination is done in terms of packing numbers in place of covering numbers – however, they are equivalent up to a factor of 2 (Anthony and Bartlett (1999)).

**Literature, Approaches, and Result**

1. Concept of Metric Entropy: The concept of metric entropy of a set has been around for some time. It seems to have been introduced by Pontriagin and Schnirelmann (1932) and was studied in detail by Kolmogorov and others (Kolmogorov and Tihomirov (1961), Lorentz, van Golitschek, and Makovoz (1996)).
2. Entropy Numbers and Linear Operator Spectrum: The use of metric entropy to say something about linear operators was carried out independently by several people. Prosser (1966) seems to have been the first to make the idea explicit.
3. Entropy Rates and Eigenvalue Asymptotics: In particular, Prosser (1966) proved a number of results concerning the asymptotic rate of decrease of the entropy numbers in terms of the asymptotic behavior of the eigenvalues. A similar result is implicit in Shannon’s famous paper (Shannon (1948)) where he considered the effect of different convolutional operators on the entropy of an ensemble. Prosser’s work led to several studies on convolutional operators (Prosser and Root (1968), Jagerman (1969), Akashi (1990), Koski, Persson, and Peetre (1994)).
4. -entropies of Linear Operators and Stochastic Processes: The connection between Prosser’s -entropy of an operator and Kolmogorov’s -entropy of stochastic processes was established in Akashi (1986). Additional work studied covering numbers (Triebel (1970), Carl and Stephani (1990)) and entropy numbers in the context of operator ideals (Pietsch (1980)), Carl (1981)).
5. Banach Spaces and Uniform Convergence: Connections between the local theory of Banach spaces and uniform convergence of empirical means has been noted before (e.g., Pajor (1985)). Gurvits (1997) related the Rademacher-type of a Banach space to the fat-shattering dimensions of linear functionals on that space, and hence via the key result in Alon, Ben-David, Cesa-Bianchi, and Haussler (1997) to the covering number of the induced class (Williamson, Scholkopf, and Smola (1999)).
6. Operator Type and Entropy Number Decay: The equivalence of the type of an operator (or of the space that it maps to), and the rate of decay of its entropy numbers has been established independently by Koltchinskii (1988, 1991), Defant and Junge (1993), and Junge and Defant (1993) (albeit under different formulations – Koltchinskii’s work is motivated by probabilistic considerations).

**References**

* Akashi, S. (1986): Characterization of -entropy in Gaussian Processes *Kodai Mathematical Journal* **9** 58-67.
* Akashi, S. (1990): The Asymptotic Behavior of -entropy of a Compact Positive Operator *Journal of Mathematical Analysis and Applications* **153** 250-257.
* Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997): Scale-sensitive Dimensions, Uniform Convergence, and Learnability *Journal of the Association of the Computing Machinery* **44 (4)** 615-631.
* Anthony, M. (1997): Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC Model and its Variants *Neural Computational Survey* **1** 1-47.
* Anthony, M., and P. L. Bartlett (1999): *Artificial Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge UK.
* Carl, B. (1981): Entropy Numbers of Diagonal Operators with an Application to the Eigenvalue Problems *Journal of Approximation Theory* **32** 135-150.
* Carl, B., and I. Stephani (1990): *Entropy, Compactness, and the Approximation of Operators* **Cambridge University Press** Cambridge UK.
* Defant, M., and M. Junge (1993): Characterization of Weak Type by the Entropy Distribution of -nuclear Operators *Studia Math.* **107 (1)** 1-14.
* Gurvits, L. (1997): A Note on Scale-sensitive Dimension of Linear Bounded Functionals in Banach Spaces, in: *Algorithmic Learning Theory ALT-97 (Lecture Notes in Artificial Intelligence)* (M. Li and A. Marouka, editors) **1316** 352-363 **Springer-Verlag** Berlin, Germany.
* Jagerman, D. (1969): -entropy and Approximation of Band-limited Functions *SIAM Journal of Applied Mathematics* **17 (2)** 362-377.
* Junge, M., and M. Defant (1993): Some Estimates of Entropy Numbers *Israeli Journal of Mathematics* **84** 417-433.
* Kolmogorov, A. N., and V. M. Tihomirov (1961): -entropy and -capacity of Sets in Function Spaces *Amer. Math. Soc. Transl. (2)* **17** 277-364.
* Koltchinskii (1988): Operators of Type and Metric Entropy *Teoriya Veroyatnosteyi Matematicheskaya Statistika* **38** 69-76, 135.
* Koltchinskii (1991): Entropic Order of Operators in Banach Spaces and the Central Limit Theorem *Theory of Probability and Its Applications* **36 (2)** 303-315.
* Koski, T., L. E. Persson, and J. Peetre (1994): -entropy, -rate, and Interpolation Spaces revisited with an Application to Linear Communication Channels *Journal of Mathematical Analysis and Applications* **186** 265-276.
* Lorentz, G. G., M. van Golitschek, and Y. Makovoz (1996): *Constructive Approximation: Advanced Problems* **Springer-Verlag** Berlin Germany.
* Pajor, A. (1985): *Sous-Espaces des Espaces de Banach* **Hermann** Paris, France.
* Pontriagin, L. S., and L. G. Schnirelmann (1932): Sur une Propriete Metrique de la Dimension *Annals of Mathematics* **33** 156-162.
* Pietsch, A. (1980): *Operator Ideals* **North-Holland** Amsterdam.
* Prosser, R. T. (1966): The -entropy and -capacity of certain Time Varying Channels *Journal of Mathematical Analysis and Applications* **16** 553-573.
* Prosser, R. T., and W. L. Root (1968): The -entropy and -capacity of certain Time Invariant Channels *Journal of Mathematical Analysis and Applications* **21** 233-241.
* Shannon, C. E. (1948): A Mathematical Theory of Communication *Bell System Technical Journal* **27** 379-423, 623-656.
* Smola, A. J., R. C. Williamson, S. Mika, and B. Scholkopf (1999): Regularized Principal Manifolds, in: *Proceedings of the 4th European Workshop on Computational Learning Theory (EUROCOLT ’99)* 214-229.
* Smola, A. J., A. Elisseff, B. Scholkopf, and R. C. Williamson (2000): Entropy Numbers for Convex Combinations and mlps, in: *Advances in large Margin Classifiers* (A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors) **MIT Press** Cambridge MA.
* Triebel, H. (1970): Interpolationseigenschaften von Entropie- und Durchmesseridealen Kompakter Operatoren *Studia Math.* **34** 89-107.
* Williamson, R. C., B. Scholkopf, and A. J. Smola (1999): [A Maximum Margin Miscellany](http://www.citeulike.org/user/mlbook/article/619676).
* Williamson, R. C., J. Shawe-Taylor, B. Scholkopf, and A. J. Smola (1999): Sample-based Generalization Bounds *NeuroCOLT2 Technical Report Series* **NC-TR-1999-055**.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2000): Entropy Numbers of Linear Function Classes, in: *Proceedings of the 13th Annual Conference on Computational Learning Theory* *(N. Cesa-Bianchi and S. Goldman, editors)* **ACM** New York.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2001): Generalization Performance of Regularization Networks and Support Vector Machines via Entropy Numbers of Compact Operators *IEEE Transactions on Information Theory* **47 (6)** 2516-2532.

**Generalization Bounds via Uniform Convergence**

**Basic Uniform Convergence Bounds**

1. Generalization Performance of Learning Machines: The generalization performance of learning machines can be bounded via uniform convergence results presented in Vapnik and Chervonenkis (1981) and Vapnik (1982) (Anthony (1997) and Kulkarni, Lugosi, and Venkatesh (1998) contain reviews). The capacity of the hypothesis class is often expressed in terms of the covering numbers.
2. Classification and Regression Covering Numbers: Results for both classification and regression are known. For the sake of concreteness, below is a quote suitable for regression, proved in Alon, Ben-David, Cesa-Bianchi, and Haussler (1997). Bartlett and Shawe-Taylor (1999) contain results of classifier performance in terms of covering numbers. We first set

to denote the *empirical means* of on the sample .

1. Regression Uniform Convergence Bound: Let be a class of functions from to , and let be a distribution over . Then for all

and

where denotes the probability with respect to the sample drawn i.i.d. from , and the expectation with respect to the first, and a second sample also drawn i.i.d. from , i.e.

(Alon, Ben-David, Cesa-Bianchi, and Haussler (1997)).

1. Use of the Regression Bound Lemma: To be able to use the above lemma, one usually makes use of the fact that for any

**Loss Function Induced Classes**

1. Application to Loss Function: The above result can be used to give a generalization error result by applying it to the loss-function-induced class. The following Lipschitz-condition lemmas on loss function, which are an improved version of the lemmas in Bartlett, Long, and Williamson (1996), is useful in this regard (a similar result appears in Anthony and Bartlett (1999)).
2. Loss Function Lemmas - Background: Let be a class of functions from to , with

and

a loss function. Let the following conditions/definitions hold:

1. Lipschitz Loss Condition Lemma: Suppose satisfies the Lipschitz condition

for all

Then for all

and

1. Approximate Lipschitz Loss Condition Lemma: Suppose that for some

satisfies the “approximate Lipschitz condition”

for all

then for all

1. Lipschitz Loss Lemma Proof: We show that, for any sequence of pairs in

and any functions and , if the restrictions of and to are close, then the restrictions of and to are closer. Thus, given a cover of we can construct a cover of that is no bigger.

1. Lipschitz Loss Condition Bound:
2. Approximate Lipschitz Loss Condition Bound: This condition is useful when the exact form of the loss function is unknown, happens to be discontinuous, or is badly behaved in some other way

for

1. Polynomial Loss Function Bound: For either of the Lipschitz loss conditions above, if we assume a loss function of the type

with

we have

in particular

for

and therefore

**Standard Form of Uniform Convergence**

1. The Standard Form: One can readily combine the uniform convergence results with the above results to get overall bounds on generalization performance. A typical uniform convergence result takes the form

where is the empirical risk, and is the expected risk of

(Vapnik (1998), Anthony and Bartlett (1999)).

