Statistical Learning Library in DRIP

Lakshmi Krishnamurthy

v0.60 13 April 2015

Overview

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).

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.
- 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.
- 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.
- Natarajan, B. (1991): *Machine Learning: A Theoretical Approach* **Morgan Kauffman** San Mateo, CA.
- Scholkopf, B. and A. Smola (2002): *Learning with Kernels* **MIT Press** Cambridge, MA.
- 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.

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 $a\mathbb{I}_{x\geq a} \leq x$ or $a\mathbb{I}_{x>a} < x$. Here x is the "bad metric", and a is the bounds (or cut-off) on the "bad metric".
2. Bounds Setting Techniques:
    a. Markov Inequality
    b. Jensen's Convexity Inequality
    c. Cauchy's Joint Bounding Inequality
    d. Sample Independence Inequality
    e. Local Taylor Expansion Inequality
    f. Local Extrema (mini-max) Inequality
    g. Cross Function (e.g., Hypothesis) Union Bounding Inequality
    h. 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: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.
2. Maximized Function Space Inequality: $E\left[\underset{f \in \Im}{sup}(\ldots)\right] \geq \underset{f \in \Im}{sup} E[\ldots]$, since the former automatically filters out and retains only the quadrature maxima.
3. Markov Inequality: Applies to the situation where x, a are strictly non-decreasing. Starting from $a\mathbb{I}_{x \geq a} \leq x$ or $(a\mathbb{I}_{x>a} < x)$, and taking expectations on both sides, we get $P(x \geq a) \leq \frac{E(x)}{a}$ (or $P(x > a) < \frac{E(x)}{a}$). This upper bounds the probability of a "bad outcome".
4. Chebyshev Inequality: Here we set $E(x) \rightarrow \{x - E(x)\}^2$. Then $P(x \geq a) \leq \frac{E(x)}{a} \rightarrow P(\{x - E(x)\}^2 \geq a^2) \leq \frac{E(\{x-E(x)\}^2)}{a^2} \rightarrow P(\{x - E(x)\} \geq a) \leq \frac{Var(x)}{a^2}$. This is the Chebyshev's inequality.
5. Generalized Moment Bounds: Set $E(x) \rightarrow \{x - E(x)\}^m$. Using the treatment above, we get $P(\{x - E(x)\} \geq a) \leq \frac{Moment_{2k}(x)}{a^{2k}}; k = 1,2, \ldots, \infty$. Here $Moment_{2k}(x) = E(\{x - E(x)\}^{2k})$.
6. Chernoff Bound: Here $E(x) \rightarrow e^x$, so $P(x \geq a) = P(e^x \geq e^a) \leq e^{-a}E[e^x]$. In practice, an additional coefficient t that allows for optimization is inserted, so $P(x \geq a) \leq e^{-at}E[e^{xt}]$ for $a, t > 0$.
7. Chernoff vs. Moment Bounds: While both Chernoff bounds and generalized moment bounds allow for customization and/or optimization using the t parameter and the k 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 X and Y have finite second moments ($\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$), then $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$.
    a. Consider 2 sequences of the random variables of X and Y - $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$. Now, $\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i y_j - x_j y_i\right)^2 \geq 0$ for obvious reasons. Expanding, we get $\sum_{i=1}^{n}\sum_{j=1}^{n} {x_i}^2 {y_j}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} {x_j}^2 {y_i}^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j y_i y_j \Longrightarrow 2\left(\sum_{i=1}^{n} {x_i}^2\right)\left(\sum_{j=1}^{n} {y_j}^2\right) - 2\left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{j=1}^{n} x_j y_j\right) \geq 0$. Since $\sum_{i=1}^{n} {x_i}^2 = \mathbb{E}[X^2]$ and $\sum_{i=1}^{n} {y_i}^2 = \mathbb{E}[Y^2]$ and $\sum_{i=1}^{n} x_i y_i = \mathbb{E}[XY]$, the result follows.
2. Chebyshev-Cantelli Inequality: Let $t \geq 0$. Then $\mathbb{P}[X - \mathbb{E}[X] \geq t] \leq \frac{Var(X)}{Var(X)+t^2}$.
    a. Proof => Re-cast $X - \mathbb{E}[X] \to X$, $Var(X)$ does not alter. For all t, $t = \mathbb{E}[t - X] \leq \mathbb{E}[(t - X)\mathbb{I}_{X<t}]$ where $\mathbb{I}$ denotes the indicator function. Thus, for $t \geq 0$, from the Cauchy-Schwartz inequality, $t^2 \leq \mathbb{E}[(t - X)^2]\mathbb{E}\left[{\mathbb{I}_{X<t}}^2\right] = \mathbb{E}[(t - X)^2]\mathbb{P}(X < t) = [Var(X) + t^2]\mathbb{P}(X < t)$, that is Then $\mathbb{P}(X < t) \geq \frac{t^2}{Var(X)+t^2} \Longrightarrow \mathbb{P}(X \geq t) \leq \frac{Var(X)}{Var(X)+t^2}$.

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 f and g both be non-decreasing (or non-increasing) real-valued functions defined on the real line. If X is a real-valued random variable, then $\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$. If f is non-increasing and g is non-decreasing (or vice-versa), then $\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$.
    a. Proof => Let the random variable Y be distributed as X and be independent of it. If f and g are non-decreasing, then $[f(X) - f(Y)][g(X) - g(Y)] \geq 0$ so that

$\mathbb{E}\{[f(X) - f(Y)][g(X) - g(Y)]\} \geq 0$. Expand this expectation to obtain the first inequality. The proof of the second inequality is similar.

3. Multi-variate Association Inequality: An important generalization of the Chebyshev's association inequality is described as follows. A real-valued function f defined on $\mathbb{R}^n$ is said to be non-decreasing (non-increasing) in each variable while keeping all the other variables fixed at any value.
4. Harris' Inequality: Let $f: \mathbb{R}^n \to \mathbb{R}$ be a non-decreasing functions. Let $X_1, .., X_n$ be independent real-valued random variables and define the random variable $X = (X_1, .., X_n)$ taking values in $\mathbb{R}^n$. Then $\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$. Similarly, if f is non-increasing and g is non-decreasing, then $\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)]\mathbb{E}[g(X)]$.
    a. Proof Step #1 => As before, it suffices to prove the first inequality. We proceed by induction. For $n = 1$ the statement is just Chebyshev's association inequality.
    b. Proof Step #2 => Now suppose that the statement is true for $m < n$. Then $\mathbb{E}[f(X)g(X)] = \mathbb{E}\{\mathbb{E}[f(X)g(X)|X_1, .., X_{n-1}]\} \geq \mathbb{E}\{\mathbb{E}[f(X)|X_1, .., X_{n-1}]\mathbb{E}[g(X)|X_1, .., X_{n-1}]\}$, because, given $X_1, .., X_{n-1}$, both f and g are non-decreasing functions of the n^{th} variable. Now it follows by independence that the functions f', g': $\mathbb{R}^{n-1} \to \mathbb{R}^n$ defined by $f'(x_1, \dots, x_{n-1}) = \mathbb{E}[f(X)|X_1 = x_1, .., X_{n-1} = x_{n-1}]$ and $g'(x_1, \dots, x_{n-1}) = \mathbb{E}[g(X)|X_1 = x_1, .., X_{n-1} = x_{n-1}]$ are non-decreasing functions, so by the induction hypothesis, $\mathbb{E}[f'(X_1, .., X_{n-1})g'(X_1, .., X_{n-1})] \geq \mathbb{E}[f'(X_1, .., X_{n-1})]\mathbb{E}[g'(X_1, .., X_{n-1})] = \mathbb{E}[f(X)]\mathbb{E}[g(X)]$.

Alternate Bounds

1. Moment Bounds vs. Chernoff Bounds: The moment bounds for tail probabilities are always better Chernoff bounds for tail probabilities. More precisely, let X be a non-negative random variable and set $t > 0$. The best moment bound for tail probability is $\min_q \mathbb{E}[X^q]t^{-q}$ where

the minimum is taken over all positive integers q. The best Chernoff bound is $\inf_{s>0} \mathbb{E}\left[e^{s(X-t)}\right]$. It can be shown that $\min_{q} \mathbb{E}[X^q]t^{-q} \leq \inf_{s>0} \mathbb{E}\left[e^{s(X-t)}\right]$.

2. First/Second Moments on Integer-valued Distributions: If X is a non-negative integer-valued random variable then $\mathbb{P}[X \neq 0] \leq \mathbb{E}[X]$. Further, it can be shown that $\mathbb{P}[X \neq 0] \leq \frac{Var(X)}{Var(X)+\{\mathbb{E}[X]\}^2}$.
3. Sub-Gaussian Bounds: We say that a random variable X has a sub Gaussian distribution if there exists a constant $c > 0$ such that for all > 0 $\mathbb{E}[e^{sX}] \leq e^{cs^2}$. It can be shown that there exists a universal constant K such that if X is sub-Gaussian, then for every possible integer q, $\{\mathbb{E}[{X_+}^q]\}^{\frac{1}{q}} \leq K\sqrt{cq}$.
4. Sub-Gaussian Bounds - Converse: Let X be a random variable such that there exists a constant $c > 0$ such that $\{\mathbb{E}[{X_+}^q]\}^{\frac{1}{q}} \leq \sqrt{cq}$ for every positive integer q. Then X is a sub-Gaussian – more precisely, for every $s > 0$, $\mathbb{E}[e^{sX}] \leq \sqrt{2}e^{\frac{1}{6}}e^{\frac{ces^2}{2}}$.
5. Sub-Exponential Bounds: We say that a random variable has a sub-exponential distribution if there exists a constant $c > 0$ such that for all $0 < s < \frac{1}{c}$, $\mathbb{E}[e^{sX}] \leq \frac{1}{1-cs}$. If is sub-exponential, then for every positive integer q $\{\mathbb{E}[{X_+}^q]\}^{\frac{1}{q}} \leq \frac{4c}{e}q$.
6. Sub-Exponential Bounds - Converse: Let X be a random variable such that there exists a constant $c > 0$ such that $\{\mathbb{E}[{X_+}^q]\}^{\frac{1}{q}} \leq cq$ for every positive integer q. Then X is sub exponential – more precisely, for any $0 < s < \frac{1}{c}$, $\mathbb{E}[e^{sX}] \leq \frac{1}{1-cs}$.

Non-Moment Based Bounding of Sum of Independent Random Variables - Hoeffding Bound

1. Motivation – Sum of Independent Random Variables: Chebyshev's inequality and independence immediately imply $\mathbb{P}\{|S_n - \mathbb{E}[S_n]| \geq \varepsilon\} \leq \frac{\mathbb{E}[|S_n-\mathbb{E}[S_n]|^2]}{\varepsilon} = \frac{\sum_{i=1}^{n}\mathbb{E}[|X_i-\mathbb{E}[X_i]|^2]}{\varepsilon}$.

Writing $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[|X_i - \mathbb{E}[X_i]|^2], = \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{n\varepsilon^2}$. This is referred to as the ***Weak Law of Large Numbers***.

2. Inadequacy of the Weak Law: Recall that, under some additional regularity conditions, the central limit theorem states that $\mathbb{P}\left\{\sqrt{\frac{n}{\sigma^2}}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\right| \geq \varepsilon\right\} \longrightarrow 1 - \Phi(\varepsilon) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{\varepsilon^2}{2}}}{\varepsilon}$ from which, within a certain range of parameters, we expect something like $\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i] \geq \varepsilon\right\} \approx e^{-\frac{n\varepsilon^2}{2\sigma^2}}$.
3. General Bounding Technique:
    a. 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.
    b. Bounding individual Errors => Use Markov bounds in conjunction with moment/exponential bounds to bound the individual empirical errors.
    c. 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).
4. 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).
5. Principle of Hoeffding's Bounding:
    a. Hoeffding's inequality bounds the individual variable exponential probability in terms of another.
    b. This bound, along with "exponential Chernoff scaling", results in a convex optimization problem that may be optimized for an optimal "free parameter".
6. Hoeffding's Lemma: Suppose X is a real variable with zero mean such that $Prob(X \in [a, b]) = 1$. Then $E[e^{sX}] \leq e^{\frac{1}{8}s^2(a-b)^2}$. Note that if $a \neq 0$ and $b \neq 0$, then $a < 0$ and $b > 0$. However, if $a = 0$ or $b = 0$, it is easy to see that the inequality follows.
7. Proof of Hoeffding's Lemma:

a. Exploit e^{sX} Convexity => $e^{sx} \leq \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$. Applying expectation to both sides, $E[e^{sX}] \leq \frac{b-E[X]}{b-a}e^{sa} + \frac{E[X]-a}{b-a}e^{sb}$. Using $E[X] = 0$, the LHS becomes $\left(-\frac{a}{b-a}\right)e^{sa}\left[-\frac{b-a}{a} - 1 + e^{s(b-a)}\right] = \left[1 - \theta + \theta e^{s(b-a)}\right]e^{-s\theta(b-a)}$ where $\theta = -\frac{a}{b-a} > 0$.

b. Substitution => Let $u = s(b-a)$ and define $\varphi(u) = -\theta u + log(1 - \theta + \theta e^{u})$ and $(1 - \theta + \theta e^{u}) = \theta\left[-\frac{b}{a} + e^{u}\right] \geq 0$ if $\theta \geq 0, \frac{b}{a} \geq 0$. Thus, $E[e^{sX}] \leq E\left[e^{\varphi(u)}\right]$.

c. Invoke Taylor's Expansion Inequality => By Taylor's theorem, for every real u there exists a v between 0 and u such that $\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2}u^2\varphi''(v)$. Note that $\varphi(0) = 0$, $\varphi'(0) = 0$, and $\varphi''(0) = \frac{\theta e^{v}}{1-\theta+\theta e^{v}}\left[1 - \frac{\theta e^{v}}{1-\theta+\theta e^{v}}\right] = t(1-t)$. Recognizing that $t > 0$, $t(1-t) \leq \frac{1}{4}$.

d. Bring it together => $\varphi(u) \leq \frac{1}{2}u^2\frac{1}{4} \leq \frac{1}{8}s^2(b-a)^2$ => $E[e^{sX}] \leq E\left[e^{\frac{1}{8}s^2(b-a)^2}\right]$.

8. Hoeffding's Inequality Statement: Let $X_1, \ldots, X_n$ be independent random variables such that $Prob(X_i \in [a_i, b_i]) = 1; 1 \leq i \leq n$. Set $S_n = X_1 + \cdots + X_n$. Then $Prob(S_n - E[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}\left(b_i{}^2 - a_i{}^2\right)}}$.

a. Proof Step #1: For $s, t \geq 0$, the Chernoff inequality and the independence of X_i implies $Prob\{S_n - E[S_n] \geq t\} \leq Prob\left\{e^{s(S_n - E[S_n])} \geq e^{st}\right\} \leq e^{-st}E\left\{e^{s(S_n - E[S_n])}\right\} = e^{-st}\prod_{i=1}^{n}E\left\{e^{s(S_n - E[S_n])}\right\} \leq e^{-st}\prod_{i=1}^{n}e^{\frac{1}{8}s^2(b-a)^2} = e^{-st+\frac{1}{8}s^2\sum_{i=1}^{n}(b_i-a_i)^2}$.

b. Proof Step #2: To get the best possible upper bound, we find the minimum of the RHS as a function of s. Defining $g(t) = -st + \frac{1}{8}s^2\sum_{i=1}^{n}(b_i - a_i)^2$, $g(t)$ achieves its minimum at $\frac{4}{\sum_{i=1}^{n}(b_i-a_i)^2}$. Substituting this value of s establishes the inequality.

9. Boundedness of Hoeffding:

a. Recall that the Hoeffding inequality may only be applied for bounded (albeit negative), zero-mean random variables.

b. 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).

10. 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 σ^2 is replaced by $\frac{1}{4}\sum_{i=1}^{n}(b_i - a_i)^2$. Thus, Hoeffding's inequality ignores the information on the variance of X_i, thus an alternate approach that accommodates this becomes necessary.
11. Inequality of Absolute Value Deviation: The absolute value deviation around mean is looser:
    a. $Prob(S_n - E[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$.
    b. $Prob(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$.
    c. For ERM, we are primarily interested in absolute value deviations, so this could be another reason why optimized moment bounds (including optimized factorial moment bounds for integer valued random variables) offer tighter bounds.

Moment Based Bounds

1. Moment-Bound Series: Without loss of generality, assume that $\mathbb{E}[X_i]$ for all $i = 1, \ldots, n$.We look for the bounds of $\mathbb{E}[e^{sX_i}]$. Setting ${\sigma_i}^2 = \mathbb{E}\left[{X_i}^2\right]$ and $F_i = \sum_{r=2}^{\infty}\frac{s^{r-2}\mathbb{E}[{X_i}^r]}{r!{\sigma_i}^2}$, and using $e^{sx} = 1 + sx + \sum_{r=2}^{\infty}\frac{s^r x^r}{r!}$ and $\mathbb{E}[X_i] = 0$, we have $\mathbb{E}[e^{sX_i}] = 1 + s\mathbb{E}[X_i] + \sum_{r=2}^{\infty}\frac{s^r\mathbb{E}[{X_i}^r]}{r!} = 1 + s^2{\sigma_i}^2F_i \leq e^{s^2{\sigma_i}^2F_i}$.
2. Bounding Individual X_i: Assume that each X_i is bounded such that $|X_i| \leq c$. Then for each $r \geq 2$, $\mathbb{E}[{X_i}^r] \leq c^{r-2}{\sigma_i}^2$. Thus, $F_i \leq \sum_{r=2}^{\infty}\frac{s^{r-2}c^{r-2}{\sigma_i}^2}{r!{\sigma_i}^2} = \frac{1}{(sc)^2}\sum_{r=2}^{\infty}\frac{(sc)^r}{r!} = \frac{e^{sc}-1-sc}{(sc)^2}$. Thus, we have obtained $\mathbb{E}[e^{sX_i}] \leq e^{s^2{\sigma_i}^2\frac{e^{sc}-1-sc}{(sc)^2}}$.
3. Bounding X_i Moments: The above expression for F_i may be bounded differently, in other ways. Possible options are:
    a. Applying the moment bounds considered earlier along with potential sub-Gaussian-ness.

b. Applying the moment bounds considered earlier along with potential sub-exponential-ness.
c. Applying extraneous "noise variance" criterion such as the Massart's or the Tsybakov's noise conditions.

4. Sum of Independent Variables: Using the Chernoff bound for bounding the sums of independent random variables we get $\mathbb{P}\{\sum_{i=1}^{n} X_i \geq \varepsilon\} \leq e^{s^2\sigma_i{}^2\frac{e^{sc}-1-sc}{(sc)^2}-s\varepsilon}$. Optimizing the choice of s, we find that the upper bound is minimized by $s = \frac{1}{c}\log\left[1+\frac{tc}{n\sigma^2}\right]$.
5. Bennett's Inequality (Bennett (1962)): Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean, and assume $|X_i| \leq c$ with probability one. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var\{X_i\}$. Then, for any $\varepsilon > 0$, $\mathbb{P}\{\sum_{i=1}^{n} X_i \geq \varepsilon\} \leq e^{-\frac{n\sigma^2}{c^2}h\left(\frac{ct}{n\sigma^2}\right)}$ where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.
    a. Bound for Real-Valued Regression => The bound $|X_i| \leq c$ may also be applied for real-valued regression in a piece-wise constant sense where we expect the basis spline functions to be bounded.
6. Bernstein Inequality Motivation: The meaning behind Bennett's inequality is best seen if we do additional bounding. Applying the basis inequality $h(u) = \frac{u^2}{2+\frac{2u}{3}}$ for $u \geq 0$ (which may be seen by comparing the derivatives of both sides), we obtain the classical (but more loose) inequality of Bernstein (Bernstein (1946)).
    a. Statement => Under the same conditions as Bennett's inequality, for any $\varepsilon > 0$, $\mathbb{P}\{\sum_{i=1}^{n} X_i \geq \varepsilon\} \leq e^{-\frac{n\varepsilon^2}{2\sigma^2+\frac{2c\varepsilon}{3}}}$.
7. Bernstein's Inequality vs. Central Limit Theorem Gaussian Bounds: Except for the term $\frac{2c\varepsilon}{3}$ 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.
8. Bernstein's Inequality vs. Poisson Bounds: If $\sigma^2 > \varepsilon$, then the upper bound given by Bernstein's inequality behaves as $e^{-n\varepsilon}$ instead of the $e^{-n\varepsilon^2}$ guaranteed by Central Limit Theorem/Hoeffding. This may be intuitively explained by recalling that a $Binomial\left(n, \frac{\lambda}{n}\right)$

distribution can be approximated, for large n, by a $Poisson(\lambda)$ distribution whose tail decreases as $e^{-\lambda}$.

Alternate Moment-Based Bounds of Independent Random Variables

1. [0, 1] Bounded Random Variables: Let $X_1, \dots, X_n$ be independent random variables taking values in $[0, 1]$. Denoting $m = \mathbb{E}[S_n]$, for any $t \geq m$, Chernoff bounding can be used to show that $\mathbb{P}\{S_n \geq t\} \leq \left(\frac{m}{t}\right)^t \left(\frac{n-m}{n-t}\right)^{n-t}$.
    a. Corollary => From the above, it can be seen that $\mathbb{P}\{S_n \geq t\} \leq \left(\frac{m}{t}\right)^t e^{t-m}$, and for all $\varepsilon > 0$, $\mathbb{P}\{S_n \geq m(1+\varepsilon)\} \leq e^{-mh(\varepsilon)}$ where h is the function defined in Bennett's inequality. Finally, $\mathbb{P}\{S_n \leq m(1-\varepsilon)\} \leq e^{-\frac{m\varepsilon^2}{2}}$ (Karp (1988), Hagerup and Rub (1990)).
2. Poisson Random Variable: Comparing $\mathbb{P}\{S_n \geq t\} \leq \left(\frac{m}{t}\right)^t e^{t-m}$ with the best Chernoff bound for the tail of a Poisson random variable $Y = Poisson(m)$ reveals that $\mathbb{P}\{Y \geq t\} \leq \inf_{s>0} \frac{\mathbb{E}[e^{sY}]}{e^{st}} = \left(\frac{m}{t}\right)^t e^{t-m}$.
3. Sampling with Replacement: Let $\mathcal{X}$ be a finite set with $\mathcal{N}$ elements, and let $X_1, \dots, X_n$ be a random sample without replacement from $\mathcal{X}$ and $Y_1, \dots, Y_n$ a random sample with replacement from $\mathcal{X}$. For any convex real-valued function f, it may be shown that $\mathbb{E}[f(\sum_{i=1}^n X_i)] \leq \mathbb{E}[f(\sum_{i=1}^n Y_i)]$. In particular, by taking $f(x) = e^x$, we see that all inequalities derived from sums of independent random variables Y_i using Chernoff bounding remain true for the sum of X_i'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*.
- Marton, K. (1986): A Simple Proof of the Blowing-up Lemma *IEEE Transactions on Information Theory* **32** 445-446.
- Marton, K. (1996a): Bounding d-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.

Binomial Tails

1. Setup: We recall that the functions we consider are binary-valued. Thus, for a fixed function f, the distribution of $P_n f$ is just a binomial law of parameters Pf and n (since we are assuming n i.i.d. random variables $f(\mathbb{Z}_i)$ which can be either 0 or 1, and are equal to 1 with probability $\mathbb{E}[f(\mathbb{Z}_i)] = Pf$).
2. Formulation: Denoting $p = Pf$, we can have an exact expression for the deviations of $P_n f$ from Pf as $\mathbb{P}[Pf - P_n f \geq \varepsilon] = \sum_{k=0}^{\lfloor n(p-\varepsilon) \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$. 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.
3. Alternate Upper Bounds: The following quantities are an upper bound on $\mathbb{P}[Pf - P_n f \geq \varepsilon]$:
    a. Hoeffding => $e^{-2n\varepsilon^2}$
    b. Bernstein => $e^{-\frac{n\varepsilon^2}{2p(1-p)+\frac{2\varepsilon}{3}}}$
    c. Bennett => $e^{-\frac{np}{1-p}\left\{\left(1-\frac{\varepsilon}{p}\right)\log\left(1-\frac{\varepsilon}{p}\right)+\frac{\varepsilon}{p}\right\}}$
    d. Exponential => $\left(\frac{1-p}{1-p-\varepsilon}\right)^{n(1-p-\varepsilon)} \left(\frac{p}{p+\varepsilon}\right)^{n(p+\varepsilon)}$
4. Extremes of Binomial Distribution: From the above bounds, using inversion, we can say that, roughly speaking, small deviations of $Pf - P_n f$ have a Gaussian behavior of the form $e^{-\frac{n\varepsilon^2}{2p(1-p)}}$ (i.e., Gaussian with variance $p(1-p)$) while large deviations have a Poisson behavior of the form $e^{-\frac{3n\varepsilon}{2}}$.
5. 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., $\frac{1}{4}$), hence the term $e^{-2n\varepsilon^2}$.
6. Empirical Bernstein Bound: Each function $f \in \mathfrak{I}$ has a different variance, i.e., $Pf(1-Pf) \leq Pf$. For each $f \in \mathfrak{I}$, using Bernstein's inequality, with a probability of at least $1-\delta$, $Pf - P_n f \leq \sqrt{\frac{2Pf \log\frac{1}{\delta}}{n}} + \frac{2\log\frac{1}{\delta}}{3n}$.

7. Intuition Behind Talagrand's Inequality: The Gaussian part $\sqrt{\frac{2Pf\log\frac{1}{\delta}}{n}}$ dominates when Pf is not too small, or if n is large enough, and it depends on Pf. Thus, Talagrand's inequality strives to combine Bernstein's inequality with the union bound and symmetrization.

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 $\mathcal{X}$ be some set, and let $g : \mathcal{X}^n \to \mathbb{R}$ be a measurable function of n variables. We derive inequalities for the random variable $\mathcal{Z} = g(X_1, \dots, X_n)$ and its expected value $\mathbb{E}[\mathcal{Z}]$ where $X_1, \dots, X_n$ are arbitrary independent (but not necessarily identically distributed) random variables in $\mathcal{X}$.
3. Notation: The main Efron-Stein inequality follows from the next result – the *Martingale Differences Sum Inequality*. To simplify the notation, we use $\mathbb{E}_i$ for the expected value with respect to the variable X_i, that is, $\mathbb{E}_i = \mathbb{E}[\mathcal{Z}|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]$.

Martingale Differences Sum Inequality

1. Statement: $Var[\mathcal{Z}] \leq \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathbb{E}_i[\mathcal{Z}])^2]$, i.e., $Var[\mathcal{Z}]$ is less than or equal to the sum of variances of $\mathcal{Z}$ along each random variable dimension.
2. Proof:
    a. The proof is based on the elementary properties of conditional expectations. Recall that if X and Y are arbitrary bounded random variables, then $\mathbb{E}[XY] = \mathbb{E}\big[\mathbb{E}[XY|Y]]\big] = \mathbb{E}\big[Y\mathbb{E}[X|Y]\big]$.

b. Introduce $\mathcal{V} = \mathcal{Z} - \mathbb{E}[\mathcal{Z}]$ where the expectation is across all $X_1, \ldots, X_n$, and define $\mathcal{V}_i = \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_i] - \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}]$ for $i = 1, \ldots, n$. Under this notation, note that $\mathcal{Z} = \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_n]$ and $\mathcal{Z} = \mathbb{E}[\mathcal{Z}|\Theta]$ where Θ refers to a NULL variable set.

c. Clearly $\mathcal{V} = \sum_{i=1}^{n} \mathcal{V}_i$, since individual $\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_i]$ telescope out. Further, note that $\mathcal{V}_i$ and $\mathcal{V}_j$ are martingales in X_i and X_j, respectively. Thus, $\mathcal{V}$ is written as a sum of martingale differences.

d. Now $Var[\mathcal{Z}] = \mathbb{E}[(\sum_{i=1}^{n} \mathcal{V}_i)^2] = \mathbb{E}\left[\sum_{i=1}^{n} {\mathcal{V}_i}^2\right] + 2\mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \mathcal{V}_i \mathcal{V}_j\right] = \mathbb{E}\left[\sum_{i=1}^{n} {\mathcal{V}_i}^2\right]$, since, for any $i > j$, $\mathbb{E}\left[\mathcal{V}_i \mathcal{V}_j\right] = \mathbb{E}\left[\mathbb{E}\left[\mathcal{V}_i \mathcal{V}_j | X_i, \ldots, X_j\right]\right] = \mathbb{E}\left[\mathcal{V}_j \mathbb{E}\left[\mathcal{V}_i | X_i, \ldots, X_j\right]\right] = 0$, as $\mathcal{V}_i$ is a martingale.

e. To bound $\mathbb{E}\left[\sum_{i=1}^{n} {\mathcal{V}_i}^2\right]$, note that by independence of X_i, $\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}] = \mathbb{E}[\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n]|X_1, \ldots, X_i]$, and therefore ${\mathcal{V}_i}^2 = \{\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_i] - \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}]\}^2 = \{\mathbb{E}[\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_i] - \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}]|X_1, \ldots, X_i]\}^2 \leq \mathbb{E}[\{\mathbb{E}[\mathcal{Z}|X_1, \ldots, X_i] - \mathbb{E}[\mathcal{Z}|X_1, \ldots, X_{i-1}]\}^2|X_1, \ldots, X_i] = \mathbb{E}[(\mathcal{Z} - \mathbb{E}_i[\mathcal{Z}])^2|X_1, \ldots, X_i]$, 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 $X'_1, \ldots, X'_n$ form an independent copy of $X_1, \ldots, X_n$, and set $\mathcal{Z}'_i = g(X_1, \ldots, X'_i, \ldots X_n)$. Then $Var[\mathcal{Z}] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}'_i)^2]$.
2. Proof of the Inequality: The statement follows from the Martingale Differences Sum Inequality by using the elementary fact that if X and Y are independent and identically distributed random variables, then $Var[X] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(X - Y)^2]$, and therefore $\mathbb{E}[(\mathcal{Z} - \mathbb{E}_i[\mathcal{Z}])^2] = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}'_i)^2]$.
3. Relation to the Case of Sum of Independent Random Variables: Observe that in the case when $\mathcal{Z} = \sum_{i=1}^{n} X_i$ 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.

4. 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 X, $Var[X] \leq \mathbb{E}[(X-a)^2]$ for any constant $a \in \mathbb{R}$. Using this fact, we have for every $i = 1, \dots, n$, $\mathbb{E}[(\mathcal{Z} - \mathbb{E}_i[\mathcal{Z}])^2] \leq \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}_i)^2]$ where $\mathcal{Z}_i = g_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots X_n)$ for arbitrary measurable functions $g_i: \mathcal{X}^{n-1} \to \mathbb{R}$ of $n-1$ variables. Taking expectations, we get an extension of the Efron-Stein inequality: $Var[\mathcal{Z}] \leq \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}_i)^2]$.

Bounded Differences Inequality

1. Functions with Bounded Differences: We say that function $g: \mathcal{X}^n \to \mathbb{R}$ has the *Bounded Differences Property* if for some non-negative constants $c_1, \dots, c_n$ $\sup_{\substack{x_1, \dots, x_n \\ x'_i \in \mathcal{X}}} |g(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)| \leq c_i$, for $1 \leq i \leq n$. In other words, if we change the i^{th} variable of g keeping all the others fixed, the value of the function cannot change by more than c_i.
2. Bounded Differences Inequality – Statement: If g has the bounded differences property with constants $c_1, \dots, c_n$, then $Var[\mathcal{Z}] \leq \sum_{i=1}^{n} {c_i}^2$. This bound comes in particularly handy when the direct estimation of the variance becomes involved, but the bound is obtained effortlessly.
    a. Bounded Differences with Probability Bounds => If the random variables $X_1, \dots, X_n$ are independent and binary $\{0, 1\}$-valued with $\mathbb{P}\{X_i = 1\} = \mathcal{P}_i$ and that g has the bounded differences property with constants $c_1, \dots, c_n$ then it is easy to show that $Var[\mathcal{Z}] \leq \sum_{i=1}^{n} {c_i}^2 p_i(1-p_i)$.
3. Empirical Estimation of Multivariate Martingale Differences Inequality: For a sequence of size s with n multivariates, the empirical estimation goes as $\mathbb{O}(s^n)$. Thus, even for a modest sample of size $s = n = 10$, martingale differences empirical estimates can be $\sim 10^{10}$, i.e., the computational demands blow up very, very fast.

4. 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 n numbers $x_1, \dots, x_n \in [0, 1]$, 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 $g(x_1, \dots, x_n)$ be this number. The behavior of $\mathcal{Z} = g(X_1, \dots, X_n)$ when $X_1, \dots, X_n$ are independent has been extensively studied (Rhee and Talagrand (1987), Rhee (1993), Talagrand (1995)). By changing one of the x_i's, the value of $g(x_1, \dots, x_n)$ cannot change by more than one, so we have $Var[\mathcal{Z}] \leq \frac{n}{2}$. However, sharper bounds may be established using Talagrand's convex distance inequality discussed later.
5. 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 $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be 2 sequences of coin flips. Define $\mathcal{Z}$ as the length of longest common sub-sequence that appears in both sequences, i.e., $\mathcal{Z} = max\{k: X_{i_1} = Y_{j_1}, \dots, X_{i_k} = Y_{j_k}, where\ 1 \leq i_1 < \cdots < i_k \leq \text{n}, \text{and } 1 \leq j_1 < \cdots < j_k \leq \text{n}\}$.

a. Behavior of the Expectation => The behavior of $\mathbb{E}[\mathcal{Z}]$ has been investigated in many papers. It is known that $\frac{\mathbb{E}[\mathcal{Z}]}{n}$ converges to some number γ whose value is unknown, but conjectured to be $\frac{2}{1+\sqrt{2}}$, and is known to fall between 0.75796 and 0.83763.

b. Variance of $\mathcal{Z}$ => Changing one bit cannot change the length of the longest common sub-sequence by more than one, so $\mathcal{Z}$ satisfies the bounded differences inequality with $c_i = 1$. Thus $Var[\mathcal{Z}] \leq \frac{n}{2}$ (Steele (1986)). Thus, by Chebyshev's inequality, with large probability, $\mathcal{Z}$ is within a constant times $\sqrt{n}$ of its expected value. In other words, it is strongly concentrated around the mean, which means that the results on $\mathbb{E}[\mathcal{Z}]$ really tell us about the behavior of the longest common sub-sequence of two random strings.