1. Exponents in the Standard Form: The exponent depends on the setting. For regression, can be set to . However, in agnostic learning in general

(Kearns, Schapire, and Sellie (1994)), except if the class is convex, in which case can be set to (Lee, Bartlett, and Williamson (1998)).

1. Sample Complexity: The generalization bounds produced above are typically used by setting the right hand side to , and solving for

is called the sample complexity.

1. Learning Curve Bound: Another way to use the above result is as a learning curve bound , where

It is to be noted that the determination of is quiet convenient in terms of , the dyadic entropy number associated with the covering number .

1. The Dyadic Entropy Number: Setting the right hand side in the standard uniform convergence form to , we have

Thus

holds. Therefore the use of or is a convenient thing to do for locating learning curves.

**References**

* Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1997): Scale-sensitive Dimensions, Uniform Convergence, and Learnability *Journal of the Association of the Computing Machinery* **44 (4)** 615-631.
* Anthony, M. (1997): Probabilistic Analysis of Learning in Artificial Neural Networks: The PAC Model and its Variants *Neural Computational Survey* **1** 1-47.
* Anthony, M., and P. L. Bartlett (1999): *Artificial Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge UK.
* Bartlett, P., and J. Shawe-Taylor (1999): Generalization Performance of Support Vector Machines and other Pattern Classifiers, in: *Advances in Kernel Methods – Support vector Learning* (B. Scholkopf, C. J. C. Burges, and A. J. Smola (editors)) 43-54 **MIT Press** Cambridge MA.
* Bartlett, P., P. Long, and R. C. Williamson (1999): Fat-shattering and the Learnability of Real-valued Functions *Journal of Computers and System Science* **52 (3)** 434-452.
* Kearns, M. J., R. E. Schapire, and L. M. Sellie (1994): Towards Efficient Agnostic Learning *Machine Learning* **17 (2)** 115-141.
* Kulkarni, S. R., G. Lugosi, and S. S. Venkatesh (1998): Learning Pattern Classification – A Survey *IEEE Transactions on Information Theory* **44** 2178-2206.
* Lee, W. S., P. L. Bartlett, and R. C. Williamson (1998): The Importance of Convexity in Learning with Squared Loss *IEEE Transactions on Information Theory* **44** 1974-1980.
* Vapnik, V. N., and A. Y. Chervonenkis (1981): Necessary and Sufficient Conditions for the Uniform Convergence of the Means to their Expectations *Theory of Probability and its Applications* **26 (3)** 532-553.
* Vapnik, V. N. (1982): *Estimation of Dependencies from Empirical Data* **Springer-Verlag** New York.
* Vapnik, V. N. (1998): *Statistical learning Theory* **Wiley** New York.

**Section VI: Covering Number Estimation for Kernel Machines**

**Kernel Machines**

**Introduction**

1. Definition: Kernel machines perform a mapping from an input space onto a feature space (Aizerman, Braverman, and Rozonoer (1964)), construct regression functions and/or decision boundaries based on this mapping, and use constraints in the feature space for capacity control (Nilsson (1965)).
2. Support Vector Machines: Support Vector Machines, which have evolved as a full new class of learning algorithms, solving problems in pattern recognition, regression estimation, and operator inversion (Vapnik (1995)), are a well-known example of kernel machines.
3. Feature Maps: SV machines, like most kernel-based methods, possess the nice property of defining the feature map in a manner that allows its computation implicitly at little additional cost.
4. Generalization Performance of Kernel Machines: Williamson, Smola, and Scholkopf (2001) show how bounds on covering numbers can be obtained by employing relatively standard methods, and hence estimate a bound on their generalization performance. Similar approach may be applied to bound regularization networks (Girosi, Jones, and Poggio (1993)), or certain unsupervised algorithms (Scholkopf, B., A. Smola, and K. R. Muller (1998)).

**SVM – Capacity Control**

1. Maximum Margin of Separation: In order to perform pattern recognition using linear hyper-planes, often a maximum margin of separation between the classes is sought, as this leads to good generalization ability independent of dimensionality (Vapnik and Chervonenkis (1974), Vapnik (1995), Shawe-Taylor, Bartlett, Williamson, and Anthony (1996)).
2. Maximum Margin Classification: It can be shown that for a separable training data

the maximum margins are achieved by minimizing subject to the constraints

for

The decision functions then take on the form

1. Maximum Margin Linear Regression: Likewise, a linear regression

can be estimated from the data

by locating the flattest function that approximates the data with some margin of error. In this case one minimizes subject to

where the parameter

plays the role of the margin, although not in the space of inputs , but in the space of the outputs .

1. Maximum Margins - Generalization: In both classification and regression, generalizations for the non-separable and the non-realizable cases exist, using various types of cost functions (Cortes and Vapnik (1995), Vapnik (1995), Smola and Scholkopf (1998)).

**Non-linear Kernels**

1. Dot-Products in High-Dimensional Spaces: To be able to apply the above technique to a general class of *non-linear* functions, one can use kernels computing dot products in high dimensional spaces non-linearly related to input (Aizerman, Braverman, and Rozonoer (1964)).
2. The Kernel Trick: Under certain conditions on the kernel (these are listed later), there exists a non-linear map into a reproducing kernel Hilbert space (e.g., Saitoh (1988)) such that computes the dot product in , i.e.
3. The Kernel Algorithm and its Applications: Thus, given any algorithm that can be expressed in terms of dot products exclusively, one can construct a non-linear version of it by substituting a kernel for the dot-product. Examples of such machines include SV pattern recognition (Boser, Guyon, and Vapnik (1992)), SV regression estimation (Vapnik (1995)), and kernel principal component analysis (Shawe-Taylor, Bartlett, Williamson, and Anthony (1996)).
4. Extension of the Maximum Margin Philosophy: By using the kernel trick for SV machines, the maximum margin idea is thus extended to a large variety of functions classes (e.g., radial basis function networks, polynomial networks, neural networks), which, in the case of regression estimation, comprise functions written as kernel expansions

where

1. Regularization Properties: It has been noticed that the different kernels can be characterized by their regularization properties (Smola, Scholkopf, and Muller (1998)). SV machines are regularization networks minimizing the regularized risk

with a regularization parameter

and a regularization operator over the set of functions of the form shown above, provided that and are inter-related through

To this end, is chosen as the Green’s function of where is the adjoint of .

1. Kernel Selection: While the analysis above provides insight into the regularization properties of the SV kernels, it does not settle the issue of how to select a kernel for a given learning problem, and how using a specific kernel might influence the performance of an SV machine.

**Generalization Performance of Regularization Networks**

1. Introduction: Williamson, Smola, and Scholkopf (2001) show that the properties of the spectrum of the kernel can be used to estimate the generalization error of the associated class of the learning machines.
2. Tuning the Generalization Error: The kernel is not just a means of broadening the class of functions alone, e.g., by rendering the non-separable data separable in a feature space non-linearly related to the input space. It is a constructive handle by which to control the generalization error.
3. Direct Bounding of the Covering Number: Williamson, Smola, and Scholkopf (2001) show how to directly bound the covering number of interest rather than make use of a combinatorial dimension (such as the Vapnik Chervonenkis (VC) dimension or the fat-shattering dimension) and subsequently applying a general result relating such dimensions to covering numbers.
4. Covering Numbers Construction: The covering numbers can be bound directly by viewing the values induced by the relevant class of functions as the image of a unit ball under a particular compact operator.

**Covering Number Determination Steps**

1. Step #1 - Entropy Numbers and Kernel Spectrum: First, formulate the generalization error in terms of the Entropy Numbers. In particular, it is important to relate the bounding entropy numbers in terms of the spectrum of a given kernel.
2. Step #2 - Function Class Covering Numbers: Identify the relation between the covering numbers of function classes and the entropy number of suitably defined operators. In particular, establish an upper bound on the entropy numbers in terms of the size of the weight vector in the feature space and eigenvalues of the kernel used. Apply standard techniques in case of kernels such as

which do not have a discrete spectrum.

1. Step #3 - Eigenvalue Decay Rate: Well-established results on the entropy numbers obtained for given rates of decay of eigenvalues (and their extension to multiple dimensions) can be employed.
2. Step #4 - Bringing it all together: Williamson, Smola, and Scholkopf (2001) show how the steps above along with the results may be glued in together to obtain overall bounds on the generalization error. While they deal with the eigenvalues of translation-invariant (i.e., convolutional) kernels, the general theory is not restricted to them.

**Challenges Presenting Master Generalization Error**

1. Problem Specificity: The particular statistical result one needs to use may be very specific to the problem at hand.
2. SRM Weakness: While SRM helps establish the existence of good generalization bounds those are necessary and sufficient (Vapnik (1982)), many of the results obtained are in a form which, while quite amenable to ready computation on a computer, do not provide much observational insight, except in an asymptotic sense. More explicit formulas suitable for SRM have been developed in Guo, Bartlett, Shawe-Taylor, and Williamson (1999).
3. Additional Extraneous Inputs: Applications such as classification may need margins to be estimated in a data-dependent fashion, thereby requiring additional “luckiness” arguments (Shawe-Taylor, Bartlett, Williamson, and Anthony (1998)) to apply bounds.

**References**

* Aizerman, M. A., F. M. Braverman, and L. I. Rozonoer (1964): Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning *Automation Remote Control* **25** 821-837.
* Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992): A Training Algorithm for Optimal Margin Classifiers, in: *5th Annual ACM Workshop on COLT (D. Haussler, editor)* **ACM Press** Pittsburgh PA.
* Cortes, C., and V. N. Vapnik (1995): Support Vector Networks *Machine Learning* **20** 273-297.
* Girosi, F., M. Jones, and T. Poggio (1993): Prior, Stabilizers, and Basis Functions: From Regularization to Radial, Tensor, and Additive Splines *A. I. Memo 1430* **Massachusetts Institute of Technology** Cambridge MA.
* Guo, Y., P. L. Bartlett, J. Shawe-Taylor, and R. C. Williamson (1999): Covering Numbers for Support Vector Machines, in: *Proceedings of the 12th Annual Conference on Computational Learning Theory* 267-277 **ACM** New York.
* Nilsson, N. J. (1965): *Learning Machines: Foundations of Trainable Pattern Classifying Systems* **McGraw-Hill** New York.
* Saitoh, S. (1988): *Theory of Reproducing Kernels and its Applications* **Longman** Harlow UK.
* Scholkopf, B., A. Smola, and K. R. Muller (1998): Non-linear Component Analysis as a Kernel Eigenvalue Problem *Neural Computation* **10** 1299-1319.
* Shawe-Taylor, J., P. L. Bartlett, R. C. Williamson, and M. Anthony (1996): A Framework for Structural Risk Minimization, in: *Proceedings of the 9th Annual Conference on the Computational Learning Theory* 68-76 **ACM** New York.
* Shawe-Taylor, J., P. L. Bartlett, R. C. Williamson, and M. Anthony (1998): Structural Risk Minimization over Data-dependent Hierarchies *IEEE Transactions on Information Theory* **44** 1926-1940.
* Smola, A. J., and B. Scholkopf (1998): On a Kernel-based Method for Pattern Recognition, Regression, Approximation, and Operator Inversion *Algorithmica* **22** 211-231.
* Vapnik, V. N., and A. Chervonenkis (1974): *Theory of Pattern Recognition (in Russian)* **Nauka** Moscow.
* Vapnik, V. N. (1982): *Estimation of Dependencies from Empirical Data* **Springer-Verlag** New York.
* Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory* **Springer-Verlag** New York.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2001): Generalization Performance of Regularization Networks and Support Vector Machines via Entropy Numbers of Compact Operators *IEEE Transactions on Information Theory* **47 (6)** 2516-2532.

**Entropy Numbers for Kernel Machines**

**Mercer Kernels**

1. Introduction: Here we will mainly consider machines whose mapping into the feature space is defined by Mercer kernels as they are easier to deal with using functional analytics methods. More general kernels are considered in Smola, Elisseff, Scholkopf, and Williamson (2000). Such machines have become very popular due to the success of SV machines.
2. Goals: Our goal is to make statements of the shape of the image of the input space under the feature map . The version of the Mercer’s theorem we use here is a special case of the theorem proven in Konig (1986). In what follows, we assume is a finite measure space, i.e.
3. Mercer Theorem - Notation: Suppose

is a symmetric kernel, i.e.

such that the integral operator

is positive. Let

be the eigen-function of associated with the eigenvalue

and normalized by

Finally, suppose is continuous for all

1. Mercer’s Condition: Using the notation above, we can specify Mercer’s conditions:

for

and

holds for all

* 1. Finally, the series converges absolutely and uniformly for all

1. Assumption of Continuity of : Note that if is compact and is continuous, then is continuous (Ash (1965)). Alternatively, if is translation-invariant, then ’s are scaled cosine functions, and thus continuous. Therefore the assumption that are continuous is not very restrictive.
2. Impact of Mercer’s Second Condition: From the second condition it follows that there exists some constant

depending on such that

for all

and

1. Impact of Mercer’s Third Condition: From the third condition it follows that corresponds to a dot-product in , i.e.

with

1. Spatial Span of : From the above argument one can see that lives not only in , but in an axis parallel parallelepiped with lengths 2. Also we assume, without loss of generality, that the sequence is sorted in a non-increasing order.

**Equivalent Kernels**

1. The Equivalent Kernel Lemma: Denote by a compact set, and by

a Mercer kernel. Then for any and a surjective map

the kernel

also satisfies Mercer’s condition, and, moreover, the eigenvalues and the coefficient of the integral operator

can be used equivalently in any application of .

1. Proof of the Lemma: The first part of the claim, namely that also satisfies Mercer’s condition, follows immediately from the construction of . For the second claim, note that due to the fact that is surjective for any distribution on there must exist an equivalent distribution on . Therefore we can always consider the problem as being one on from the start.
2. Impact of the Lemma: Note that since and were chosen arbitrarily, we can optimize over them. In particular, this means that we could construct diffeomorphisms

and look for the function such that the eigenvalues and are as small as possible.

1. Dependence on : Note that the measure need have nothing to do with the distribution of our sample. However, the Equivalent Kernel lemma shows that the specific bounds we obtain *will* depend on since that will affect , , and . The question of the optimal to use and how it may be chosen if one knows (the distribution from which are chosen) requires a full treatment on in its own right, but it is common for to be a Lebesgue measure.

**Mapping Into**

1. Motivation: It will be useful to consider maps that map into balls of some radius centered at the origin. The following proposition shows that the class of all these maps is determined by the elements of and the sequence of eigenvalues .
2. Bounding of - Statement: Let be a diagonal map

with

Then maps into a ball of finite radius centered at the origin if and only if

1. Sufficiency of : Suppose

and let

For any

Hence

1. Necessity of : Suppose is not in . Hence the sequence

is unbounded. Now define

Then

due to the normalization condition on . However, as

there exists a set of non-zero measure such that

for all

Combining the expression for from the sufficiency condition with the one for , we obtain

for all

and all . Since is unbounded for a set with non-zero measure in , we can see that

**Corrigenda to the Mercer Conditions**

1. Redefining : It has been pointed out that the condition #2 of Mercer’s theorem, i.e.

and

needs adjustment, and therefore redefining as

Note that most practically used kernels still have

1. Impact of the Re-definition: This means that only the sufficiency part of the proposition for mapping onto remains valid, which is all we need for bounds on the entropy numbers. All of the upper bounds on the entropy numbers still hold as long as redefined as above stays finite.

**Unit Ball Mapping Scaling Operator**

1. Motivation for the Construction of : Once we know that is contained in the parallelepiped described above, we can use the result of the above proposition to construct a mapping from the unit ball in to an ellipsoid such that

as listed in the schematic below.

1. Metric Space Mapping Schematic:
   1. Remember that
2. Use of : The operator will be useful for computing the entropy numbers of concatenations of operators. Knowing the inverse will allow us to compute the forward operator, and that -can be used to bound the covering numbers of the class of functions.
3. Approach towards constructing : We thus seek an operator

such that

This means that

will be such that

The latter can be ensured by constructing such that

with

where and are chosen with respect to a specific kernel, and where

From the necessary/sufficient propositions above, it follows that all those operators for which

will satisfy the criterion

We call such scaling/inverse operators *admissible*.

**Unit Bounding Operator Entropy Numbers**

1. Computing the Entropy Numbers of : Here we bound the entropy number for the operator , and use it to obtain bounds on the entropy numbers for kernel machines like SV machines. We make use of the following theorem due to Gordon, Konig, and Schutt (1987) in the form stated by Carl and Stephani (1990).
2. Diagonal Operator Bounds: Let

be a non-increasing sequence of non-negative numbers, and let

for

be the diagonal operator from into itself, generated by the sequence , where

Then for all

1. Minimizing the Entropy Numbers by Choosing : We can exploit the freedom in choosing to minimize the entropy number, as shown below. This will be a key ingredient in the calculation of the covering numbers of kernel machines and SV classes.
2. Scaling Operators Lemma: Let

be a Mercer kernel with eigenvalues . Choose

for

such that

and define

with

Then

1. Use of the Scaling Operator Lemma: The scaling operator lemma follows immediately by identifying in the diagonal bounds statement with in the scaling operator lemma. We can optimize over all possible choices of to obtain the optimal entropy number for . It turns out that the optimum for (i.e., the infimum over ) is attainable when is a Mercer kernel (Guo, Bartlett, Shawe-Taylor, and Williamson (1999)). Thus, we can minimize the RHS of the scaling operator lemma, as shown below.
2. Optimal Entropy Number for : There exists an defined by

with

that satisfies

**The SVM Operator**

1. The SVM Feature Space: To recap, the hypothesis that an SVM generates may be expressed as

where both and are defined in the feature space

and

1. Applying the Kernel Trick: The kernel trick, as introduced in Aizerman, Braverman, and Rozonoer (1964), was successfully employed by Boser, Guyon, and Vapnik (1992) and Cortes and Vapnik (1995) to extend the optimal margin hyper-plane classifier to what is now known as the SV machine. Ignoring for now, we consider the class

Note that depends implicitly on since does.

1. Covering Numbers for : We seek the covering numbers for induced by the kernel in terms of the parameter which is the inverse of the size of the margin in the feature space as defined by the dot-product in (see Vapnik and Chervonenkis (1974) and Vapnik (1995) for details).
2. SVM Operator for SV Classes – Definition: We refer to those hypotheses classes that have length constraint on the feature space to be *SV Classes*. We define the operator as

where

and the operator is defined by

with

for all . We then seek to compute the entropy numbers of the SV operators.

**Maurey’s Theorem**

1. Introduction: Maurey’s theorem is useful for computing the entropy numbers in terms of and . The original theorem was extended by Carl (1985), and formulated in the form used below by Carl and Stephani (1990).
2. Statement of the Theorem: Let

where is a Hilbert Space. Then there exists a constant

such that for all

1. vs. Input Dimensionality: Carl and Stephani (1990) state an additional condition, namely, that

It turns out that for

and even tighter bound holds, and so it is not incorrect to state the result as above (Williamson, Smola, and Scholkopf (2000)). This tighter bound is of little value in learning theory applications, as it corresponds to determining the -covering number for an extremely small for which

1. Estimate for the Constant : Williamson, Smola, and Scholkopf (2000) provide proof of a fairly tight bound for the constant of

however, there is reason to believe that should be , the constant obtained for identity maps from to .