6. Uniform Deviations of Indicator Functions: Let $X_1, \ldots, X_n$ be i.i.d. random variables taking their values in some set $\mathcal{X}$, and let $\mathcal{A}$ be a collection of sub-sets of $\mathcal{X}$. Let μ be a distribution of X_1, i.e., $\mu_n(\mathcal{A}) = \mathbb{P}\{X_1 \in \mathcal{A}\}$, and let μ_n denote the empirical distribution $\mu_n(\mathcal{A}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \in \mathcal{A}}$. The quantity of interest is $\mathcal{Z} = \underset{A \in \mathcal{A}}{sup} |\mu_n(A) - \mu(A)|$.

   a. Variance Bound of $\mathcal{Z}$ => If $\underset{n \to \infty}{Limit} \mathbb{E}[\mathcal{Z}] = 0$ for every distribution of X_i, then $\mathcal{A}$ is called a *Uniform Glivenko-Cantelli Class*, and Vapnik and Chervonenkis (1971) provide a combinatorial characterization of such classes. But regardless of $\mathcal{A}$, by changing a single X_i, $\mathcal{Z}$ can change by at most $\frac{1}{n}$, so regardless of the behavior of $\mathbb{E}[\mathcal{Z}]$, we always have $Var[\mathcal{Z}] \leq \frac{1}{2n}$ (in other words, $\mathcal{Z}$ need not be uniform). More information on the behavior of $\mathcal{Z}$ 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).

   b. 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 $\mathcal{Z}$. We do this in a slightly more general framework of the empirical process.

   c. Efron-Stein Bound Formulation => Let $\mathcal{F}$ be a class of real-valued functions (no boundedness is assumed), and define $\mathcal{Z} = g(X_1, \ldots, X_n) = \underset{f \in \mathcal{F}}{sup} \sum_{j=1}^{n} f(X_j)$. Observe that by symmetry, the Efron-Stein bound may be written as $Var(\mathcal{Z}) \leq$

$\frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(\mathcal{Z}-\mathcal{Z}'_i)^2]=\sum_{i=1}^{n}\mathbb{E}\left[(\mathcal{Z}-\mathcal{Z}'_i)^2\mathbb{I}_{\mathcal{Z}'_i<\mathcal{Z}}\right]$. Let $f^*\in\mathcal{F}$ denote the function that achieves the supremum in the definition of $\mathcal{Z}$, that is, $\mathcal{Z}=\sum_{j=1}^{n}f^*(X_j)$. Then, clearly, $(\mathcal{Z}-\mathcal{Z}'_i)^2\mathbb{I}_{\mathcal{Z}'_i<\mathcal{Z}}\le[f^*(X_i)-f^*(X'_i)]^2$ and therefore $Var(\mathcal{Z})\le\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}\{f(X_j)-f(X'_j)\}^2\right]\le 4\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{j=1}^{n}f(X_j)^2\right]$.

d. Efron-Stein Bound vs. Simple Bounded Differences => For functions $f\in\mathcal{F}$ taking values in $[0,1]$, from just the bounded differences property we derived $Var[\mathcal{Z}]\le 2n$. The new bound may be a significant improvement whenever the maximum of $\sum_{j=1}^{n}f(X_j)^2$ of the functions in $\mathcal{F}$ is small. More importantly, in deriving the new bounds we have not assumed any boundedness of the functions in $\mathcal{F}$.

e. 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).

7. First Passage Time in Oriented Percolation: Consider a directed graph such that the weight X_i is associated with the edge e_i such that the X_i are non-negative independent random variables with a second moment $\mathbb{E}[X_i^2]=\sigma_i^2$. Let v_1 and v_2 be fixed vertices of the graph. We are interests in the total weight of the path from v_1 to v_2 with minimum weight. Set $\mathcal{Z}=\min_{\mathbb{P}}\sum_{e_i\in\mathbb{P}}X_i$ where the minimum is taken over all the paths $\mathbb{P}$ from v_1 to v_2.

a. Denote the optimal path by $\mathbb{P}^*$. By replacing X_i with X'_i, we can see that the total minimum weight can only increase if the edge e_i is on $\mathbb{P}^*$, and therefore $(\mathcal{Z}-\mathcal{Z}'_i)^2\mathbb{I}_{\mathcal{Z}'_i<\mathcal{Z}}\le[X_i-X'_i]^2\mathbb{I}_{e_i\in\mathbb{P}^*}$. Thus, $Var[\mathcal{Z}]\le\mathbb{E}\left[\sum_i X'_i\mathbb{I}_{e_i\in\mathbb{P}^*}\right]\le\sigma^2\mathbb{E}\left[\sum_i\mathbb{I}_{e_i\in\mathbb{P}^*}\right]\le\sigma^2 L$ where L is the length of the longest path between v_1 and v_2.

8. 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).

a. Efron-Stein Formulation => Let $\mathcal{F}$ denote a class of $[0,1]$-valued functions on some space $\mathcal{X}$. For simplicity of exposition we assume that $\mathcal{F}$ is finite, although the results

below remain true as long as the measurability issues are taken care of. Given an i.i.d. sample $\mathcal{D}_n = \{\langle X_i, Y_i\rangle\}_{i\leq n}$ of n pairs of random variables $\langle X_i, Y_i\rangle$ taking values in $\mathcal{X} \times \{0,1\}$, for each $f \in \mathcal{F}$ we define the empirical loss $\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} l[f(X_i), Y_i]$ where the loss functional l is defined on $\{0,1\}^2$ by $l(y, y') = \mathbb{I}_{y\neq y'}$.

b. Efron-Stein Bound – Steps => In non-parametric classification and learning theory, it is common to select an element $f \in \mathcal{F}$ by minimizing the empirical loss. The quantity of interest in this section is the minimal empirical loss $\hat{\mathcal{L}} = \inf_{f \in \mathcal{F}} \mathcal{L}_n(f)$. The bounded Efron-Stein bound implies that $Var\left[\hat{\mathcal{L}}\right] \leq \frac{1}{2n}$. However, a more careful application of the Efron-Stein inequality reveals that $\hat{\mathcal{L}}$ may be much more concentrated than predicted by this simple inequality. Getting tight results for the fluctuations of $\hat{\mathcal{L}}$ provides better insight into the calibration of penalties in certain model selection methods.

c. Tighter Bounding of the Variance => Let $\mathcal{Z} = n\hat{\mathcal{L}}$ and let $\mathcal{Z}'_i$ be defined as in the Efron-Stein inequality, that is, $\mathcal{Z}'_i = \inf_{f \in \mathcal{F}} \{\sum_{j\neq i} l\left[f\left(X_j\right), Y_j\right] + l[f(X_i), Y_i]\}$ where $\langle X'_i, Y'_i\rangle$ is independent of D_n and has the same distribution as $\langle X_i, Y_i\rangle$. The convenient form of the Efron-Stein inequality to use is the following: $Var[\mathcal{Z}] \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}'_i)^2] = \sum_{i=1}^{n} \mathbb{E}\left[(\mathcal{Z} - \mathcal{Z}'_i)^2 \mathbb{I}_{\mathcal{Z}'_i > \mathcal{Z}}\right]$.

d. Tighter Efron-Stein Bound => Let f^* denote a possibly non-unique minimizer of the empirical risk so that $\mathcal{Z} = \sum_{i=1}^{n} l[f(X_i), Y_i]$. The key observation is that $(\mathcal{Z} - \mathcal{Z}'_i)^2 \mathbb{I}_{\mathcal{Z}'_i > \mathcal{Z}} = \{l[f^*(X_i), Y_i] - l[f^*(X'_i), Y'_i]\}^2 \mathbb{I}_{\mathcal{Z}'_i > \mathcal{Z}} = l[f^*(X'_i), Y'_i]\mathbb{I}_{l[f^*(X_i), Y_i]=0}$. Thus, $\sum_{i=1}^{n} \mathbb{E}\left[(\mathcal{Z} - \mathcal{Z}'_i)^2 \mathbb{I}_{\mathcal{Z}'_i > \mathcal{Z}}\right] \leq \sum_{i: l[f^*(X_i), Y_i]=0} \mathbb{E}_{X'_i, Y'_i}\left[l[f^*(X'_i), Y'_i]\right] \leq n\mathbb{E}[\mathcal{L}(f^*)]$ where $\mathbb{E}_{X'_i, Y'_i}$ denotes expectation with respect to the variables X'_i and Y'_i, and for each $f \in \mathcal{F}$, $\mathcal{L}(f) = \mathbb{E}\left[l[f(X), Y]\right]$ is the true expected of f. Therefore the Efron-Stein inequality implies that $Var\left[\hat{\mathcal{L}}\right] \leq \frac{\mathbb{E}[\mathcal{L}(f^*)]}{n}$.

e. Improvement over the Naïve Efron-Stein Bound => The result above is a significant improvement over the bound $\frac{1}{2n}$ whenever $\mathbb{E}[\mathcal{L}(f^*)]$ is much smaller than $\frac{1}{2}$. This is

very often the case. For example, we have $\mathcal{L}(f^*) = \hat{\mathcal{L}} - [\mathcal{L}_n(f^*) - \mathcal{L}(f^*)] \leq \frac{\mathcal{Z}}{n} + \sup_{f \in \mathcal{F}} \frac{\mathcal{L}_n(f) - \mathcal{L}(f)}{n}$, so that we obtain $Var[\hat{\mathcal{L}}] \leq \frac{\mathbb{E}[\hat{\mathcal{L}}]}{n} + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{\mathcal{L}_n(f) - \mathcal{L}(f)}{n}\right]$.

f. Bounding $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{\mathcal{L}_n(f) - \mathcal{L}(f)}{n}\right]$ => In most cases of interest, $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{\mathcal{L}_n(f) - \mathcal{L}(f)}{n}\right]$ may be bounded by a constant term depending on $\mathcal{F}$ times $n^{-\frac{1}{2}}$ (Lugosi (2002)), and the second term on the RHS is of the order of $n^{-\frac{3}{2}}$. Boucheron, Lugosi, and Massart (2003) detail exponential concentration inequalities for $\hat{\mathcal{L}}$.

9. Kernel Density Estimation: Let $X_1, \ldots, X_n$ be i.i.d. samples drawn according to some unknown density function f on the real line. The density is estimated by the kernel estimate $f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right)$ where $h > 0$ is a smoothing parameter, and $\mathcal{K}$ is a non-negative function with $\int \mathcal{K} = 1$. The performance of the estimate is measured by the L_1 error $\mathcal{Z} = g(X_1, \ldots, X_n) = \int |f(x) - f_n(x)| dx$.

   a. Bounding the Kernel Density Estimate => It is easy to see that $|g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x'_i, \ldots, x_n)| \leq \frac{1}{nh} \int \left|\mathcal{K}\left(\frac{x - X_i}{h}\right) - \mathcal{K}\left(\frac{x - X'_i}{h}\right)\right| dx$, so without further work, we get $Var[\mathcal{Z}] \leq \frac{2}{n}$.

   b. L_1 Bounds – Bounding the L_1 Error => It is known that for every f, $\sqrt{n}\mathbb{E}[g] \to \infty$ (Devroye and Gyorfi (1985)) which implies, by Chebyshev's inequality, that for every $\varepsilon > 0$, $\mathbb{P}\left\{\left|\frac{\mathcal{Z}}{\mathbb{E}[\mathcal{Z}]} - 1\right| \geq \varepsilon\right\} = \mathbb{P}\{|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]| \geq \varepsilon \mathbb{E}[\mathcal{Z}]\} \leq \frac{Var[\mathcal{Z}]}{\varepsilon^2 (\mathbb{E}[\mathcal{Z}])^2} \to 0$ as $n \to \infty$. That is, $\frac{\mathcal{Z}}{\mathbb{E}[\mathcal{Z}]} \to 1$ in probability, or in other words, $\mathcal{Z}$ is *relatively stable*. This means that the random L_1-error behaves like its expected value (Devroye (1988, 1991)). For more on the behavior of the L_1-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 $g: \mathcal{X}^n \to \mathbb{R}$ has the self-bounding property if there exists $g: \mathcal{X}^{n-1} \to \mathbb{R}$ such that for all $x_1, \dots, x_n \in \mathcal{X}$ and all $= 1, \dots, n$ $0 \leq g(x_1, \dots, x_n) - g_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \leq 1$, and also $\sum_{i=1}^{n}[g(x_1, \dots, x_n) - g_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)] \leq g(x_1, \dots, x_n)$.
    a. 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 $\sum_{i=1}^{n}[g(x_1, \dots, x_n) - g_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)]^2 \leq g(x_1, \dots, x_n)$.
    b. Variance Bound for Self-Bounding Functions => Using i) the extended Efron-Stein inequality $Var[\mathcal{Z}] \leq \sum_{i=1}^{n} \mathbb{E}[(\mathcal{Z} - \mathcal{Z}_i)^2]$, ii) setting $\mathcal{Z} = g(x_1, \dots, x_n)$, and $\mathcal{Z}_i = g(x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n)$, applying the observation above, iii) finally, using the squared form above, we get $Var[\mathcal{Z}] \leq \mathbb{E}[\sum_{i=1}^{n}[g(x_1, \dots, x_n) - g_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)]] \leq \mathbb{E}[g(x_1, \dots, x_n)] \leq \mathbb{E}[\mathcal{Z}]$. It turns out that in many cases, the obtained bound form $Var[\mathcal{Z}] \leq \mathbb{E}[\mathcal{Z}]$ is a significant improvement over what may be obtained by the bounded differences inequality.
2. Relative Stability: Bounding the variance of $\mathcal{Z}$ in many cases by its expected value implies the relative stability of $\mathcal{Z}$. A sequence of non-negative random variables $\mathcal{Z}_n$ is said to be relatively stable if $\frac{\mathcal{Z}}{\mathbb{E}[\mathcal{Z}]} \to 1$ in probability. This property guarantees that the random fluctuations of $\mathcal{Z}_n$ around its expectation are of negligible size when compared to the expectation, and therefore the most important information about the size of $\mathcal{Z}_n$ is given by $\mathbb{E}[\mathcal{Z}_n]$. As seen above, if $\mathcal{Z}_n$ has the self-bounding property, then, by Chebyshev's inequality, for all $\varepsilon > 0$, $\mathbb{P}\left\{\left|\frac{\mathcal{Z}_n}{\mathbb{E}[\mathcal{Z}_n]} - 1\right| \geq \varepsilon\right\} \leq \frac{Var[\mathcal{Z}_n]}{\varepsilon^2(\mathbb{E}[\mathcal{Z}_n])^2} = \frac{1}{\varepsilon^2}$. Thus, for relative stability, it suffices to have $\mathbb{E}[\mathcal{Z}_n] \to \infty$.
3. Empirical Processes: A typical example of self-bounding functions is the supremum of non-negative empirical processes. Let $\mathcal{F}$ be a class of functions taking values in $[0, 1]$, and consider $\mathcal{Z} = g(X_1, \dots, X_n) = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i)$. A special case of the above is mentioned in the example of uniform deviations.

a. Formulation => Define $g_i = g'$ for $i = 1, \ldots, n$ with $g(x_1, \ldots, x_n) = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n-1} f(X_i)$, so that $\mathcal{Z}_i = \sup_{f \in \mathcal{F}} \sum_{j=1, j \neq i}^{n} f(X_j)$. Let $f^* \in \mathcal{F}$ be a function for which $\mathcal{Z} = \sum_{j=1}^{n} f^*(X_j)$, one obviously has $0 \leq \mathcal{Z} - \mathcal{Z}_i \leq f^*(X_i) \leq 1$, and therefore $\sum_{i=1}^{n} (\mathcal{Z} - \mathcal{Z}_i) \leq \sum_{i=1}^{n} f^*(X_i) = \mathcal{Z}$.

b. Variance Bounding => Here we have assumed that the supremum is always achieved. The modification of the argument for the general case is straightforward. This by the variance bound, we get $Var[\mathcal{Z}] \leq \mathbb{E}[\mathcal{Z}]$. Note that the bounded differences inequality implies $Var[\mathcal{Z}] \leq \frac{n}{2}$. In some applications, $\mathbb{E}[\mathcal{Z}]$ may be significantly smaller than $\frac{n}{2}$, and the improvement is essential.

4. <u>Rademacher Averages</u>: A less trivial example for the self-bounding functions is the Rademacher complexity. Let $\mathcal{F}$ be a class of functions taking values in $[0, 1]$. If $\sigma_1, \ldots, \sigma_n$ denote independent symmetric $\{-1, +1\}$-valued random variables independent of the X_i's (the so-called Rademacher variables), we define the *Conditional Rademacher Average* as $\mathcal{Z} = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{j=1}^{n} \{\sigma_j f(X_j) | X_1^n\} \right]$. Thus, the expected values are taken with respect to the Rademacher variables σ_i, and $\mathcal{Z}$ is a function therefore of X_i's - this makes $\mathcal{Z}$ data dependent.

a. Learning Class Complexity => Quantities like $\mathcal{Z}$ 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)).

b. Variance Bounding => It is immediately apparent that $\mathcal{Z}$ has the bounded differences property, and therefore the bounded differences inequality implies that $Var[\mathcal{Z}] \leq \frac{n}{2}$. However, this bound may be improved by observing that $\mathcal{Z}$ also has the self-bounding property, and therefore $Var[\mathcal{Z}] \leq \mathbb{E}[\mathcal{Z}]$. Indeed, defining $\mathcal{Z} = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{j=1}^{n} \{\sigma_j f(X_j) | X_1^n\} \right]$, it is easy to see that $0 \leq \mathcal{Z} - \mathcal{Z}_i \leq 1$ and $\sum_{i=1}^{n} (\mathcal{Z} - \mathcal{Z}_i) \leq \mathcal{Z}$.

c. 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 $\mathcal{F}$ is a relatively small class of functions), $\mathbb{E}[\mathcal{Z}]$ may be bounded

by something like $Cn^{\frac{1}{2}}$ where the constant C depends on the class $\mathcal{F}$ (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)).
    a. Definition and Setup => Assume that we have a property $\mathcal{P}$ defined over a union of finite products of a set $\mathcal{X}$, that is, a sequence of sets $\mathcal{P}_1 \subset \mathcal{X}$, $\mathcal{P}_2 \subset \mathcal{X} \times \mathcal{X}$, …, $\mathcal{P}_n \subset \mathcal{X}^n$. We say that $(x_1, \dots, x_m) \in \mathcal{X}^m$ satisfies the property $\mathcal{P}$ if $(x_1, \dots, x_m) \in \mathcal{P}_m$.We assume that $\mathcal{P}$ is hereditary in the sense that if $(x_1, \dots, x_m)$ satisfies $\mathcal{P}_m$, so does any sub-sequence $\left(x_{i_1}, \dots, x_{i_k}\right)$ of $(x_1, \dots, x_m)$. The function g_n that maps any tuple $(x_1, \dots, x_n)$ to the size of the largest sub-sequence satisfying $\mathcal{P}$ is the *Configuration Function* associated with $\mathcal{P}$.
    b. Bound => Let g_n be a configuration function, and let $\mathcal{Z} = g_n(X_1, \dots, X_n)$ where $X_1, \dots, X_n$ are independent random variables. Then $Var[\mathcal{Z}] \leq \mathbb{E}[\mathcal{Z}]$.
    c. Bound Derivation => In order to apply the self-bounding function bound, it suffices to show that any configuration function is self-bounding. Let $\mathcal{Z}_i = g_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. The condition $0 \leq \mathcal{Z} - \mathcal{Z}_i \leq 1$ is trivially satisfied. On the other hand, assume that $\mathcal{Z} = k$ and let $\left(X_{i_1}, \dots, X_{i_k}\right) \subset (X_1, \dots, X_n)$ be a sub-sequence of cardinality k such that $g_k\left(X_{i_1}, \dots, X_{i_k}\right) = k$. Clearly if the index i is such that $i \notin \{i_1, \dots, i_k\}$, then $\mathcal{Z} = \mathcal{Z}_i$, and therefore $\sum_{i=1}^{n}(\mathcal{Z} - \mathcal{Z}_i) \leq \mathcal{Z}$ is also satisfied, which concludes the proof.
2. Number of Distinct Values in a Discrete Sample: Let $X_1, \dots, X_n$ be independent, identically distributed random variables taking values on a set of positive integers such that $\mathbb{P}\{X_1 = k\} = \mathcal{P}_k$, and let $\mathcal{Z}$ denote the number of distinct values taken by these n random

variables. Then we may write $\mathcal{Z} = \sum_{i=1}^{n} \mathbb{I}_{\{X_1 \neq X_i, \ldots, X_{i-1} \neq X_i\}}$, and thus the expected value of $\mathcal{Z}$ is computed easily as $\mathbb{E}[\mathcal{Z}] = \sum_{i=1}^{n} \sum_{j=1}^{\infty} (1 - p_j)^{i-1} p_j$.

a. Basic Bounds => It is easy to see that $\frac{\mathbb{E}[\mathcal{Z}]}{n} \to 0$ as $n \to \infty$; in fact the variance of $\mathcal{Z}$ may also be computed explicitly. Clearly $\mathcal{Z}$ satisfies the bounded differences property with $c_i = 1$, so the bounded differences inequality implies $Var[\mathcal{Z}] \leq \frac{n}{2}$, so, by Chebyshev's inequality $\frac{\mathcal{Z}}{n} \to 0$ in probability.

b. Configuration Function Bounds => On the other hand, it is obvious that $\mathcal{Z}$ is a configuration function associated with the property of "distinctness", and by the Bounding of the Configuration Function we have $Var[\mathcal{Z}] \leq \mathbb{E}[\mathcal{Z}]$, which is a significant improvement since $\frac{\mathbb{E}[\mathcal{Z}]}{n} \to 0$ as $n \to \infty$.

3. <u>VC Dimension</u>: 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)).

a. Setup => Let $\mathcal{A}$ be an arbitrary collection of subsets of $\mathcal{X}$, and let ${x_1}^n = (x_1, \ldots, x_n)$ be a vector n points of $\mathcal{X}$. Define the *Trace* of $\mathcal{A}$ by $Tr({x_1}^n) = \{A \cap [x_1, \ldots, x_n] : A \in \mathcal{A}\}$. The *Shatter Coefficient* (or *Vapnik-Chervonenkis Growth Function*) of $\mathcal{A}$ in ${x_1}^n$ is $\mathcal{T}({x_1}^n) = |Tr({x_1}^n)|$, the size of the trace. $\mathcal{T}({x_1}^n)$ is the number of different sub-sets of the n -point set generated by intersecting it with the elements of $\mathcal{A}$.

b. Formal Definition => A subset $(x_{i_1}, \ldots, x_{i_k})$ of $(x_1, \ldots, x_n)$ is said to be *shattered* if $\mathcal{T}(x_{i_1}, \ldots, x_{i_k}) = 2^k$. The VC Dimension $\mathcal{D}({x_1}^n)$ of $\mathcal{A}$ (with respect to ${x_1}^n$) is the cardinality of the largest subset of ${x_1}^n$.

c. Configuration Function Bounding => From the definition, it is obvious that $g_n({x_1}^n) = \mathcal{D}({x_1}^n)$ is a configuration function associated with the property of "shatteredness", and therefore if $X_1, \ldots, X_n$ are independent random variables, the $Var[\mathcal{D}({x_1}^n)] \leq \mathbb{E}[\mathcal{D}({x_1}^n)]$.

4. <u>Increasing Sub-sequences</u>: Consider a vector ${x_1}^n = (x_1, \ldots, x_n)$ of n different numbers in $[0, 1]$. The positive integers $i_1 < i_2 < \cdots < i_m$ form an *increasing sub-sequence* if $x_{i_1} < x_{i_2} < \cdots < x_{i_m}$ (where $i_1 \geq 1$ and $i_m \leq n$).

a. Configuration Function Bound => Let $L(x_1{}^n)$ denote the length of the longest increasing sub-sequence. $g_n(x_1{}^n) = L(x_1{}^n)$ is clearly a configuration function associated with the "increasing sequence" property, and therefore if $X_1, \ldots, X_n$ are independent random variables such that they are all different with probability one (it suffices if every X_i has an absolutely continuous distribution), then $Var[L(x_1{}^n)] \leq \mathbb{E}[L(x_1{}^n)]$.

b. $\mathbb{E}[L(x_1{}^n)]$ Estimate => If the X_i's are uniformly distributed in $[0, 1]$, it is known that $\mathbb{E}[L(x_1{}^n)] \approx 2\sqrt{n}$ (Logan and Shepp (1977), Groeneboom (2002)). A tighter, but more difficult result, obtained implies that $Var[L(x_1{}^n)] \leq \mathcal{O}\left(n^{\frac{1}{3}}\right)$ (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 $L(X)$.

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 L_1-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 X be a random variable taking values in a countable set $\mathcal{X}$ with distribution $\mathbb{P}(X = x) = p(x), x \in \mathcal{X}$. The *entropy* of X is defined by $\mathcal{H}(X) = \mathbb{E}[-\log p(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$ where log denotes the natural logarithm and

$0 \log 0 = 0$. If X, Y is a pair of discrete random variables taking values in $\mathcal{X} \times \mathcal{Y}$, the *Joint Entropy* $\mathcal{H}(X,Y)$ of X and Y is defined as the entropy of the pair (X,Y). The *Conditional Entropy* $\mathcal{H}(X \mid Y)$ is defined as $\mathcal{H}(X \mid Y) = \mathcal{H}(X,Y) - \mathcal{H}(Y)$.

3. Conditional Entropy Inequality: $\mathcal{H}(X \mid Y) = \mathcal{H}(X,Y) - \mathcal{H}(Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log p(x,y) + \sum_{y \in \mathcal{Y}} p(y) \log p(y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y)[\log p(x,y) - \log p(y)] = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(y)} = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log p(x|y)$. Thus $\mathcal{H}(X \mid Y) \geq 0$.
4. 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 $X, Y, and\ Z$, $\mathcal{H}(X,Y|Z) = \mathcal{H}(Y|Z) + \mathcal{H}(X|Y,Z)$. This may be easily seen by adding $\mathcal{H}(Z)$ 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 $X_1, \ldots, X_n$, $\mathcal{H}(X_1, \ldots, X_n) = \mathcal{H}(X_1) + \mathcal{H}(X_2|X_1) + \mathcal{H}(X_3|X_1,X_2) + \cdots + \mathcal{H}(X_n|X_1, \ldots, X_{n-1})$.
5. Kullback-Leibler Divergence: Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability distributions over a countable set $\mathcal{X}$ with a probability mass functions p and q. Then the *Kullback-Leibler Divergence* or the *Relative Entropy* of $\mathcal{P}$ and $\mathcal{Q}$ is $\mathcal{D}(\mathcal{P}||\mathcal{Q}) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$.
6. The Relative Entropy Inequality: Since $\log x \leq x - 1$, $\mathcal{D}(\mathcal{P}||\mathcal{Q}) = -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \geq -\sum_{x \in \mathcal{X}} p(x) \left[\frac{q(x)}{p(x)} - 1\right] = -\sum_{x \in \mathcal{X}} [q(x) - p(x)] = 0$, thus $\mathcal{D}(\mathcal{P}||\mathcal{Q}) \geq 0$, and equals zero if and only if $\mathcal{P} = \mathcal{Q}$.
    a. Consequence of the relative entropy inequality => If $\mathcal{X}$ is a finite set with $\mathcal{N}$ elements in it, and X is a random variable with distribution $\mathcal{P}$, and we take $\mathcal{Q}$ to be the uniform distribution over $\mathcal{X}$, then $\mathcal{D}(\mathcal{P}||\mathcal{Q}) = \sum_{x \in \mathcal{X}} p(x) \log p(x) - \sum_{x \in \mathcal{X}} p(x) \log q(x) = \log \mathcal{N} - \mathcal{H}(X)$ (since $q(x) = \frac{1}{\mathcal{N}}$), and therefore the entropy of X never exceeds the logarithm of the cardinality of its range.
7. Conditioning the Relative Entropy: Consider a pair of random variables with joint distribution $\mathcal{P}_{X,Y}$ and marginal distributions $\mathcal{P}_X$ and $\mathcal{P}_Y$. Then $\mathcal{D}\left(\mathcal{P}_{X,Y}||\mathcal{P}_X, \mathcal{P}_Y\right) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$. Observing that $p(x|y) = \frac{p(x,y)}{p(y)}$, $\mathcal{D}\left(\mathcal{P}_{X,Y}||\mathcal{P}_X, \mathcal{P}_Y\right)\mathcal{D}\left(\mathcal{P}_{X,Y}||\mathcal{P}_X, \mathcal{P}_Y\right) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x|y)}{p(y)} =$

$-\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x|y) = -\sum_{x\in\mathcal{X}} p(x,y)\log p(x) + \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x|y) = \mathcal{H}(X) - \mathcal{H}(X|Y)$.

a. Consequence of Conditioning the Relative Entropy => Thus $\mathcal{D}\big(\mathcal{P}_{X,Y}||\mathcal{P}_X,\mathcal{P}_Y\big) = \mathcal{H}(X) - \mathcal{H}(X|Y)$, and the non-negativity of the relative implies that $\mathcal{H}(X) \geq \mathcal{H}(X|Y)$, that is, conditioning reduces entropy. It is similarly easy to see that this fact remains true for conditional entropies as well, that $\mathcal{H}(X|Y) \geq \mathcal{H}(X|Y,Z)$.

8. Han's Inequality (Han (1978)): Let $X_1,\dots,X_n$ be discrete random variables. Then $\mathcal{H}(X_1,\dots,X_n) \leq \frac{1}{n-1}\sum_{i=1}^{n}\mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$.

a. Proof => For any $i = 1,\dots,n$, by the definition of conditional entropy and the fact that conditioning reduces entropy, $\mathcal{H}(X_1,\dots,X_n) = \mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n) + \mathcal{H}(X_i|X_1,\dots,X_{i-1},X_{i+1},\dots,X_n) \leq \mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n) + \mathcal{H}(X_i|X_1,\dots,X_{i-1})$. Summing both sides, we get $n\mathcal{H}(X_1,\dots,X_n) \leq \sum_{i=1}^{n}\mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n) + \mathcal{H}(X_1) + \mathcal{H}(X_2|X_1) + \mathcal{H}(X_3|X_1,X_2) + \cdots + \mathcal{H}(X_n|X_1,\dots,X_{n-1}) = \sum_{i=1}^{n}\mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n) + \mathcal{H}(X_1,\dots,X_n)$. Re-arranging we get $\mathcal{H}(X_1,\dots,X_n) \leq \frac{1}{n-1}\sum_{i=1}^{n}\mathcal{H}(X_1,\dots,X_{i-1},X_{i+1},\dots,X_n)$.

9. Han's Inequality for Relative Entropy:

a. Setup => Here we follow the layout of Massart (2000b). Let $\mathcal{X}$ be a countable set, and let $\mathcal{P}$ and $\mathcal{Q}$ be probability distributions on $\mathcal{X}^n$ such that $\mathcal{P} = \mathcal{P}_1 \times \dots \times \mathcal{P}_n$ is a product measure. We denote the elements of $\mathcal{X}^n$ by $x_1^n = (x_1,\dots,x_n)$ and denote $x^{(i)} = (x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)$ for the $n-1$ vector obtained by leaving out the i^{th} component of x_1^n.

b. Marginal Distributions of $\mathcal{P}$ and $\mathcal{Q}$ => Denote by $\mathcal{P}^{(i)}$ and $\mathcal{Q}^{(i)}$ the marginal distributions of $x^{(i)}$ according to $\mathcal{P}$ and $\mathcal{Q}$, that is, $\mathcal{Q}^{(i)}\big[x^{(i)}\big] = \sum_{x\in\mathcal{X}}\mathcal{Q}(x_1,\dots,x_{i-1},x,x_{i+1},\dots,x_n)$ and $\mathcal{P}^{(i)}\big[x^{(i)}\big] = \sum_{x\in\mathcal{X}}\mathcal{P}(x_1,\dots,x_{i-1},x,x_{i+1},\dots,x_n) = \sum_{x\in\mathcal{X}}\mathcal{P}_1(x_1)\dots\mathcal{P}_{i-1}(x_{i-1})\mathcal{P}_i(x_i)\mathcal{P}_{i+1}(x_{i+1})\dots\mathcal{P}_n(x_n)$. Just as we used $\mathcal{P}$ to represent a uniform measure earlier, here we use $\mathcal{P}$ to represent an *independence* product measure, and will be used later.

c. Statement => $\mathcal{D}(\mathcal{Q}||\mathcal{P}) \geq \frac{1}{n-1}\sum_{i=1}^{n}\mathcal{D}\left(\mathcal{Q}^{(i)}||\mathcal{P}^{(i)}\right)$ or, equivalently, $\mathcal{D}(\mathcal{Q}||\mathcal{P}) \leq \sum_{i=1}^{n}\left[\mathcal{D}(\mathcal{Q}||\mathcal{P}) - \mathcal{D}\left(\mathcal{Q}^{(i)}||\mathcal{P}^{(i)}\right)\right]$.

d. Proof:

i. Re-casting Han's Inequality => Han's inequality may be directly re-cast as stating $\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{Q}(x_1^n) \geq \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x^{(i)}\in\mathcal{X}^{n-1}}\mathcal{Q}^{(i)}\left(x^{(i)}\right)\log\mathcal{Q}^{(i)}\left(x^{(i)}\right)$.

ii. Re-casting the Relative Entropy => Since $\mathcal{D}(\mathcal{Q}||\mathcal{P}) = \sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{Q}(x_1^n) - \sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}(x_1^n)$ and $\mathcal{D}\left(\mathcal{Q}^{(i)}||\mathcal{P}^{(i)}\right) = \sum_{x^{(i)}\in\mathcal{X}^{n-1}}\left[\mathcal{Q}^{(i)}\left(x^{(i)}\right)\log\mathcal{Q}^{(i)}\left(x^{(i)}\right) - \mathcal{Q}^{(i)}\left(x^{(i)}\right)\log\mathcal{P}^{(i)}\left(x^{(i)}\right)\right]$, it is sufficient to show that $\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}(x_1^n) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x^{(i)}\in\mathcal{X}^{n-1}}\mathcal{Q}^{(i)}\left(x^{(i)}\right)\log\mathcal{P}^{(i)}\left(x^{(i)}\right)$. This is easily seen from the product property of $\mathcal{P}$ - we have $\mathcal{P}(x_1^n) = \mathcal{P}^{(i)}\left(x^{(i)}\right)\mathcal{P}_i(x_i)$ for all i, as well as $\mathcal{P}(x_1^n) = \prod_{i=1}^{n}\mathcal{P}_i(x_i)$, therefore $\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}(x_1^n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\left[\log\mathcal{P}^{(i)}\left(x^{(i)}\right) + \log\mathcal{P}_i(x_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}^{(i)}\left(x^{(i)}\right) + \frac{1}{n}\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}(x_1^n)$.

iii. Re-arrangement => Re-arranging the above, we obtain $\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}(x_1^n) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x_1^n\in\mathcal{X}^n}\mathcal{Q}(x_1^n)\log\mathcal{P}^{(i)}\left(x^{(i)}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\sum_{x^{(i)}\in\mathcal{X}^{n-1}}\mathcal{Q}^{(i)}\left(x^{(i)}\right)\log\mathcal{P}^{(i)}\left(x^{(i)}\right)$ where we have used the defining property of $\mathcal{Q}^{(i)}$.

Tensorization of the Entropy

1. Introduction: We now use the results above to prove our main exponential concentration inequality. As before, we let $X_1, \ldots, X_n$ be independent random variables, and investigate the concentration properties of $Z = g(X_1, \ldots, X_n)$.
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 $Var[Z] \leq \sum_{i=1}^{n}\mathbb{E}[\mathbb{E}_i[Z^2] -$

$(\mathbb{E}_i[Z])^2]$, we generalize the martingale differences inequality to a large class of convex functions ϕ (by exploiting Jensen's inequality for convex functions) as $\mathbb{E}[\phi(Z)] - \phi(\mathbb{E}[Z]) \leq \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}_i[\phi(Z)] - \phi(\mathbb{E}_i[Z])]$. Using $\phi(x) = x^2$ recovers the original martingale differences inequality.