1. Maurey’s Theorem for Entropy Numbers (instead of Dyadic Entropy): The re-statement of Maurey’s theorem in terms of

under the assumptions above, becomes

1. Carl and Stephani’s Lemma (Carl and Stephani (1990)): Let , , and be Banach spaces, with

and

Then for all

* 1. Note that b) and c) follow directly from a), and the fact that

for all

and

for all

**Bounds for SV Classes**

1. Setup: Let be a Mercer kernel, let be induced from

and let

where is given from

and

Let be defined from

and suppose

for

1. The SV Bounds: Then, the entropy numbers of satisfy the following inequalities ( here is the same as specified earlier):
2. Optimal Bound for : The above set of bound provide several options for bounding . The optimal inequality bound to use depends on the rate of decay of the eigenvalues of . The result gives effective bounds on since
3. Factorization of to the Upper Bound : We will us the following factorization of to upper-bound .
4. Description of the Factorization Schematic: The statement a) follows from the definition of . The fact that b) – d) commutes, as in e), stems from the fact that since is diagonal, it is self-adjoint, and so for any
5. Alternate Representation of : Instead of computing the entropy number of

directly, which is difficult or wasteful, as the bound on does not take into account that

but simply assumes that

for some

we will represent as

This is more efficient as we constructed such that

filling a larger proportion of it than does.

1. Proof of the SV Bounds: By construction of , and due to the Cauchy-Schwartz inequality, we have

Thus, applying the Carl and Stephani’s lemma to the factorization of this version of , and using the Maurey’s theorem, we get the SV bounds.

**Asymptotic Rates of Decay for**

1. Introduction: Eventually we seek asymptotic rates of decay for – in fact, we provide non-asymptotic results with explicitly evaluable constants. It is thus of some interest to give overall asymptotic rates of decay of in terms of . By asymptotic here we mean asymptotic in ; this corresponds to asking how scales as

for fixed .

1. Rates Bounds on : Let be a Mercer kernel and suppose that is the scaling operator defined for

such that

and define

with

such that for

and

* 1. If

for some

then for fixed

* 1. If

for some

then for fixed

1. Proof Step #1 Splitting the Index : From Maurey’s theorem we know that

Using the 3rd SV Class bound, ignoring the constants and ignoring the constants and assuming is fixed, we split the index in the following way:

with

1. Proof Step #2 - Bounds for the First Case: For the first case, this yields
2. Proof Step #3 - Bounds for the Second Case: For the first case, this yields
3. Impact of the Rate Bounds on : These bounds show that, in the first case, Maurey’s theorem allows an exponent in the entropy number of , whereas in the 2nd it affords none, since the entropy numbers decay so fat anyway. Maurey’s theorem may still help in that case for non-asymptotic .
4. Improvements over Maurey’s Theorem: In a nutshell, we can always obtain rates of convergence better than those due to Maurey’s theorem because we are not dealing with *arbitrary* mappings into infinite-dimensional spaces. In fact, for logarithmic dependence of on , the effect of kernel is so strong that it completely dominates the for arbitrary Hilbert spaces. An example of such a kernel is .

**References**

* Aizerman, M. A., F. M. Braverman, and L. I. Rozonoer (1964): Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning *Automation Remote Control* **25** 821-837.
* Ash, R. (1965): *Information Theory* **Interscience** New York.
* Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992): A Training Algorithm for Optimal Margin Classifiers, in: *5th Annual ACM Workshop on COLT (D. Haussler, editor)* **ACM Press** Pittsburgh PA.
* Carl, B. (1985): Inequalities of the Bernstein-Jackson Type and the Degree of Compactness of Operators in Banach Spaces *Ann. Inst. Fourier* **35 (3)** 79-118.
* Carl, B. and I. Stephani (1990): *Entropy, Compactness, and the Approximation of Operators* **Cambridge University Press** Cambridge UK.
* Cortes, C., and V. N. Vapnik (1995): Support Vector Networks *Machine Learning* **20** 273-297.
* Gordon, Y., H. Konig, and C. Schutt (1987): Geometric and Probabilistic Estimates for Entropy and Approximation Numbers of Operators *Journal of Approximation Theory* **49** 219-239.
* Guo, Y., P. L. Bartlett, J. Shawe-Taylor, and R. C. Williamson (1999): Covering Numbers for Support Vector Machines, in: *Proceedings of the 12th Annual Conference on Computational Learning Theory* 267-277 **ACM** New York.
* Konig, H. (1986): *Eigenvalue Distribution of Compact Operators* **Birkhauser** Basel Switzerland.
* Smola, A. J., A. Elisseff, B. Scholkopf, and R. C. Williamson (2000): Entropy Numbers for Convex Combinations and mlps, in: *Advances in Large Margin Classifiers (A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors)* **MIT Press** Cambridge MA.
* Vapnik, V. N., and A. Chervonenkis (1974): *Theory of Pattern Recognition (in Russian)* **Nauka** Moscow.
* Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory* **Springer-Verlag** New York.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2000): Entropy Numbers of Linear Function Classes, in: *Proceedings of the 13th Annual Conference on Computational Learning Theory* *(N. Cesa-Bianchi and S. Goldman, editors)* **ACM** New York.

**Discrete Spectra of Convolution Operators**

**Kernels with Compact/Non-compact Support**

1. Introduction: The results presented above show that if one knows the eigenvalue sequence of a compact operator, one can bound its entropy numbers. While it is always possible to assume that the data fed into the SV machine have bounded support, the same cannot be said of any kernel; in fact, a commonly used kernel is

which has non-compact support.

1. Integral Operator Induced Continuous Spectrum: As in the kernel above, the induced integral operator

has a continuous spectrum (i.e., a non-denumerable infinity of eigenvalues), and, thus, is not compact (Ash (1965)). The question then arises is whether we can make use of such kernels in SV machines and still obtain generalization error bounds of the form above.

1. Convolution Kernels with Compact Support: It is well-known that the eigenvalue decay of any convolution operator defined on a compact set via a kernel having compact support can decay no faster than

(Widom (1963)), and thus if one seeks rapid decay of eigenvalues (and, therefore, small entropy numbers), one must seek convolution kernels with non-compact support.

1. Kernels with Compact Support: We first consider the case where the support of

for some

We also further suppose that the data points satisfy

1. Kernels with Compact Support applied to SV Hypothesis: If is a convolution kernel, i.e.

then the SV hypothesis can be written as

for

where is the -periodic extension of - analogously

**Eigenvalues of**

1. The -dimensional Fourier Transform Operator: The -dimensional Fourier transform is defined by

Then its inverse is defined by

From the above, is an isometry on .

1. The Approach: We now relate the eigenvalues of to the Fourier transform of . We do so for the case of

and state the general cases later.

1. Fourier Transform of the Symmetric Convolution Kernel - Statement: Let

be a symmetric convolution kernel, let

denote the Fourier transform of , and denote the -periodical kernel derived from (also assume that exists). Then has a representation as a Fourier series with

and

Moreover, the eigenvalues of satisfy

for

and

1. Proof Step #1 - Fourier Series Coefficients of : Clearly, the Fourier series coefficients of exist (as exists), with

and therefore, by the definition of and the existence of , we conclude

1. Proof Step #2 - The Orthogonal Basis: This and the fact that

form an orthogonal basis in proves that

1. Proof Step #3 - Choice of the Trigonometric Basis Function: Furthermore, we are interested in real-valued basis functions for . The functions

and

for all

satisfy

and form an eigen-system of the integral operator defined by with the corresponding eigenvalues . Finally, one can see that

by computing the maximum over

and

1. Extension to non-compact Support: Thus, even though may not be compact, can be - if

for example. The above Fourier transform can be applied whenever we can form from . Clearly

for some

suffices to ensure that the sum

converges.

1. Discrete, Periodic Spectrum: In conclusion, in order to obtain a discrete spectrum, one needs to use a periodic kernel. For a given problem, one can always periodize a non-periodic kernel in a way that changes the eventual hypothesis in an arbitrarily small way. The results that we produce below can then be applied.

**Choosing**

1. Asymptotics of : From the Riemann-Lebesgue lemma, it is clear that for integrable of bounded variations (surely any kernel one would choose would satisfy that criterion), one has
2. Tradeoff in the Choice of : There is a tradeoff in choosing in that, for large enough , is a decreasing function of (at least as fast as ), and thus, from

we get that

is an increasing function of . This suggests that one should choose a small value of . But a small will lead to high empirical error (as the kernel “wraps around” and its localization properties are lost, i.e., its becomes high) and large .

1. An Approach for picking : There are several approaches for picking a value for . One obvious way is to *a priori* pick some

and choose the smallest such that

for all

Thus, one would obtain a hypothesis uniformly within of where

**Extension to -dimension**

1. The Statement: Assume that is -periodic in each direction

we get

for radially symmetric , and finally for the eigen-functions we get

1. Kernel Bandwidth Choice: Here we examine how a difference choice for the bandwidth of the kernel, i.e., after letting

affects the eigen-spectrum of the corresponding operator. We have

hence scaling a kernel by means more densely spaced eigenvalues in the spectrum of the integral operator .

**References**

* Ash, R. (1965): *Information Theory* **Interscience** New York.
* Widom, H: (1963): Asymptotic Behavior of Eigenvalues of certain Integral Operators *Transactions of the American Mathematical Society* **109** 278-295.

**Covering Numbers for Given Decay Rates**

**Asymptotic/Non-asymptotic Decay of Covering Numbers**

1. Introduction: Here we show how the asymptotic behavior of

where is the scaling operator introduced earlier, depends on the eigenvalues of .

1. Comparison with Prosser’s Analysis: A similar analysis has been carried out by Prosser (1966) in order to compute the entropy numbers of integral operators. However, all of his operators mapped into . Furthermore, while the analysis here states the propositions as asymptotic results as Prosser (1966) does, the proofs provide non-asymptotic details with explicit constants.
2. Need to work with Sorted Eigenvalues: Note that we need to sort the eigenvalues in a non-increasing manner due to the requirements of the proposition for , i.e.

If the eigenvalues were unsorted, one could obtain far too small numbers for the geometrical mean of .

1. Non-degeneracy of Eigenvalues: Many 1D kernels have non-degenerate systems of eigenvalues in which case it is straight-forward to explicitly calculate the geometric means of the eigenvalues. While all of the instances we treat here are convolution kernels, i.e.

there is nothing in the formulations of the propositions themselves that requires this. When we consider the -dimensional case, we shall see that with rotationally invariant kernels, degenerate systems of eigenvalues are generic.