3. 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)).
4. Usage for the Relative Entropy Function: The function of interest in our case is $\phi(x) = x \log x$ - 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 Z on $\mathcal{X}^n$ and the distribution of X_1^n. Hence the name "Tensorization of the entropy inequality" (Ledoux (1997)).
5. Tensorization of Entropy Theorem - Statement: Let $\phi(x) = x \log x$ for $x > 0$. Let $X_1, \dots, X_n$ be independent random variables taking values in $\mathcal{X}$ and let f be a positive-valued function on $\mathcal{X}^n$. Letting $Y = f(X_1, \dots, X_n)$ we have $\mathbb{E}[\phi(Y)] - \phi(\mathbb{E}[Y]) \leq \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}_i[\phi(Y)] - \phi(\mathbb{E}_i[Y])]$.
    a. Proof Step #1 – Outline => We only prove the statement for discrete random variables $X_1, \dots, X_n$. 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 Y, then it is also true for cY where c is a positive constant. Hence, without loss of generality, we assume that $\mathbb{E}[Y] = 1$.
    b. Proof Step #2 – The Tensorial Measure => We set the probability measure $\mathcal{Q}$ on $\mathcal{X}^n$ to be $\mathcal{Q}(x_1^n) = f(x_1^n)\, \mathcal{P}(x_1^n)$ where $\mathcal{P}$ denotes the distribution of $X_1^n = [X_1, \dots, X_n]$. Then clearly, $\mathbb{E}[\phi(Y)] - \phi(\mathbb{E}[Y]) = \mathbb{E}[\phi(Y)] - 0 = \mathbb{E}[Y \log Y] = \mathcal{D}(\mathcal{Q}||\mathcal{P})$. Using Han's inequality for relative entropy, $\mathcal{D}(\mathcal{Q}||\mathcal{P})$ cannot exceed $\sum_{i=1}^{n}\left[\mathcal{D}(\mathcal{Q}||\mathcal{P}) - \mathcal{D}\left(\mathcal{Q}^{(i)}||\mathcal{P}^{(i)}\right)\right]$. However, straightforward calculation shows that $\sum_{i=1}^{n}\left[\mathcal{D}(\mathcal{Q}||\mathcal{P}) - \mathcal{D}\left(\mathcal{Q}^{(i)}||\mathcal{P}^{(i)}\right)\right] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}_i[\phi(Y)] - \phi(\mathbb{E}_i[Y])]$, and the original statement of inequality follows.
6. 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 $Y = e^{sZ}$. Denoting the moment generating function of Z by $F(s) = \mathbb{E}[e^{sZ}]$, the LHS of the tensorization of entropy inequality becomes $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] = s\frac{dF(s)}{ds} - F(s)\log F(s)$.

a. Approach Behind the Ledoux's Entropy Method => The strategy then is to derive upper bounds for $\frac{dF(s)}{ds}$, 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 $Z_i = g_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ where g_i is some function over $\mathcal{X}^n$. 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)).
2. Relative Entropy Inequality Lemma: Let Y be a positive random variable. Then, for any $u > 0$, $\mathbb{E}[Y\log Y] - (\mathbb{E}[Y])\log\mathbb{E}[Y] \leq \mathbb{E}[Y\log Y - Y\log u - (Y-u)]$.
    a. Proof => Since for any > 0 $\log x \leq x - 1$, we have $\log\frac{u}{\mathbb{E}[Y]} \leq \frac{u}{\mathbb{E}[Y]} - 1$, hence $(\mathbb{E}[Y])\log\frac{u}{\mathbb{E}[Y]} \leq u - \mathbb{E}[Y]$, upon re-arranging which, you retrieve the original statement.
3. Logarithmic Sobolev Inequality: Denote $\psi(x) = e^x - x - 1$. Then $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^n \mathbb{E}[\psi(-s\{Z - Z_i\})e^{sZ}]$.
    a. Proof => We bound each term on the RHS of the tensorial entropy inequality. Note that the relative entropy inequality implies that if Y_i is a positive function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, then $\mathbb{E}_i[Y\log Y] - (\mathbb{E}_i[Y])\log\mathbb{E}_i[Y] \leq \mathbb{E}_i[Y\log Y - Y\log Y_i] - \mathbb{E}_i[Y - Y_i]$. Applying the above inequality to the variables $Y = e^{sZ}$ and $Y_i = e^{sZ_i}$, one gets $\mathbb{E}_i[Y\log Y] - (\mathbb{E}_i[Y])\log\mathbb{E}_i[Y] \leq \sum_{i=1}^n \mathbb{E}[\psi(-s\{Z - Z_i\})e^{sZ}]$.

4. Symmetrized Logarithmic Sobolev Inequality (Massart (2000a)): If $\psi(x)$ is defined as in the logarithmic Sobolev inequality, then $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^{n}\mathbb{E}[\psi(-s\{Z - Z'_i\})e^{sZ}]$. Moreover, denote $\tau(x) = x(e^x - 1)$. Then, for all $s \in \mathbb{R}$, $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^{n}\mathbb{E}\left[\tau(-s\{Z - Z'_i\})\mathbb{I}_{Z>Z'_i}e^{sZ}\right]$, and $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^{n}\mathbb{E}\left[\tau(s\{Z - Z'_i\})\mathbb{I}_{Z<Z'_i}e^{sZ}\right]$.
    a. Proof => The first inequality is proved exactly as in the logarithmic Sobolev inequality, by nothing that, just like Z_i, Z'_i is also independent of X_i. To prove the 2nd and the 3rd inequalities, write $\psi(-s\{Z - Z'_i\})e^{sZ} = \psi(-s\{Z - Z'_i\})\mathbb{I}_{Z>Z'_i}e^{sZ} + \psi(+s\{Z'_i - Z\})\mathbb{I}_{Z<Z'_i}e^{sZ}$. By symmetry, the conditional expectation of the 2nd term may be written as $\mathbb{E}\left[\psi(+s\{Z'_i - Z\})\mathbb{I}_{Z<Z'_i}e^{sZ}\right] = \mathbb{E}\left[e^{sZ'_i}\psi(s\{Z - Z'_i\})\mathbb{I}_{Z>Z'_i}\right] = \mathbb{E}\left[e^{sZ}e^{-s(Z-Z'_i)}\psi(s\{Z - Z'_i\})\mathbb{I}_{Z>Z'_i}\right]$. Summarizing, we have $\mathbb{E}_i[\psi(-s\{Z - Z'_i\})e^{sZ}] = \mathbb{E}_i\left[\psi(-s\{Z - Z'_i\}) + e^{-s(Z-Z'_i)}\psi(s\{Z - Z'_i\})e^{sZ}\mathbb{I}_{Z>Z'_i}\right]$. The 2nd inequality of the theorem follows simply by noting that $\psi(x) = e^x\psi(-x) = x(e^x - 1) = \tau(x)$. 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 C such that, almost surely, $\sum_{i=1}^{n}(Z - Z'_i)^2\mathbb{I}_{Z>Z'_i} \leq \mathcal{C}$. Then, for all $t > 0$, $\mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq e^{-\frac{t^2}{4\mathcal{C}}}$.
    a. Proof Step #1 – Application of the Symmetrized Logarithmic Sobolev Inequality => Observe that for $x > 0$, $\tau(-x) \leq x^2$, and therefore, for any $s > 0$, the symmetrized logarithmic inequality implies $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \mathbb{E}\left[e^{sZ}\sum_{i=1}^{n}(Z - Z'_i)^2\mathbb{I}_{Z<Z'_i}\right] \leq s^2\mathcal{C}\mathbb{E}[e^{sZ}]$ where we have used the assumptions of the theorem.

b. Proof Step #2 – Bounding of the Expectation => Denoting the moment generating function of Z by $F(s) = \mathbb{E}[e^{sZ}]$, the above inequality may be re-written as $s\frac{dF(s)}{ds} - F(s)\log F(s) \leq s^2\mathcal{C}F(s)$. After diving both sides by $s^2F(s)$, we observe that the LHS is just a derivative of $H(s) = \frac{\log F(s)}{s}$, that is, we obtain the inequality $\frac{dH(s)}{ds} \leq \mathcal{C}$. By l'Hospital's rule, we note that $\underset{s \to 0}{Lim} H(s) = \frac{F'(0)}{F(0)} = \mathbb{E}[Z]$, so by integrating the above inequality, we get $H(s) \leq \mathbb{E}[Z] + s\mathcal{C}$, or, in other words, $F(s) \leq e^{s\mathbb{E}[Z]+s^2\mathcal{C}}$.

c. Proof Step #3 – Apply Markov's Inequality and Chernoff Bounds => By Markov's inequality $\mathbb{P}[Z - \mathbb{E}[Z] \geq t] \leq F(s)e^{-s\mathbb{E}[Z]-st} \leq e^{-s^2\mathcal{C}-st}$. The choice of $s = \frac{t}{2\mathcal{C}}$ makes the upper bound $e^{-\frac{t^2}{4\mathcal{C}}}$. Replace Z by $-Z$ to obtain the same upper bound for $\mathbb{P}[Z \leq \mathbb{E}[Z] - t]$.

3. Two-sided Bounded Differences Inequality: It is clear from the proof above that under the condition $\sum_{i=1}^{n}(Z - Z_i')^2\mathbb{I}_{Z>Z_i'} \leq \mathcal{C}$ one has a two-sided inequality $\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2e^{-\frac{t^2}{4\mathcal{C}}}$. An immediate corollary of this is that functions following the bounded-square differences inequality also satisfy sub-Gaussian tail probability.
4. McDiarmid Bounded Differences Inequality: Under the conditions of bounded square differences, McDiarmid (1989) showed that $\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2e^{-\frac{2t^2}{\mathcal{C}}}$. This may be seen by writing Z 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.
5. Bounded Square Differences Inequality - Statement: Assume that the function g satisfies the bounded square differences assumption with constants $\mathcal{C}_1, \ldots, \mathcal{C}_n$, then $\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2e^{-\frac{t^2}{4\mathcal{C}}}$ where $\mathcal{C} = \sum_{i=1}^{n}\mathcal{C}_i{}^2$.
6. Usage of the Bounded Square Differences Inequality: We remark here that the constant $\mathcal{C}$ 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

$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq \frac{\mathcal{C}}{2t^2}$, 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.

7. Relaxing the Bounded Differences Square Bound: The differences square bound maybe relaxed in the following way to produce expected data-dependent tail bounds. If $\mathbb{E}\left[\sum_{i=1}^{n}(Z - Z_i')^2 \mathbb{I}_{Z>Z_i'} | X_1^n\right] \leq \mathcal{C}$, then for all $t > 0$, $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{4\mathcal{C}}}$. (This can be seen by using the Association Inequality for the monotonic multi-variate functions $(Z - Z_i')^2$ and e^{sZ}). Further, if $\mathbb{E}\left[\sum_{i=1}^{n}(Z - Z_i')^2 \mathbb{I}_{Z<Z_i'} | X_1^n\right] \leq \mathcal{C}$, we can see that $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\frac{t^2}{4\mathcal{C}}}$.
8. Example - Hoeffding's Inequality in Hilbert Space: As a simple illustration of the power f 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 $X_1, \ldots, X_n$ are independent zero-mean random variables taking values in a separable Hilbert space such that $\|X_i\| \leq \frac{\mathcal{C}_i}{2}$ with probability one, then for all $t \geq 2\sqrt{\mathcal{C}}$, $\mathbb{P}[\|\sum_{i=1}^{n} X_i\| \geq t] \leq e^{-\frac{t^2}{2\mathcal{C}}}$ where $\mathcal{C} = \sum_{i=1}^{n} {\mathcal{C}_i}^2$.
    a. Proof => The above follows simply by observing that, by triangle inequality $Z = \|\sum_{i=1}^{n} X_i\|$ satisfies the bounded differences inequality with constants $\mathcal{C}_i$, and therefore $\mathbb{P}[\|\sum_{i=1}^{n} X_i\| \geq t] = \mathbb{P}\left[\|\sum_{i=1}^{n} X_i\| - \mathbb{E}[\|\sum_{i=1}^{n} X_i\|] \geq t - \mathbb{E}[\|\sum_{i=1}^{n} X_i\|]\right] \leq e^{-\frac{2(t - \mathbb{E}[\|\sum_{i=1}^{n} X_i\|])^2}{\mathcal{C}}}$. The proof is completed by observing that, by independence, $\mathbb{E}[\|\sum_{i=1}^{n} X_i\|] \leq \sqrt{\left\{\mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|\right]\right\}^2} = \sqrt{\sum_{i=1}^{n} \mathbb{E}\left[\left\|{X_i}^2\right\|\right]} \leq \sqrt{\mathcal{C}}$. Application of the Markov/Chebyshev inequality produces the final result.
9. 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 g be bounded – all that is required is that $\sup_{\substack{x_1, \ldots, x_n \\ x'_1, \ldots, x'_n}} \sum_{i=1}^{n} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)|^2 \leq \sum_{i=1}^{n} {\mathcal{C}_i}^2$, 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.

10. 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 A be a symmetric real matrix whose entries $X_{i,j}$ $1 \le i \le j \le n$ are independent random variables with absolute value bounded by 1. We seek to show that if $Z = \lambda_1$ is the largest eigenvalue of A, then $\mathbb{P}[Z \ge \mathbb{E}[Z] + t] \le e^{-\frac{t^2}{16}}$. The property of the largest eigenvalue we need is that if $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ is an eigenvector corresponding to the largest eigenvalue λ_1 with $\|v\| = 1$, then $\lambda_1 = v^T A v = \sup_{u:\, \|u\| = 1} u^T A u$.
    a. Using the Bounded Square Differences Theorem => To use the bounded square differences inequality, consider the symmetric matrix $A'_{i,j}$ obtained by replacing $X_{i,j}$ in A by the independent copy $X'_{i,j}$ while keeping all other variables fixed. Let $Z'_{i,j}$ denote the largest eigenvalue of the obtained matrix. Then, using the above-mentioned property of the largest eigenvalue, $\left(Z - Z'_{i,j}\right)\mathbb{I}_{Z>Z'_{i,j}} \le \left(v^T A v - v^T A'_{i,j} v\right)\mathbb{I}_{Z>Z'_{i,j}} = v^T\left(A - A'_{i,j}\right)v\mathbb{I}_{Z>Z'_{i,j}} = \left[v_i v_j\left(X_{i,j} - X'_{i,j}\right)\right] \le 2\left|v_i v_j\right|$. Therefore, $\sum_{1\le i\le j\le n}\left(Z - Z'_{i,j}\right)^2 \mathbb{I}_{Z>Z'_{i,j}} \le \sum_{1\le i\le j\le n} 4\left|v_i v_j\right|^2 \le [4\sum_{i=1}^n {v_i}^2] = 4$. The result now follows from the bounded differences inequality.
    b. Comparison with Efron-Stein => Note that by the Efron-Stein inequality, we also have $Var[Z] \ge 4$. 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 $t > 0$, $\mathbb{P}[Z \le \mathbb{E}[Z] - t] \le e^{-\frac{t^2}{16(e-1)}}$.
    c. 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 Z is the k^{th} largest eigenvalue (or the k^{th} smallest eigenvalue), then the upper bound becomes $e^{-\frac{t^2}{16k^2}}$, although it is not clear whether the factor $\frac{1}{k^2}$ 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 $Var[Z] \leq \mathbb{E}[Z]$. Based on the logarithmic Sobolev inequality, we may now obtain exponential concentration inequality bounds (Boucheron, Lugosi, and Massart (2000), Massart (2000a)).
2. 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:
    a. $h(u) = (1+u)\log(1+u) - u \; \forall \; u \geq -1$
    b. $\underset{u \geq -1}{sup}\,[uv - h(v)] = e^v - v - 1$
3. Logarithmic Sobolev Inequality for Self-Bounding Functions: Assume that g satisfies the self-bounding property. Then for every $s \in \mathbb{R}$, $\log \mathbb{E}\left[e^{s(Z-\mathbb{E}[Z])}\right] \leq \psi(s)\mathbb{E}[Z]$. Moreover for every $t > 0$, $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\mathbb{E}[Z]h(\frac{t}{\mathbb{E}[Z]})}$, and for every $0 < t < \mathbb{E}[Z]$ $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\mathbb{E}[Z]h(-\frac{t}{\mathbb{E}[Z]})}$.
    a. Corollary from the re-cast => By recalling that $h(u) = \frac{u^2}{2+\frac{2u}{3}}$ for $u \geq 0$ (we have already used this in the proof for Bernstein's inequality), and observing that $h(u) \geq \frac{u^2}{2}$ for $u \leq 0$, we obtain the following two corollaries:
        i. For every $t < 0$, $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z]+\frac{2t}{3}}}$
        ii. For every $0 < t < \mathbb{E}[Z]$, $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z]}}$.
4. Proof of Logarithmic Sobolev Inequality for Self-Bounding Functions:
    a. Step #1 – Applying the Logarithmic Sobolev Inequality => Since the function $\psi(s)$ is convex with $\psi(0) = 0$ for any s and any $u \in [0,1]$, $\psi(-su) \leq u\psi(-s)$. Thus, since $(Z - Z_i) \in [0,1]$, we have that, for every s $\psi\left(-s(Z - Z_i)\right) \leq (Z - Z_i)\psi(-s)$. Therefore the combination of the logarithmic inequality and the condition $\sum_{i=1}^n (Z - Z_i) \leq Z$ implies that $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log \mathbb{E}[e^{sZ}] \leq \mathbb{E}[\psi(-s)e^{sZ}\sum_{i=1}^n (Z - Z_i)] \leq \psi(-s)\mathbb{E}[Ze^{sZ}]$.

b. Step #2 – Variable Transformation => Introduce $\tilde{Z} = Z - \mathbb{E}[Z]$ and define for any s, $\tilde{F}(s) = \mathbb{E}\left[e^{s\tilde{Z}}\right]$. The inequality above then becomes $[s - \psi(-s)]\frac{\tilde{F}'(s)}{\tilde{F}(s)} - \log \tilde{F}(s) \leq \psi(-s)\mathbb{E}[Z]$, which, on writing $G(s) = \log \tilde{F}(s)$, implies $[1 - e^{-s}]\frac{dG(s)}{ds} - G(s) \leq \psi(-s)\mathbb{E}[Z]$.

c. Step #3 – Inequality for $G(s)$ => Observe that the function $G_0(s) = \psi(s)\mathbb{E}[Z]$ is a solution to the ODE $[1 - e^{-s}]\frac{dG(s)}{ds} - G(s) \leq \psi(-s)\mathbb{E}[Z]$. Here we intend to show that $G(s) \leq G_0(s)$. In fact, if $G_1(s) = G(s) - G_0(s)$, then $[1 - e^{-s}]\frac{dG_1(s)}{ds} - G_1(s) \leq 0$. Thus, defining $\tilde{G}(s) = \frac{G_1(s)}{e^s - 1}$, we have $[1 - e^{-s}][e^s - 1]\frac{d\tilde{G}(s)}{ds} \leq 0$. This clearly indicates that $\frac{d\tilde{G}(s)}{ds}$ is non-positive, and therefore $\tilde{G}(s)$ is non-increasing.

d. Step #4 – Proof that $G(s) \leq G_0(s)$ => Now, since $\tilde{Z}$ is centered, $\left.\frac{dG_1(s)}{ds}\right|_{s=0} = 0$. Using the fact that $\frac{s}{e^s - 1} \to 1$ as $s \to 0$, we conclude that $\frac{d\tilde{G}(s)}{ds} \to 0$ as $s \to 0$. This shows that $\tilde{G}(s)$ is non-positive over $(0, \infty)$ and non-negative over $(-\infty, 0)$, hence $G_1(s)$ is non-positive everywhere, therefore $G(s) \leq G_0(s)$. This proves the first inequality.

e. Step #5 – Tail Probabilities Calculation => The proof of inequalities for the tail probabilities maybe completed by Chernoff Bounding:

   i. $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\sup_{s>0}[ts - \psi(s)\mathbb{E}[Z]]}$

   ii. $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\sup_{s<0}[-ts - \psi(s)\mathbb{E}[Z]]}$

f. Step #6 – Reduced Form Bounds => The final step of the proof is done using the following easy-to-check and well-known relations:

   i. $\sup_{s>0}\left[ts - \psi(s)\mathbb{E}[Z]\right] = \mathbb{E}[Z]h\left(\frac{t}{\mathbb{E}[Z]}\right)$ for $t > 0$

   ii. $\sup_{s<0}\left[-ts - \psi(s)\mathbb{E}[Z]\right] = \mathbb{E}[Z]h\left(-\frac{t}{\mathbb{E}[Z]}\right)$ for $0 < t < \mathbb{E}[Z]$

Combinatorial Entropy

1. VC Entropy: The VC Entropy is closely related to the VC dimension discussed earlier. Let $\mathcal{A}$ be an arbitrary collection of subsets of $\mathcal{X}$, and let $x_1^n = (x_1, \ldots, x_n)$ be a vector of n points of $\mathcal{X}$. Recall that the *shatter coefficient* is defined as the size of the trace of $\mathcal{A}$ on x_1^n, that is, $\mathcal{T}(x_1^n) = |Trace(x_1^n)| = |\{A \cap [x_1, \ldots, x_n] : A \in \mathcal{A}\}|$. The VC entropy is defined as the log of the shatter coefficient, that is, $h(x_1^n) = log_2 \mathcal{T}(x_1^n)$.
2. VC Entropy Self-Bounding Property: The VC Entropy has the self-bounding property.
    a. Proof Step #1 – Setup => We need to show that there exists a function h' of $n-1$ variables such that for all $i = 1, \ldots, n$, writing $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, the following true self-bounding conditions are valid:
        i. $0 \leq h(x_1^n) - h'\left(x^{(i)}\right) \leq 1$, and
        ii. $\sum_{i=1}^n \left[h(x_1^n) - h'\left(x^{(i)}\right)\right] \leq h(x_1^n)$.
    b. Proof Step #2 – Outline => We define h' in the natural way, that is, as the entropy based on the $n-1$ points in its arguments. Then, clearly for any $h(x_1^n) \geq h'\left(x^{(i)}\right)$, 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.
    c. Proof Step #3 – Invoking Uniform Outcome Distribution => Consider the uniform distribution over the outcome set $Trace(x_1^n)$. This defines a random vector $Y = (Y_1, \ldots, Y_n) \in \mathcal{Y}^n$. Then clearly $h(x_1^n) = log_2 |Trace(x_1^n)| = \frac{1}{\log 2} \mathcal{H}(Y_1, \ldots, Y_n)$, where $\mathcal{H}(Y_1, \ldots, Y_n)$ is the joint entropy of $Y_1, \ldots, Y_n$. 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 $i \leq n$, $h'\left(x^{(i)}\right) \geq \frac{1}{\log 2} \mathcal{H}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$.
    d. Proof Step #4 – Invoking Han's Inequality => Since by Han's inequality for joint entropy $\mathcal{H}(Y_1, \ldots, Y_n) \leq \frac{1}{n-1} \sum_{i=1}^n \mathcal{H}(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have $\sum_{i=1}^n \left[h(x_1^n) - h'\left(x^{(i)}\right)\right] \leq h(x_1^n)$ as desired.
3. VC Entropy Probabilistic Bounds: Let $X_1, \ldots, X_n$ be a set of independent random variables taking values in $\mathcal{X}$ and let $Z = h(x_1^n)$ denote the random VC entropy. Then the following are true:

a. $Var[Z] \leq \mathbb{E}[Z]$.

b. For every $t < 0$, $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z]+\frac{2t}{3}}}$.

c. For every $0 < t < \mathbb{E}[Z]$, $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z]}}$.

d. Moreover, for the random shatter coefficient $\mathcal{T}(x_1^n)$, we have $\mathbb{E}[log_2\mathcal{T}(x_1^n)] \leq log_2\mathbb{E}[\mathcal{T}(x_1^n)] \leq log_2 e\mathbb{E}[log_2\mathcal{T}(x_1^n)]$.

4. 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 $s = \log 2$ in the first inequality of the exponential inequality for self-bounding functions. This last statement shows that the expected VC entropy $\mathbb{E}[log_2\mathcal{T}(x_1^n)]$ and the annealed VC Entropy $log_2\mathbb{E}[\mathcal{T}(x_1^n)]$ are tightly connected regardless of the class of sets $\mathcal{A}$ and the distribution of X_i's.

5. 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 $\frac{1}{n}log_2\mathbb{E}[\mathcal{T}(x_1^n)]$ converges to zero.

6. Bounded Supremum of the Self Bounding Exponential Inequality: Let C and a denote two positive real numbers and denote $h_1(x) = x + 1 - \sqrt{2x+1}$. It is easy to show that the local supremum is given from $\sup_{\lambda \in \left[0, \frac{1}{a}\right]} \left(\lambda t - \frac{C\lambda^2}{1-a\lambda}\right) = \frac{2C}{a^2} h\left(\frac{at}{2C}\right) \geq \frac{t^2}{2(2C+at)}$, and that the supremum is attained at $\lambda = \frac{1}{a}\left[1 - \frac{1}{\sqrt{1+\frac{at}{C}}}\right]$.

7. Unbounded Supremum of the Self-Bounding Exponential Inequality: Further, $\sup_{\lambda \in [0,\ \infty)} \left(\lambda t - \frac{C\lambda^2}{1+a\lambda}\right) = \frac{2C}{a^2} h\left(-\frac{at}{2C}\right) \geq \frac{t^2}{4C}$ if $t < \frac{C}{a}$, and the supremum is attained at $\lambda = \frac{1}{a}\left[\frac{1}{\sqrt{1-\frac{at}{C}}} - 1\right]$.

8. 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.

9. Combinatorial Entropy - Setup: Let $x_1^n = (x_1, \ldots, x_n)$ be an n-vector of elements $x_i \in \mathcal{X}_i$ to which we associate a set $Trace(x_1^n) \subset \mathcal{Y}^n$ of n-vectors whose component are elements of a possibly different set $\mathcal{Y}$. We assume that for each $x \in \mathcal{X}^n$ and $i \leq n$, the set $Trace\left(x^{(i)}\right) = Trace(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ is a projection of $Trace(x_1^n)$ along the i^{th} co-ordinate, that is, $Trace\left(x^{(i)}\right) = \left\{y^{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) \in \mathcal{Y}^{n-1} : \exists\, y_i \in \mathcal{Y} : (y_1, \ldots, y_n) \in Trace(x_1^n)\right\}$. The associated combinatorial entropy is $h(x_1^n) = log_b|Trace(x_1^n)|$ where b is an arbitrary positive number ≥ 2.
10. 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 $h(x_1^n) = log_b|Trace(x)|$ is a combinatorial entropy such that for all $x \in \mathcal{X}^n$ and $i \leq n$, $h(x_1^n) - h\left(x^{(i)}\right) \leq 1$ then h has the self-bounding property.
11. Combinatorial Entropy - Exponential Bounds Generalization: Assume that $h(x_1^n) = log_b|Trace(x)|$ is a combinatorial entropy such that for all $x \in \mathcal{X}^n$ and $i \leq n$, $h(x_1^n) - h\left(x^{(i)}\right) \leq 1$. If $X_1^n = (X_1, \ldots, X_n)$ is a vector of n-independent random variables taking values in $\mathcal{X}$, then the random combinatorial entropy $Z = h(x_1^n)$ satisfies $\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z] + \frac{2t}{3}}}$ for $t < 0$, and $\mathbb{P}[Z \leq \mathbb{E}[Z] - t] \leq e^{-\frac{t^2}{2\mathbb{E}[Z]}}$ for $0 < t < \mathbb{E}[Z]$. Moreover, $\mathbb{E}[log_b|Trace(X_1^n)|] \leq log_b \mathbb{E}[|Trace(X_1^n)|] \leq \frac{b-1}{\log b}\mathbb{E}[log_b|Trace(X_1^n)|]$.
12. Combinatorial Entropy Example - Increasing Sub-sequences: Recall the setup of the example of the increasing sub-sequences earlier, and let $N(x_1^n)$ denote the number of different increasing sub-sequences of x_1^n. Observe that $log_2 N(x_1^n)$ is a combinatorial entropy. This is easy to see by considering $\mathcal{Y} = \{0, 1\}$, and by assigning, to each increasing sub-sequence $i_1 < i_2 < \cdots < i_m$ of x_1^n a binary n-vector $y_1^n = (y_1, \ldots, y_n)$ such that $y_j = 1$ if and only if $j = i_k$ for some $k = 1, \ldots, m$ (i.e., the indices appearing in the increasing sub-sequence are marked by 1).
    a. Exponential Bounds => Now that the conditions for combinatorial entropy are met, $Z = log_2 N(x_1^n)$ satisfies all the inequalities associated with the combinatorial entropy exponential inequality. This result improves a concentration inequality obtained by Frieze (1991) for $log_2 N(x_1^n)$.

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 $\sum_{i=1}^{n}(Z-Z'_i)^2$ is controlled by different functions of Z. 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 $e^{-\frac{t^2}{\sigma^2}}$ where σ^2 is the corresponding Efron-Stein upper bound of $Var[Z]$. 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 a and b such that $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z>Z'_i} \le aZ+b$. Then, for $s \in \left(0,\frac{1}{a}\right)$, $\log \mathbb{E}\left[e^{s(Z-\mathbb{E}[Z])}\right] \le \frac{s^2}{1-as}(a\mathbb{E}[Z]+b)$, and for all $t>0$, $\mathbb{P}(Z \ge \mathbb{E}[Z]+t) \le e^{-\frac{t^2}{4a\mathbb{E}[Z]+4b+2at}}$.
    a. Reduction to Expectations Bound => Let $s>0$. Just as in the first step of the proof for the Bounded Square Differences Exponential Bounds, we use the fact that for $x>0$, $\tau(-x) \le x^2$, and therefore by the Symmetrized Logarithmic Inequality we have $s\mathbb{E}[Ze^{sZ}]-\mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \le \mathbb{E}\left[e^{sZ}\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z>Z'_i}\right] \le s^2(a\mathbb{E}[Ze^{sZ}]+b\mathbb{E}[e^{sZ}])$, where at the final step we have used the assumptions of our postulate.
    b. Inequality ODE for $H(s)$ => Denoting, again, $F(s)=\mathbb{E}[e^{sZ}]$, the above inequality becomes $s\frac{dF(s)}{ds}-F(s)\log F(s) \le as^2\frac{dF(s)}{ds}$. After scaling both sides by $s^2F(s)$, we notice that the LHS is just the derivative of $H(s)=\frac{1}{s}\log F(s)$, so we obtain $\frac{dH(s)}{ds} \le a\frac{d\log F(s)}{ds}+b$.
    c. Bounding the Expectation and Extracting Exponential Inequality => Using the fact that $\underset{s\to 0}{Limit}\, H(s)=\frac{1}{F(s)}\frac{d\log F(s)}{ds}\Big|_{s=0}=\mathbb{E}[Z]$ and $0\log F(0)=0$, and integrating the inequality, we obtain $H(s) \le \mathbb{E}[Z]+a\log F(s)+bs$, or, if $s<\frac{1}{a}$,

$\log \mathbb{E}\left[e^{s(Z-\mathbb{E}[Z])}\right] \leq \frac{s^2}{1-as}(a\mathbb{E}[Z]+b)$, 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.

4. Distinction Between the Upper and the Lower Bounds: Bounds for the lower tail $\mathbb{P}(Z \geq \mathbb{E}[Z]+t)$ maybe more easily derived by exploiting the association inequalities seen before, under much more general conditions on $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i}$. Notice the difference between this and the quantity $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z>Z'_i}$ appearing in the upper bound above.
5. Generalized Self-Bounding Lower Tail: Assume that for some non-decreasing function $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i} \leq g(Z)$. Then for all $t>0$, $\mathbb{P}(Z \leq \mathbb{E}[Z]-t) \leq e^{-\frac{t^2}{4\mathbb{E}[g(Z)]}}$.
    a. Proof Step #1 - Applying the Symmetrized Logarithmic Sobolev Inequality => To prove the lower tail inequalities, we obtain the upper bounds for $F(s)=\mathbb{E}[e^{sZ}]$ with $s<0$. By the third inequality of the symmetrized logarithmic Sobolev inequality, $s\mathbb{E}[Ze^{sZ}]-\mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\tau\left(s(Z-Z'_i)\right)\mathbb{I}_{Z<Z'_i}\right]$. By using $s<0$ and $\tau(-x)\leq x^2$ for $x>0$, the RHS above becomes $\sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}s^2(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i}\right] = s^2\mathbb{E}\left[e^{sZ}\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i}\right] \leq s^2\mathbb{E}[e^{sZ}g(Z)]$.
    b. Prof Step #2 – Apply Chebyshev's Association Inequality => Since $g(Z)$ is a non-decreasing function, and e^{sZ} is a decreasing function of Z, Chebyshev's association inequality implies that $\mathbb{E}[e^{sZ}g(Z)] \leq \mathbb{E}[e^{sZ}]\mathbb{E}[g(Z)]$.
    c. Proof Step #3 – Extraction of the Lower Tail => Scaling both sides of the above inequality by $s^2F(s)$ and writing $H(s)=\frac{1}{s}\log F(s)$, we obtain $\frac{dH(s)}{ds} \leq \mathbb{E}[g(Z)]$. Integrating in the interval $[s,0)$, we obtain $F(s) \leq e^{s^2\mathbb{E}[g(Z)]+s\mathbb{E}[ZX]}$. Finally, applying Markov's inequality and carrying out the optimization results in the statement of the inequality.
6. Data-Dependent Lower Tail: Assume that $Z=g(X_1^n)=g(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are independent real-valued random variables, and that g is a non-decreasing function of each variable. Suppose there exists another non-decreasing function $f:\mathbb{R}^n \to \mathbb{R}$ such that $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i} \leq f(X_1^n)$. Then, for all $t>0$, it can be shown, following above, that $\mathbb{P}(Z \leq \mathbb{E}[Z]-t) \leq e^{-\frac{t^2}{4\mathbb{E}[f(X_1^n)]}}$.

7. Alternate Approach to the Lower Tail Bound: The next result is useful when one is interested in lower tail-bounds, but $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z<Z'_i}$ is difficult to handle. In certain situations the right symmetrization pivot, i.e., $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z>Z'_i}$, is easier to bound. In such a situation, however, we need the additional guarantee that $|Z-Z'_i|$ remains bounded. Without loss of generality, we assume that this bound is 1.
8. Lower Tail Bound Using the Right Symmetrization Pivot: Assume that there exists a non-decreasing function g such that $\sum_{i=1}^{n}(Z-Z'_i)^2\mathbb{I}_{Z>Z'_i} \leq g(Z)$, and for any value of X_1^n and X_1', $|Z-Z'_i| \leq 1$. Then the following are true:
    a. $\log \mathbb{E}\left[e^{-s(Z-\mathbb{E}[Z])}\right] \leq s^2 \frac{\tau(\kappa)}{\kappa^2}\mathbb{E}[g(Z)]$ for all $\kappa > 0$ and $s \in \left[0, \frac{1}{\kappa}\right]$
    b. $\mathbb{P}(Z \leq \mathbb{E}[Z]-t) \leq e^{-\frac{t^2}{4(e-1)\mathbb{E}[g(Z)]}}$ for all $t > 0$ and $t \leq (e-1)\mathbb{E}[g(Z)]$
9. Proof Step #1 - Using the Right Symmetrization Pivot: The key observation is that the function $\frac{\tau(x)}{x^2} = \frac{e^x-1}{x}$ is increasing if $x > 0$. Choose $\kappa > 0$. Thus, for $s \in \left[-\frac{1}{\kappa}, 0\right]$, the right symmetrization pivot of the symmetrized logarithmic Sobolev inequality implies that $s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}]\log\mathbb{E}[e^{sZ}] \leq \sum_{i=1}^{n}\mathbb{E}\left[e^{sZ}\tau\left(-s(Z-Z'_i)\right)\mathbb{I}_{Z>Z'_i}\right]$ where at the last step, we have used the assumption of our postulate.
10. Proof Step #2 - Use of the Association Inequality: Just as in the proof for generalized self-bounding functions lower tail inequality, we bound $\mathbb{E}[e^{sZ}g(Z)]$ by $\mathbb{E}[e^{sZ}]\mathbb{E}[g(Z)]$. For the rest of the proof, we proceed as in the proof for self-bounding functions lower tail inequality. Here we have taken $\kappa = 1$.

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 ϕ-entropies and ϕ-Sobolev Inequalities* **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 d-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 ϕ-mixing Processes *Annals of Probability* **28** 416-461.
- Shamir, E., and J. Spencer (1987): Sharp Concentration of Chromatic Number on Random Graphs $g_{n,p}$ *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 $\sqrt{n}$ 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 $X_1, \dots, X_n$ taking their values in a measurable set $\mathcal{X}$, and denote a vector of these variables by $X_1^n = (X_1, \dots, X_n)$ taking its values in X^n.
2. Hamming Distance: Let $\mathcal{A} \subset \mathcal{X}^n$ be an arbitrary measurable set, and write $\mathbb{P}[\mathcal{A}] = \mathbb{P}[X_1^n \in \mathcal{A}]$. The *Hamming Distance* $d(x_1^n, y_1^n)$ between the vectors $x_1^n, y_1^n \in X_1^n$ is defined as the number of co-ordinates in which x_1^n and y_1^n differ. Introduce $d(x_1^n, \mathcal{A}) = \min_{y_1^n \in \mathcal{A}} d(x_1^n, y_1^n)$ as the Hamming distance between set $\mathcal{A}$ and the point x_1^n.
3. Hamming Distance Inequality: For any $t > 0$, $\mathbb{P}\left[d(x_1^n, \mathcal{A}) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}[\mathcal{A}]}}\right] \leq e^{-\frac{2t^2}{n}}$.
    a. Proof Step #1 – Bounded Differences Property => Observe that the function $g(x_1^n) = d(x_1^n, \mathcal{A})$ cannot change more than 1 by altering one component of x_1^n, that is, it follows the bounded differences property with constants $c_1 = \cdots = c_n = 1$. Thus, by

the bounded square differences inequality, with the optimal constants provided by the McDiarmid's inequality, $\mathbb{P}[\mathbb{E}[d(x_1^n, \mathcal{A})] - d(x_1^n, \mathcal{A}) \geq t] \leq e^{-\frac{2t^2}{n}}$.

b. Proof Step #2 - Inequality for $\mathbb{E}[d(x_1^n, \mathcal{A})]$ => By taking $t = \mathbb{E}[d(x_1^n, \mathcal{A})]$, the LHS becomes $\mathbb{P}[d(x_1^n, \mathcal{A}) \leq 0] \leq \mathbb{P}(\mathcal{A})$, so the above inequality implies $\mathbb{E}[d(x_1^n, \mathcal{A})] \leq \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}[\mathcal{A}]}}$. By using the bounded differences inequality again, we obtain $\mathbb{P}\left[d(x_1^n, \mathcal{A}) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{\mathbb{P}[\mathcal{A}]}}\right] \leq e^{-\frac{2t^2}{n}}$ as desired.

4. <u>*t-Blowup* of the Set $\mathcal{A}$</u>: Observe that on the RHS we have a measure of the complement of the *t-Blowup* of the set $\mathcal{A}$, that is, the measure of the set of points whose Hamming distance from $\mathcal{A}$ is at least t. If we consider a set, say with $\mathbb{P}[\mathcal{A}] = \frac{1}{10^6}$, we see something very surprising; the measure of the set of points whose Hamming distance to $\mathcal{A}$ is more than $10\sqrt{n}$ is smaller than e^{-108}! In other words, product measures are concentrated on extremely small *sets* – hence the name concentration of measure.
5. <u>Recovering the Bounded Differences Inequality</u>: Observe that the bounded differences inequality may also be derived from the above theorem. Indeed, if we consider a function g on $\mathcal{X}^n$ having the bounded differences property with constants $c_1 = \cdots = c_n = 1$ (for simplicity), then we may let $\mathcal{A} = \{x_1^n \in \mathcal{X}^n; g(x_1^n) \leq \mathbb{M}[Z]\}$ where $\mathbb{M}[Z]$ denotes the median of the random variable $Z = g(X_1, \dots, X_n)$. Then clearly $\mathbb{P}[\mathcal{A}] \geq \frac{1}{2}$, so the above inequality implies $\mathbb{P}\left[Z - \mathbb{M}[Z] \geq t + \sqrt{\frac{n}{2}\log 2}\right] \leq e^{-\frac{2t^2}{n}}$. (Note that the Hamming distance in this case - of $Z = g(X_1, \dots, X_n)$ - can be represented by $Z - \mathbb{M}[Z]$).
6. <u>Reduction to the Bounded Differences Form</u>: This has the same form as the bounded differences inequality, except that the expected value of Z is replaced by its mean. This difference is usually negligible, since $|\mathbb{E}[Z] - \mathbb{M}[Z]| \leq \mathbb{E}[|Z - \mathbb{M}[Z]|] = \int_0^\infty \mathbb{P}[|Z - \mathbb{M}[Z]| \geq t]dt$, so whenever the deviation of Z from its mean is small, its expected value must also be close to the mean.

   a. Mean-Median Closeness Statement => Let Z be a random variable with median $\mathbb{M}[Z]$ such that there exist positive constants a and b such that for all $t > 0$, $\mathbb{P}[|Z - \mathbb{M}[Z]| \geq t] \leq ae^{-\frac{t^2}{b}}$. Then $|\mathbb{E}[Z] - \mathbb{M}[Z]| \leq \frac{a\sqrt{\pi b}}{2}$.

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 X_1^n from the set $\mathcal{A}$ to be measured by a *Weighted Hamming Distance* $d_\alpha(x_1^n, \mathcal{A}) = \inf_{y_1^n \in \mathcal{A}} d_\alpha(x_1^n, y_1^n) = \inf_{y_1^n \in \mathcal{A}} \sum_{i:x_i \neq y_i} |\alpha_i|$ where $\alpha = (\alpha_1, \ldots, \alpha_n)$ is a vector of non-negative numbers.
4. 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 α, $\mathbb{P}\left[d_\alpha(X_1^n, \mathcal{A}) \geq t + \sqrt{\frac{\|\alpha\|^2}{2} \log \frac{1}{\mathbb{P}[\mathcal{A}]}}\right] \leq e^{-\frac{2t^2}{\|\alpha\|^2}}$ where $\|\alpha\|^2 = \sum_{i=1}^n {\alpha_i}^2$ denotes the Euclidean norm of α (recall that $\|\alpha\|^2$ serves the role of $c = \sum_{i=1}^n {c_i}^2$ in the bounded square differences inequality).
5. Reduction to Euclidean Norm: For all vectors α with unit norm $\|\alpha\|^2 = 1$, $\mathbb{P}\left[d_\alpha(X_1^n, \mathcal{A}) \geq t + \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}[\mathcal{A}]}}\right] \leq e^{-2t^2}$. Thus, denoting $u = \sqrt{\frac{1}{2} \log \frac{1}{\mathbb{P}[\mathcal{A}]}}$, for any $t \geq u$, we get $\mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t] \leq e^{-2(t-u)^2}$.

6. Reduction to the Talagrand Form: On the one hand, if $u = \sqrt{2 \log \frac{1}{\mathbb{P}[\mathcal{A}]}}$, $\mathbb{P}[\mathcal{A}] \leq e^{-\frac{t^2}{2}}$. On the other hand, since $(t-u)^2 \geq \frac{t^2}{4}$ for $t \geq 2u$, for any $t \geq \sqrt{2 \log \frac{1}{\mathbb{P}[\mathcal{A}]}}$ the inequality implies $\mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t] \leq e^{-\frac{t^2}{2}}$. Thus, for all $t > 0$, we have $\mathbb{P}[\mathcal{A}].\mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t] \leq \min(\mathbb{P}[\mathcal{A}], \mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t]) \leq e^{-\frac{t^2}{2}}$.
7. Talagrand Supremum Form: The main impact of the Talagrand's inequality is that the above inequality remains that even if the supremum over $\|\alpha\|$ with unit norm is taken with probability, i.e., $\sup_{\|\alpha\|: \|\alpha\| = 1} \mathbb{P}[\mathcal{A}].\mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t] \leq \sup_{\|\alpha\|: \|\alpha\| = 1} \min(\mathbb{P}[\mathcal{A}], \mathbb{P}[d_\alpha(X_1^n, \mathcal{A}) \geq t]) \leq e^{-\frac{t^2}{2}}$.
8. Relation of $\|\alpha\|$ to Chernoff/Hoeffding Type Optimization Parameters: To make above statement more precise, we introduce, for any $x_1^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, the *convex distance* of x_1^n from the set $\mathcal{A}$ by $d_\tau(x_1^n, \mathcal{A}) = \sup_{\alpha \in [0,\infty)^n: \|\alpha\| = 1} d_\alpha(x_1^n, \mathcal{A})$. 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 $\mathcal{A} \subseteq \mathcal{X}^n$ with $\mathbb{P}[X_1^n \in \mathcal{A}] \geq \frac{1}{2}$ and $t > 0$, $\min(\mathbb{P}[\mathcal{A}], \mathbb{P}[d_\tau(X_1^n, \mathcal{A}) \geq t]) \leq e^{-\frac{t^2}{4}}$.
2. Step #1 - Saddle Point Setup: Define the random variable $Z = d_\tau(X_1^n, \mathcal{A})$. First we observe that $d_\tau(x_1^n, \mathcal{A})$ can be expressed as a saddle-point. Let $\mathcal{M}(\mathcal{A})$ denote the set of probability measure on $\mathcal{A}$. Then $d_\tau(x_1^n, \mathcal{A}) = \sup_{\alpha: \|\alpha\| \leq 1} \inf_{\nu \in \mathcal{M}(\mathcal{A})} \sum_j \alpha_j \mathbb{E}_\nu \left[\mathbb{I}_{x_j \neq Y_j}\right]$ (where Y_1^n is distributed according to ν) $d_\tau(x_1^n, \mathcal{A}) = \inf_{\nu \in \mathcal{M}(\mathcal{A})} \sup_{\alpha: \|\alpha\| \leq 1} \sum_j \alpha_j \mathbb{E}_\nu \left[\mathbb{I}_{x_j \neq Y_j}\right]$, thereby the saddle point is achieved.

3. 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 $f(x, y)$ denotes a function $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is convex and lower semi-continuous with respect to x, concave and upper semi-continuous with respect to y, where $\mathcal{X}$ is convex and compact, then $\underset{x}{inf}\ \underset{y}{sup}\ f(x,y) = \underset{y}{sup}\ \underset{x}{inf}\ f(x,y)$. Additional details relating to checking the conditions of Sion's theorem are available in Boucheron, Lugosi, and Massart (2003).
4. Step #3 - Applying the Saddle Point: Let now $(\hat{\alpha}, \hat{\nu})$ be a saddle point for x_1^n. We have $Z_i' = \underset{\nu \in \mathcal{M}(\mathcal{A})}{inf}\ \underset{\alpha}{sup}\ \sum_j \alpha_j \mathbb{E}_\nu \left[\mathbb{I}_{x_j{}^{(i)} \neq Y_j}\right] \geq \underset{\nu \in \mathcal{M}(\mathcal{A})}{inf}\ \sum_j \hat{\alpha}_j \mathbb{E}_\nu \left[\mathbb{I}_{x_j{}^{(i)} \neq Y_j}\right]$ where $x_j{}^{(i)} = x_j$ if $j \neq i$ and $x_j{}^{(i)} = x_j{}'$. Let $\hat{\nu}$ denote the distribution on $\mathcal{A}$ that achieves the infimum in the latter expression. Now we have $Z = \underset{\nu}{inf}\ \sum_j \hat{\alpha}_j \mathbb{E}_\nu \left[\mathbb{I}_{x_j{}^{(i)} \neq Y_j}\right] \leq \sum_j \hat{\alpha}_j \mathbb{E}_{\hat{\nu}} \left[\mathbb{I}_{x_j{}^{(i)} \neq Y_j}\right]$. Hence we get $Z - Z_i' \leq \sum_j \hat{\alpha}_j \mathbb{E}_{\hat{\nu}} \left[\mathbb{I}_{x_j \neq Y_j} - \mathbb{I}_{x_j{}^{(i)} \neq Y_j}\right] \leq \hat{\alpha}_i \mathbb{E}_{\hat{\nu}} \left[\mathbb{I}_{x_i \neq Y_i} - \mathbb{I}_{x_i{}^{(i)} \neq Y_i}\right] \leq \hat{\alpha}_i$.
5. Step #4 - Use of the Bounded Square Differences Inequality: From above, we can see that $\sum_{i=1}^n (Z - Z_i')^2 \mathbb{I}_{Z > Z_i'} \leq \sum_{i=1}^n \hat{\alpha}_i{}^2 = 1$. Thus, by the bounded square differences inequality (or, more precisely, using its generalized expectations), for any $t > 0$, $\mathbb{P}[d_\tau(X_1^n, \mathcal{A}) - \mathbb{E}[d_\tau(X_1^n, \mathcal{A})] \geq t] \leq e^{-\frac{t^2}{4}}$.
6. Step #5 - Application of the Equivalent Bounded Square Differences Inequality for the Upper Tail: Applying the above to the lower tail, we get $\mathbb{P}[d_\tau(X_1^n, \mathcal{A}) - \mathbb{E}[d_\tau(X_1^n, \mathcal{A})] \leq -t] \leq e^{-\frac{t^2}{4(e-1)}}$, which by taking $t = \mathbb{E}[d_\tau(X_1^n, \mathcal{A})]$, implies $\mathbb{E}[d_\tau(X_1^n, \mathcal{A})] \leq \sqrt{4(e-1)\log\frac{1}{\mathbb{P}[\mathcal{A}]}}$. Thus, this means $\mathbb{P}\left[d_\tau(X_1^n, \mathcal{A}) - \sqrt{4(e-1)\log\frac{1}{\mathbb{P}[\mathcal{A}]}} \geq t\right] \leq e^{-\frac{t^2}{4}}$.
7. Step #6 - Reduction to the Talagrand Form: Now, if $0 < u < \sqrt{4\log\frac{1}{\mathbb{P}[\mathcal{A}]}}$, then $\mathbb{P}[\mathcal{A}] \leq e^{-\frac{u^2}{4}}$. On the other hand, if $u \geq \sqrt{4\log\frac{1}{\mathbb{P}[\mathcal{A}]}}$, then $\mathbb{P}[d_\tau(X_1^n, \mathcal{A}) \geq u] \leq \mathbb{P}\left[d_\tau(X_1^n, \mathcal{A}) -\right.$

$\sqrt{4(e-1)\log\frac{1}{\mathbb{P}[\mathcal{A}]}} \geq u\left[1-\sqrt{\frac{e-1}{e}}\right]\Big] \leq e^{-\frac{u^2\left[1-\sqrt{\frac{e-1}{e}}\right]^2}{4}}$ where the second inequality follows from the upper-tail inequality above.

8. Step #7 - Bringing it all together: In conclusion, for all $u > 0$ $\min(\mathbb{P}[\mathcal{A}], \mathbb{P}[d_\tau(X_1^n, \mathcal{A}) \geq u]) \leq e^{-\frac{u^2\left[1-\sqrt{\frac{e-1}{e}}\right]^2}{4}}$ 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 $g(x_1^n)$ denote the minimum number of bins of size 1 into which the numbers $x_1, \ldots, x_n \in [0,1]$ can be packed. We consider the random variable $Z = g(X_1^n)$ where $X_1, \ldots, X_n$ are independent, and take values in $[0,1]$.
2. Convex Distance Inequality for Bin Packing: For each $t > 0$, $\mathbb{P}[|Z - \mathbb{M}[Z]| \geq t + 1] \leq 8e^{-\frac{t^2}{16\left(2\mathbb{E}\left[\sum_{i=1}^n X_i{}^2\right]+t\right)}}$.
3. Proof Step #1 - Symmetrization of Bin Packing: We first establish (and this is the only specific property of g that we use in the proof) that for any $x_1^n, y_1^n \in [0,1]^n$, $g(x_1^n) \leq g(y_1^n) + 2\sum_{i:x_i \neq y_i} x_i + 1$. To see this, it suffices to show that those x_i for which $x_i \neq y_i$ can be packed into at most $2\sum_{i:x_i \neq y_i} x_i + 1$ 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.
4. Proof Step #2 - Recasting this into the Hamming Distance Metric: Denoting by $\alpha = \alpha(x_1^n) \in [0,\infty)^n$ the unit vector $\frac{x_1^n}{\|x_1^n\|}$, we clearly have $\sum_{i:x_i \neq y_i} x_i = \|x_1^n\| \sum_{i:x_i \neq y_i} \alpha_i = \|x_1^n\| d_\alpha(x_1^n, y_1^n)$.
5. Proof Step #3 - Symmetrization Bounding Using Convex Distance: Let a be a positive number, and define the set $\mathcal{A}_a = \{y_1^n : g(y_1^n) \leq a\}$. Then, by the argument above and using

the definition of convex distance, for each $x_1^n \in [0,1]^n$ there exists a $y_1^n \in \mathcal{A}_a$ such that $g(x_1^n) \leq g(y_1^n) + 2\sum_{i:x_i \neq y_i} x_i + 1 \leq a + 2\|x_1^n\| d_\alpha(x_1^n, y_1^n) + 1$, from which we conclude that for each $a > 0$, $Z \leq a + 2\|x_1^n\| d_\alpha(x_1^n, y_1^n) + 1$.

6. Proof Step #4 - Applying the Hamming Distance Metric: From the above, we have for any $t > 0$, $\mathbb{P}[Z \geq a + 1 + t] \leq \mathbb{P}\left[Z \geq a + 1 + t\frac{2\|X_1^n\|}{2\sqrt{2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t}}\right] + \mathbb{P}\left[\|X_1^n\| \geq 2\sqrt{2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right] + t}\right]$.
7. Proof Step #5 - Application of the Bernstein Inequality: Given that $X_1, \ldots, X_n$ are independent random variables taking values in $[0,1]$, it may be shown that, by applying the Bernstein inequality to the sum $\sum_{i=1}^n {X_i}^2$, $\mathbb{P}\left[\sqrt{\sum_{i=1}^n {X_i}^2} \geq \sqrt{2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right] + t}\right] \leq e^{-\frac{3}{8}\left(\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t\right)}$. Using this along with the Hamming metric in Step #7, we get $\mathbb{P}[Z \geq a + 1 + t] \leq \mathbb{P}\left[d_\tau(X_1^n, \mathcal{A}_a) \geq \frac{t}{2\sqrt{2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t}}\right] + e^{-\frac{3}{8}\left(\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t\right)}$.
8. Proof Step #6 - Acquiring the Bounds: To obtain the desired inequality, we incorporate two difference choices of a on the inequality above. To derive the bound for the upper tail, we take $a = \mathbb{M}[Z]$. Then $\mathbb{P}[\mathcal{A}_a] \geq \frac{1}{2}$, and the convex distance inequality yields $\mathbb{P}[Z \geq \mathbb{M}[Z] + 1 + t] \leq 2\left\{e^{-\frac{t^2}{16\left(2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t\right)}} + e^{-\frac{3}{8}\left(\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t\right)}\right\} \leq 4e^{-\frac{t^2}{16\left(2\mathbb{E}\left[\sum_{i=1}^n {X_i}^2\right]+t\right)}}$. We obtain a similar equality in the same way for $\mathbb{P}[Z \geq \mathbb{M}[Z] - 1 - t]$ by taking $a = \mathbb{M}[Z] - 1 - t$.

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).

Standard SLT Framework

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:
    a. Learnability/Learn Feasibility/Learn Complexity
    b. Learning Speed/Performance
    c. Learn Formulation Criterion
    d. Learn Basis/Representation Choice
    e. Probabilistic Learning Framework and Viability
4. 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:
    a. Positive Results => Showing that a certain class of functions is learnable in polynomial time.
    b. Negative Results => Showing that a certain class of functions cannot be learned in polynomial time.
5. Non Polynomial Time Results: Negative results often rely on commonly believed, but as yet unproven assumptions such as:
    a. Computational Complexity - $P \neq NP$
    b. Cryptographic – One-way functions exist
6. 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.
7. Computational Learning Theory Approaches:
    a. Probably Approximately Correct Learning (PAC) proposed by Valiant (1984)
    b. VC Theory proposed by Vapnik and Chervonenkis (1971)

c. Bayesian Inference
d. Algorithmic learning theory, from the work of Gold (1967)
e. Online machine learning algorithms, from the work of Nick Littlestone

8. 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 (mis-classified samples).
    a. Probably/Approximately => This model uses $P(x, y)$, which is the "probably" part, and within that chooses the low error, the "approximately" part. Since $P(x, y)$ 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 X 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 $\{0,1\}^n$. In the interval problem, this would be $X = \mathbb{R}$ where $\mathbb{R}$ denotes the set of all real numbers.
3. PAC Concept and the Concept Class: A concept is the subset $c \in X$. One concept is the set of all patterns in bits $X = \{0,1\}^n$ that encode the picture of the letter P. An example concept from the second example is the set of all numbers between $\frac{\pi}{2}$ and 10. A concept class C is a set of concepts over X. 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).
4. The PAC Procedure: Let $EX(c, D)$ be a procedure that draws an example x using a probability distribution D and gives the correct label $c(x) = \begin{cases} 1 & x \in C \\ 0 & Else \end{cases}$.
5. PAC Learnability and the Learning Algorithm: Say that there is an algorithm A that, given access to $EX(c, D)$ alongwith inputs ε and δ that, with a probability of at least $1 - \delta$, A outputs a hypothesis h that has error less than or equal to ε with examples drawn from X using the distribution D. If there is an algorithm for every concept $c \in C$, for every distribution D over X, and for all $0 < \varepsilon < \frac{1}{2}$ and $0 < \delta < \frac{1}{2}$, then C is PAC learnable (or distribution-free PAC learnable). Further, we can also say that A is a PAC learning algorithm for C.
6. Efficient PAC Algorithm: An algorithm runs in time t if it draws at most t samples and requires at most t time steps. A concept class is efficiently PAC learnable if it is PAC learnable by an algorithm that runs in time polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, 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):
    a. Which learning tasks can be performed by computers in general (producing positive/negative results)?
    b. What kind of assumptions do we have to make to ensure that such machine learning tasks are successful?
    c. What key properties does a learning algorithm need to satisfy in order to be successful?
    d. What performance guarantees can we provide on certain learning algorithms?

The Setup

1. Empirical Error Determinants: In general, the empirical error depends upon:
    a. The Joint Probability Distribution $P(x, y)$
    b. The Sample Size n
    c. 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 n to within a tolerance δ depends on:
    a. The tolerance δ
    b. The sample size n
    c. The Joint Probability Distribution $P(x, y)$
    d. The Function Complexity – the complexity of the function proxying $P(x, y)$
3. SLT Objectives:
    a. No assumptions on $P(x, y)$
    b. Labels can be non-deterministic so as to be able to accommodate label noise and overlapping label classes
    c. Sampling i.i.d.
    d. $P(x, y)$ is unknown at the time of learning
4. Use of $P(x, y)$ 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 $P(x, y)$, i.e., only using sample size, function space complexity, and tolerance.
5. Loss and Risk: The loss function l is the cost of misclassifying one observation point, and is written as $l[X, Y, f(X)] = \begin{cases} 1, & f(X) \neq Y \\ 0, & f(X) = Y \end{cases}$. Risk of a function is the average of the loss over the data points generated according to the underlying distribution $P(x, y)$, so the distribution DOES factor in $R(f) = E\{l[X, Y, f(X)]\}$. Empirical risk, however, is simply a miscalculation counter.
6. Bayes' Classifier: $f_{Bayes}(x) = \begin{cases} -1, & P(Y = 1|X = x) \geq 0.5 \\ +1, & P(Y = 1|X = x) < 0.5 \end{cases}$. Almost certainly, Bayes' classifier does not possess zero empirical risk. Further, it has explicit dependence on $P(x, y)$ (or $P(y|x)$).
7. SLT Problem Statement: Given a set of training points $(X_1, Y_1), \ldots, (X_n, Y_n)$ that have been drawn from an unknown distribution $P(x, y)$, and given a loss function l, how can we construct a function $f: X \to Y$ that has its risk $R(f)$ 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 $\{\Im_d : d = 1,2, \dots\}$ 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., $f_n = \underset{f \in \Im_d, d \in \mathbb{N}}{\arg min} R_n(f) + pen(d,n)$. The penalty $pen(d,n)$ gives preference to those models whose estimation error is small, and thus measures the size/capacity of the model.
4. Standardized Regularization: This easier to implement approach consists in choosing of a large model $\Im$ (possibly dense in continuous functions, for example), and to define on $\Im$ a regularizer, usually a norm $\|f\|$. One then minimizes the regularized empirical risk $f_n = \underset{f \in \Im}{\arg min} R_n(f) + \lambda\|f\|^2$. Most existing and successful methods can be thought of as a variant of the regularization method.
5. Regularization by Hypothesis Restriction: Regularization can also occur by restricting the hypothesis space. A common example is the restriction of H to linear functions – thereby reducing the problem to that of linear regression. H may also be reduced to polynomials of degree p, exponentials, or bounded functions on L_1. 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.

6. Tikhonov Regularization: This consists of minimizing $\frac{1}{n}\sum_{i=1}^{n} L[h(x_i), y_i] + \gamma \|f\|_H{}^2$ where γ 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)).
7. 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 $\mathfrak{I}$.

Normalized Regularization Details

1. Regularization from Probability: Given a probability distribution π (referred to as a prior) defined on $\mathfrak{I}$, one can use $-\log \pi(f)$ as the regularizer. In case $\mathfrak{I}$ is countably/uncountably infinite/continuous, we use the density associated with $\pi(f)$ instead.
2. "Prior" from the Regularizer: Reciprocally, from the regularizer of the form $\|f\|^2$, if there exists a measure μ on $\mathfrak{I}$ such that $\int e^{-\lambda\|f\|^2} d\mu(f) < \infty$ for some $\lambda > 0$, then one may construct a prior corresponding to this regularizer.
    a. As an example, if $\mathfrak{I}$ is a set of hyper-planes in $\mathbb{R}^d$ going through the origin, $\mathfrak{I}$ can be identified with $\mathbb{R}^d$, and taking μ as the Lebesgue measure, it is possible to go from Euclidean norm regularizer to spherical Gaussian measure on $\mathbb{R}^d$ as a prior.
3. 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.
4. Posterior: From this type of normalized regularizer (or its posterior), we can construct another probability distribution ρ on $\mathfrak{I}$ (typically called the posterior) as $\rho(f) = \frac{e^{-\gamma R_n(f)}}{\mathbb{Z}(\gamma)}\pi(f)$ where $\gamma \geq 0$ is a free parameter and $\mathbb{Z}(\gamma)$ is a normalization factor.
5. Uses of the Posterior:

a. MAP as the Regularization Process => If we maximize ρ, we recover the original regularization framework as $\underset{f \in \Im}{\arg max}\ \rho(f) = \underset{f \in \Im}{\arg max}\ [\gamma R_n(f) - \log \pi(f)]$ where the regularizer is $-\frac{\log \pi(f)}{\gamma}$ (note that maximizing $\gamma R_n(f) - \log \pi(f)$ is equivalent to minimizing $R_n(f) - \frac{\log \pi(f)}{\gamma}$).

b. Randomization of Predictions => ρ can also be used to randomize the predictions. In this case, before computing the predicted labels for the input x, one samples a function f according to ρ and computes $f(x)$. This procedure is called Gibbs classification.

c. Bayesian Averaging => Another way in which the ρ constructed above can be used is by taking the expected prediction of the functions in $\Im$; $f_n(x) = sign\left[\mathbb{E}_\rho\{f(x)\}\right]$. This is called Bayesian averaging.

References

- Computational Learning Theory (Wiki): *Wikipedia Entry for Computational Learning Theory*.
- Gold, E. (1967): Language Identification in the Limit, *Information and Control* **10 (5)**: 447-474.
- Kearns, M., and U. Vazirani (1994): *An Introduction to Computational Learning* **MIT Press**.
- 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*

- Probably Approximately Correct Learning (Wiki): *Wikipedia Entry for Probably Approximately Correct Learning*.
- 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*.
- 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.
- 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 f_n generalizes well if the difference $\left|R(f_n) - R_{Emp}(f_n)\right|$ between the overall risk $R(f_n)$ and the empirical risk $R_{Emp}(f_n)$ is small. This does NOT mean that the classifier has a small overall error R_{Emp}; it simply means that the empirical error $R_{Emp}(f_n)$ is a good estimate of the true error $R(f_n)$ (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 f_n, it is a property of the hypothesis space $\Im$ (von Luxburg and Scholkopf (2008)).

Types of Consistency

1. The Setup: Let $(X_i, Y_i)_{i \in N}$ be an infinite sequence of training points that have been drawn independently from some probability distribution $P(x, y)$. Let l be a loss function. For each $n \in N$, let f_n be a classifier constructed by some learning algorithm on the basis of the first n training points.
2. Consistency with respect to $\Im$ and $P(x, y)$: The learning algorithm is called consistent with respect to $\Im$ and $P(x, y)$ if the risk $R(f_n)$ converges in probability to the risk $R(f_\Im)$ of the best classifier in $\Im$, that is, for all $\varepsilon > 0$, $P[R(f_n) - R(f_\Im) > \varepsilon] \to 0$ as $n \to \infty$.
3. Bayes' Consistent: The learning algorithm is called Bayes' consistent with respect to $P(x, y)$ if the risk $R(f_n)$ converges to $R\left(f_{Bayes}\right)$ of the Bayes' classifier, that is, for all $\varepsilon > 0$, $P\left[R(f_n) - R\left(f_{Bayes}\right) > \varepsilon\right] \to 0$ as $n \to \infty$.

4. Universal Consistency: The learning algorithm is called universally consistent with respect to $\mathfrak{I}$ (respectively universally Bayes' consistent) if it is consistent with respect to $\mathfrak{I}$ (respectively Bayes' consistent) for all probability distributions $P(x, y)$.
5. 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: $R(f_n) - R\left(f_{Bayes}\right) = [R(f_n) - R(f_{\Im})] + \left[R(f_{\Im}) - R\left(f_{Bayes}\right)\right]$. $[R(f_n) - R(f_{\Im})]$ is called the Variance or the Estimation Error. Traditionally this has occupied much of the focus of SLT. $\left[R(f_{\Im}) - R\left(f_{Bayes}\right)\right]$ 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 $P(x, y)$.
2. Setup: We employ the terminology of Vijayakumar (2007). Let the data specified $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ be derived from the true model $y = f(x) + \epsilon$ where ϵ is a random error with $E(\epsilon) = 0$. Using this we train a model $g(x|D)$ to approximate f. We also set $g(x_i) = g_i$ and $f(x_i) = f_i$.
3. Mean Squared Error: $MSE = \frac{1}{N}\sum_{i=1}^{N}[y_i - g(x_i|D)]^2$. The quantity of interest is the expectation of the MSE across the different realizations of the data: $E[MSE] = \frac{1}{N}\sum_{i=1}^{N} E[y_i - g(x_i|D)]^2$
4. MSE Expansion: Re-writing $\frac{1}{N}\sum_{i=1}^{N}[y_i - g_i]^2$ as $\frac{1}{N}\sum_{i=1}^{N}[y_i - f_i + f_i - g_i]^2$, we recognize that $y_i - f_i$ is the noise term ϵ above, and $f_i - g_i$ is simply the proxying error. Thus, $E[y_i - g_i]^2 = E[y_i - f_i]^2 + E[f_i - g_i]^2 + 2E\{[y_i - f_i][f_i - g_i]\} = E[\epsilon^2] + E[f_i - g_i]^2 + 2\{E[y_i f_i] - E[f_i^2] - E[y_i g_i] + E[f_i g_i]\}$. Given that f is deterministic, $E[y_i f_i] - E[f_i^2] = 0$. Likewise, $E[y_i g_i] = E[g_i(f_i + \varepsilon)] = E[f_i g_i] + E[\varepsilon g_i] = E[f_i g_i]$ if we assume that $E[\varepsilon g_i] = 0$. Thus the last term also vanishes. Therefore $E[y_i - g_i]^2 = E[\epsilon^2] + E[f_i - g_i]^2$.
5. MSE Reduction: $E[f_i - g_i]^2 = E[f_i - E[g_i] + E[g_i] - g_i]^2 = E\left[f_i - E[g_i]\right]^2 + E[E[g_i] - g_i]^2 + 2\{E([f_i - E[g_i]][E[g_i] - g_i])\}$. Here $\{E([f_i - E[g_i]][E[g_i] - g_i])\} = E\{f_i E[g_i]\} - E\{E[g_i]E[g_i]\} - E\{f_i g_i\} + E\{g_i E[g_i]\} = f_i E[g_i] - \{E[g_i]\}^2 - f_i E[g_i] + \{E[g_i]\}^2 = 0$. Thus $E[f_i - g_i]^2 = E\left[f_i - E[g_i]\right]^2 + E[E[g_i] - g_i]^2$.
6. MSE Final Form: $E[y_i - g_i]^2 = E[\epsilon^2] + E\left[f_i - E[g_i]\right]^2 + E[E[g_i] - g_i]^2$. Here $E[\epsilon^2]$ is the irreducible sample error (assuming infinite sample), $E\left[f_i - E[g_i]\right]^2$ is the square of the

bias term, and $E[E[g_i] - g_i]^2$is the model variance term. Cast in another manner, $E[y_i - g_i]^2 = Mean\ Squared\ Error = Irreducible\ Error + Bias^2 + Model\ Variance.$

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:
    a. Generalized linear models can be regularized to increase their bias.
    b. In neural nets, deeper models with more layers will have stronger variance, this regularization (as in GLM) is applied.
    c. In k-nearest neighbor models, a high value of k leads to low variance.
    d. In Instance based learning, regularization can be achieved by varying the mixture of prototypes and examplars (Gagliardi (2011)).
    e. 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 $d: \chi \times \chi \to \mathbb{R}$ that assigns a distance value $d(X, X')$ to each pair of training points (X, X'). Then $NN(X)$, the nearest

neighbor to X among all the training points is $NN(X) = \underset{X' \in \{X_1, \ldots, X_n\}}{\arg min}\{d(X, X') \leq d(X, X") \vee X" \in \{X_1, \ldots, X_n\}$ (Stone (1977), Scholkopf and Smola (2002)). The corresponding classifier f_n is $f_n(X) = Y_i$ where $X_i = NN(X)$.