1. Explicit Decay of : We shall consider the specific cases where decays asymptotically in some polynomial or exponential degree. In this case, we choose the sequence for which we can evaluate

explicitly. In what follows, by the eigenvalues of kernel we refer to the sorted eigenvalues of the induced integral operator .

**Polynomial Eigenvalue Decay**

1. Eigenvalue Decay Statement: Let be a Mercer kernel with eigenvalues

for some

Then, for any

we have

An example of such a kernel is

1. Proof Step #1 - Bounding : Since

there exists some

with

In this case, all sequences

with

lead to an admissible scaling property. Thus, one has

where is the Riemann’s zeta function. We recall that one can bound by

where is Euler’s constant.

1. Proof Step #2 - Bounding : We now evaluate

The Gamma function can be bound as follows: for

1. Proof Step #3 - Bounds on : Thus, one may bound as

and

1. Proof Step #4 - Re-casting : In order to avoid unneeded technicalities, we replace

by

This is no problem when computing the upper bound, but it is an issue for the lower bound. In fact, as is shown in Williamson, Shawe-Taylor, Scholkopf, and Smola (1999)

for that particular where

is actually obtained. Hence, the maximum quotient

which in the present case is , determines the value by which the bound has to be lowered in order to obtain a true lower bound. Therefore, since on is within a constant factor of of its corresponding values on the integer domain (the biggest discrepancy being at ), we may safely ignore this concern.

1. Proof Step #5 - Estimating : Now we can compute

The maximum of this argument is obtained for

hence the above supremum holds for all

which is fine, since we want to compute the bounds on as

1. Proof Step #6 - Lower Bounds on : For lower bounds on we obtain

This shows that is always bounded from below by

1. Proof Step #7 - Upper Bounds on : Computation of the upper bound is slightly more involved, since one has to evaluate

Clearly, for any fixed

we are able to get a rate of

thus the statement follows. For practical purposes, a good approximation to can be found as

by computing the derivative of the argument of the with respect to , and dropping all terms independent of and . However, numerical minimization of the arguments to is more advisable when small values of are crucial.

**Summation and Integration of Non-decreasing Functions**

1. Summation and Integration in - Lemma: Suppose

is an integrable, non-increasing function. Then the following inequality holds for any

1. Proof of the Lemma: The proof relies on the fact that due to the monotonicity of and the decomposition of the integral . The lemma is a direct consequence thereof.

**Exponential Polynomial Decay**

1. Introduction: The exponential polynomial decay covers a wide range of practically used kernels, namely, those with polynomial decay in their eigenvalues. For instance, the Gaussian kernel

has exponential quadratic decay in . The damped harmonic oscillator kernel

is another example, this time with just exponential linear decay in its eigenvalues.

1. Exponential Polynomial Decay Bound - Proposition: Suppose is a Mercer kernel with

for some

Then

1. Proof of the Proposition - Step #1: Bounding => Since

there exists some

with

As in the polynomial decay scenario, we use the series

By applying the lemma on summation and integration on , we have that, for any

1. Proof Step #2: Bounding the Diagonal Entries of => We apply a similar bound to the product of the first entries of the scaling operator :
2. Proof Step #3: Bounding the Suprema => Next we compute

Differentiating the exponent with respect to leads to

and thus

1. Proof Step #4: Replacing the Suprema Domain => Replacing the domain from

to

is not a problem when it comes to computing the upper bounds on . As for the lower bounds, again, a similar reasoning to that in the polynomial decay would have to be applied. Thus, the problem reduces to bounding the quotient

where is the variable for which

is attained. However, the quotient can only be bounded by

Fortunately, this is of lower order than the remaining terms, hence it will not change the *rate* of the lower bounds.

1. Proof Step #5: Lower Bound of =>

Hence a lower bound on the rate of is

1. Proof Step #6: Upper Bound on =>

This may be evaluated numerically. However, it can be seen that for any fixed

the rate of can be bounded by

which shows that the obtained rates are tight.

**References**

* Prosser, R. T. (1966): The -entropy and -capacity of certain Time-Varying Channels *Journal of Mathematical Analysis and Applications* **16** 553-573.
* Williamson, R. C., J. Shawe-Taylor, B. Scholkopf, and A. J. Smola (1999): Sample-based Generalization Bounds *NeuroCOLT2 Technical Report Series* **NC-TR-1999-055**.

**Kernels for High Dimensional Data**

**Introduction**

1. Translationally and Rotationally Invariant Kernels: Things get more complicated in higher dimension. For simplicity, we restrict ourselves to translation-invariant kernels.
2. Constructing High-Dimension Kernels: There are two simple ways to construct kernels in

with

First, one could construct kernels by

This choice in general will lead to preferred directions in the input space, as these kernels are not rotationally invariant in general. The second approach consist in setting

This approach leads to translationally invariant kernels that are also rotationally invariant.

1. Computing the Regularization Operators: In what follows, we exploit the 2nd approach to compute the regularization operators, and their corresponding Green’s functions. It is quiet straightforward to generalize our exposition to the rotationally asymmetric case.

**Kernel Fourier Transforms**

1. The Regularization Operator: We define the regularization operators by

for some non-negative function converging to for

It can be shown that for the kernel to be Green’s function of , i.e.

we need

(Smola, Scholkopf, and Muller (1998)).

1. Radially Symmetric Kernels: For radially symmetric kernels, i.e.

we can carry out the integration on a sphere to obtain a Fourier transform which is also radially symmetric, i.e.

where

and is the Hankel transform over the positive real line (Sneddon (1972)).

1. The Hankel Transform: The Hankel transform is defined as

Here is the Bessel function of the first kind defined by

Note that

i.e.

(in ) due to the Hankel inversion theorem (Sneddon (1972)).

**Degenerate Kernel Bounds**

1. Fourier Transforms For arbitrary Kernels: Computing the Fourier transform for an arbitrary kernel will give us its continuous spectrum. As discussed earlier, we are interested in the discrete spectra of integral kernels defined on . This means that the eigenvalues are defined on the grid with
2. Degeneracy of Rotationally Invariant Kernels: If is rotationally invariant, so is , and therefore there are repeated eigenvalues at

Consequently, we have degeneracies on the point spectrum of the integral operator given by (or respectively) as all with equal length will have the same eigenvalues.

1. Modified Diagonal Operator Bounds: In order to deal with the degeneracies in the eigenvalues efficiently, we modify slightly the Diagonal Operator Bounds. The degenerate Diagonal Operator Bounds allows proper account to be taken of the multiplicity of the eigenvalues, and thus allow for a more refined calculation of the desired entropy numbers.
2. Degenerate Diagonal Operator Bounds: Let

be an increasing sequence with

and

be a non-increasing sequence of non-negative numbers with

for

and

for

and let

for

be the diagonal operator from to itself generated by the sequence , where

Then for all

1. Proof:
   1. Proof Approach Outline => The first part (i.e., the lower bound) follows directly from the Diagonal Operator Bounds lower bound, as it is a weaker statement than the original one. The upper bound is established by mimicking the proof in Carl and Stephani (1990). We define

and show that for all there is an index such that

* 1. Induction Initial Phase => We choose an index such that

and thus

Moreover, we have

because of the monotonicity of

and combining the both above, we get

Using the definition of , we thus conclude that

If this happens to be the case for , we have

which proves the theorem.

* 1. Induction Extension to the Sectional Operator => If the above is not the case, there exists an index such that

Hence the corresponding sectional operator

with

is of rank , and the image of the closed unit ball of is isometric to the subset

of . In any case, is a pre-compact set of .

* 1. Entropy Number of the Sectional Operator => So let be a maximal set of elements in , with

for

The maximality of this system guarantees that

and thus

* 1. Estimation of => In order to get an estimate, for , we split the operator into two parts

which allows us to bound

Using

and the bound

above we arrive at

* 1. Establishing => The final step is to show that

as then by substituting into the definition of yields the result. This is achieved by a comparison of volumes. Consider the sets

as subsets of the space , which is possible since

and

These sets are obviously pairwise disjoint.

* 1. -dimensional Euclidean Volumes Comparison => On the other hand, we have

as

Now a comparison of the -dimensional Euclidean volumes provides

and therefore

Using the definition of , this yields

**Covering Numbers for Degenerate Systems**

1. Optimal Entropy Numbers for Degenerate Systems: Let

be a Mercer kernel, and let be defined by the Scaling Operator Lemma with the additional restriction that the coefficients have to match the degeneracy of , i.e.

for

and

for

Then one can choose such that

1. Computational Impact of Degenerate :

for the degenerate case is slightly more tighter than that for the non-degenerate case, as the supremum effectively has to be carried out over only a subset of .

1. Estimation of the Multiplicity of Degeneracy: Here we compute the degree of multiplicity that occurs for different indexes . For this purpose, consider shells of radius in centered at the origin, i.e., , which contains a non-zero number of elements in . Denote the corresponding radii by and let be the number of elements on these shells. Observe that

only when

Thus

The determination of is a classical problem that is completely solved by -series (Grosswald (1985)).

1. Occupation Numbers of Shells - Theorem: Let the formal power series be defined by

Then

1. Use of the -series: While there do exist closed form asymptotic formulas for , they are inordinately complicated, and are of little use for our purposes (Grosswald (1985)). The -series theorem does allow one to calculate exactly; this helps us construct an index of eigenvalues that satisfies the required ordering (at least for non-increasing functions ) for us to get the main result.
2. Optimal Entropy Numbers Using Degeneracy Multiplicity: Let

be a Mercer kernel with eigenvalues given by a radially symmetric non-increasing function on a lattice, i.e.

with

and let be defined by the Scaling Operator Lemma with the additional restriction that the coefficients have to match the degeneracy of , i.e.

Then

1. Impact of Using the Degeneracy Multiplicity in Entropy Numbers: Note that while computing the optimal entropy using the degeneracy multiplicity as shown above may appear straightforward, it cannot be obtained from the Proposition for Computing Optimal Entropy Numbers for non-degenerate systems directly, as there the supremum would have to be evaluated over instead of

**Bounds for Kernels in**

1. Overview: In this section we examine some examples of the eigenvalue sequences for kernels typically used in SV machines. These can be used to evaluate the optimal entropy numbers for those degenerate systems using their degeneracy multiplicity. We begin by computing the Fourier/Hankel transform for these kernels - recalling that

in all the cases here.