2. kNN Bayes' Risk vs. f_n Risk: It is easy to construct instances to see where the Bayes' risk for 1NN is quiet different from $R_{1NN}(f_n)$. However, by expanding the nearest neighbor out to kNN, we can see the difference between $R_{kNN}(f_n)$ and $R_{kNN}(f_{Bayes})$ decreases, thereby improving the consistency.
3. Stone (1977) kNN Consistency Theorem: Let f_n be the kNN classifier constructed on n sample points. If $n \to \infty$ and $k \to \infty$ such that $\frac{k}{n} \to 0$, then $R_{kNN}(f_n) \to R_{kNN}(f_{Bayes})$ for all probability distributions $P(x, y)$. Therefore, kNN classification thus constructed can be made universally consistent.
    a. Interpretation => Essentially, one has to allow the sample size k of the neighborhood under consideration to grow with the sample size n in a controlled way. For e.g., if the parameter k grows slowly with n, say, $k \approx \log n$, then kNN becomes universally Bayes' consistent.
4. Intuition behind the kNN Function Complexity: Treating k as the hypothesis variant dimension in the space of functions $\Im_{kNN}$, the extremes to observe are: a) 1NN, where potentially each node can switch its ± 1 label depending upon its neighbors, and b) kNN with $k = n$, the sample size, which has only one label. In this situation, for a location independent $P(x, y)$, $R(f_{Bayes}) = R(f_{kNN})$ (Devroye, Gyorfi, and Lugosi (1996)).

References

- Bias-variance dilemma (Wiki): *Wikipedia Entry for 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*

- 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*
- 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*.
- 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).

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 $f_n \in \Im$ that minimizes the empirical risk of the given finite sample is called the ERM induction principle: $f_n \coloneqq \underset{f \in \Im}{\arg min} R_{emp}(f)$.
4. Principle Caveats: With every sample set change, f_n will vary. Ideally, f_{ayes} should be part of $\Im$, so that as $n \to \infty$, $f_n \to f_{Bayes}$. Otherwise $f_n \to f_\Im$ but $f_\Im \neq f_{Bayes}$, even as $n \to \infty$; this produces sizeable estimation error $[R(f_n) - R(f_\Im)]$ as well as approximation error $\left[R(f_\Im) - R\left(f_{Bayes}\right)\right]$.

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 X and Y and would like to learn a function (called a hypothesis) $h_i: X \rightarrow Y$ which outpts an object $y \in Y$ given $x \in X$. We have a set of training examples $(x_1, y_1), \ldots, (x_N, y_N)$ where $x_i \in X$ is the input and $y_i \in X$ is the corresponding response, and we wish to get the response predictor hypothesis $h(x_i)$.
3. Training Set Generation: Formally we assume that there is a joint probability distribution $P(x, y)$ over X and Y, and the training set consists of N instances drawn i.i.d. from $P(x, y)$. Note that the assumptions of joint probability allows us to model uncertainties in the predictors (e.g., noise from data) because y is not a deterministic function of x, but rather a random variable with a conditional distribution $P(y|x)$ for a fixed x.
4. ERM Loss Function: We also assume that we are given a non-negative real-valued loss-function $L(\hat{y}, y)$ that quantifies how different the prediction $\hat{y}$ of the hypothesis is from the true outcome y. The risk associated with the hypothesis is then defined as the expectation of the loss function: $R(h) = E\{L[h(x), y]\} = \int L[h(x), y] \mathrm{d}\, P(x, y)$. A commonly used loss function for classification is the $0 - 1$ loss function $L(\hat{y}, y) = I(\hat{y} \neq y)$ where $I(\ldots)$ is the indicator notation.
    a. Expectation Maximization Insight => The least squares loss function corresponds to maximizing the joint normal probability $P(x, y)$. If $P(x, y)$ 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).
5. The Goal of the Computational Learning Algorithm: To find the hypothesis h^* among the fixed class of hypotheses H that minimizes the risk $R(h)$: $h^* = \underset{h \in H}{\arg min}\, R(h)$

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 $R(h)$ cannot be computed since $P(x, y)$ 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 $P(x, y)$): $R_{Emp}(h) = \frac{1}{N}\sum_{i=1}^{N} L[h(x_i), y_i]$ (See Empirical Risk Minimization (Wiki)).
3. Problem Statement: Empirical Risk Minimization Principle states that the learning algorithm should choose a hypothesis h^* that minimizes the empirical risk: $h^* = \underset{h \in H}{\arg min}\, R_{Emp}(h)$. 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 $n \to \infty$ 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 $\mathcal{P}$-Donsker; LLN over hypothesis space is, likewise, called Glivenko-Cantelli. Thus, one implies the other.
2. Glivenko-Cantelli Literature: The question of how $\underset{f \in \Im}{sup}[P(f) - P_n(f)]$ 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).
3. ERM LLN: LLN simply states that $\frac{1}{n}\sum_{i=1}^{n} \xi_i \to E(\xi)$ as $n \to \infty$, where ξ_i and ξ are drawn from the same distribution. LLN is valid under very mild conditions, and is easily extensible to ERM as follows: $\frac{1}{n}\sum_{i=1}^{n} l[X_i, Y_i, f(X_i, Y_i)] \to E\{l[X, Y, f(X, Y)]\}$ as $n \to \infty$. Again, (X_i, Y_i) is drawn from the same distribution that generates $f(X, Y)$.

4. 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 $P(x, y)$, and is quiet conservative). Applied to ERM, it is: $P\left[\left|R_{Emp}(f) - R(f)\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2}$.
5. 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 $P(x, y)$ 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., $VCdim \to \infty$). This makes them terrible learners.
2. 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 $f \in \Im$ that the learning machine could choose.

2. The ERM Approach: For each function $f \in \Im$ LLN tells us that as $n \to \infty$, $R_{Emp}(f) \to R(f)$. However, this does not necessarily mean that the ER minimizer function f_n produces a risk that approaches that of $f_\Im$ as $n \to \infty$. The latter is what results in statistical/functional class consistency. For this to be true, we explicitly require that $R_{Emp}(f) \to R(f)$ as $n \to \infty$ for all $f \in \Im$ (Scholkopf and Smola (2002)).
3. Sample Size producing Uniform Convergence: Of course, from LLN, there will exist some large enough n such that $\left|R_{Emp}(f) - R(f)\right| \geq \varepsilon$ for all $f \in \Im$ (we write this as $\sup_{f \in \Im} \left|R_{Emp}(f) - R(f)\right| \geq \varepsilon$). Ideally we hope that this $n \to \infty$ is not too large.
4. f_n ERM Consistency Workout: For the hypothesis class estimation error consistency, we require that $|R(f_n) - R(f_\Im)|$ be bounded. Now, $|R(f_n) - R(f_\Im)| = \left|R(f_n) - R_{Emp}(f_n) + R_{Emp}(f_n) - R_{Emp}(f_\Im) + R_{Emp}(f_\Im) - R(f_\Im)\right| \leq \left|R(f_n) - R_{Emp}(f_n)\right| + \left|R(f_\Im) - R_{Emp}(f_\Im)\right| + \left|R_{Emp}(f_n) - R_{Emp}(f_\Im)\right| \leq \left|R(f_n) - R_{Emp}(f_n)\right| + \left|R_{Emp}(f_n) - R_{Emp}(f_\Im)\right|$

ERM Complexity

1. ERM Complexity Analysis: ERM for a classification problem with a $0-1$ 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 $0-1$ 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 $P(x, y)$ (thereby not being agnostic anymore).
    a. Loss Functions for Classification => The $0-1$ 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: $L[h(x), y] = [-yh(x)]_+$

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*.
- 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* **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 $R(f)$ in $\sup_{f \in \mathfrak{J}} \left| R_{Emp}(f) - R(f) \right| \geq \varepsilon$ (since $R(f)$ is not an observable) by $R_{Emp}(f)$, which is an observable. To this end, we introduce a new copy $(X_i', Y_i')_{i=1,\ldots,n}$ of our original sample, and refer to this as our ghost sample. This all expectations for $R(f)$ now occur over (X_i', Y_i').
2. 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.
3. Vapnik Chervonenkis Symmetrization Lemma: For $m\varepsilon^2 \geq 2$, we have $\mathbb{P}\left[\sup_{f \in \mathfrak{J}} \left| R(f) - R_{Emp}(f) \right| \geq \varepsilon\right] \leq 2\mathbb{P}\left[\sup_{f \in \mathfrak{J}} \left| R'_{Emp}(f) - R_{Emp}(f) \right| \geq \frac{\varepsilon}{2}\right]$. The $\mathbb{P}$ 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 n each – the original and the ghost samples – thereby an i.i.d. distribution of size $2n$.
4. Transformation onto the Outcome Space: Using the lemma, we can now convert the probability bounds estimation in the infinite sample case, i.e., $R(f)$, to samples over the dual sequence of n and n', i.e., $2n$ sample outcomes. Remember that the set (n, n') can only assume a finite number of outcomes, i.e., in the case of binary classifiers, at most $2^{n+n'} = 2^{2n}$ outcomes.
5. 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 2^{2n} 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 $\varepsilon > 0$ such that $n\varepsilon^2 \geq 2$, $\mathbb{P}\left[\sup_{f \in \Im} (P - P_n)f \geq \varepsilon\right] \leq 2\mathbb{P}\left[\sup_{f \in \Im} (P'_n - P_n)f \geq \frac{\varepsilon}{2}\right]$ where P'_n is the corresponding empirical measure taken over an independent ghost sample $z'_1, \ldots, z'_n$.
3. #1 - Applying Indicator Functions over the Split Subsamples: Let f_n be the function that achieves the supremum (note that it depends on $z_1, \ldots, z_n$). Using $\wedge$ to represent the conjunction of the 2 events we get $1_{(P-P_n)f_n > \varepsilon} 1_{(P-P'_n)f_n < \frac{\varepsilon}{2}} = 1_{(P-P_n)f_n > \varepsilon \wedge (P'_n - P)f_n > -\frac{\varepsilon}{2}} \leq 1_{(P'_n - P_n)f_n < \frac{\varepsilon}{2}}$
4. #2 - Expectation over the Ghost Sample: Taking expectations over the ghost samples $\mathbb{P}'$ gives $1_{(P-P_n)f_n > \varepsilon} P'\left[(P - P'_n)f_n < \frac{\varepsilon}{2}\right] \leq P'\left[(P'_n - P_n)f_n > \frac{\varepsilon}{2}\right]$.
5. #3 - Applying Chebyshev Inequality to the Ghost Sample: By Chebyshev's inequality $P'\left[(P - P'_n)f_n \geq \frac{\varepsilon}{2}\right] \leq \frac{4Var(f_n)}{n\varepsilon^2} \leq \frac{1}{n\varepsilon^2}$, since the random variable in the range $[0, 1]$ has a variance $\leq \frac{1}{4}$. Thus, switching from $\geq \frac{\varepsilon}{2}$ to $< \frac{\varepsilon}{2}$, $1_{(P-P_n)f_n > \varepsilon}\left[1 - \frac{1}{n\varepsilon^2}\right] \leq P'\left[(P'_n - P_n)f_n > \frac{\varepsilon}{2}\right]$.
6. #4 Expectation over the Original Sample: Taking expectation with respect to the original sample ($\mathbb{P}$), and observing that $n\varepsilon^2 \geq 2$, we recover the statement we set out to prove, i.e., the Symmetrization lemma.
7. Applying Hoeffding's Union Bound: $\mathbb{P}\left[\sup_{f \in \Im} (P - P_n)f \geq \varepsilon\right] \leq 2\mathbb{P}\left[\sup_{f \in \Im} (P'_n - P_n)f \geq \frac{\varepsilon}{2}\right] = 2\mathbb{P}\left[\sup_{f \in \Im_{z_1,\ldots,z_n,z'_1,\ldots,z'_n}} (P'_n - P_n)f \geq \frac{\varepsilon}{2}\right] \leq 2S_{\Im}(2n)\mathbb{P}\left[(P'_n - P_n)f \geq \frac{\varepsilon}{2}\right] \leq 4S_{\Im}(2n)e^{-\frac{n\varepsilon^2}{8}}$. Inversion of this proves the VC theorem for risk bound using the growth function; For $n\varepsilon^2 \geq 2$, $\delta > 0$ and, with probability at least $1 - \delta$, $R(f) - R_n(f) \leq 2\sqrt{2\frac{\log S_{\Im}(2n) + \log\frac{2}{\delta}}{n}}$.

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 $Prob\left[\sup_{f \in \Im}\left|R_{Emp}(f) - R(f)\right| \geq \varepsilon\right]$.
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: $Prob\left[\sup_{f \in \Im}\left|R_{Emp}(f) - R(f)\right| \geq \varepsilon\right] \leq \sum_{i=1}^{m} Prob\left[\left|R_{Emp}(f_i) - R(f_i)\right| \geq \varepsilon\right] \leq 2me^{-2n\varepsilon^2}$ where m is the number of (discrete) functions in the function space $\Im$, and n is the number of data points. Note that we seek to optimize the modulus $\left|R_{Emp}(f_i) - R(f_i)\right|$.
4. Hoeffding for the Function Space: We start from $Prob(|S_n - E[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$. Note that $S_n \to R_{Emp}(f)$, and $E[S_n] \to R(f)$. Further $t \to n\varepsilon$, and $b_i - a_i \to 1$. Therefore $\sum_{i=1}^{n}(b_i - a_i)^2 = n^2$. Thus $Prob\left[\left|R_{Emp}(f_i) - R(f_i)\right| \geq \varepsilon\right] \leq 2e^{-2n\varepsilon^2}$.
5. Discrete Union Bound - Convergence: Obviously, as $n \to \infty$, $Prob\left[\sup_{f \in \Im}\left|R_{Emp}(f) - R(f)\right| \geq \varepsilon\right] \to 0$. This ensures statistical sample consistency. Extension to the continuous function space case requires a few more techniques, including replacing m with a more general capacity measure.
6. Refined Union Bound and the Countable Case #1: For each $f \in \Im$, for each $\delta > 0$ (possibly depending on f, which we write as $\delta(f)$), Hoeffding's inequality says that $\mathbb{P}\left[Pf - P_n f >\right.$

$\sqrt{\frac{\log\frac{1}{\delta(f)}}{2n}}\Bigg]$. Hence, if we have a countable set $\Im$, the union bound immediately yields

$\mathbb{P}\left[\exists\, f \in \Im: Pf - P_n f > \sqrt{\frac{\log\frac{1}{\delta(f)}}{2n}}\right] \leq \sum_{f\in\Im} \delta(f)$.

7. Countably Infinite Case: Choosing $\delta(f) = \delta p(f)$, $\sum_{f\in\Im} p(f) = 1$, this makes the RHS above equal to δ, and we get the following result; with probability at least $1-\delta$, $\forall f \in \Im$: $Pf - P_n f < \sqrt{\frac{\log\frac{1}{p(f)} + \log\frac{1}{\delta}}{2n}}$. We notice that if $\Im$ is finite in N, taking a uniform p results in $\log N$ as before.
8. ERM Consistency Bounds: Applying the bounds separately to each counterpart above, we see that $|R(f_n) - R(f_\Im)| \leq 2 \sup_{f\in\Im} \left|R_{Emp}(f) - R(f)\right|$. Expressing this in terms of probabilities using LLN produces the following: $P[|R(f_n) - R(f_\Im)| \geq \varepsilon] \leq P\left[\sup_{f\in\Im} \left|R_{Emp}(f) - R(f)\right| \geq \frac{\varepsilon}{2}\right]$.
9. Necessary and Sufficient Condition for ERM Convergence: Vapnik and Chervonenkis (1971), Devroye, Gyorfi, and Lugosi (1996), and Mendelson (2003) contain the details. $P\left[\left|R(f) - R_{Emp}(f)\right| \geq \varepsilon\right] \to 0$ as $n \to \infty$ for all $\varepsilon > 0$ is a necessary and sufficient condition for the consistency of ERM with respect to $\Im$.

Shattering Coefficient

1. Motivation: The infinite function space case may now be reduced to $N \leq 2^{2n}$ (over the original and the ghost) subset. Of course, in many classifiers, the full spectrum 2^{2n} will never be produced, so N is used as the corresponding outcome "capacity measure", i.e., the outcome spectrum of the hypothesis set.
2. Definition: Let $\mathbb{Z}_n \coloneqq [(X_1, Y_1), \ldots, (X_n, Y_n)]$ be a given sample set. Denote by $\left|\Im_{\mathbb{Z}_n}\right|$ the cardinality of $\Im$ when restricted to $\{x_1, \ldots, x_n\}$, i.e., the number of f in $\Im$ that can be

distinguished by their values on $\{x_1, \dots, x_n\}$. The shattering coefficient $\mathcal{N}(\mathfrak{I}, n)$ is defined as $\mathcal{N}(\mathfrak{I}, n) = max\left\{\left|\mathfrak{I}_{\mathbb{Z}_n}\right| : x_i \in \{x_1, \dots, x_n\}\right\}$.

3. Intuition: Shattering coefficient is simply a count of the number of ways the function space can classify the input pattern set. When $\mathcal{N}(\mathfrak{I}, n) = 2^k$, this means that there a sample of size k on which all possible separations can be achieved. Then $\mathfrak{I}$ is said to shatter $\{x_1, \dots, x_n\}$ into k points.
4. Continuous $\mathfrak{I}$ Supremum Bound: It is easy to see that $\mathbb{P}\left[\sup_{f \in \mathfrak{I}} \left|R(f) - R_{Emp}(f)\right| \geq \varepsilon\right] \leq 2\mathcal{N}(\mathfrak{I}, 2n)e^{-\frac{n\varepsilon^2}{4}}$. Convergence therefore depends on the growth of $2\mathcal{N}(\mathfrak{I}, 2n)$ with n.
5. Polynomial Growth of the Shattering Coefficient: If $\mathcal{N}(\mathfrak{I}, 2n) \leq (2n)^k$, then $2\mathcal{N}(\mathfrak{I}, 2n)e^{-\frac{n\varepsilon^2}{4}} = 2e^{klog(2n)-\frac{n\varepsilon^2}{4}}$. Thus $2\mathcal{N}(\mathfrak{I}, 2n)e^{-\frac{n\varepsilon^2}{4}} \to 0$ as $n \to \infty$.
6. Shattering Coefficient growth as 2^{2n}: In this case, $2\mathcal{N}(\mathfrak{I}, 2n)e^{-\frac{n\varepsilon^2}{4}} = 2e^{n\left[2log2-\frac{\varepsilon^2}{4}\right]}$, 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.
7. 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 $\frac{\log \mathcal{N}(\mathfrak{I}, 2n)}{n} \to 0$.

Empirical Risk Generalization Bound

1. Probabilistic Bounds: Expressing the probability bound the other way, with a probability of at least $1 - \delta$, $f \in \mathfrak{I}$ satisfies $R(f) \leq R_{Emp}(f) + \sqrt{\frac{4}{n}[\log \mathcal{N}(\mathfrak{I}, 2n) - \log \delta]}$. For polynomial growth functions, $\sqrt{\frac{\log \mathcal{N}(\mathfrak{I}, 2n)}{n}} \to 0$ as $n \to \infty$, therefore it converges. However, for $\mathcal{N}(\mathfrak{I}, 2n) \sim 2^{2n}$, $\sqrt{\frac{\log \mathcal{N}(\mathfrak{I}, 2n)}{n}} \to 2$, therefore that DOES NOT converge.

2. Union Bound Contribution - Inversion: To recap, for a specific f, with probability at least $1-\delta$, $R(f) \leq R_n(f) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$. For all $f \in \mathfrak{I}$, with probability of at least $1-\delta$, $R(f) \leq R_n(f) + \sqrt{\frac{\log N + \log\frac{2}{\delta}}{2n}}$. The extra $\log N$ term contribution from the union bound may also be interpreted as the number of additional bits needed to encode the hypothesis space $\mathfrak{I}$. 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)).
3. 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 ε, and in the way they measure capacity (Devroye, Gyorfi, and Lugosi (1996), Vapnik (1998)).
4. Improvements over Classical Bounds: There are several things that can be improved:
    a. Hoeffding's inequality only uses the boundedness of the functions, not their variance.
    b. The union bound is as bad as if all the functions in the class were independent (i.e., if $f_1(\mathbb{Z})$ and $f_2(\mathbb{Z})$ were independent).
    c. The supremum over $\mathfrak{I}$ of $R(f) - R_n(f)$ is not necessarily what the algorithm would choose, so that upper bounding $R(f_n) - R_n(f_n)$ 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 $\mathbb{R}^2$ data space into separate classes using a straight line. A generalization would be linear classifiers in $\mathbb{R}^d$.
2. Setup: Given a set of training points and a classifier f_n that would perfectly separate them, we define the *Margin of the Classifier* as the smallest distance of any training point to the separating line f_n. A similar margin can be defined in $\mathbb{R}^d$ for an arbitrary dimension d.
3. VC Dimension: The VC dimension of a class $\mathfrak{I}_\rho$ of linear classifiers with all having a margin of at least ρ can be essentially bounded by the ratio of the radius R of the smallest sphere

enclosing the data points with the margin ρ, that is, $\mathcal{VC}(\Im_\rho) \leq min\left\{d, \frac{4R^2}{\rho^2}\right\} + 1$ (Vapnik (1995)).

4. Capacity: Thus, larger the ρ, 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)).
5. $\mathbb{R}^d$ Definition: Assume that the data lies in a ball of radius R in $\mathbb{R}^d$. Consider the set $\Im_\rho$ of linear classifiers with a margin of at least ρ. Assume that we are given n training examples. Use $\nu(f)$ to denote the fraction of the training examples with margin smaller than ρ, or which are wrongly classified by a classifier $f \in \Im$. Then, with probability of at least $1 - \delta$, the true error of any $f \in \Im_\rho$ can be bounded by $R(f) \leq \nu(f) + \sqrt{\frac{c}{n}\left[\frac{R^2}{\rho^2}(\log n)^2 + \log\frac{1}{\delta}\right]}$ 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.

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:
    a. 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?
    b. Non-asymptotic theory of the rate of convergence of a learning process => How fast is the rate of convergence of the learning process?
    c. 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?
    d. Theory of construction learning machines => How can one construct the algorithms that control the generalization ability?

Empirical Processes

1. Background: Let $X_1, \dots, X_n$ be the random elements defined on a measurable space (X, A). Define the empirical measure $\mathbb{P} = \frac{1}{n}\sum_{i=1}^{n} \delta(X_i)$ where δ stands for the Dirac notation. Further, using the notation of van der Vaart and Wellner (2000), denote $\mathbb{Q}f = \int f d\mathbb{Q}$ for any probability measure $\mathbb{Q}$.
2. The Empirical Measure Map: Let $\mathfrak{I}$ be a class of measurable functions $f: X \to \mathbb{R}$. The empirical measure above induces a map from $\mathfrak{I}$ to $\mathbb{R}$ given by $f \to \mathbb{P}_n f$. Notice the

similarity between the success counting measure $\mathbb{P}_n$ and the $0-1$ empirical loss function.

3. Empirical Process Theory: Let $\|\mathbb{Q}\|_{\mathfrak{F}} = sup\{|\mathbb{Q}f|: f \epsilon \mathfrak{F}\}$ where sup is the supremum operator on the set, representing the least upper bound of $|\mathbb{Q}f|$. Empirical process theory aims at identifying the classes $\mathfrak{F}$ for which the following statements hold. In both cases we assume that the underlying distribution of the data $\mathbb{P}$ is unknown in practice.
    a. $\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{F}} \to 0$, i.e., this is the uniform law of large numbers
    b. $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G} \in l^{\infty}(\mathfrak{F})$, i.e., this is the generalized uniform central theorem limit.
4. Glivenko-Cantelli and $\mathbb{P}$-Donsker Classes: If the law of large numbers can be explicitly established across all elements of $\mathfrak{F}$, the class $\mathfrak{F}$ is called Glivenko-Cantelli. If the validity of the central limit theorem (under the assumption $\sup_{f \in \mathfrak{F}} |f(x) - \mathbb{P}f| < \infty \; \forall \; x$) can be established, then the class $\mathfrak{F}$ is called $\mathbb{P}$-Donsker. Obviously a $\mathbb{P}$-Donsker class is Glivenko-Cantelli in probability by the application of the Slutsky's theorem.
5. Determination of $\mathfrak{F}$: The main challenge here is to determine $\mathfrak{F}$ that satisfies the LLN and the CLT criteria bove (under regularity conditions) for all $f \in \mathfrak{F}$. Intuitively, $\mathfrak{F}$ cannot be too large, and the geometry of $\mathfrak{F}$ will play an important role.
6. Size of $\mathfrak{F}$: One way of measuring how big $\mathfrak{F}$ is is by using covering numbers. The covering number $N(\varepsilon, \mathfrak{F}, \|.\|)$ is the minimum number of balls $\{g : \|g - f\| < \varepsilon\}$ needed to cover $\mathfrak{F}$ fully (obviously assuming there is an underlying norm on $\mathfrak{F}$). The entropy number is the logarithm of the covering number.
7. Sufficiency conditions for $\mathfrak{F}$ using Glivenko-Cantelli: A class $\mathfrak{F}$ is Glivenko-Cantelli if it is $\mathbb{P}$-measurable inside of an envelope $\mathbb{F}$ such that $\mathbb{P}^*\mathbb{F} < \infty$ and satisfies $\sup_{\mathbb{Q}} \left[\varepsilon \|\mathbb{F}\|_{\mathbb{Q}}, \mathfrak{F}, L_1(\mathbb{Q})\right] < \infty$ for every $\varepsilon > 0$.
8. Sufficiency Conditions for $\mathfrak{F}$ being $\mathbb{P}$-Donsker: This criterion is a version of the Dudley's theorem. If $\mathfrak{F}$ belongs to the class of functions such that $\int_0^{\infty} \sup_{\mathbb{Q}} \sqrt{logN\left[\varepsilon \sqrt{\int |\mathbb{F}|^2 d\mathbb{P}}, \mathfrak{F}, L_2(\mathbb{Q})\right]}\, d\varepsilon < \infty$, then $\mathfrak{F}$ is $\mathbb{P}$-Donsker for every probability measure $\mathbb{P}$ such that $\mathbb{P}^*\mathbb{F}^2 < \infty$.

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 $(\mathbb{P}_n - \mathbb{P})f = \frac{1}{n}\sum_{i=1}^{n}[f(X_i) - \mathbb{P}f]$. Here we establish the connection between the empirical processes above and its symmetrized equivalent $\mathbb{P}_n^0 f = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)$ where ε_i is an independent random variable that can be ± 1 with a probability 0.5 each. This symmetrized process, therefore, is a Rademacher process, conditional on the data X_i. Therefore this process is a sub-Gaussian process by the Hoeffding's inequality.
3. The Symmetrization Inequality: The Symmetrization bounds the LLN criterion of the empirical process theory. For every non-decreasing convex $\Phi : \mathbb{R} \to \mathbb{R}$ and a class of measurable functions $\Im$, $E\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\Im}) \leq E\Phi(2\|\mathbb{P}_n^0\|_{\Im})$.
4. Conceptual idea behind the proof by Symmetrization: The proof of the symmetrization lemma lies on introducing independent copies of the original variables X_i (sometimes referred to as the ghost sample) and replacing the innder expectation of $\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\Im})$ with these copies. After an application of Jensen's inequality, different signs could be introduced (hence the name symmetrization) without changing the expectation.
5. Proof Steps:
    a. Introduction of the Ghost Sample => The empirical measure $\mathbb{P}f$ is replaced by $Ef(Y_i)$ – thereby providing the motivation for introducing the ghost sample variables $Y_1, \ldots, Y_n$ as independent copies of $X_1, \ldots, X_n$. For fixed values of $X_1, \ldots, X_n$ one has $\|\mathbb{P}_n - \mathbb{P}\|_{\Im} = \sup_{f \in \Im} \frac{1}{n}|\sum_{i=1}^{n}[f(X_i) - Ef(Y_i)]| \leq E_Y \sup_{f \in \Im} \frac{1}{n}|\sum_{i=1}^{n}[f(X_i) - f(Y_i)]|$.

b. Apply Jensen's Inequality => Apply the convex operator Φ to both sides: $\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{I}}) \leq E_Y \Phi\left\{\frac{1}{n}\|\sum_{i=1}^{n}[f(X_i) - f(Y_i)]\|_{\mathfrak{I}}\right\}$. We are able to switch the locations of E_Y and Φ, because Φ is convex, and we have applied Jensen's inequality to convert this to an inequality.

c. Take expectation with respect to X: $E\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{I}}) \leq E_X E_Y \Phi\left\{\frac{1}{n}\|\sum_{i=1}^{n}[f(X_i) - f(Y_i)]\|_{\mathfrak{I}}\right\}$. Given that only X is the variate on the RHS, the expectation simply reduces to E_X on RHS (and E on LHS).

d. Sign Perturbation Step => Note that adding a minus sign ahead of $[f(X_i) - f(Y_i)]$ does not alter the RHS, because it is a symmetric function of and . Thus the RHS remains the same under sign perturbation: $E_X E_Y \Phi\left\{\frac{1}{n}\|\sum_{i=1}^{n} e_i[f(X_i) - f(Y_i)]\|_{\mathfrak{I}}\right\}$ for a random Rademacher sequence e_i given as $(e_1, e_2, \ldots, e_n) \in \{-1, +1\}$. Therefore $E\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{I}}) \leq E_\varepsilon E\Phi\left\{\frac{1}{n}\|\sum_{i=1}^{n} \varepsilon_i[f(X_i) - f(Y_i)]\|_{\mathfrak{I}}\right\}$.

e. Triangle Inequality and Convexity of Φ => $E\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{I}}) \leq \frac{1}{2}E_\varepsilon E\Phi\left\{2\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathfrak{I}}\right\} + \frac{1}{2}E_\varepsilon E\Phi\left\{2\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right\|_{\mathfrak{I}}\right\}$. Since the 2 right hand terms are equal, we get $E\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{I}}) \leq E_\varepsilon E\Phi\left\{2\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathfrak{I}}\right\} \leq E\Phi(2\|\mathbb{P}_n^0\|_{\mathfrak{I}})$.

6. Proving Empirical CLT's Using Symmetrization: Proof for the CLT's proceed analogously to the LNN. First use symmetrization to pass the process to $\mathbb{P}_n^0$, 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 $\mathfrak{I}$, $\mathcal{VC}(\mathfrak{I})$, is defined as the largest number n such that there exists a sample of size n which is shattered by $\mathfrak{I}$; i.e., $\mathcal{VC}(\mathfrak{I}) = max\{n \in \mathbb{N} : \left|\mathfrak{I}_{\mathbb{Z}_n}\right| = 2^n \; for \; some \; \mathbb{Z}_n\}$. 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 C be a concept class over an infinite space X, i.e., a set of functions from X to $\{0,1\}$ (both X and C may be infinite). For any $S \subseteq X$, $C(S)$ denotes the set of all labels or dichotomies on S that are induced or realized by C, i.e., if $S = \{x_1, \dots, x_m\}$, then $C(S) \subseteq \{0,1\}^m$, and $C(S) = \{h(x_1), \dots, h(x_m) | c \in C\}$. For any natural number m, we consider $C\lceil m \rceil$ to be the number of ways to split m points using concepts in C, that is $C\lceil m \rceil = max\{|C(S)|;\ |S| = m;\ S \subseteq X\}$.
2. Shattering of C by S: If $|C(S)| = 2^{|S|}$, then S is said to be *shattered* by C.
3. VC Dimension of C: The *Vapnik-Chervonenkis* dimension of C, denoted as $\mathcal{VCDim}(C)$, is the cardinality of the largest set S shattered by C. If arbitrarily large finite sets can be shattered by C, then $\mathcal{VCDim}(C) = \infty$.

VC Dimension Examples

1. Computing the VC Dimension: In order to show that the VC Dimension of a class is at least d, we must find some shattered set of size d. In order to show that the VC Dimension is at most d, we must show that no set of size $d + 1$ is shattered.
2. VC Dimension Examples:
    a. Thresholds on a Number Line => Let C 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 class 2 can be shattered, since it is impossible to choose a threshold such that x_1 is labeled positive and x_2 is labeled negative. Hence $\mathcal{VCDim}(C) = 1$.
    b. Intervals on a Real Line => Here a sample of size 2 is shattered, nut no sample of size 3 is shattered, since no concept can satisfy a sample whose middle point is negative and the outer points are positive. Hence the $\mathcal{VCDim}(C) = 2$.
    c. k Non-intersecting Intervals => Let C be the concept class of k non-intersecting intervals on the real line. A sample of size $2k$ shatters (simply treating each pair of points as a separate case of the previous example), but no sample of size $2k + 1$ shatters, since if the sample points are alternated negative/positive, starting with a positive point, the positive points cannot all be covered by only k intervals. Hence $\mathcal{VCDim}(C) = 2k$.
    d. Linear Separators in $\mathbb{R}^2$ => Here 3 points can be shattered, but not 4, so $\mathcal{VCDim}(C) = 3$. In general, one can show that the $\mathcal{VCDim}(C)$ of the class of linear separators in $\mathbb{R}^n$ is $n + 1$.
    e. Axis-aligned Rectangles => The class of axis-aligned rectangles in a plane has $\mathcal{VCDim}(C) = 4$. 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 $\mathbb{R}$, Vapnik (1995)).
2. Popper's Dimension of a Theory: As stated in Popper (1959), if there exists, for a theory t, a field of singular (but not necessarily basic) statements such that, for some number d, the theory cannot be falsified for any d-tuple of the field, although it can be falsified by certain $(d+1)$-tuples, then we call d 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 d, 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 $\mathbb{R}^k$ do not cut all sub-sets of $k+2$ 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 **Association of 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*.
- 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 $\Im$ be a function class with a finite VC dimension d. Then $\mathcal{N}(\Im, n) \leq \sum_{i=0}^{d} \binom{n}{i}$ for all $n \epsilon \mathbb{N}$. In particular, for all $n \geq d$, we have $\mathcal{N}(\Im, n) \leq \left(\frac{en}{d}\right)^{d}$.
3. Nature of $\mathcal{N}(\Im, n)$: $\mathcal{N}(\Im, n)$ 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 n, and the VC dimension.
4. Implication of Sauer's Lemma: Thus, for $n \geq d$, the shattering coefficient behaves like a polynomial in the sample size n. Once we know that the VC dimension is finite, the shattering coefficients grow polynomially, and this implies ERM consistency. The converse is true too.
5. ERM Consistency Statement: ERM is consistent with respect to $\Im$ if and only if $\mathcal{VC}(\Im)$ is finite.

Derivation of Sauer Lemma Bounds

1. Statement: If $\mathcal{VCDim}(C) = d$, the for all m, $C[m] \leq \Phi_d(m)$, where $\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}$.