1. Gaussian Radial Basis Functions: For Gaussian radial basis functions in -dimensions, we have

and correspondingly

1. Exponential Radial Basis Functions: In this case

so

Thus, in the case of

we recover the case of damped harmonic oscillator (below) in the frequency domain.

* 1. Eigenvalue Decay => In general, we get a decay in terms of eigenvalues like . Moreover, one can conclude from this that the Fourier transform of , viewed itself as a kernel, i.e.

yields the initial kernel as its corresponding power spectrum in the Fourier domain.

1. Damped Harmonic Oscillator: Another way to generalize the Harmonic oscillator, this time in a way that does not depend on the dimensionality , is to set
   1. Fourier Transform and Asymptotics => Following Watson (1958), we get

where is the Bessel function of the 2nd kind, defined by

(Sneddon (1972)). It is possible to upper bound by utilizing the asymptotic representation

(e.g., Gradshteyn and Ryzhik (1981), equation (8.451.6)), and we get exponential decay of the eigenvalues.

1. Bounds Computation Steps - Recap: Using a) the theorem for the Occupation Numbers of Shells, b) the Optimal Entropy Numbers using Degeneracy Multiplicity, and c) the -dimensional Eigenvalue Calculation Lemma, one may compute the entropy numbers numerically for a particular kernel, and for a particular set of parameters.
2. Conservative Nature of the Obtained Bounds: The obtained bounds may seem too loose and conservative from a practical point of view. However, as the ultimate goal is to use the obtained bounds for model selection, it is desirable to obtain as tight bounds as possible, esp. in the constants.
3. Computation of more Precise Bounds: If much more precise bounds may be obtained by some not too expensive numerical calculations, it is definitely worthwhile to use that instead of the theoretically nice but insufficiently tight upper bounds. The computational effort required to calculate these quantities is typically negligible in comparison to training the actual machine.

**Impact of the Fourier Transform Decay on the Entropy Numbers**

1. Introduction: Notwithstanding the caveats of the section above, in order to get a feeling for the decay of the Fourier transform of the kernel on the entropy numbers of the operator, we conclude with the following general result, and its eventual proof.
2. Eigenvalue Polynomial Exponential Decay in : For kernels

with

where

the entropy number of the corresponding scaling operators satisfies

1. Proof:
   1. Assumption of Continuous Eigenvalues => We will completely ignore the fact that we are dealing with a countable set of eigenvalues in a lattice, and replace all summations by integrals. Of course this is not accurate, but will still give the correct rates for the entropy numbers.
   2. Estimate Corresponding Infinitesimal Volumes and Number Density => We denote

as the size of the unit cell, i.e.

as the density of the lattice points in the frequency space as seen earlier. Then we get for infinitesimal volumes and the number of eigen-points in frequency space

and therefore

(here denotes the volume of the dimensional unit sphere) leading to

* 1. Eigenvalue Decay of the Scaling Operator => We introduce a scaling operator whose eigenvalue decays as

for

It is straightforward to check that all these values lead to both useful and admissible scaling operators. We now estimate the individual terms in the expression for the Optimal Entropy Numbers using the Degeneracy Multiplicity:

* 1. Estimation of =>
  2. Estimating => Next we have
  3. Estimating =>
  4. Computing => Combining all of the above leads to

Computing

yields

and therefore

* 1. Optimizing over => Already from the above one can observe the rate bounds on . What remains to be done is the computation of the infimum over . This can be done by differentiating the final expression with respect to . Defining

which leads to the optimality condition on

with

can be solved numerically.

**References**

* Gradshteyn, I. H., and I. M. Ryzhik (1981): *Tables of Integrals, Series, and Products* **Academic Press** New York.
* Grosswald, F. (1985): *Representation of Integers as Sums of Squares* **Springer-Verlag** New York.
* Smola, A. J., B. Scholkopf, and A. J. Muller (1998): The Connection between Regularization Operators and Support Vector Kernels *Neural Networks* **11** 637-649.
* Sneddon, I. H. (1972): *The Use of Integral Transforms* **McGraw Hill** New York.
* Watson, G. N. (1958): *A Treatise on the Theory of Bessel Functions, 2nd edition* **Cambridge University Press** Cambridge UK.

**Regularization Networks Entropy Numbers Determination - Practice**

**Introduction**

1. Motivation: Williamson, Smola, and Scholkopf (2001) have shown how to connect properties known about mappings into feature spaces with bounds on covering numbers. Exploiting the geometric structure of the feature space map, they relate properties of the kernel inducing the feature space to the covering numbers of the class of functions implemented by the SV machines based on such kernels.
2. Usage in Model Selection: The application of the results of Williamson, Smola, and Scholkopf (2001) (e.g., for model selection using structural risk minimization) is somewhat limited, but certainly doable. For instance, Guo, Bartlett, Shawe-Taylor, and Williamson (1999) apply these results to study the performance of SV machines for pattern classification.

**Custom Application of the Kernel Machines Entropy Numbers**

1. Choosing the Kernel and the Bandwidth: While the results do not aid in the selection of the kernel machines per se, they offer suggestions on the choice of bandwidth. Since the Fourier transform of the kernel goes as

scaling the kernel by implies more densely spaced eigenvalues in the spectrum of the integral operator .

1. Choosing the Kernel Period : Since

its eigenvalue increases with , thereby suggesting a small choice of . However, small results in large empirical error, as the kernel looses its localization properties. Williamson, Smola, and Scholkopf (2001) suggest an approach to pick an optimally satisfactory based on the above trade-off.

1. Bound on : These bounds are computed using either the Optimal Scaling Operator Lemma for the unidimensional case, or the adjustment to the Optimal Scaling Operator Lemma for Degeneracy Multiplicity (along with the theorem for the Occupation Numbers of Shells) for the multidimensional case.
2. Bound on : These are estimated using the combination of the Maurey’s theorem, the Carl and Stephani’s theorem, and the customization applied for kernel machines in Williamson, Smola, and Scholkopf (2001).
3. Taking Account of : The key observation is that given a class with known , one can bound where

as follows. Suppose that is an -cover for , and that the elements of are uniformly bounded by (this implies a limit on as well as a uniform bound on the elements of ). Then

is an -cover for , and thus

Obviously this will only be “noticeable” for those classes whose covering numbers grow very slowly (i.e., polynomial in ).

1. Taking Account of the Loss Functions: While much of the treatment of Williamson, Smola, and Scholkopf (2001) focusses on the regularization component of the regularized error, the corresponding treatment of the empirical error considers loss functions in terms of their Lipschitz-continuous and Lipschitz-continuous-light equivalents. Applying one of the standard empirical bounds in terms of their covering numbers, one may extract the corresponding uniform convergence results.

**Extensions to the Operator-Theoretic Viewpoint for Covering Numbers**

1. Overall Bounds of Learning Machines: The overall bounds for SV classes, via a more involved argument, can be somewhat simplified (Guo, Bartlett, Shawe-Taylor, and Williamson (1999)). The general approach can be applied to various other learning machines such as convex combinations of basis functions and multi-layer networks (Smola, Elisseff, Scholkopf, and Williamson (2000)).
2. Tighter Statistical Bounds and Unsupervised Learning Extensions: When combined with an appropriate statistical argument (Shawe-Taylor and Williamson (1999)), the approach yields bounds on the generalization that depend strongly on the particular sample observed (Williamson, Shawe-Taylor, Scholkopf, and Smola (1999)). These methods can also be applied to some problems of unsupervised learning (Smola, Williamson, Mika, and Scholkopf (1999)).
3. Extensions to non- Norms: The results of Williamson, Smola, and Scholkopf (2001) hinge on the measurement of the size of the weight vector using the norm. Williamson, Smola, and Scholkopf (2000) show the impact of different norms for measuring the size of , as well as a number of related results.

**References**

* Guo, Y., P. L. Bartlett, J. Shawe-Taylor, and R. C. Williamson (1999): Covering Numbers for Support Vector Machines, in: *Proceedings of the 12th Annual Conference on Computational Learning Theory* 267-277 **ACM** New York.
* Shawe-Taylor, J., and R. C. Williamson (1999): Generalized Performance of Classifiers in Terms of Observed Covering Numbers, in: *Proceedings of the 4th European Conference on Computational Learning Theory (EUROCOLT ’99)* 274-284.
* Smola, A. J., R. C. Williamson, S. Mika, and B. Scholkopf (1999): Regularized Principal Manifolds, in: *Proceedings of the 4th European Workshop in Computational Learning Theory (EUROCOLT ’99)* 214-229.
* Smola, A. J., A. Elisseff, B. Scholkopf, and R. C. Williamson (2000): Entropy Numbers for Convex Combinations and mlps, in: *Advances in Large Margin Classifiers (A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors)* **MIT Press** Cambridge MA.
* Williamson, R. C., J. Shawe-Taylor, B. Scholkopf, and A. J. Smola (1999): Sample-based Generalization Bounds *NeuroCOLT2 Technical Report Series* **NC-TR-1999-055**.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2000): Entropy Numbers of Linear Function Classes, in: *Proceedings of the 13th Annual Conference on Computational Learning Theory (N. Cesa-Bianchi and S. Goldman, editors)* 309-319.
* Williamson, R. C., A. J. Smola, and B. Scholkopf (2001): Generalization Performance of Regularization Networks and Support Vector Machines via Entropy Numbers of Compact Operators *IEEE Transactions on Information Theory* **47 (6)** 2516-2532.

**Section VII: Alternate Statistical Learning Approaches**

**Minimum Description Length Approach**

**Motivation**

1. MDL Function Coding Philosophy: The MDL approach is based on the following notion of simplicity: A function object is simple if it only needs a small number of bits to code it.
2. MDL Function Coding Approach: A naïve function coding approach maintains a simple input-to-output table scheme (either direct or through ranges). More sophisticated coding schemes us the theory of data coding and compression to uncover data regularity.