2. Shattering Cardinality Polynomial Limit: We estimate $\Phi_d(m)$ in the limit of d. Note that $m > d$, so $0 \leq \frac{d}{m} < 1$. Thus $\left(\frac{d}{m}\right)^d \sum_{i=0}^{d}\binom{m}{i} \leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{d}{m}\right)^i \leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^i = \left[1+\frac{d}{m}\right]^m \leq e^d$. Thus, for $m > d$ we have $\Phi_d(m) \leq \left(\frac{em}{d}\right)^d$.
3. General Sauer Complexity Results: Sauer's lemma can be used to get closed form expressions on sample complexity. For the treatment below, C is an arbitrary hypothesis space, D an arbitrary unknown fixed probability distribution, and c^* an arbitrary unknown target function.
4. Shattering Cardinality Bound - Recap: For any $\varepsilon, \delta > 0$, if we draw a sample S from D of size m satisfying $2C[2m]2^{-\frac{m\varepsilon}{2}} \leq \delta$, then with probability of at least $1-\delta$, all hypothesis in C with $err_D(h) > \varepsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.
5. Sample Size Bound: For any $\varepsilon, \delta > 0$, if we draw a sample S from D of size m satisfying $m \geq \frac{8}{\varepsilon}\left[d\log\frac{16}{\varepsilon} + \log\frac{2}{\delta}\right]$, then with probability of at least $1-\delta$, all hypothesis in C with $err_D(h) > \varepsilon$ are inconsistent with the data, i.e., $err_S(h) \neq 0$.
    a. Step #1 => From the shatter cardinality bound we've just seen, we get $2C[2m]2^{-\frac{m\varepsilon}{2}} \leq \delta \rightarrow 2^{\frac{m\varepsilon}{2}} \geq \frac{2C[2m]}{\delta} \rightarrow m \geq \frac{4}{\varepsilon}\log\frac{2C[2m]}{\delta}$.
    b. Step #2 => Applying Sauer's lemma, and recognizing that $C[2m] \leq \left(\frac{em}{d}\right)^d$, we can reduce the inequality for m to $m \geq \frac{4}{\varepsilon}\left[d\log\frac{2me}{d} + \log\frac{2}{\delta}\right] = \frac{4}{\varepsilon}\left[d\log m + d\log\frac{2e}{d} + \log\frac{2}{\delta}\right]$.
    c. Step #3 => We use Taylor expansion to estimate the bound $\log m$. For $\alpha, x > 0$, we get $\log x \leq \alpha x - \log\alpha - 1$, so $\frac{4d}{\varepsilon}\log m \leq \frac{4d}{\varepsilon}\left[\frac{\varepsilon}{8d}m + \log\frac{8d}{\varepsilon} - 1\right] = \frac{m}{2} + \frac{4d}{\varepsilon}\log\frac{8d}{e\varepsilon}$.
    d. Step #4 => Re-cast the above bounding for $\log m$ to get $m \geq \frac{m}{2} + \frac{4d}{\varepsilon}\log\frac{8d}{e\varepsilon} + \frac{4d}{\varepsilon}\log\frac{2e}{d} + \frac{4}{\varepsilon}\log\frac{2}{\delta} \rightarrow m \geq \frac{8}{\varepsilon}\left[d\log\frac{16}{\varepsilon} + \log\frac{2}{\delta}\right]$.

Sauer Lemma ERM Bounds

1. Growth Function - Definition: The growth function is the maximum number of ways into which n points can be classified by the function class $\Im$: $S_{\Im}(n) = \underset{z_1, \ldots, z_n}{sup} \left|\Im_{z_1,\ldots,z_n}\right|$.
2. Risk Bound Using Growth Function: For any δ, with probability at least $1 - \delta$, $R(f) - R_n(f) \leq 2\sqrt{2\frac{\log S_{\Im}(2n) + \log\frac{2}{\delta}}{n}}$.
3. Bounding with Sauer's Lemma: Let $\Im$ be a class of functions with a finite VC dimension h. Then, for all $n \in \mathbb{N}$, $S_{\Im}(n) \leq \sum_{i=0}^{h}\binom{n}{i}$, and for all $n \geq h$, $S_{\Im}(n) \leq \left(\frac{en}{h}\right)^h$. This implies, that with probability at least $1 - \delta$, $\forall f \in \Im\ R(f) - R_n(f) \leq 2\sqrt{2\frac{h\log\frac{en}{h} + \log\frac{2}{\delta}}{n}}$. Thus, this indicates that the difference between the TRUE and the EMPIRICAL risk is at most of the order of $\sqrt{\frac{h\log n}{n}}$.
4. Behavior of the Growth Function: The growth function, which is exponential in the sample size n as long as $n \leq h$, where h is the VC Dimension, becomes polynomial in n for $n > h$.

VC Index

1. Motivation and Setup: Consider a collection $\mathbb{C}$ of subsets of the sample space χ. The collection of sets $\mathbb{C}$ is said to pick out a certain subset of finite set $S = \{x_1, \ldots, x_n\} \in \chi$ if $S = S \cap C$ from $C \in \mathbb{C}$. $\mathbb{C}$ is said to shatter S if it picks out each of its 2^n subsets. The VC-index $V(\mathbb{C})$ of $\mathbb{C}$ is the smallest number n for which NO SET of size n is shattered by $\mathbb{C}$ (in fact, the VC-index is similar the $VC_{Dimension} + 1$ for an appropriately chosen classifier set).
2. Bounds on the VC Class Subset Number: Sauer's lemma then states that the number $\Delta_n(\mathbb{C}, x_1, \ldots, x_n)$ of subsets picked out by a VC class $\mathbb{C}$ satisfies $\underset{x_1, \ldots, x_n}{max} \Delta_n(\mathbb{C}, x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathbb{C})-1}\binom{n}{j} \leq \left(\frac{ne}{V(\mathbb{C})-1}\right)^{V(\mathbb{C})-1}$. This is polynomial in the number of subsets (i.e.,$n^{V(\mathbb{C})-1}$) rather than exponential. Intuitively, this means that finite VC index implies that $\mathbb{C}$ has an apparent simplistic structure.

3. 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 $f: X \to \mathbb{R}$, the sub-graph is a subset of $\chi \times \mathbb{R}$ such that $\{(x, t): t < f(x)\}$. A collection of is called a VC sub-graph class if all the sub-graphs form a VC-class.
4. Covering Number for the $0-1$ Loss Function: Consider a set of indicator functions $I_C = \{1_C: C \in \mathbb{C}\}$ in $L_1(\mathbb{Q})$ for the discrete empirical measure $\mathbb{Q}$ (or any type of probability measure $\mathbb{Q}$ - discrete or not). It can be shown that for any ≥ 1 $N\left(\varepsilon, I_C, L_r(\mathbb{Q})\right) \leq k\, V(\mathbb{C})(4e)^{V(\mathbb{C})}\left(\frac{1}{\varepsilon}\right)^{r(V(\mathbb{C})-1)}$.
5. Entropy Number for the Symmetric Convex Hypotheses Hull: Next consider the symmetric convex hull of a set $\Im_{CONVEX}$, which is a collection of functions of the form $\sum_{i=1}^{m} \alpha_i f_i$ with $\sum_{i=1}^{m} |\alpha_i| \leq 1$. Then, if $N\left[\varepsilon \sqrt{\int |\mathbb{F}|^2 d\mathbb{P}}, \Im_{CONVEX}, L_2(\mathbb{Q})\right] < C\left(\frac{1}{\varepsilon}\right)^{\frac{2V}{V+2}}$.
6. Convergence of the Convex Hull: The most important consequence of the above is that the power of $\frac{1}{\varepsilon}$ - $\frac{2V}{V+2}$ is strictly less than 2, which is just enough so that entropy integral is guaranteed to converge, therefore the class $\Im_{CONVEX}$ is going to be $\mathbb{P}$-Donsker.
7. VC Index of a Sub-graph Class: Any finite dimensional vector space $\Im$ of measurable functions $f: X \to \mathbb{R}$ is a VC-subgraph of index smaller than or equal to $dim(\Im) + 2$.
8. 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 $\Upsilon = \{0, 1\}$. The function $f: X \to \Upsilon$ is called the classifier. Similar to before, we define the shattering coefficient (also referred to as growth coefficient) as $S(\Im, \mathrm{n}) = \underset{x_1, \ldots, x_n}{max} |\{f(x_1), f(x_2), \ldots, f(x_n)\}, f \in \Im|$
2. Shattering Coefficient Relation: In terms of the previous section the shattering coefficient is precisely $\underset{x_1, \ldots, x_n}{max} \Delta_n(\mathbb{C}, x_1, \ldots, x_n)$ with $\mathbb{C}$ being a collection of all the sets as laid out above.

Further, using Sauer's lemma, it can be shown that $S(\mathfrak{I}, \mathrm{n})$ is going to be a polynomial in n provided that class $\mathfrak{I}$ has a finite VC dimension, or equivalently, the collection $\mathbb{C}$ has a finite VC index.

3. VC Indicator Empirical Risk: Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be the data set. As always, we assume that P_{XY} is unknown, thus we work on bounding the $0-1$ indicator empirical risk: $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} I[f(X_n) \neq Y_n]$.
4. VC Inequality Bounds: For binary classification and $0-1$ loss function, we have the following generalization bounds:
   a. $\mathbb{P}\left(\sup_{f \in \mathfrak{I}} \left|\hat{R}_n(f) - R(f)\right| > \varepsilon\right) \leq 8\, S(\mathfrak{I}, \mathrm{n}) e^{-\frac{n\varepsilon^2}{32}}$
   b. $E\left(\sup_{f \in \mathfrak{I}} \left|\hat{R}_n(f) - R(f)\right|\right) \leq 2\sqrt{\frac{\log S(\mathfrak{I}, \mathrm{n}) + \log 2}{n}}$
5. Implication of the VC Inequality Bounds: The impact of the VC inequality relation is that, as the sample size increases, provided $\mathfrak{I}$ has a finite VC dimension, the empirical $0-1$ risk becomes a good proxy for the expected $0-1$ risk. Note that both RHS of the two inequalities will converge to 0 provided $S(\mathfrak{I}, \mathrm{n})$ grows polynomially in n.
6. Connection between the Classifier Framework and the Empirical Process Framework: With the empirical classifier framework, on deals with the modified empirical process $\left|\hat{R}_n(f) - R(f)\right|_{\mathfrak{I}}$, 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.

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 Z and a fixed distribution $D_{|z}$, let $S = \{z_1, \dots, z_m\}$ be a set of samples drawn from . Let $f \in \Im$ be a class of functions $f: Z \to \mathbb{R}$.
3. Rademacher Complexity - Definition: The *Empirical Rademacher Complexity* of $\Im$ is defined to be $\hat{\mathcal{R}}_m(\Im) = E_\sigma \left[\sup_{f \in \Im} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$ where $\sigma_1, \dots, \sigma_m$ are independent random variables chosen from $\{-1, +1\}$. We refer to these as Rademacher variables. The *Rademacher Complexity* is defined as $\mathcal{R}_m(\Im) = E_D\left[\hat{\mathcal{R}}_m(\Im)\right]$.
4. Empirical Rademacher Complexity - Intuition: The sup intuitively measures, for a given set S and the Rademacher vector $\sigma_1, \dots, \sigma_m$, the maximum correlation between $f(z_i)$ and σ_i over all $f \in \Im$. The expectation over σ, i.e., the empirical Rademacher complexity of $\Im$ measures the ability of functions from $\Im$ (when applied to the fixed set S) to fit random noise.
5. Rademacher Complexity – Intuition: The Rademacher complexity of $\Im$ measures the expected noise-fitting-ability of $\Im$ over all the data sets $S \epsilon Z^m$ that could be drawn according to the distribution $D_{|z}$.
6. Sample Space Generalization: The Rademacher complexity can be defined even more generally on sets $A \subseteq \mathbb{R}^m$ by making the sup over $a \in A$ (in place of $f \in \Im$), and replacing each $f(z_i)$ with a_i. Of course, setting $A \equiv \Im(S) = \{f(z) | \, f \in \Im, z \in S\}$ 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:
    a. Use *concentration* to relate $\sup_{f \in \mathcal{F}} (P - P_n)f$ to its expectation.
    b. Use *symmetrization* to relate the expectation to the Rademacher average.
    c. Use *concentration* again to relate the unconditional Rademacher average to its conditional one.
3. Symbology and Steps: Denote the empirical average over a sample S as $\hat{E}_s[f(z)] = \frac{1}{|S|}\sum_{z \in S} f(z)$. Thus for a function f, the definition of sup leads to $E_D[f(z)] - \hat{E}_s[f(z)] \leq \sup_{h \in \mathfrak{J}}\{E_D[f(z)] - \hat{E}_s[f(z)]\}$. We set $\varphi(S) = \sup_{h \in \mathfrak{J}}\{E_D[f(z)] - \hat{E}_s[f(z)]\}$, and try to bound it by using the McDiarmid inequality.
    a. McDiarmid Bounded Differences Inequality: Let $x_1, \ldots, x_n$ be independent random variables taking on values in set A, and let $c_1, \ldots, c_n$ be positive real constants. If $\varphi: A^n \to \mathbb{R}$ satisfies $\sup_{x_1, \ldots, x_n, x'_i \in \mathfrak{J}} |\varphi(x_1, \ldots, x_i, \ldots, x_n) - \varphi(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i$, then, for $1 \leq i \leq n$, $Prob\{\varphi(x_1, \ldots, x_i, \ldots, x_n) - E[\varphi(x_1, \ldots, x_i, \ldots, x_n)] \geq \varepsilon\} \geq e^{-\frac{2\varepsilon^2}{\sum_{i=1}^{n} {c_i}^2}}$.
4. Supremum Bounds: $\varphi(S)$ as defined above satisfies $\sup_{z_1, \ldots, z_m, z'_i \in \mathfrak{J}} |\varphi(z_1, \ldots, z'_i, \ldots, z_m) - \varphi(z_1, \ldots, z'_i, \ldots, z_m)| \leq \frac{1}{m}$ where m is the number of data points.
    a. Proof Setup => Let $S = \{z_1, \ldots, z_i, \ldots, z_m\}$ and $S' = \{z_1, \ldots, z'_i, \ldots, z_m\}$. We then define $|\varphi(S) - \varphi(S')| = \left| \sup_{h \in \mathfrak{J}}\{E_D[f(z)] - \hat{E}_s[f(z)]\} - \sup_{h \in \mathfrak{J}}\{E_D[f(z)] - \hat{E}_{s'}[f(z)]\} \right|$.

b. S Optimizer Function => Letting $h^* \in \mathfrak{J}$ be the maximizing function for the supremum in $\varphi(S)$, $|\varphi(S) - \varphi(S')| \leq \left|E_D[h^*(z)] - \hat{E}_s[h^*(z)] - \underset{h \in \mathfrak{J}}{sup}\{E_D[f(z)] - \hat{E}_{s'}[f(z)]\}\right|$.

c. S' Optimizer Function => By definition of the supremum, h^* can also, at best, maximize $\varphi(S')$, so $E_D[h^*(z)] - \hat{E}_{s'}[h^*(z)] \leq \underset{h \in \mathfrak{J}}{sup}\{E_D[f(z)] - \hat{E}_{s'}[f(z)]\}$.

d. h Value Realization => Using the above 2 inequalities, and noting the negative sign ahead of $\varphi(S')$, $|\varphi(S) - \varphi(S')| \leq \left|E_D[h^*(z)] - \hat{E}_s[h^*(z)] - E_D[h^*(z)] + \hat{E}_{s'}[h^*(z)]\right| = \left|\hat{E}_{s'}[h^*(z)] - \hat{E}_s[h^*(z)]\right| = \frac{1}{m}|\sum_{z \in S} h^*(z) - \sum_{z \in S'} h^*(z)|$

e. Bounds Estimator => Since S and S' differ in only element j, the above becomes $|\varphi(S) - \varphi(S')| \leq \frac{1}{m}\left|h^*(z_j) - h^*(z'_j)\right| \leq \frac{1}{m}$ where the ;last stpe follows from the fact that $h^*: \mathbb{Z} \to [a, a+1]$, so $\underset{z_j, z'_j \in \mathbb{Z}}{sup}\left|h^*(z_j) - h^*(z'_j)\right| = 1$.

5. <u>McDiarmid Bounds on the Inequality $\varphi(S)$</u>: Using the bound $|\varphi(S) - \varphi(S')| \leq \frac{1}{m}$, we apply McDiarmid's inequality to get $\mathbb{P}\{\varphi(S) - E[\varphi(S')] \geq t\} \geq e^{-\frac{t^2}{\sum_{i=1}^{m}\left(\frac{1}{m}\right)^2}} = e^{-mt^2}$. Setting this probability to be less than δ and solving for t, we get $t \geq \sqrt{\frac{\log\frac{1}{\delta}}{m}}$.

6. <u>First ERM Bounds</u>: Thus, with a probability of at least $1 - \delta$, $E_D[h(z)] - \hat{E}_s[h(z)] \leq E_S\left\{\underset{h \in \mathfrak{J}}{sup}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\} + \sqrt{\frac{\log\frac{1}{\delta}}{m}}$. 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.

   a. 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 S (plus a bound that depends on the probability), i.e., $E_D[h(z)] - \hat{E}_S[h(z)] \leq E_S\left\{\underset{h \in \mathfrak{J}}{sup}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\} \to E_D[h(z)] - \hat{E}_s[h(z)] \leq E_S\left\{\underset{h \in \mathfrak{J}}{sup}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\} + \sqrt{\frac{\log\frac{1}{\delta}}{m}}$

7. Estimation of the Sample Bound for the Supremum: To estimate $E_S\left\{\sup_{h \in \mathfrak{J}}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\}$, we do the following steps:
    a. Draw a "ghost sample" $\tilde{S} = \{\tilde{z}_1, \ldots, \tilde{z}_m\}$ that is i.i.d. and independent of $S = \{z_1, \ldots, z_m\}$.
    b. Get $E_D[h(z)]$ as an expectation over these ghost samples.
    c. Observe that $E_{\tilde{S}}\left[\hat{E}_{\tilde{S}}[h(z)]|S\right] = E_D[h(z)]$ and that $E_{\tilde{S}}\left[\hat{E}_S[h(z)]|S\right] = \hat{E}_S[h(z)]$.
8. Sample Expectation of the Supremum: $E_S\left\{\sup_{h \in \mathfrak{J}}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\} = E_S\left\{\sup_{h \in \mathfrak{J}} E_{\tilde{S}}\left[\hat{E}_{\tilde{S}}[h(z)] - \hat{E}_S[h(z)]|S\right]\right\} = E_S\left\{\sup_{h \in \mathfrak{J}} E_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^{m}[h(\tilde{z}_i) - h(z_i)|S]\right]\right\}$
9. 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: $E_S\left\{\sup_{h \in \mathfrak{J}} E_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^{m}[h(\tilde{z}_i) - h(z_i)|S]\right]\right\} \leq E_{S,\tilde{S}}\left\{\sup_{h \in \mathfrak{J}}\left[\frac{1}{m}\sum_{i=1}^{m}[h(\tilde{z}_i) - h(z_i)|S]\right]\right\}$. Remember that $E_{S,\tilde{S}}[\ldots] \to E_S\left[E_{\tilde{S}}[\ldots]\right]$ since $S, \tilde{S}$ are independent.
10. Advantage of the Rademacher Formulation: There are 2 main advantages to introducing Rademacher variables:
    a. Multiplying each term in the summation $[h(\tilde{z}_i) - h(z_i)|S]$ by a Rademacher variable σ_i independent of $S, \tilde{S}$ does not alter the expectation;
    b. Negating a Rademacher variable uniformly does not alter its distribution.
11. Reduction using Rademacher Variables: $E_{S,\tilde{S}}\left\{\sup_{h \in \mathfrak{J}}\left[\frac{1}{m}\sum_{i=1}^{m}[h(\tilde{z}_i) - h(z_i)|S]\right]\right\} = E_{\sigma,S,\tilde{S}}\left\{\sup_{h \in \mathfrak{J}}\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i[h(\tilde{z}_i) - h(z_i)|S]\right]\right\} \leq E_{\sigma,S,\tilde{S}}\left\{\sup_{h \in \mathfrak{J}}\left[\frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(z_i) + \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\tilde{z}_i)\right]\right\} = 2E_{\sigma,S}\left\{\sup_{h \in \mathfrak{J}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right\} = 2\mathcal{R}_m(\mathfrak{J})$
12. Rademacher Uniform Convergence Bound: $\boldsymbol{E_D[h(z)] - \hat{E}_s[h(z)] \leq E_S\left\{\sup_{h \in \mathfrak{J}}\left[E_D[h(z)] - \hat{E}_s[h(z)]\right]\right\} + \sqrt{\frac{\log\frac{1}{\delta}}{m}} \leq 2\mathcal{R}_m(\mathfrak{J}) + \sqrt{\frac{\log\frac{1}{\delta}}{m}}}$.
13. Empirical Rademacher Bounds Supremum Estimation: For $\hat{\mathcal{R}}_m(\mathfrak{J}) = E_\sigma\left[\sup_{f \in \mathfrak{J}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$, it is bounded precisely the same way as before using the McDiarmid bounded difference inequality as: $\left|\hat{\mathcal{R}}_{m,\mathbb{Z}}(\mathfrak{J}) - \hat{\mathcal{R}}_{m,\mathbb{Z}'}(\mathfrak{J})\right| \leq \frac{1}{m}$. 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.

14. Rademacher Average Uniform Convergence Bound #2: Thus, the second application of the McDiarmid' inequality (now using confidence limits of $\frac{\delta}{2}$ for each) bound $\hat{\mathcal{R}}_m(\Im)$ in terms of its expectation $\mathcal{R}_m(\Im)$ gives $\boldsymbol{E_D[h(z)] - \hat{E}_s[h(z)] \leq 2\hat{\mathcal{R}}_m(\Im) + 3\sqrt{\frac{\log\frac{2}{\delta}}{m}}}$.
15. 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 s, apply Jensen's inequality, and optimize over s. In other words, to compute $E_\sigma\left\{\underset{a \in A}{sup}\frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right\}$, start with $sE_\sigma\left\{\underset{a \in A}{sup}\frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right\}$.
16. Bounding Rademacher Complexity – Steps: $e^{sE_\sigma\left\{\underset{a\in A}{sup}\sum_{i=1}^{m}\sigma_i a_i\right\}} \leq E_\sigma\left[e^{s\left\{\underset{a\in A}{sup}\sum_{i=1}^{m}\sigma_i a_i\right\}}\right] = E_\sigma\left[\underset{a \in A}{sup}e^{s\{\sum_{i=1}^{m}\sigma_i a_i\}}\right] \leq \sum_{a\in A}E_\sigma\left[e^{s\{\sum_{i=1}^{m}\sigma_i a_i\}}\right] = \sum_{a\in A}E_\sigma[\prod_{i=1}^{m}e^{s\sigma_i a_i}] = \sum_{a\in A}\prod_{i=1}^{m}E_\sigma[e^{s\sigma_i a_i}]$.
17. Hoeffding's Application Criterion: The last step above exploits the fact that the σ_i's are independent. Now we can apply Hoeffding's inequality since $E_\sigma[\sigma_i a_i] = 0$ and $\sigma_i a_i \in [\alpha, \beta]$, where $\beta - \alpha = 2a_i$.
18. Application of Hoeffding's Inequality: $e^{sE_\sigma\left\{\underset{a\in A}{sup}\sum_{i=1}^{m}\sigma_i a_i\right\}} \leq \sum_{a\in A}\prod_{i=1}^{m}E_\sigma[e^{s\sigma_i a_i}] \leq \sum_{a\in A}\prod_{i=1}^{m}e^{\frac{s^2(2a_i)^2}{8}} = \sum_{a\in A}e^{\frac{s^2}{2}\sum_{i=1}^{m}a_i{}^2} \leq |A|e^{\frac{s^2}{2}R^2}$ where $|A|$ is the cardinality of the space A, and $R^2 = \underset{a \in A}{sup}\sum_{i=1}^{m}a_i{}^2$.
19. Optimization around s: Taking log of both sides above we get $E_\sigma\left\{\underset{a \in A}{sup}\frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right\} \leq \frac{\log|A|}{s} + s\frac{R^2}{2}$, which has a maximum at $s = \frac{\sqrt{2\log|A|}}{R}$.
20. Rademacher Complexity Bounds: Substituting this value for s back into the previous bound, and dividing both sides by m gives $\boldsymbol{\hat{\mathcal{R}}_m(\Im) = E_\sigma\left[\underset{a \in A}{sup}\frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right] \leq \frac{R\sqrt{2\log|A|}}{m}}$.
21. Application to VC Theory Finite Concept Class: For any finite concept class $\mathcal{H} \subseteq \{h: X \to [-1, +1]\}$ and data points $S = \{x_1, \ldots, x_m\}$, we can take $A = \{h(x_1), \ldots, h(x_m) | h \in$

$\mathcal{H}\}$. Thus, $|A| \to |\mathcal{H}|$, and $R = \sqrt{\underset{a \in A}{sup} \sum_{i=1}^{m} {a_i}^2} = \sqrt{m}$ (setting each $a_i = 1$). Thus, in this case, $\hat{\mathcal{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log|\mathcal{H}|}{m}}$.

22. Application to the VC Theory - The General Case: In general, irrespective of whether $\mathcal{H}$ is finite or not, we can take $|A| = \mathcal{H}[S]$, the set of distinct labels of points in S using concepts in $\mathcal{H}$. Then $|A| = \mathcal{H}[m]$, the shatter coefficient of $\mathcal{H}$ on m points, and $\hat{\mathcal{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log \mathcal{H}[m]}{m}}$. By Sauer's lemma, $|\mathcal{H}[m]| \leq m^d$, where d is the VC dimension of $\mathcal{H}$. Thus the above may be further simplified into $\hat{\mathcal{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log m}{m}}$.
23. Rademacher Average on the Loss Class vs. Hypothesis Class: $\mathcal{R}(\mathcal{L}) = \mathbb{E}_\sigma \left[\underset{f \in \Im}{sup} \frac{1}{n} \sum_{i=1}^{n} \sigma_i 1_{f(X_i) \neq Y_i}\right] = \mathbb{E}_\sigma \left[\underset{f \in \Im}{sup} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \frac{1}{2}\{1 - Y_i f(X_i)\}\right] = \frac{1}{2} \mathbb{E}_\sigma \left[\underset{f \in \Im}{sup} \frac{1}{n} \sum_{i=1}^{n} \sigma_i Y_i f(X_i)\right] = \frac{1}{2} \mathcal{R}(\Im)$. Using the 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 $\mathcal{N}(\mathcal{F}, \mathcal{Z}_1^n) \coloneqq \left|\mathcal{F}_{\mathcal{Z}_1, \ldots, \mathcal{Z}_n}\right|$.
2. VC Entropy Definition: The annealed VC entropy is defined as $\mathcal{H}_{\mathcal{F}}(n) = \log \mathbb{E}[\mathcal{N}(\mathcal{F}, \mathcal{Z}_1^n)]$. Using this definition, for any δ, with probability at least $1 - \delta$, we now show that, by extending the VC theory, $R(f) - R_n(f) \leq 2\sqrt{2 \frac{\mathcal{H}_{\mathcal{F}}(n) + \log \frac{2}{\delta}}{n}}$.
3. Distribution Dependent Bounds Estimator: As before, we begin with the Symmetrization lemma, so we have to upper bound the quantity $\mathcal{J} = \mathbb{P}_{\mathcal{X}} \left[\underset{f \in \mathcal{F}_{\mathcal{Z}_1^n, \mathcal{Z}'^n_1}}{sup} \quad (P'_n - P_n) f \geq \frac{\varepsilon}{2}\right]$

4. Distribution Dependent Rademacher Bounds: Using Rademacher variables σ_i, we note that $(P'_n - P_n)f$ and $\frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(\mathcal{Z}_i) - f(\mathcal{Z}'_i)]$ have the same distribution, since changing one σ_i corresponds to exchanging $\mathcal{Z}_i$ and $\mathcal{Z}'_i$.
5. Distribution Dependent Bounds - Union Bounding: Applying Rademacher variables as shown above, we get $\mathcal{I} \leq \mathbb{E}_{\mathcal{X}}\left\{\mathbb{P}_\sigma\left[f \in \sup_{\mathcal{F}_{\mathcal{Z}_1^n, \mathcal{Z}'^n_1}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(\mathcal{Z}_i) - f(\mathcal{Z}'_i)] \geq \frac{\varepsilon}{2}\right]\right\}$, and union bounding followed by the application of the distribution-dependent shattering leads to $\mathcal{I} \leq \mathbb{E}_{\mathcal{X}}\left\{\mathcal{N}(\mathcal{F}, \mathcal{Z}_1^n, \mathcal{Z}'^n_1)_f \in \sup_{\mathcal{F}_{\mathcal{Z}_1^n, \mathcal{Z}'^n_1}} \mathbb{P}_\sigma\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(\mathcal{Z}_i) - f(\mathcal{Z}'_i)] \geq \frac{\varepsilon}{2}\right]\right\}$.
6. Distribution Dependent Bounds - Final Step: Noting that $\sigma_i[f(\mathcal{Z}_i) - f(\mathcal{Z}'_i)] \in [-1, 1]$, and applying Hoeffding's inequality finally gives $\mathcal{I} \leq \mathbb{E}_{\mathcal{X}}\left\{\mathcal{N}(\mathcal{F}, \mathcal{Z}_1^n, \mathcal{Z}'^n_1)e^{-\frac{n\varepsilon^2}{8}}\right\}$. Finally, observe that just as $\log[\mathbb{E}\{\mathcal{N}(\mathcal{F})\}] \to \mathcal{S}_{\mathcal{F}}(2n)$ before, we now have $\log[\mathbb{E}\{\mathcal{N}(\mathcal{F}, \mathcal{Z}_1^n, \mathcal{Z}'^n_1)\}] \to \mathcal{H}_{\mathcal{F}}(2n)$. 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., $\mathcal{R}(\mathcal{F}) \leq 2\sqrt{\frac{\log \mathcal{N}(\mathcal{F})}{n}}$), the dependence on n 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.
2. Dudley-Haussler Entropy Bound: Using the chaining technique, one has the following so-called Dudley's entropy bound: $\mathcal{R}_n(\mathcal{F}) \leq 2\frac{\mathcal{C}}{\sqrt{n}}\int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{F}, \varepsilon, n)}\, d\varepsilon$. As a consequence, along with the Haussler's upper bound, we get the following result: $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{K}\sqrt{\frac{h}{n}}$. We can, with this approach, thus remove the unnecessary $\log n$ 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 d-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 $r \geq 0$ for the class $\Im$ is defined as $\mathcal{R}_n(\mathcal{F}, r) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n(f) \right]$.
2. 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 $\psi : \mathbb{R} \to \mathbb{R}$ is a sub-root if:
    a. ψ is non-decreasing
    b. ψ is non-negative, and
    c. $\frac{\psi(r)}{\sqrt{r}}$ is non-increasing.

    These characteristics ensure that the sub-root function is continuous, and has a unique (non-zero) fixed-point r^* satisfying $\psi(r^*) = r^*$.

2. Star-Hull Definition: Let $\mathcal{F}$ be a set of functions. Its star-Hull is defined as $* \mathcal{F} = \{\alpha f : f \in \mathcal{F}, \alpha \in [0,1]\}$.
3. 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 $\mathcal{F}$, $\mathcal{R}_n(* \mathcal{F}, r)$ is a sub-root.
4. 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 $\mathcal{F}$ and $* \mathcal{F}$. 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 $\mathcal{F}$ be a class of bounded functions (i.e., $f \in [-1,1]$), and r^* be the fixed point of $\mathcal{R}_n(* \mathcal{F}, r)$. There exists a constant $c > 0$ such that, with probability at least $1 - \delta$, $\forall f \in \mathcal{F}\ Pf - P_n f \leq c\left[\sqrt{r^* Var\, f} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right]$. If, in addition, the functions in $\mathcal{F}$ satisfy $Var\, f \leq c(Pf)^{\alpha}$, then one obtains, with probability at least $1 - \delta$, $\forall f \in \mathcal{F}\ Pf \leq c\left[P_n f + (r^*)^{\frac{1}{2-\alpha}} + \frac{\log\frac{1}{\delta} + \log\log n}{n}\right]$.
2. 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, $\sup_{f \in \mathcal{F}} Pf - P_n f \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} Pf - P_n f\right] + c\sqrt{\sup_{f \in \mathcal{F}} \frac{Var\, f}{n}} + \frac{c'}{n}$ for some c, c'.
3. 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

$\mathcal{F}_k = \{f : Var\ f\ \in [x_k, x_{k+1}]\}$. Note that although $Pf - P_n f$ depends on both the suprema across the expectation as well as the variance, here we classify the function space only on the variance.

4. 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, $\sup_{f \in \mathcal{F}_k} Pf - P_n f \le \mathbb{E}\left[\sup_{f \in \mathcal{F}_k} Pf - P_n f\right] + c\sqrt{\sup_{f \in \mathcal{F}_k} \frac{Var\ f}{n}} + \frac{c'}{n}$.
5. 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, $Pf - P_n f \le 2\mathcal{R}_n\left(\mathcal{F}, \sup_{f \in \mathcal{F}_k} Var\ f\right) + c\sqrt{\sup_{f \in \mathcal{F}_k} \frac{Var\ f}{n}} + \frac{c'}{n}$.
6. Proof Step #5 - Express Local Rademacher Average in terms of the Fixed Point: We then "solve" the above inequality for $Pf - P_n f$. Things are simple if $\mathcal{R}_n\left(\mathcal{F}, \sup_{f \in \mathcal{F}_k} Var\ f\right)$ 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 $Pf - P_n f \le 2\sqrt{r^* Var\ f} + c\sqrt{\sup_{f \in \mathcal{F}_k} \frac{Var\ f}{n}} + \frac{c'}{n}$.
7. Proof Step #6 - Relation Between the Variance and the Expectation: Finally we obtain the relationship between the variance and the expectation $Var\ f \le c(Pf)^\alpha$ and "solve" the inequality in Pf to get the original result.

Local Rademacher Average – Consequences

1. Fast Rate: An important example in this case is where the class $\mathcal{F}$ is of finite VC dimension h. In that case, one has $\mathcal{R}(\mathcal{F}, r) \le \mathcal{C}\sqrt{\frac{rh\log n}{n}}$ so that $r^* \le \mathcal{C}'\frac{rh\log n}{n}$. As a consequence, under the Tsybakov noise condition, we obtain a rate of convergence of $P_n f$ to Pt as $\mathcal{O}\left(\frac{1}{n^{\frac{1}{2-\alpha}}}\right)$. It is

important to note that, in this case, the rate of convergence of $P_n f$ to Pf is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. So we obtain a fast rate by looking at the relative error. These fast rates can be obtained provided $t \in \mathcal{F}$ (but it is not necessary that $R^* = 0$). This requirement can be removed if one uses structural risk minimization or regularization.

2. 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 $\mathcal{R}_n(\mathcal{F}, r) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}: Pf^2 \leq r} R_n(f) \right]$. The result is the same as before (but with different constants) under the same conditions. With probability at least $1 - \delta$, $Pf \leq \mathcal{D}\left[P_n f + (r_n{}^*)^{\frac{1}{2-\alpha}} + \frac{\log\frac{1}{\delta} + \log\log n}{n} \right]$ where $r_n{}^*$ is the fixed-point of a sub-root upper bound of $\mathcal{R}_n(\mathcal{F}, r)$.
3. Conditional Local Rademacher Averages - Asymptotic Noise Behavior: Hence, we get improved rates when the noise is well-behaved, and these rates interpolate between $n^{-\frac{1}{2}}$ and n^{-1}. However, in general it is not possible to estimate the parameters $\mathcal{D}$ and α that enter in the noise condition.
4. 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 $\mathcal{R}_n(* \mathcal{F}, r)$, each function $f \in \mathcal{F}$ with $Pf^2 \geq r$ is considered at a scale $\frac{r}{Pf^2}$.