**Coding Approaches**

1. MDL Coding Practice: Given a function space and a set of training points, we try to pick out the function

that minimizes the following expression:

where is the length of the code used to encode the function

1. : The term denotes the length of the code needed to express the given training data with the aid of the function . The idea is simple; if fits the training data well, this part of the code will be very small.
2. MDL vs. Classical SLT: corresponds to the training error the function makes on the data. The term measures the complexity of . Thus MDL and SLT are both very similar; the both sum up the training error and the complexity term. However, in SLT, the complexity term is only a function of , whereas in MDL it is a function of .
3. Building Universal Function Codes: One way to approach the above is to decompose into subsets

Then we encode the elements of each subset with a fixed length code that assigns each member of the same length. If is finite with elements, it is possible to encode each member

using bits. If is infinite, one may use the VC dimension (or some complexity measure) to encode the function.

1. MDL Coding Complexity: The *coding complexity* of can be defined as the smallest code length one can achieve for encoding the functions

It is uniquely defined, and is *universal* in the sense that it does not rely on a specific scheme.

1. Function Coding Technique: Given a function

we encode our data in several steps; we first encode the index of the function class, then we use a uniform code to encode which element of the function is, and finally code the data with help of The code length then becomes

The goal of this segmented approach is, of course, to choose the hypothesis that minimizes this code.

* 1. Function Class Code => Usually for one chooses a uniform code of integers, as long as we are dealing with finitely many classes . Thus, this entity is typically constant across all classes.
  2. => This term is identical for all

It does not distinguish between different functions within , but only from with functions that come from different ’s.

* 1. => This term explains how well the given can explain the training data.

**MDL Analyses**

1. MDL for a Fixed Function Class : If we are given just one fixed function class (without splitting it down into sub-classes ), the code length essentially depends on some measure of complexity of plus an additional term that accounts for how well fits the data. In this setting, it is easy to prove that the learning bounds of MDL compare as well with SLT.
2. Compression Coefficients SLT: As shown in Vapnik (1995), the MDL approach outlined above is closely related to the classical SLT approaches based on compression bounds.
3. MDL and the PAC-Bayesian Approach: In more advanced MDL approaches one assigns different code lengths to different elements of – in this case the approach is more closely related to the PAC-Bayesian approach.
4. Infinite Data Limit: Just as in classical SLT (but under certain other assumptions), MDL can be performed in a consistent way, i.e., in the infinite data limit, the approach can find the correct model for the data (Grunwald (2007), Rissanen (2007)).

**References**

* Grunwald, P. (2007): *The Minimum Description Length Principle* **MIT Press** Cambridge, MA.
* Rissanen, J. (2007): *information and Complexity in Statistical Modeling* **Springer** New York.
* Vapnik, V. (1995): *The Nature of Statistical Learning Theory* **Springer Verlag** New York.

**Bayesian Methods**

**Bayesian and Frequentist Approaches**

1. Bayesian Approach - Principle: As opposed to the agnostic approach taken in classical SLT, we assume that the underlying data distribution comes from some class of probability distribution . This probability class is described by one/more parameters, i.e., it is of the form

where is the set of parameters values, and the corresponding distribution parameters (e.g., if it is normal, this may be the mean/variance).

1. Goal of the Frequentist Approach: The main goal of the frequentist approach is to infer the correct parameter set of the underlying distribution. Classification etc. is then performed on the data set using the inferred parameters.
   1. Maximum Likelihood Estimate => The frequentist approach de facto estimates the likelihood of the data given the parameter realization, i.e., . The MLE computes the parameter that maximizes this likelihood, i.e.

Obviously MLE only estimates the confidence of the parameters given the data, not the probability that parameters are correct in any absolute sense.

**Bayesian Approaches**

1. Motivation: Here we define a distribution which, for each parameter , encodes how likely we find that this is a good parameters to describe our problem. The important point is that this prior distribution is defined *before* we get to see the data points from which we like to learn.
2. Literature:
   1. Cox (1961) introduces the fundamental axioms that allow expressing beliefs using probability calculus.
   2. Jaynes (2003) addresses philosophical/practical issues with these axioms.
   3. O’Hagan (1994) provides a very general treatment.
   4. Bayesian methods in Machine Learning approaches are available in Tipping (2003) and Bishop (2006).
3. Computation of the Posterior: As in the frequentist approach we do compute , the likelihood term. In addition, combining with the prior, we can also compute the posterior distribution as
4. MAP Estimator: The MAP approach involves choosing the that maximizes the posterior probability
5. Bayesian Approaches for Finite Samples: On a finite sample, usage of the prior helps bias towards solutions more likely (thereby, no washing out of the prior, see Berger (1985)). Here it is more appropriate to indicate that the Bayesian method is more convenient means of updating our beliefs about the solutions before looking at the data.
6. MAP re-cast in SLT terms:

Like SLT, is the term that describes the model fit to the data.

1. as a measure of Model Complexity: If there are more models, then the probability gets stretched across the entire parameter space, thereby trimming/limiting the contribution due to each , thus increasing . From coding theory, this indicates that, as the model complexity increases, greater is . This is consistent with the concept of model complexity of SLT.

**References**

* Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis* **Springer Verlag** New York.
* Bishop, C. (2006): *Pattern Recognition and Machine Learning* **Springer**.
* Cox, R. T. (1961): *The Algebra of Probable Inference* **Johns Hopkins University Press** Baltimore.
* Jaynes, E. T. (2003): *Probability Theory – The Logic of Science* **Cambridge University Press** Cambridge.
* O’Hagan, A. (1994): Bayesian Inference, in: Volume 2B, *Kendall’s Advanced Theory of Statistics* **Arnold** London.
* Tipping, M (2003): Bayesian Inference: An Introduction to Principles and Practice in Machine Learning, in: *Advanced Lectures in Machine Learning* (editors O. Bosquet, U. von Luxburg, and G. Ratsch) 41-62 **Springer**.

**Knowledge Based Bounds**

**Places to incorporate Bounds**

1. The Purpose: The purpose is to get tighter bounds into the pessimistic bounds above, as well as for accounting for the issues raised in the *NFL Theorems*, which states that learning is impossible unless we make assumptions on the underlying distribution.
2. Incorporating Prior Knowledge into Data Space: This may be done using a topology based distance metric, for e.g., and the closeness of this metric signals the likelihood of the output. The classifier inputs will then be coded using these metric transforms (e.g., as in kNN).
3. Prior Knowledge via the Underlying Probability Distribution: This simply an alternate way of specifying

However, this *hits* our assumption of distribution agnosticity (we are OK with that in this section, given that we try to cater to NFT).

1. Encoding Prior Knowledge via the Loss Function : The loss function encodes the learning goals, so can be customized in different manner using different weights (e.g., on daily points or “error types”).
2. Example: In many problems “false positives” and “false negatives” have differential costs associated. For instance, in spam filtering, the cost of accidentally labeling the spam as “not spam” is not that high, whereas the cost of labeling a non-spam as a spam is high.

**Prior Knowledge into the Function Space**

1. Prior Knowledge into the Choice of : Here the assumptions of usefulness of a classifier are encoded by the appropriate choice of

Generally, this is not very useful, as it still leaves significant wiggle room in the choice – the exception where the encoding via becomes important is in Bayesian approaches.

1. Shortcomings of the Capacity Measures: While both

and the data points go together into the construction of the capacity/complexity measure, they do not typically weight the contribution of a specific (except in a fairly cumbersome manner). The focus is on how to incorporate the knowledge into the bounds rather than simply counting the function in , either via knowledge a *priori* or *a posteriori*.

1. Classifier Function Specific Prior Knowledge: The Bayesian approach is one way to incorporate the classifier function specific prior knowledge. A prior distribution expresses our belief on how to place emphasis on the classifier function space. However, should be chosen well before the statistical inference process is started.
2. PAC Bayesian Bounds: With the incorporation of the Bayesian priors into the classical SLT framework, it can be shown that, with a probability of at least

This is the simplest among the PAC bounds (Boucheron, Bousquet, and Lugosi (2005)), and does not involve the capacity term for ; instead, it penalizes the individual functions according to .

* 1. Data Agnosticism of PAC Bounds => While the PAC Bayesian approach outlined above accounts for fixing the function agnosticism challenge, it is data sample agnostic, and therefore produces conservative bounds without differentiating among the different types of data.

1. Luckiness Framework: Here, both the bounds as well as the capacity are allowed to depend on the actual data at hand, thus varying with different “groups” of data. This approach was first introduced in statistics under the name “conditional confidence sets” (Keifer (1977)), and then expanded under the classical SLT framework by Shawe-Taylor, Bartlett, Williamson, and Anthony (1998) and Herbrich and Williamson (2002).

**References**

* Boucheron, S., O. Bousquet, and G. Lugosi (2005): Theory of Classification *ESAIM: Probability and Statistics* **9** 323-375.
* Herbrich, R., and R.C. Williamson (2002): Learning and Generalization: Theoretical Bounds, in *Handbook of Brain Theory and Neural Networks* (editor: M Arbib).
* Keifer, J. (1977): Conditional Confidence Statements and Confidence Estimators *Journal of the American Statistical Association* **72 (360)** 789-808.
* Shawe-Taylor, J., P. Bartlett, R. Williamson, and M. Anthony (1998): Structural Risk Minimization over Data-dependent Hierarchies *IEEE Transactions on Information Theory* **44 (5)** 1926-1940.

**Section VIII: Bayes’ Consistency and No Free Lunch Theorems**

**Approximation Error and Bayes’ Consistency**

**Motivation**

1. Estimation vs. Approximation Error Optimization - Contradictory Goals: In general, the estimation error may be made small by reducing the complexity of the hypothesis space. However, the approximation error requires a widening of the function class!
2. Approaches for Tackling Bayes’ Consistency: If

for some known, small function space , the approximation error becomes zero, and the Bayes’ consistency criterion then reduces to consistency with respect to . If the above is not the case, then the only other known method is to ensure that the nested function spaces contain “*in the limit*”.