Normalized ERM

Background

1. The Concept: The main idea is to consider the ratio $\frac{Pf-P_nf}{\sqrt{Pf}}$ where $f \in \{0,1\}$ and $Var\ f \leq Pf^2, Pf$.
2. Motivation: The motivation for considering the above normalized variant is that the fluctuations of the variants from $\Im$ are more "uniform". Hence, the supremum in $\sup_{f \in \Im} \frac{Pf-P_nf}{\sqrt{Pf}}$ 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 Pf (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 $\delta > 0$, with probability at least $1-\delta$, $\forall f \in \Im\ \frac{Pf-P_nf}{\sqrt{Pf}} \leq 2\sqrt{\frac{\log \mathcal{S}_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n}}$, and also with probability at least $1-\delta$, $\forall f \in \Im\ \frac{Pf-P_nf}{\sqrt{P_nf}} \leq 2\sqrt{\frac{\log \mathcal{S}_{\mathcal{F}}(2n)+\log\frac{4}{\delta}}{n}}$.
2. Proof Step #1 - Using Symmetrization: The first step uses a variant of the symmetrization lemma: $\mathbb{P}\left[\sup_{f \in \Im} \frac{Pf-P_nf}{\sqrt{Pf}} \geq \varepsilon\right] \leq 2\mathbb{P}\left[\sup_{f \in \Im} \frac{P'_nf-P_nf}{\sqrt{\frac{P'_nf+P_nf}{2}}} \geq \varepsilon\right]$.

3. Proof Step #2 - Use of Rademacher Variables: In this step we use randomization using Rademacher variables, i.e., $\mathbb{P}\left[\sup_{f \in \Im} \frac{P\prime_n f - P_n f}{\sqrt{\frac{P\prime_n f + P_n f}{2}}} \geq \varepsilon\right] = \mathbb{E}\left\{\mathbb{P}_\sigma\left[\sup_{f \in \Im} \frac{\frac{1}{n}\sum_{i=1}^{n} \sigma_i [f(\mathbb{Z}_i) - f(\mathbb{Z}\prime_i)]}{\sqrt{\frac{P\prime_n f + P_n f}{2}}} \geq \varepsilon\right]\right\}$.
4. 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, A, B, and C, $A \leq B + C\sqrt{A} \rightarrow A \leq B + C^2 + \sqrt{B}C$, we get $\forall f \in \Im\ Pf - P_n f \leq 2\sqrt{P_n f \frac{\log \mathcal{S}_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}} + 4\frac{\log \mathcal{S}_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}$.
2. Ideal, Noiseless Minimizer: In such a situation, there is no noise (i.e., $Y = t(X)$ almost surely), and $t \in \Im_n$. Denoting by f_n the empirical minimizer, we have $R^* = 0$ as well as $R_n(f_n) = 0$ in this case. Thus, in the situation where $\Im$ is a function space with VC dimension h we get $R(f_n) \approx \mathcal{O}\left[\frac{h \log n}{n}\right]$. This clearly demonstrates that we de facto interpolate between the best rate of convergence $\mathcal{O}\left[\frac{h \log n}{n}\right]$ and the worst rate of convergence $\mathcal{O}\left[\sqrt{\frac{h \log n}{n}}\right]$ (the $\log n$ factor cannot be removed in this case).
3. Corresponding ERM Bound: Casting $Pf \rightarrow R(f)$ in the approach above, we get $R(f_n) \leq R^* + 2\sqrt{R^* \frac{\log \mathcal{S}_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}} + 4\frac{\log \mathcal{S}_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}$, thus when $R^* = 0$, $t \in \Im$ and the convergence is $\frac{1}{n}$, while if $R^* > 0$, the rate contains contribution from the $\frac{1}{\sqrt{n}}$ term. Therefore it is not possible to obtain a rate with a power of n between $-\frac{1}{2}$ and -1.

4. Challenges with Denormalized Bounding: The main challenge with the Denormalized ERM approach is that the factor of the square root term R^* (which arises out of the denormalization term $\sqrt{Pf}$) is not a convenient entity to use here, since it does not vary with n. However, if we use $R(f_n) - R^*$ as the corresponding entity under the square root, this converges to zero with increasing n (see later on the relative error classes). However, the denormalized approach cannot be used for $f - f^*$, 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 $\mathcal{P}(x, y) = \mathcal{P}(x) \times \mathcal{P}(y \mid x)$. We define the regression function as $\eta(x) = \mathbb{E}[Y|X = x] = 2\mathbb{P}[Y = 1|X = x] - 1$, and the target function (i.e., the Bayes' classifier) as $t(x) = sign[\eta(x)]$.
2. Noise Level: We define the noise level $s(x)$ as $s(x) = \min\{\mathbb{P}[Y = 1|X = x], 1 - \mathbb{P}[Y = 1|X = x]\}$. In practice, a more useful definition (one that does not involve $\mathbb{P}$) is $s(x) = \frac{1-\eta(x)}{2}$. In the deterministic case, $s(x) = 0$ almost surely. Thus, the Bayes' risk is $\mathcal{R}^* = \mathbb{E}[s(x)]$.
3. Noise Condition Motivation: To improve the treatment/results above, we seek a refinement that requires us to make assumptions about the noise function $s(x)$ (in the ideal case, $s(x) = 0$ everywhere, thus $R^* = 0$ and $Y = t(X)$). So we use the above-mentioned metrics to quantify how well-behaved the noise function is.
4. Range of the Noise and the Regression Function: From a classification point-of-view, the favorable classifier situation occurs when $\eta(x)$ is not too close to zero. Indeed, $\eta(x) = 0$ means that the noise is a maximum at that x (and the corresponding $s(x) = \frac{1}{2}$); here the label is completely undetermined, as any prediction results in an error with a probability of $\frac{1}{2}$.

Types of Noise Conditions

1. Type 1 - Massart Noise Condition: For some c, assume $|\eta(x)| > \frac{1}{c}$ almost surely (c 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 $\frac{1}{2}$.

2. Type 2 - Tsybakov Noise Condition: Let $\alpha \in [0,1]$ and assume that one of the following equivalent conditions is satisfied:
    a. $\exists\, c > 0, \forall f \in \{-1,1\}^{\mathcal{X}}, \mathbb{P}[f(X)\eta(X) \leq 0] \leq c[R(f) - R^*]^{\alpha}$
    b. $\exists\, c > 0, \forall \mathcal{A} \subset \mathcal{X}, \int_{\mathcal{A}} dP(x) \leq c\left[\int_{\mathcal{A}} |\eta(x)| dP(x)\right]^{\alpha}$
    c. $\exists\, B > 0, \forall \varepsilon \geq 0, \mathbb{P}[|\eta(X)| \leq \varepsilon] \leq B\varepsilon^{\frac{\alpha}{1-\alpha}}$. This condition is the easiest to interpret, as it indicates that $\eta(x)$ is close to the critical value 0 with low probability.
3. Equivalence of the Tsybakov Conditions:
    a. Equivalence between a) and c) => It is easy to check that $R(f) - R^* = \mathbb{E}\left[|\eta(x)| 1_{f\eta \leq 0}\right]$. For each function f, there exists a set $\mathcal{A}$ such that $1_{\mathcal{A}} = 1_{f\eta \leq 0}$.
    b. Equivalence between b) and c) => Let $\mathcal{A} = \{x : |\eta(x)| \leq \varepsilon\}$. $\mathbb{P}[|\eta(x)| \leq \varepsilon] = \int_{\mathcal{A}} dP(x) \leq c\left[\int_{\mathcal{A}} |\eta(x)| dP(x)\right]^{\alpha} \leq c\varepsilon^{\alpha}\left[\int_{\mathcal{A}} dP(x)\right]^{\alpha} \Longrightarrow \mathbb{P}[|\eta(x)| \leq \varepsilon] \leq c^{\frac{1}{1-\alpha}}\varepsilon^{\frac{\alpha}{1-\alpha}} = B\varepsilon^{\frac{\alpha}{1-\alpha}}$ where $B = c^{\frac{1}{1-\alpha}}$.
    c. Equivalence between Tsybakov Conditions a) and c) => We write $R(f) - R^* = \mathbb{E}\left[|\eta(x)| 1_{f\eta \leq 0}\right] \geq \varepsilon\mathbb{E}\left[1_{f\eta \leq 0} 1_{|\eta(x)| > \varepsilon}\right] = \varepsilon\mathbb{P}[|\eta(x)| > \varepsilon] - \varepsilon\mathbb{E}\left[1_{f\eta > 0} 1_{|\eta(x)| > \varepsilon}\right] \geq \varepsilon\left[1 - B\varepsilon^{\frac{\alpha}{1-\alpha}}\right] - \varepsilon\mathbb{P}[f\eta > 0] = \varepsilon\left[\mathbb{P}[f\eta \leq 0] - B\varepsilon^{\frac{\alpha}{1-\alpha}}\right]$. Setting $\varepsilon = \left\{\frac{(1-\alpha)\mathbb{P}[f\eta \leq 0]}{B}\right\}^{\frac{\alpha}{1-\alpha}}$ finally gives $\mathbb{P}[f\eta \leq 0] \leq \frac{B^{1-\alpha}}{(1-\alpha)\alpha^{\alpha}}[R(f) - R^*]^{\alpha}$, and, setting $c = \frac{B^{1-\alpha}}{(1-\alpha)\alpha^{\alpha}}$ recovers a).
4. Restriction on the Range of α: The parameter α has to be in the range $[0,1]$. Otherwise we end up with the opposite inequality $R(f) - R^* = \mathbb{E}\left[|\eta(x)| 1_{f\eta \leq 0}\right] \leq \mathbb{E}\left[1_{f\eta \leq 0}\right] = \mathbb{P}[f\eta \leq 0]$ which is incompatible with condition a) if $\alpha > 1$ (i.e., this is a special case).
5. Equivalence of the Tsybakov and the Massart Cases: When $\alpha = 0$ the Tsybakov condition becomes invalid, and when $\alpha = 1$ 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 $\widetilde{\mathfrak{I}} = \{(x, y) \mapsto f(x, y) - 1_{t(x)\neq y} : f \in \mathfrak{I}\}$. This relationship will allow exploiting the Bernstein-type inequalities applied to this latter class.
2. Massart and Tsybakov Noise Relative Error: Under Massart's condition, one has $\mathbb{E}\left[\left(1_{f(x)\neq y} - 1_{t(x)\neq y}\right)^2\right] \leq c[R(f) - R^*]$, or equivalently, for $f \in \widetilde{\mathfrak{I}}$, $var\, f \leq Pf^2 \leq cPf$. Under Tsybakov's condition, this becomes, for $f \in \mathfrak{I}$, $\mathbb{E}\left[\left(1_{f(x)\neq y} - 1_{t(x)\neq y}\right)^2\right] \leq c[R(f) - R^*]^\alpha$, and for $f \in \widetilde{\mathfrak{I}}$, $var\, f \leq Pf^2 \leq c(Pf)^\alpha$.
3. Bernstein Inequality on Tsybakov Condition: In the finite case, with $|\mathfrak{I}| = N$, one can applye the Bernstein inequality to $\widetilde{\mathfrak{I}}$ and get (in conjunction with the finite union bound), with a probability of at least $1 - \delta$, for all $f \in \mathfrak{I}$, $R(f) - R^* \leq R_n(f) - R_n(t) + \sqrt{\frac{8c[R(f)-R^*]^\alpha \log\frac{N}{\delta}}{n}} + \frac{4\log\frac{N}{\delta}}{3n}$.
4. Bernstein-Tsybakov Bound - $t \in \mathfrak{I}$ Case: When $t \in \mathfrak{I}$, f_n is the minimizer of the empirical error (hence $R_n(f) \leq R_n(t)$), one has $R(f) - R^* \leq c\left[\frac{\log\frac{N}{\delta}}{n}\right]^{\frac{1}{2-\alpha}}$ which is always better than $n^{-\frac{1}{2}}$ for $\alpha > 0$, and is valid even if $R^* > 0$.

Alternate Generalization Bounds and Capacity Measures

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 $1 - \delta$, $R(f) \leq R_{emp}(f) + Capacity\ (\mathfrak{I}) + Confidence\ (\delta)$. In the simplest case, the capacity term only depends on $\mathfrak{I}$ and the confidence term on the underlying probability. Obviously, as is to be expected, these bounds are worst case bounds on *bad behavior*.
2. 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.
3. Alternate Capacity/Complexity Measures: Some alternate capacity/complexity measures are:
    a. Covering Numbers (all variants)
    b. VC Dimension
    c. Rademacher Complexity
    d. MDL Coded Size
    e. Parametrized Bayesian Priors/Posteriors

Covering and Entropy Numbers

Nomenclature - Normed Spaces

1. l_p^d Spaces: For $d \in \mathbb{N}$, $\mathbb{R}^d$ denotes the d-dimensional space of vectors $\vec{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$. We define spaces l_p^d as follows: as vector spaces, they are identical to $\mathbb{R}^d$, in addition, they are endowed with p-norms; for $0 < p < \infty$, $\|\vec{x}\|_{l_p^d} \doteqdot \|\vec{x}\|_p = \left[\sum_{j=1}^d |x_j|^p\right]^{\frac{1}{p}}$; and for $p = \infty$, $\|\vec{x}\|_{l_\infty^d} \doteqdot \|\vec{x}\|_\infty = \max_{j = 1, \ldots, d} |x_j|$. Note that a different normalization of the l_p^d norm is used in some papers in learning theory (e.g., Talagrand (1996)). For $0 < p \leq \infty$, $l_p \doteqdot l_p^\infty$.
2. l_∞^d of $f \in \mathcal{F}$ with respect to $\vec{X}_m$: Given m points $\vec{x}_1, \ldots, \vec{x}_m \in l_p^d$, we use the short-hand $\vec{X}_m = (\vec{x}_1, \ldots, \vec{x}_m)$. Suppose $\mathcal{F}$ is a class of functions $f : \mathbb{R}^d \to \mathbb{R}$. The l_∞^d norm with respect to $\vec{X}_m$ of $f \in \mathcal{F}$ is defined as $\|f\|_{l_\infty^{\vec{X}_m}} \doteqdot \max_{i = 1, \ldots, m} |f(\vec{x}_i)|$. Likewise, $\|f\|_{l_p^{\vec{X}_m}} = \|[f(\vec{x}_1), \ldots, f(\vec{x}_m)]\|_{l_p^m}$.
3. Integral Norm over a σ-algebra Space: Given some set $\mathcal{X}$ with a σ-algebra, a measure μ on $\mathcal{X}$, some $0 < p < \infty$, and a function $f : \mathcal{X} \to \mathbb{R}$, we define $\|f\|_{\mathcal{L}_p(\mathcal{X},\mathbb{R})} \doteqdot [\int |f(x)|^p d\mu(x)]^{\frac{1}{p}}$ if the integral exists, and $\|f\|_{\mathcal{L}_\infty(\mathcal{X},\mathbb{R})} \doteqdot ess \sup_{x \in \mathcal{X}} f(x)$. For $0 < p < \infty$, we let $\mathcal{L}_p(\mathcal{X}, \mathbb{R}) \doteqdot \left\{f : \mathcal{X} \to \mathbb{R} : \|f\|_{\mathcal{L}_p(\mathcal{X},\mathbb{R})} < \infty\right\}$. We also let $\mathcal{L}_p(\mathcal{X}) \doteqdot \mathcal{L}_p(\mathcal{X}, \mathbb{R})$.

Covering, Entropy, and Dyadic Numbers

1. ϵ-Covering Numbers: If $\mathcal{S}$ is a set and d is a metric on $\mathcal{S}$, then the ϵ-covering number of $M \subset \mathcal{S}$ with respect to the metric d denoted $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the smallest number of elements of an ϵ-cover for $\mathcal{F}$ using the metric d.

2. n^{th} Entropy Number of a Set $M \subset \mathcal{S}$: Given a metric space $\mathcal{E} = (\mathcal{S}, d)$ we also write the covering number as $\mathcal{N}(\epsilon, \mathcal{F}, \mathcal{E})$. The n^{th} entropy number of a set $M \subset \mathcal{E}$, for $n \in \mathbb{N}$, is $\epsilon_n(M) \doteqdot \inf\{\epsilon > 0;\ \mathcal{N}(\epsilon, \mathcal{F}, \mathcal{E}) \leq n\}$.
3. Operator Norm: Let $\mathfrak{L}(\mathcal{E}, \mathcal{F})$ be the set of all bounded operators $\mathcal{T}$ between the normed spaces $(\mathcal{E}, \|\cdot\|_{\mathcal{E}})$ and $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$, i.e., operators such that the image of the closed unit ball $\mathcal{U}_{\mathcal{E}}\{x \in \mathcal{E} :\ \|x\|_{\mathcal{E}} \leq 1\}$ is bounded. The smallest such bound is called the *Operator Norm*.
4. Entropy Numbers of an Operator: The *Entropy Numbers of an Operator* $\mathcal{T} \in \mathfrak{L}(\mathcal{E}, \mathcal{F})$ are defined as $\epsilon_n(\mathcal{T}) = \epsilon_n\big(\mathcal{T}(\mathcal{U}_{\mathcal{E}})\big)$. Note that $\epsilon_n(\mathcal{T}) = \|\mathcal{T}\|$, and that $\epsilon_n(\mathcal{T})$ is certainly well-defined for all $n \in \mathbb{N}$ if $\mathcal{T}$ is a compact operator, i.e., for any $\epsilon > 0$ there exists a finite cover of $\mathcal{T}(\mathcal{U}_{\mathcal{E}})$ with open ϵ balls over $\mathcal{X}$.
5. Dyadic Entropy Numbers of an Operator: The dyadic entropy numbers of an operator are defined by $e_n(\mathcal{T}) = \epsilon_{2^{n-1}}(\mathcal{T})$, $n \in \mathbb{N}$. 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).
6. Banach Spaces: $\mathcal{E}$ and $\mathcal{F}$ will always be Banach spaces (for instance, l_p^d spaces with $p \leq 1$). In some cases they may be *Hilbert Spaces* $\mathcal{H}$, Banach spaces endowed with a dot-product $\langle ., . \rangle_{\mathcal{H}}$ giving rise to its norm via $\|\vec{x}\|_{\mathcal{H}} = \sqrt{\langle \vec{x}, \vec{x} \rangle_{\mathcal{H}}}$.

Background and Overview of Results

1. Introduction: Covering numbers 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. Setup: We start by endowing a function class $\Im$ with the following random metric: $d_n(f, f') = \frac{1}{n}|\{f(\mathcal{Z}_i) - f'(\mathcal{Z}_i); i = 1, \ldots, n\}|$. This is the normalized Hamming distance of the projections on the sample. Given such a metric, we say that $f_1, \ldots, f_n$ *covers* $\Im$ at a radius ε if $\Im \subset \bigcup_{i=1}^{N} \mathcal{B}(f_i, \varepsilon)$. The term *cover* is used analogous to the term *space* in function space.

3. Definition: The covering number of $\mathfrak{I}$ at a radius ε with respect to d_n, denoted by $\mathcal{N}(\mathfrak{I}, \varepsilon, N)$, is the minimum size of a cover of radius ε. Note that it does not matter if we apply this definition to the original class $\mathfrak{I}$ or the loss class $\mathcal{L}$, since $\mathcal{N}(\mathfrak{I}, \varepsilon, N) = \mathcal{N}(\mathcal{L}, \varepsilon, N)$.
4. Growth of the Covering Function: The covering numbers characterize the size of the function class measured by the metric d_n. The rate of growth of the logarithm of $\mathcal{N}(\mathfrak{I}, \varepsilon, N)$, called the metric entropy, is related to the classical concept of the vector dimension. Indeed, if $\mathfrak{I}$ is a compact set in a d-dimensional Euclidean space, $\mathcal{N}(\mathfrak{I}, \varepsilon, N) \approx \varepsilon^{-d}$.
5. Finite Covering Numbers: When the covering numbers are finite, it is possible to approximate the class $\mathfrak{I}$ by a finite set of functions (which cover $\mathfrak{I}$). This again allows the use of finite union bound, provided we can relate the behavior of all functions in $\mathfrak{I}$ to that of the functions in the cover.
6. Finite Covering Number Sample Bound: $\mathbb{P}[\exists\, f \in \mathfrak{I}: R(f) - R_n(f) > \varepsilon] \leq 8\mathbb{E}[\mathcal{N}(\mathfrak{I}, \varepsilon, n)]e^{-\frac{n\varepsilon^2}{128}}$.
7. Covering Numbers for Real Valued Functions: Covering Numbers can also be defined for real-valued functions – see treatment below.
8. Covering Numbers and the VC Dimension: Notice that, because functions in $\mathfrak{I}$ can only take 2 values, for all $\varepsilon > 0$, $\mathcal{N}(\mathfrak{I}, \varepsilon, n) \leq \left|\mathfrak{I}_{\mathbb{Z}_1^n}\right| = \mathcal{N}(\mathfrak{I}, \mathbb{Z}_1^n)$. Hence the VC entropy corresponds to the log covering number at the minimal scale, which implies $\mathcal{N}(\mathfrak{I}, \varepsilon, n) \leq h \log \frac{en}{h}$, but one can have a considerably better result.
9. Haussler's Bound: Let $\mathfrak{I}$ be a class of VC dimension h. Then for all $\varepsilon > 0$, all n, and any sample, $\mathcal{N}(\mathfrak{I}, \varepsilon, n) \approx Ch(4e)^h \varepsilon^{-h}$. The interesting aspect of this bound is that it does not depend on the sample size n.
10. 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.

Covering Number for Real-Valued Function Classes

1. Definition: Let $\mathcal{F}$ be a sub-set of a metric space $(\mathcal{X}, \rho)$. For a given $\epsilon > 0$, the metric covering number $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ is defined as the smallest number of sets/balls of radius ϵ whose union contains $\mathcal{F}$. In what follows, ρ is omitted if the context is clear.
2. Functions of Bounded Variation under the $\mathcal{L}_1$ Metric: Here our intention is to find bounds on the covering numbers of functions of bounded variations under the $\mathcal{L}_1$ metric. Specifically, let $\mathcal{F}_1$ be the set of all functions on $[0, T]$ taking values in $\left[-\frac{V}{2}, +\frac{V}{2}\right]$ with total variation of at most $V > 0$. It is natural to use the same V for both the range and the variation, since a bound on the variation of a function implies a bound on its range.
3. Tight Bounds on $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1)$: Our goal is to find tight bounds on $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1)$ 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)).
4. 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 $\log VT$ constant in Birge (1987) rather than VT that we obtain here.
5. 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 $\mathcal{L}_1(dP)$ metrics, where P 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)).
6. 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 $\max_P \log \mathcal{N}\left(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)\right)$ 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 $\epsilon \leq \frac{VT}{12}$, $\frac{VT}{54\epsilon} \leq \log \mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq \frac{12VT}{\epsilon} \log 2$ (Bartlett, Kulkarni, and Posner (1997)).
2. Special Case of the $\mathcal{L}_\infty$ Metric: Certainly under the $\mathcal{L}_\infty$ metric, the classes of functions of bounded variation (or even the subset of functions under this class that are continuous) are not compact, hence $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_\infty) = \infty$. 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 $\left(\frac{1}{\epsilon}\right)^{\frac{1}{\epsilon}}$ in the $\mathcal{L}_\infty$ metric (Lorentz (1966)).
3. Function Bound vs. Class Bound: Let $\mathcal{F}_2$ be all functions of variation at most V that map from $[0, T]$ to $[-B, +B]$ for some $B > \frac{V}{2}$. The theorem above can be extended to $\mathcal{F}_2$ as $\frac{VT}{54\epsilon} \log 2 + \frac{eBT}{6\epsilon} \leq \log \mathcal{N}(\epsilon, \mathcal{F}_2, \mathcal{L}_1) \leq \frac{18VT}{\epsilon} \log 2 + \frac{3(2B-V)T}{8\epsilon}$ (Bartlett, Kulkarni, and Posner (1997)). Both upper and lower bounds are obtained by considering vertical shifts in $\mathcal{F}_1$ and using the proof approach outlined below.
4. Distribution-Dependent Bounds: The upper bound (with T set to 1) can be extended to give the upper bound $\log \mathcal{N}\left(\epsilon, \mathcal{F}_1, \mathcal{L}_1(dP)\right) \leq \frac{12V}{\epsilon}$ (Bartlett, Kulkarni, and Posner (1997)), i.e., a uniform bound on the covering numbers for all weighted $\mathcal{L}_1$ norms where P is an arbitrary probability distribution in $[0, T]$. The proof for this is obtained by modifying the proof presented below for an equi-probable partition of $[0, T]$ rather than a partition of equal size.

Functions of Bounded Variation – Upper Bound Proof

1. $\frac{\epsilon}{2}$-Covering of the Function Class: Define $\mathcal{I}$ to be the class of all non-decreasing functions on $[0,T]$ taking values in $[0,V]$. Let $\mathcal{G}\left(\frac{\epsilon}{2}\right)$ be any $\frac{\epsilon}{2}$-covering of $\mathcal{I}$. It is well-known that for any $f \in \mathcal{F}_1$ there exist $g,h \in \mathcal{I}$ such that $f = g - h$ (e.g., Kolmogorov and Fomin (1970)).
2. Range of the $\frac{\epsilon}{2}$-cover Class: More precisely, if $v(x)$ denotes the total variation over the interval $[0,x]$, then g and h can be taken to be $g(x) = \frac{v(x)+f(x)}{2} + \frac{V}{4}$ and $h(x) = \frac{v(x)-f(x)}{2} + \frac{V}{4}$. It is easy to show that both g and h are non-decreasing. Also, if f takes values in $\left[-\frac{V}{2}, +\frac{V}{2}\right]$, and has total variation bounded by V, then both g and h take values only in the range $[0,V]$.
3. Metric Covering Number of the $\frac{\epsilon}{2}$-cover Set: By then definition of the cover of a set, there exist $\phi_1, \phi_2 \in \mathcal{G}\left(\frac{\epsilon}{2}\right)$ such that $\|g - \phi_1\|_{\mathcal{L}_1} \leq \frac{\epsilon}{2}$, $\|h - \phi_2\|_{\mathcal{L}_1} \leq \frac{\epsilon}{2}$. This gives $\|f - (\phi_1 - \phi_2)\|_{\mathcal{L}_1} = \|g - h - (\phi_1 - \phi_2)\|_{\mathcal{L}_1} \leq \|g - \phi_1\|_{\mathcal{L}_1} + \|h - \phi_2\|_{\mathcal{L}_1} \leq \epsilon$. Thus, we can produce ϵ-covering of $\mathcal{F}_1$ by taking all pairs from $\mathcal{G}\left(\frac{\epsilon}{2}\right) \times \mathcal{G}\left(\frac{\epsilon}{2}\right)$. Hence, $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq \left[\mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{I}, \mathcal{L}_1\right)\right]^2$.
4. Cardinality of the Equi-Partition Set: We now find an upper bound on $\mathcal{N}(\epsilon, \mathcal{I}, \mathcal{L}_1)$ by producing an ϵ-covering of $\mathcal{I}$. Partition $[0,T]$ into $n_1 = \frac{T}{h_1} \geq 1$ sub-intervals of length h_1, i.e., $[0,h_1), [h_1, 2h_1), \ldots, , [(n_1 - 1)h_1, 1)$. Let $\Phi(n_1, n_2)$ be the set of all functions that are constant on these sub-intervals, non-decreasing, and taking values only in the set $\left\{\left(j - \frac{1}{2}\right)h_2 | j = 1, \ldots, n_2\right\}$ where $h_2 = \frac{V}{n_2}$. It is easy to see that cardinality of $\Phi(n_1, n_2)$, denoted by $|\Phi(n_1, n_2)|$, satisfies $|\Phi(n_1, n_2)| = \binom{n_1+n_2}{n_1} \leq 2^{n_1+n_2}$.
5. Metric Covering Number - Upper Bound: Now, note that for any $f \in \mathcal{I}$ there exists a $\phi \in \Phi(n_1, n_2)$ such that $\|f - \phi\|_{\mathcal{L}_1} \leq h_1 V + \frac{h_2 T}{2}$. To see this, consider the error to the best constant approximation to f on a sub-interval, and the additional error introduced by quantizing the range. Choosing $h_1 = \frac{\epsilon}{2V}$ and $h_2 = \frac{\epsilon}{T}$ gives $\|f - \phi\|_{\mathcal{L}_1} \leq \epsilon$. Hence, with this choice we get $\mathcal{N}(\epsilon, \mathcal{I}, \mathcal{L}_1) \leq |\Phi(n_1, n_2)| \leq 2^{n_1+n_2} = 2^{\frac{3VT}{\epsilon}}$ for $\epsilon < VT$. By using the ϵ-covering bound of $\mathcal{F}_1$ from $\mathcal{G}\left(\frac{\epsilon}{2}\right)$, we get $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq \left[\mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{I}, \mathcal{L}_1\right)\right]^2 \leq 2^{\frac{12VT}{\epsilon}}$.

Functions of Bounded Variation – Lower Bound Proof

1. Partition of $[0,T]$: Partition $[0,T]$ into n segments. Take the set Γ of all the binary $\{0,h\}$-valued functions constant over each segment. If n ≥ 2, the bounded variation and the boundedness constraints impose h $\leq \frac{V}{n}$.
2. The Difference Function: There are 2^n functions in the set Γ. For any two functions $\gamma_i, \gamma_j \in \Gamma$ define $d(i,j)$ as the number of segments on which the 2 functions have different values. It is then easy to see that $\left\|\gamma_i - \gamma_j\right\|_{\mathcal{L}_1} = d(i,j)\frac{hT}{n}$.
3. Bounding the Differences Function: Thus, $\gamma_i, \gamma_j \in \Gamma$ are close if $d(i,j) \leq \frac{n\epsilon}{hT}$. For an arbitrary $\gamma \in \Gamma$, let $C(\epsilon)$ be the number of functions in Γ that are ϵ-close to Γ. Then $C(\epsilon) = \sum_{l=0}^{\left[\left[\frac{n\epsilon}{hT}\right]\right]}\binom{n}{l}$ where $[\![\alpha]\!]$ denotes the largest number no bigger than α. The Chernoff-Okamoto inequality (Dudley (1978)) is $\sum_{l=0}^{m}\binom{n}{l}p^l(1-p)^{n-l} \leq e^{-\frac{(np-m)^2}{2np(1-p)}}$ for $p \leq \frac{1}{2}$ and $m \leq np$. Letting $p = \frac{1}{2}$ we get that $\sum_{l=0}^{\left[\left[\frac{n\epsilon}{hT}\right]\right]}\binom{n}{l} \leq 2^n e^{-\frac{2\left(\frac{n}{2}-\left[\left[\frac{n\epsilon}{hT}\right]\right]\right)^2}{n}} \leq 2^n e^{-\frac{n}{2}\left(1-\frac{2\epsilon}{hT}\right)^2}$. The same result can also be obtained by using Hoeffding's inequality.
4. Optimal Lower Bound: Since the cardinality of Γ is 2^n, it is clear that we need at least $\frac{2^n}{C(\epsilon)}$ functions for a ϵ-cover, i.e., $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \geq \frac{2^n}{C(\epsilon)} \geq e^{-\frac{n}{2}\left(1-\frac{2\epsilon}{hT}\right)^2}$. To obtain a large lower bound we want to maximize this expression subject to h $\leq \frac{V}{n}$. Equivalently, $\underset{n \geq 2, \text{h} \leq \frac{V}{n}}{max} n\left(1 - \frac{2\epsilon}{hT}\right)^2 \geq \underset{n \geq 2}{max}\left(1 - \frac{2\epsilon n}{VT}\right)^2$. If $\epsilon \leq \frac{VT}{12}$, we may choose $n = \left[\left[\frac{VT}{6\epsilon}\right]\right]$ to show that this maximum is at least $\frac{2VT}{27\epsilon} - \frac{4}{9} \geq \frac{VT}{27\epsilon}$. Thus the lower bound on the covering number is $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \geq e^{\frac{VT}{54\epsilon}}$, $\forall\, \epsilon \leq \frac{VT}{12}$.

General Function Classes

1. Introduction: Here we provide lower and upper bounds on the quantity $\max_P \log \mathcal{N}\left(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)\right)$ for the general classes $\mathcal{F}$ of $\{0,1\}$-valued functions defined on the set X, where the max is taken over all the probability distributions P on X. The bounds are given in terms of the scale-sensitive dimension $\mathcal{F}$ (Alon, Ben-David, Cesa-Bianchi, and Haussler (1993)).
2. Fat-Shattering Coefficient: Define $fat_{\mathcal{F}}(\epsilon) = max\{n: some\ x \in X^n\ is\ \epsilon - shattered\ by\ \mathcal{F}\}$, where a sequence $x \in X^n$ is ϵ-shattered by $\mathcal{F}$ if there is a sequence $r \in [0,1]^n$ such that, for all $b \in [0,1]^n$, there is an $f \in \mathcal{F}$ with $f(x_i) = \begin{cases} \geq r_i + \epsilon & if\ b_i = 1 \\ \leq r_i - \epsilon & otherwise \end{cases}$. For example, it is easy to show that, bounded variation functions $\mathcal{F}_1$ have $fat_{\mathcal{F}_1}(\epsilon) = \left[\!\left[\frac{V}{2\epsilon}\right]\!\right]$. To see the upper bound, consider an ϵ-shattered sequence, and for a suitable sequence r, find the sequence b for which the corresponding function has maximum variation. An easy construction with $r = 0$ gives the lower bound.
3. Upper/Lower Bounds - Statement: There are constants c_1 and c_2 such that, for any permissible (i.e., benign measurability condition (Pollard (1984))) class $\mathcal{F}$ of $\{0,1\}$-valued functions defined on a set X, $\frac{fat_{\mathcal{F}}(4\epsilon)}{32} \leq \max_P \log \mathcal{N}\left(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)\right) \leq c_1 fat_{\mathcal{F}}(c_2\epsilon)\left[log_2 \frac{1}{\epsilon}\right]^2$ (Bartlett, Kulkarni, and Posner (1997)).
4. Gap Between Upper and Lower Bounds: There is a gap between the lower and the upper bounds. The class $\mathcal{F}_1$ 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 n, d, let $\mathcal{F}_{d,n}$ be the class of all functions from $\{1,2,\ldots,d\}$ to $\left\{0, \frac{1}{n}, \ldots, 1\right\}$, and let P be the uniform distribution on $\{1,2,\ldots,d\}$. Then $fat_{\mathcal{F}_{d,n}}(\epsilon) = d$ for $\epsilon \leq \frac{1}{2}$, yet for $\epsilon = \frac{1}{2nd}$ we have $\log \mathcal{N}\left(\epsilon, \mathcal{F}_{d,n}, \mathcal{L}_1(dP)\right) > dlog_2 \frac{1}{2d\epsilon}$.
5. Function Range Scaling: Clearly, the statement above can be trivially extended to classes of functions that map to an arbitrary interval $[a,b]$ by scaling ϵ by a factor $b-a$.
6. 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 $fat_{\mathcal{F}}$: In the first step, we show that a bound on $fat_{\mathcal{F}}$ implies that, for any finite sequence $x = (x_1, \ldots, x_m)$ from X, there is a small subset of $\mathcal{F}$ that is a cover of the restriction of $\mathcal{F}$ to x, denoted $\mathcal{F}_{|x} = \{[f(x_1), \ldots, f(x_m)] : f \in \mathcal{F}\} \subseteq \mathcal{R}^m$. In defining the cover and the covering numbers $\mathcal{N}(\epsilon, \mathcal{F}_{|x})$, we use the scaled $\mathbb{l}_1$ metric on $\mathcal{R}^m$ defined by $\rho(a, b) = \frac{1}{m}\sum_{i=1}^{m}|a_i - b_i|$. We can also consider $\mathcal{F}_{|x}$ as a class of functions from $\{1, \ldots, m\}$ to $\mathcal{R}$; this is how we define $fat_{\mathcal{F}_{|x}}$.
2. 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 $fat_{\mathcal{F}_{|x}}$.
3. 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 P, the cover for the restriction of $\mathcal{F}$ to the sequence induces a cover for $\mathcal{F}$.