**Nested Function Spaces**

1. The Concept: As the sample size increases, so does the complexity of the function space . The standard construction is to choose the spaces such that they constitute an expanding sequence of nested function spaces, i.e.

Two conditions are needed to ensure this classifier is Bayes consistent.

1. Zero Estimation Error:

as

The estimation error bound decreases as the sample size increases, but increases as the complexity term increases. To ensure that the overall estimation error is still decreasing, the complexity term contribution should not dominate the sample size contribution, i.e., does not grow too fast as increases.

1. Zero Approximation Error:

as

This is possible only if, for some large , either is contained in , or it can be approximated by it closely enough.

1. Necessary and Sufficient Condition for Bayes’ Consistency: This comes from Devroye, Gyorfi, and Lugosi (1996). Let

be a sequence of nested function spaces, and consider the classifier set described by

Assume that for a distribution the following conditions are satisfied:

* 1. The VC Dimension of the spaces in satisfies

as

and

as

Then the sequence of classifiers is Bayes’ consistent.

1. Bayes’ Consistency – Sample: Say we choose the function space such that

for some

Then the first condition is satisfied, as

as

However if

then

as

**Regularization**

1. Penalizing Regularizer: An implicit way of working with nested function spaces is the principle of regularization. Instead of minimizing the empirical risk and the expressing the generalization ability of the resulting classifier using some capacity measure of the underlying function class , one can directly minimize the regularized risk expressed as
2. Regularizer Behavior: is called the regularizer, and it is designed to punish complexity. For instance, to punish functions with large fluctuations, one can choose a that is small for functions that are smooth/vary smoothly, and large for functions that fluctuate a lot. Thus, for linear classifiers one chooses as the inverse of the margin of a function.
3. Regularizer Trade-off Constant: weights vs. . High implies greater penalty, and one ends up choosing functions with low empirical risk. Low implies more empirical risk.
4. Regularization Principle: This principle chooses the that minimizes the regularization risk . Many popular classifiers can be cast in this regularization frame-work (Scholkopf and Smola (2002)).
5. Trade-off Factor Impact for Bayes’ Consistency: Ensuring Bayes’ consistency in the regularization framework requires that:
   1. The nested function space from contain the Bayes’ classifier
   2. The impact of should be such that, as

the ERM hypothesis space contribution dominates the penalty contribution

* 1. The regularizing penalty should still have an impact at

i.e., as

should not go to too fast (Steinwart (2005)).

1. Function Class Complexity vs Function Complexity: While ERM worries about function class complexity, regularization, de facto, is concerned with the *complexity* as measured by on an individual function.

**Achieving Zero Approximation Error**

1. Approximation Theory to achieve Zero Approximation Error: Essentially we need to ensure that

or that is approximatable from for large enough . Often simple results from approximation theory are sufficient (e.g., Cucker and Zhang (2007) work pout more sophisticated one). However, from an approximation viewpoint, we need to be able establish that if two functions are *close*, then their corresponding risk values are also close.

1. Approximation Error for Binary Classifiers: If approximates (where and are binary classifiers), and their risk is less than , so is their risk, i.e.

Thus, to prove that the approximation error of a function space is smaller than , we have to show that every

can be approximated up to in the norm by functions from .

1. Approximation Theory Application - Real Numbers: Results of the above type are abundant in the literature. For example, if is a bounded set of real numbers, it is well-known that on can approximate any measurable function on this set well by a polynomial. Hence on would choose the space as the spaces of polynomials with at most a degree of where grows slowly with . This is enough to guarantee the convergence of the approximation error.

**Rate of Convergence**

1. Definition: *Rate of Convergence* gives information about how fast a quantity converges – in particular, how large should be so as to ensure that the error is smaller than a certain quantity, i.e., for estimation error , we need

samples.

1. Factors determining Rate of Convergence: Rate of Convergence depends on many parameters of the underlying problem, but while the estimation error is independent of the underlying distribution , the convergence of the approximation error is dependent on it.
2. Nested Function Space: Unless one makes additional assumptions on the true distribution, for any fixed sequence of nested function spaces, the rate of convergence of the approximation error can be arbitrarily low.
3. Weakening of the i.i.d. Assumptions: Rates of Convergence depend strongly on the underlying assumptions. For example, even with independent (but not identical) sampling, no uniform rate of convergence for the approximation error exists. A similar situation is true even for the estimation error if we weaken the sampling assumptions, where the ’s are not i.i.d. anymore.
4. Near i.i.d. Sampling: While for nearly independent sampling such as stationary -mixing process it may still be possible to recover the results similar to above, as soon as we leave this regime it can become impossible to achieve such results (Steinwart, Hush, and Scovel (2006)). Even in the case of stationary ergodicity, we can no longer achieve universal consistency (Nobel (1999)).
5. Fast Rates: If we assume that the sample points are i.i.d., and make a few assumptions about the underlying distribution (in particular, about the label noise), the rates of convergence can be improved dramatically. This branch of learning is called *Fast Learning* (Boucheron, Bousquet, and Lugosi (2005)).
6. Consistency vs. Rate of Convergence - Reprise: Consistency is a *worst case* statement; it indicates that ultimately the algorithm will provide correct solution without systematic errors. Rates of insights provides insights on how well-behaved the algorithm is, how fast it converges (sample sizes, etc.), and which approach is most fruitful and under which situation.

**References**

* Boucheron, S., O. Bousquet, and G. Lugosi (2005): Theory of Classification *ESAIM: Probability and Statistics* **9** 323-375.
* Cucker, F., and D. X. Zhou (2007): *Learning Theory – An Approximation Viewpoint* **Cambridge University Press**.
* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Nobel, A. (1999): Limits to Classification and Regression Estimation from Ergodic Processes *The Annals of Statistics* **27 (1)** 262-273.
* Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
* Steinwart, I. (2005): Consistency of Support Vector Machines and other Regularized Kernel Classifiers *IEEE Transactions on Information Theory* **51 (1)** 128-142.
* Steinwart, I., D. Hush, and C. Scovel (2006): Learning from Dependent Observations *Technical Report LA-UR-06-3507* **Los Alamos National Laboratory**.

**No Free Lunch Theorem**

**Introduction**

1. Best Classifier across all Distributions: Under the assumption of i.i.d. sampling from an underlying distribution, is there a classifier that *on average over all probability distributions* achieves better results than any other? NFT proves that the answer is ***NO***. For the finite case proof see Ho and Pepyne (2002), for convergence rates, see Devroye, Gyorfi, and Lugosi (1996), and for the most general proof see Wolpert and Macready (1997) and Wolpert (2001).
2. Intuition Behind NFT: For any set of for which a given classifier performs better than any other, we can always construct another where the classifier performs worse than any other (Ho and Pepyne (2002)).
3. Implication: The most important ramification of NFT is: *Learning is possible ONLY in the presence of definite, non-uniform*  (uniform reduces this to a uniform Bayes’ prior scenario, which is de-facto information free). Thus, prior knowledge needs to be *injected* using one of the means above.
4. NFT vs. Consistency: NFT does not contradict the (universal/functional) consistency criterion, since, in some ways, consistency also indicates convergence of risk at

sample sizes independent of (remember translates into ). However, at finite sizes, some may be more effective than the other.

1. NFT vs. Hypothesis Space Reduction: NFT simply states that, in order to be able to learn successfully with guarantees on the behavior of the classifier, we need to make assumptions on the underlying distribution. This fits with the complexity of - more complexity, less limiting assumption on .
2. Combining NFT and the Hypothesis Reduction Paradigms: Use NFT principles to first restrict the space of probability distributions, and then use a small function class that is able to model the distributions in the class.

**Algorithmic Consistency**

1. Countably Finite Data Space: When the data space is finite, even where is no relationship between the inputs and the outputs, since repeated sampling of inputs “converge” to their *TRUE* values, simple learning-by-heart (i.e., a majority vote when the instance has been seen before, and an arbitrary prediction otherwise) would be consistent.
2. Uncountably Finite or Infinite Data Space: For the uncountable case, there is a subtle hidden assumption. To be able to define a probability measure on , one needs a -algebra on the space, which is typically the Borel -algebra. So the hidden assumption is that is a Borel measure.
3. Topological Consistency: The fact that is a Borel measure means that the topology of plays a role, and therefore the target function (the Bayes’ classifier) is a Borel measurable. This guarantees that it is possible to approximate the target function from its value (or approximate value) at a finite number of points.
4. Continuous Data Space: The algorithms that will achieve consistency are thus those that will use the topology in the sense of “generalizing” the observed values to their neighborhoods (e.g., local classifiers). In a way, measurability of the target function is one of the crudest notions of smoothness of functions.

**NFT Formal Statements**

1. Basic Statement: Also called as theorem on *Arbitrarily Close to Random Guessing* (Devroye, Gyorfi, and Lugosi (1996)); Fix some

For every

and for every classifier , there exists a distribution with Bayes’ risk such that the expected risk of is greater than .

1. Statement on Algorithmic Consistency: An algorithm is consistent if for any probability measure

almost surely.

1. No Free Lunch At All: For any algorithm, and any sequence () that converges to , there exists a probability distribution such that

and

1. VC Consistency of ERM: The ERM algorithm is consistent if for any probability measure

in probability, and

in probability.

1. VC non-trivial Consistency of ERM: The ERM algorithm is non-trivially consistent for the set and the probability distribution if for any

in probability.

**References**

* Devroye, L., L. Gyorfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition* **Springer** New York.
* Ho, Y. C., and D. L. Pepyne (2002): Simple Explanation of the No Free Lunch Theorem and its Implications *Journal of Optimization Theory and Applications* **115 (3)** 549-570.
* Wolpert, D., and W. G. Macready (1997): No Free Lunch Theorems for Optimization *IEEE Transactions on Evolutionary Computing* **1 (1)** 67-82.
* Wolpert, D. (2001): The Supervised Learning No Free Lunch Theorems *Proceedings of the 6th Online World Conference on Soft Computing and Industrial Applications*.