General Function Class Bounds – Lemmas

1. Lemma #1 - Statement: Suppose $x \in X^m$, and $\mathcal{F}$ is a set of $\{0, 1\}$-valued functions on X. Let $d = fat_{\mathcal{F}_{|x}}\left(\frac{\epsilon}{4}\right)$, where $\epsilon > 0$. If $n \geq 2dlog_2\frac{64e^2}{\epsilon \log 2}$, there is a subset $\mathcal{T}$ of $\mathcal{F}$ for which $\mathcal{T}_{|x}$ is an ϵ-cover of $\mathcal{F}_{|x}$, and $|\mathcal{T}| \leq 2\left(\frac{16}{\epsilon}\right)^{6dlog_2\frac{32en}{d\epsilon}}$.

2. Lemma #1 - Proof: This lemma is implicit in Bartlett and Long (1995), which provides a slightly better bound - by a factor of $\log d$ - than that given in Alon, Ben-David, Cesa-Bianchi, and Haussler (1993).
3. Lemma #2 - Statement: For a class $\mathcal{F}$ of $\{0,1\}$-valued functions, let $|\mathcal{F}-\mathcal{F}| = \{|f_1 - f_2|: f_1, f_2 \in \mathcal{F}\}$. Then, with the metric $\mathcal{L}_1(dP)$ on $\mathcal{F}$, $\mathcal{N}\left(\epsilon, |\mathcal{F}-\mathcal{F}|, \mathcal{L}_1(dP)\right) \leq \left[\mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{F}, \mathcal{L}_1(dP)\right)\right]^2$.
4. Lemma #2 - Proof: Take an $\frac{\epsilon}{2}$-cover $\mathcal{T}$ for $\mathcal{F}$. Then for all $f_1, f_2 \in \mathcal{F}$, pick $t_1, t_2 \in \mathcal{T}$ within of f_1 and f_2, respectively. We have $\|\|f_1 - f_2| - |t_1 - t_2|\|_{\mathcal{L}_1(dP)} \leq \|\|f_1 - t_1| + |f_2 - t_2|\|_{\mathcal{L}_1(dP)} \leq \epsilon$. It follow that $\{|t_1 - t_2| : t_1, t_2 \in \mathcal{T}\}$ is an ϵ-cover for $\mathcal{F}$.
5. Lemma #3 – Statement: For a permissible class $\mathcal{G}$ of $\{0,1\}$-valued functions on a set X, a probability distribution P on X, and an $\epsilon > 0$, $P_m\left\{x \in X^m: \exists\, g \in \mathcal{G}, \left|\frac{1}{m}\sum_{i=1}^m g(x_i) - \mathbb{E}[g]\right| > \epsilon\right\} \leq 4 \max_{x \in X^m} \mathcal{N}\left(\frac{\epsilon}{16}, \mathcal{G}_{|x}\right) e^{-\frac{m\epsilon^2}{128}}$.
6. 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 $\mathcal{G}$ that we will consider is $|\mathcal{F}-\mathcal{F}|$.

General Function Class – Upper Bounds

1. Reduction of the Probability Bound: We start by establishing the following chain of inequalities:
   a. $P_m\left\{x \in X^m: \exists\, f_1, f_2 \in \mathcal{F}, \left|\frac{1}{m}\sum_{i=1}^m |f_1(x_i) - f_2(x_i)| - \mathbb{E}[|f_1 - f_2|]\right| > \frac{\epsilon}{2}\right\}$
   b. Use Lemma #3 to get $P_m\left\{x \in X^m: \exists\, f_1, f_2 \in \mathcal{F}, \left|\frac{1}{m}\sum_{i=1}^m |f_1(x_i) - f_2(x_i)| - \mathbb{E}[|f_1 - f_2|]\right| > \frac{\epsilon}{2}\right\} \leq 4 \max_{x \in X^m} \mathcal{N}\left(\frac{\epsilon}{32}, |\mathcal{F}-\mathcal{F}|_{|x}\right) e^{-\frac{m\epsilon^2}{128\times 4}}$
   c. Use Lemma #2 to get $P_m\left\{x \in X^m: \exists\, f_1, f_2 \in \mathcal{F}, \left|\frac{1}{m}\sum_{i=1}^m |f_1(x_i) - f_2(x_i)| - \mathbb{E}[|f_1 - f_2|]\right| > \frac{\epsilon}{2}\right\} \leq 4 \max_{x \in X^m} \mathcal{N}\left(\frac{\epsilon}{64}, \mathcal{F}_{|x}\right) e^{-\frac{m\epsilon^2}{512}}$

d. Finally use Lemma #1 for $d = fat_{\mathcal{F}_{|x}}\left(\frac{\epsilon}{256}\right)$ and $m \geq 2dlog_2\frac{64e^2}{\epsilon \log 2}$ to get $P_m\left\{x \in X^m: \exists\, f_1, f_2 \in \mathcal{F}, \left|\frac{1}{m}\sum_{i=1}^m |f_1(x_i) - f_2(x_i)| - \mathbb{E}[|f_1 - f_2|]\right| > \frac{\epsilon}{2}\right\} \leq 16\left(\frac{1024}{\epsilon}\right)^{12dlog_2\frac{2048me}{\epsilon d}} e^{-\frac{m\epsilon^2}{512}}$.

2. Sample Size Bounds: It is easy to show that the probability P_m is less than 1 for $m \geq \frac{k_1 d}{\epsilon^3}$, where k_1 is a constant. In fact, a more detailed estimation will show that $m \geq \frac{k_1 d}{\epsilon^2}\left[\log\log\frac{1}{\epsilon}\right]$ will be sufficient.
3. Cover for the Restriction of $\mathcal{F}$: For the above sample size, it follows that there is a $x \in X^m$ such that any $\mathcal{T} \subseteq \mathcal{F}$ for which $\mathcal{T}_{|x}$ is an $\frac{\epsilon}{2}$–cover for $\mathcal{F}_{|x}$ is also an ϵ-cover for $\mathcal{F}$. To see this, notice that for all $f \in \mathcal{F}$ there is a $t \in \mathcal{T}$ with $\frac{1}{m}\sum_{i=1}^m |t(x_i) - f(x_i)| \leq \frac{\epsilon}{2}$, and if x is chosen in the complement of the set that generates P_m above, then $\left|\frac{1}{m}\sum_{i=1}^m |t(x_i) - f(x_i)| - \mathbb{E}[|t - f|]\right| \leq \frac{\epsilon}{2}$, so that $\mathbb{E}[|t - f|] \leq \frac{\epsilon}{2}$. In fact, for sufficiently large m, a proper cover for the restriction of $\mathcal{F}$ to almost any m-sequence induces a cover of $\mathcal{F}$.
4. The Upper Bound - Final Form: Together with Lemma #1, the above statement shows that some $\mathcal{T} \subseteq \mathcal{F}$ satisfies $log_2\mathcal{N}\left(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)\right) \leq log_2|\mathcal{T}| \leq 1 + c_1 fat_{\mathcal{F}}(c_2\epsilon)\left[log_2\frac{1}{\epsilon}\right]^2$, from which the result follows.

General Function Class – Lower Bounds

1. The Lower Bound: Suppose $fat_{\mathcal{F}}(4\epsilon) \geq d$. Then consider the uniform distribution on a 4ϵ-shattered set of size d, and consider the restriction of the class $\mathcal{F}$ to this set. The same argument as the proof for the lower bound for the functions of bounded variation gives $\mathcal{N}\left(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)\right) \geq e^{\frac{d}{32}}$.

Operator Theory Methods for Entropy Numbers

1. Classical Statistical Theory Viewpoint: Under the traditional viewpoint of the statistical learning theory, one is given a class of functions $\mathcal{F}$, and the generalization performance attainable using $\mathcal{F}$ is determined via the covering numbers of $\mathcal{F}$.
2. Uniform Covering Numbers of $\mathcal{F}$: More precisely, for some set $\mathcal{X}$, and $\vec{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$, define *Uniform Covering Numbers* of the function class $\mathcal{F}$ on $\mathcal{X}$ by $\mathcal{N}_m(\epsilon, \mathcal{F}) \doteq \sup_{\vec{x}_1, \ldots, \vec{x}_m \in \mathcal{X}} \mathcal{N}_m\left(\epsilon, \mathcal{F}, l_\infty^{\vec{X}_m}\right)$ where is the ϵ-covering number of $\mathcal{F}$ with respect to $l_\infty^{\vec{X}_m}$. Recall $\vec{X}_m = (\vec{x}_1, \ldots, \vec{x}_m)$. Many generalization bounds can be expressed in terms of $\mathcal{N}_m(\epsilon, \mathcal{F})$.
3. 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.
4. Bounding Covering Numbers: These dimensions reduce the computation of $\mathcal{N}_m(\epsilon, \mathcal{F})$ to that of a single "dimension" like quantity independent of m. 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).
5. Information Theory Viewpoint:
    a. The Kernel Operator => An alternate view is that of a function class $\mathcal{F}$ being induced by an operator T_k that depends on some kernel function k - thus $\mathcal{F}$ is the image of a base class $\mathcal{G}$ under T_k.
    b. Covering Number in terms of T_k Properties => Thus, the determination of $\mathcal{N}_m(\epsilon, \mathcal{F})$ can be done in terms of the properties of the operator T_k, as the latter plays a constructive role in controlling the complexity of $\mathcal{F}$, 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)).

Operator Theory Methods for Entropy Numbers - Literature

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. ϵ-entropies of Linear Operators and Stochastic Processes: The connection between Prosser's ϵ-entropy of an operator and Kolmogorov's ϵ-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 ϵ-entropy in Gaussian Processes *Kodai Mathematical Journal* **9** 58-67.
- Akashi, S. (1990): The Asymptotic Behavior of ϵ-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 (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.
- Anthony, M., and P. L. Bartlett (1999): *Artificial Neural Network Learning: Theoretical Foundations* **Cambridge University Press** Cambridge UK.
- 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.
- 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.
- Cesa-Bianchi, N., and D. Haussler (1996): A Graph-theoretic Generalization of the Sauer-Shelah Lemma *Internal Report 170 96* **DSI Universita di Milano**.
- Defant, M., and M. Junge (1993): Characterization of Weak Type by the Entropy Distribution of r-nuclear Operators *Studia Math.* **107 (1)** 1-14.
- 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.
- 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.
- 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.
- Jagerman, D. (1969): ϵ-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. (1956): Asymptotic Characteristics of some Completely Bounded Metric Spaces *Dokl. Acad. Nauk. SSSR* **108** 585-589.
- Kolmogorov, A. N., and V. M. Tihomirov (1961): ϵ-entropy and ϵ -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.
- Koltchinskii (1988): Operators of Type p 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): ϵ-entropy, ϵ-rate, and Interpolation Spaces revisited with an Application to Linear Communication Channels *Journal of Mathematical Analysis and Applications* **186** 265-276.
- 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.
- Lorentz, G. G., M. van Golitschek, and Y. Makovoz (1996): *Constructive Approximation: Advanced Problems* **Springer-Verlag** Berlin Germany.
- Pajor, A. (1985): *Sous-Espaces l_n^1 des Espaces de Banach* **Hermann** Paris, France.
- Pietsch, A. (1980): *Operator Ideals* **North-Holland** Amsterdam.
- Pollard, D. (1984): *Convergence of Stochastic Processes* **Springer** New York.
- Pontriagin, L. S., and L. G. Schnirelmann (1932): Sur une Propriete Metrique de la Dimension *Annals of Mathematics* **33** 156-162.
- Prosser, R. T. (1966): The ϵ-entropy and ϵ-capacity of certain Time Varying Channels *Journal of Mathematical Analysis and Applications* **16** 553-573.
- Prosser, R. T., and W. L. Root (1968): The ϵ-entropy and ϵ-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.
- Talagrand, M. (1996): The Glivenko-Cantelli Problem, 10 years later *Journal of Theoretical Probability* **9 (2)** 371-384.
- Triebel, H. (1970): Interpolationseigenschaften von Entropie- und Durchmesseridealen Kompakter Operatoren *Studia Math.* **34** 89-107.
- 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 $\mathcal{P}_m(f) \doteqdot \frac{1}{m}\sum_{i=1}^{m} f(\vec{x}_i)$ to denote the *empirical means* of f on the sample $\vec{x}_1, \ldots, \vec{x}_m$.
3. Regression Uniform Convergence Bound: Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[0, 1]$, and let $\mathcal{P}$ be a distribution over $\mathcal{X}$. Then for all $\epsilon > 0$ and $m \geq \frac{2}{\epsilon^2}$, $\mathbb{P}\left\{ \sup_{f \in \mathcal{F}} |\mathcal{P}_m(f) - \mathcal{P}(f)| > \epsilon \right\} \leq 12m.\mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon}{6}, \mathcal{F}, l_\infty^{\vec{X}_{2m}}\right)\right] e^{-\frac{m\epsilon^2}{36}}$ where $\mathbb{P}$ denotes the probability with respect to the sample $\vec{x}_1, \ldots, \vec{x}_m$ drawn i.i.d. from $\mathcal{P}$, and $\mathbb{E}$ the expectation with respect to the first, and a second sample also drawn i.i.d. from $\mathcal{P}$, i.e., $\vec{X}_{2m} = \{\vec{x}_1, \ldots, \vec{x}_{2m}\}$ (Alon, Ben-David, Cesa-Bianchi, and Haussler (1997)).
4. Use of the Regression Bound Lemma: To be able to use the above lemma, one usually makes use of the fact that for any $\mathcal{P}$, $\mathbb{E}\left[\mathcal{N}\left(\epsilon, \mathcal{F}, l_\infty^{\vec{X}_{2m}}\right)\right] < \sup_{\vec{x}_1, \ldots, \vec{x}_m \in \mathcal{X}} \mathcal{N}\left(\epsilon, \mathcal{F}, l_\infty^{\vec{X}_{2m}}\right) < \mathcal{N}_m(\epsilon, \mathcal{F})$.

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 $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[a, b]$, with $a < b$, $a, b \in \mathbb{R}$, and $l: \mathbb{R} \to [0, \infty)$ a loss function. Let the following conditions/definitions hold:
    a. $\vec{X}_m = (\vec{x}_1, \dots, \vec{x}_m)$
    b. $\vec{z}_i \doteqdot (\vec{x}_i, y_i)$
    c. $\vec{Z}_m = (\vec{z}_1, \dots, \vec{z}_m)$
    d. $l_{f|\vec{z}_j} \doteqdot l\left(f\left(\vec{x}_j\right) - y_j\right)$
    e. $l_{f|\vec{Z}_m} \doteqdot \left(l_{\mathcal{F}|\vec{z}_j}\right)_{j=1}^{m}$
    f. $l_{\mathcal{F}|\vec{Z}_m} \doteqdot \left\{l_{f|\vec{z}_j} : f \in \mathcal{F}\right\}$
    g. $\mathcal{N}\left(\epsilon, l_{|\vec{Z}_m}\right) \doteqdot \mathcal{N}\left(\epsilon, l_{\mathcal{F}|\vec{Z}_m}, l_{\infty}^{\vec{Z}_m}\right)$
3. Lipschitz Loss Condition Lemma: Suppose l satisfies the Lipschitz condition $l(\xi) - l(\xi') \leq C|\xi - \xi'|$ for all $\xi, \xi' \in [a - b, b - a]$. Then for all $\epsilon > 0$, $\max_{\vec{Z}_m \in \{\mathcal{X} \times [a,b]\}^m} \mathcal{N}\left(\epsilon, l_{|\vec{Z}_m}\right) \leq \max_{\vec{X}_m \in \mathcal{X}_m} \mathcal{N}\left(\frac{\epsilon}{C}, \mathcal{F}_{|\vec{X}_m}, l_{\infty}^{m}\right)$, and $\max_{\vec{Z}_m \in \{\mathcal{X} \times [a,b]\}^m} \mathcal{N}\left(\epsilon, l_{|\vec{Z}_m}\right) \leq \max_{\vec{X}_m \in \mathcal{X}_m} \mathcal{N}\left(\frac{\epsilon m}{C}, \mathcal{F}_{|\vec{X}_m}, l_{\infty}^{1}\right)$.
4. Approximate Lipschitz Loss Condition Lemma: Suppose that for some $C, \tilde{C} > 0$, l satisfies the "approximate Lipschitz condition" $l(\xi) - l(\xi') \leq max\left(C|\xi - \xi'|, \tilde{C}\right)$ for all $\xi, \xi' \in [a - b, b - a]$, then for all $\epsilon \geq \frac{\tilde{C}}{C}$ $\max_{\vec{Z}_m \in \{\mathcal{X} \times [a,b]\}^m} \mathcal{N}\left(\epsilon, l_{|\vec{Z}_m}\right) \leq \max_{\vec{X}_m \in \mathcal{X}_m} \mathcal{N}\left(\frac{\epsilon}{C}, \mathcal{F}_{|\vec{X}_m}, l_{\infty}^{m}\right)$.
5. Lipschitz Loss Lemma Proof: We show that, for any sequence $\vec{Z}_m$ of $(\vec{x}, y)$ pairs in $\mathcal{X} \times [a, b]$ and any functions f and g, if the restrictions of f and g to $\vec{X}_m$ are close, then the

restrictions of l_f and l_g to $\vec{Z}_m$ are closer. Thus, given a cover of $\mathcal{F}_{|\vec{X}_m}$ we can construct a cover of $l_{\mathcal{F}|\vec{Z}_m}$ that is no bigger.

6. Lipschitz Loss Condition Bound: $\frac{1}{m}\left|\sum_{j=1}^{m}\{l[g(\vec{x}_j)-y_j]-l[f(\vec{x}_j)-y_j]\}\right| \leq \frac{1}{m}\{\sum_{j=1}^{m}|l[g(\vec{x}_j)-y_j]-l[f(\vec{x}_j)-y_j]|\} \leq \frac{1}{m}\sum_{j=1}^{m} C|g(\vec{x}_j)-f(\vec{x}_j)| = \frac{C}{m}\|g(\vec{X}_m)-f(\vec{X}_m)\|_{l_\infty^1} \leq C\|g(\vec{X}_m)-f(\vec{X}_m)\|_{l_\infty^m}$.
7. 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. $\frac{1}{m}\left|\sum_{j=1}^{m}\{l[g(\vec{x}_j)-y_j]-l[f(\vec{x}_j)-y_j]\}\right| \leq \frac{C}{m}\sum_{j=1}^{m} max\left[g(\vec{x}_j)-f(\vec{x}_j),\frac{\tilde{C}}{C}\right] \leq C\epsilon$ for $\epsilon \geq \frac{\tilde{C}}{C}$.
8. Polynomial Loss Function Bound: For either of the Lipschitz loss conditions above, if we assume a loss function of the type $l(\eta)=\frac{1}{p}\eta^p$ with $p>1$, we have $\mathcal{C}=(b-a)^{p-1}$; in particular $\mathcal{C}=b-a$ for $p=2$, and therefore $\underset{\vec{Z}_m \in \{\mathcal{X}\times[a,b]\}^m}{max} \mathcal{N}(\epsilon, l_{|\vec{Z}}) \leq \underset{\vec{x} \in \mathcal{X}_m}{max} \mathcal{N}\left(\frac{\epsilon}{(b-a)^{p-1}}, \mathcal{F}_{|\vec{x}}\right)$.

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 $\mathcal{P}_m\left\{\underset{f \in \mathcal{F}}{sup}\left|\mathcal{R}_{Emp}(f)-\mathcal{R}(f)\right|>\epsilon\right\} \leq c_1(m)\mathcal{N}_m(\epsilon,\mathcal{F})e^{-\frac{m\epsilon^\beta}{c_2}}$ where $\mathcal{R}_{Emp}(f)$ is the empirical risk, and $\mathcal{R}(f)$ is the expected risk of $f \in \mathcal{F}$ (Vapnik (1998), Anthony and Bartlett (1999)).
2. Exponents in the Standard Form: The exponent β depends on the setting. For regression, β can be set to 1. However, in agnostic learning in general, $\beta=2$ (Kearns, Schapire, and Sellie (1994)), except if the class is convex, in which case β can be set to 1 (Lee, Bartlett, and Williamson (1998)).

3. Sample Complexity: The generalization bounds produced above are typically used by setting the right hand side to δ, and solving for $m \equiv m(\epsilon, \delta)$. m is called the sample complexity.
4. Learning Curve Bound: Another way to use the above result is as a learning curve bound $\bar{\epsilon}(\delta, m)$, where $\mathcal{P}_m \left\{ \sup_{f \in \mathcal{F}} \left| \mathcal{R}_{Emp}(f) - \mathcal{R}(f) \right| > \bar{\epsilon}(\delta, m) \right\} \leq \delta$. It is to be noted that the determination of $\bar{\epsilon}(\delta, m)$ is quiet convenient in terms of e_n, the dyadic entropy number associated with the covering number $\mathcal{N}(\epsilon, \mathcal{F})$.
5. The Dyadic Entropy Number: Setting the right hand side in the standard uniform convergence form to δ, we have $\delta = c_1(m)\mathcal{N}_m(\epsilon, \mathcal{F})e^{-\frac{m\epsilon^\beta}{c_2}} \rightarrow \log\left[\frac{\delta}{c_1(m)}\right] + \frac{m\epsilon^\beta}{c_2} = \log \mathcal{N}_m(\epsilon, \mathcal{F}) \rightarrow \epsilon \geq e_{\left\lfloor \log\left[\frac{\delta}{c_1(m)}\right] + \frac{m\epsilon^\beta}{c_2} + 1 \right\rfloor}$. Thus, $\bar{\epsilon}(\delta, m) = \min\left\{ \epsilon : \epsilon \geq e_{\left\lfloor \log\left[\frac{\delta}{c_1(m)}\right] + \frac{m\epsilon^\beta}{c_2} + 1 \right\rfloor} \right\}$ holds. Therefore the use of e_n or ϵ_n 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.

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)).
5. 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.
6. 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.
7. 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 $(\vec{x}_1, y_1)$, …, $(\vec{x}_m, y_m) \in \mathbb{R}^d \times \{\pm 1\}$, the maximum margins are achieved by minimizing $\|\vec{w}\|_2$ subject

to the constraints $y_j\left(\langle \vec{w}, \vec{x}_j \rangle + b\right) \geq 1$ for $j = 1, \ldots, m$. The decision functions then take on the form $f(\vec{x}) = sgn\ (\langle \vec{w}, \vec{x} \rangle + b)$.

3. Maximum Margin Linear Regression: Likewise, a linear regression $f(\vec{x}) = \langle \vec{w}.\vec{x} \rangle + b$ can be estimated from the data $(\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$ by locating the flattest function that approximates the data with some margin of error. In this case one minimizes $\|\vec{w}\|_2$ subject to $\left|f\left(\vec{x}_j\right) - y_j\right| \leq \epsilon$, where the parameter $\epsilon > 0$ plays the role of the margin, although not in the space of inputs $\vec{x}$, but in the space of the outputs y.
4. 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 do 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 k (these are listed later), there exists a non-linear map Φ into a reproducing kernel Hilbert space F (e.g., Saitoh (1988)) such that k computes the dot product in F, i.e., $k(\vec{x}, \vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_F$.
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 $f(\vec{x}) = \sum_{j=1}^{m} \alpha_j k_j(\overrightarrow{x_j}, \vec{x}) + b$ where $\alpha_j \in \mathbb{R}, j = 1, \ldots, m$.

5. 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 $R_{Reg}[f] = R_{Emp}[f] + \frac{\lambda}{2}\|\mathcal{P}f\|^2$ (with a regularization parameter $\lambda \geq 0$, and a regularization operator $\mathcal{P}$) over the set of functions of the form $f(\vec{x})$ shown above, provided that k and $\mathcal{P}$ are inter-related through $k(\overrightarrow{x_s}, \overrightarrow{x_t}) = \langle (\mathcal{P}k)(\overrightarrow{x_s}, \cdot), (\mathcal{P}k)(\overrightarrow{x_t}, \cdot) \rangle$. To this end, k is chosen as the Green's function of $\mathcal{P}^*\mathcal{P}$ where $\mathcal{P}^*$ is the adjoint of $\mathcal{P}$.
6. 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 suitable 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 $k(x, y) = e^{-(x-y)^2}$ which do not have a discrete spectrum.
3. 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.
4. 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 $k(\vec{x}, \vec{y})$ 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 $\mathcal{X}$ under the feature map $\Phi(.)$. 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 $(\mathcal{X}, \mu)$ is a finite measure space, i.e., $\mu(\mathcal{X}) < \infty$.
3. Mercer Theorem - Notation: Suppose $k \in L_\infty(\mathcal{X}, \mathcal{X})$ is a symmetric kernel (i.e., $k(x, x') = k(x', x)$) such that the integral operator T_k: $L_2(\mathcal{X}) \to L_2(\mathcal{X}) \Longrightarrow T_k[f(\cdot)] \div \int_{\mathcal{X}}^{\mathcal{X}} k(\cdot, \vec{y})\, f(\vec{y}) d\mu(\vec{y})$ is positive. Let $\psi_j \in L_\infty(\mathcal{X})$ be the eigen-function of T_k associated with the eigenvalue $\lambda_j \neq 0$ and normalized by $\|\psi_j\|_{L_2} = 1$. Finally, suppose ψ_j is continuous for all $j \in \mathbb{N}$.
4. Mercer's Condition: Using the notation above, we can specify Mercer's conditions:
   a. $\left(\lambda_j(T)\right)_j \in l_1$ for $j = 1, 2, \ldots$
   b. $\psi_j \in L_\infty(\mathcal{X})$ and $\sup_j \|\psi_j\|_{L_\infty} < \infty$
   c. $k(\vec{x}, \vec{y}) = \sum_{j \in \mathbb{N}} \lambda_j \psi_j(\vec{x}) \psi_j(\vec{y})$ holds for all $(\vec{x}, \vec{y}) \in \mathcal{X} \times \mathcal{X}$
   d. Finally, the series $\lambda_j \psi_j(\vec{x}) \psi_j(\vec{y})$ converges absolutely and uniformly for all $(\vec{x}, \vec{y}) \in \mathcal{X} \times \mathcal{X}$
5. Assumption of Continuity of ψ_j: Note that if $\mathcal{X}$ is compact and k is continuous, then ψ_j is continuous (Ash (1965)). Alternatively, if k is translation-invariant, then ψ_j are scaled cosine functions, and thus continuous. Therefore the assumption that ψ_j are continuous is not very restrictive.

6. Impact of Mercer's Second Condition: From the second condition it follows that there exists some constant $C_k \in \mathbb{R}^+$ depending on $k(\cdot,\cdot)$ such that $\left|\psi_j(\vec{x})\right| \leq C_k$ for all $j \in \mathbb{N}$ and $\vec{x} \in \mathcal{X}$.
7. Impact of Mercer's Third Condition: From the third condition it follows that $k(\vec{x},\vec{y})$ corresponds to a dot-product in l_2, i.e., $k(\vec{x},\vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_{l_2}$ with $\Phi: \mathcal{X} \to l_2 \Longrightarrow \Phi: \vec{x} \mapsto \left(\phi_j(\vec{x})\right)_j \doteq \left(\sqrt{\lambda_j}\psi_j(\vec{x})\right)_j$.
8. Spatial Span of $\Phi(\mathcal{X})$: From the above argument one can see that $\Phi(\mathcal{X})$ lives not only in l_2, but in an axis parallel parallelepiped with lengths $2C_k\sqrt{\lambda_j}$. Also we assume, without loss of generality, that the sequence $\left(\lambda_j\right)_j$ is sorted in a non-increasing order.

Equivalent Kernels

1. The Equivalent Kernel Lemma: Denote by $\mathcal{X}$ a compact set, and by $k: \mathcal{X}^2 \to \mathbb{R}$ a Mercer kernel. Then for any $\mathcal{X}'$ and a surjective map $\chi: \mathcal{X} \to \mathcal{X}'$, the kernel $k(x,x') = k[\chi(x), \chi(x')]$ also satisfies Mercer's condition, and, moreover, the eigenvalues $\lambda_i{}'$ and the coefficient $C_{k'}$ of the integral operator $T_{k'}[f(x)] \doteq \int_{\mathcal{X}'}^{\mathcal{X}'} k'(x,x')\, f(x')dx'$ can be used equivalently in any application of k.
2. Proof of the Lemma: The first part of the claim, namely that k' also satisfies Mercer's condition, follows immediately from the construction of k'. For the second claim, note that due to the fact that χ is surjective for any distribution $p(x)$ on $\mathcal{X}$ there must exist an equivalent distribution $p'(x)$ on $\mathcal{X}'$. Therefore we can always consider the problem as being one on $\mathcal{X}'$ from the start.
3. Impact of the Lemma: Note that since $\mathcal{X}$ and $\mathcal{X}'$ were chosen arbitrarily, we can optimize over them. In particular, this means that we could construct diffeomorphisms $\chi : \mathcal{X} \to \mathcal{X}'$ and look for the function χ such that the eigenvalues $\lambda_i{}'$ and $C_{k'}$ are as small as possible.
4. Dependence on μ: Note that the measure μ 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 μ since that will affect ψ_j, C_k, and $\left(\lambda_j\right)_j$. The question of the optimal μ to use

and how it may be chosen if one knows P (the distribution from which $\vec{x}$ are chosen) requires a full treatment on in its own right, but it is common for μ to be a Lebesgue measure.

Mapping Φ Into l_2

1. Motivation: It will be useful to consider maps that map $\Phi(\mathcal{X})$ into balls of some radius R centered at the origin. The following proposition shows that the class of all these maps is determined by the elements of l_2 and the sequence of eigenvalues $(\lambda_j)_j$.
2. l_2 Bounding of $\Phi(\mathcal{X})$ - Statement: Let $\mathcal{S}$ be a diagonal map $\mathcal{S}: \mathbb{R}^n \to \mathbb{R}^n \Longrightarrow \mathcal{S}: (\lambda_j)_j \mapsto \mathcal{S}(\lambda_j)_j = (s_j\lambda_j)_j$ with $s_j \in \mathbb{R}$. Then $\mathcal{S}$ maps $\Phi(\mathcal{X})$ into a ball of finite radius $R_{\mathcal{S}}$ centered at the origin if and only if $(s_j\sqrt{\lambda_j})_j \in l_2$.
3. Sufficiency of $\mathcal{S}[\Phi(\mathcal{X})] \subseteq l_2$: Suppose $(s_j\sqrt{\lambda_j})_j \in l_2$ and let $R_{\mathcal{S}}{}^2 \doteq C_k{}^2 \left\|(s_j\sqrt{\lambda_j})_j\right\|_{l_2}^2 < \infty$. For any $\vec{x} \in \mathcal{X}$, $\|\mathcal{S}[\Phi(\vec{x})]\|_{l_2}^2 = \sum_{j\in\mathbb{N}} s_j{}^2\lambda_j\left|\psi_j(\vec{x})\right|^2 \leq \sum_{j\in\mathbb{N}} s_j{}^2\lambda_j C_k{}^2 = R_{\mathcal{S}}{}^2$. Hence $\mathcal{S}[\Phi(\mathcal{X})] \subseteq l_2$.
4. Necessity of $\mathcal{S}[\Phi(\mathcal{X})] \subseteq l_2$: Suppose $(s_j\sqrt{\lambda_j})_j$ is not in l_2. Hence the sequence $(A_n)_n$: $A_n \doteq \sum_{j=1}^n s_j{}^2\lambda_j$ is unbounded. Now define $a_n(\vec{x}) \doteq \sum_{j=1}^n s_j{}^2\lambda_j\left|\psi_j(\vec{x})\right|^2$. Then $\|a_n(\vec{x})\|_{L_1(\mathcal{X})}$ due to the normalization condition on $\psi_j(\vec{x})$. However, as $\mu(\mathcal{X}) < \infty$ there exists a set $\widetilde{\mathcal{X}}$ of non-zero measure such that $a_n(\vec{x}) \geq \frac{A_n}{\mu(\mathcal{X})}$ for all $\vec{x} \in \widetilde{\mathcal{X}}$. Combining the expression for $\|\mathcal{S}[\Phi(\vec{x})]\|_{l_2}^2$ from the sufficiency condition with the one for $a_n(\vec{x})$, we obtain $\|\mathcal{S}[\Phi(\vec{x})]\|_{l_2}^2 \geq a_n(\vec{x})$ for all $n \in \mathbb{N}$ and all $\vec{x}$. Since $a_n(\vec{x})$ is unbounded for a set $\widetilde{\mathcal{X}}$ with non-zero measure in $\mathcal{X}$, we can see that $\mathcal{S}[\Phi(\mathcal{X})] \not\subseteq l_2$.

l_2 Unit Ball $\to \mathcal{E}$ Mapping Scaling Operator $\widehat{A}$

1. Motivation for the Construction of $\hat{A}$: Once we know that $\Phi(\mathcal{X})$ is contained in the parallelepiped described above, we can use the result of the above proposition to construct a mapping $\hat{A}$ from the unit ball in l_2 to an ellipsoid $\mathcal{E}$ such that $\Phi(\mathcal{X}) \subset \mathcal{E}$ as listed in the schematic below.
2. Metric Space Mapping Schematic:
    a. $\mathcal{X} \overset{\Phi}{\to} \Phi(\mathcal{X}) \subset l_2$
    b. $\Phi(\mathcal{X}) \subset l_2 \overset{\hat{\mathrm{A}}^{-1}}{\longrightarrow} U_{l_2} \subset l_2$
    c. $U_{l_2} \subset l_2 \overset{\hat{\mathrm{A}}}{\to} \mathcal{E} \subset l_2$
    d. Remember that $\Phi(\mathcal{X}) \subset l_2 \subset [\mathcal{E} \subset l_2]$
3. Use of $\hat{A}$: The operator $\hat{A}$ 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.
4. Approach towards constructing $\hat{A}$: We thus seek an operator $\hat{A} : l_2 \to l_2$ such that $\hat{\mathrm{A}}^{-1}[\Phi(\mathcal{X})] \subset U_{l_2}$. This means that $\mathcal{E} \doteqdot \hat{A}U_{l_2}$ will be such that $\Phi(\mathcal{X}) \subset \mathcal{E}$.The latter can be ensured by constructing $\hat{A}$ such that $\hat{A}$: $\left(x_j\right)_j \mapsto \left(R_{\hat{A}} \cdot a_j \cdot x_j\right)_j$ with $R_{\hat{A}}, a_j \in \mathbb{R}^+$, where a_j and C_k are chosen with respect to a specific kernel, and where $R_{\hat{A}} \doteqdot C_k \left\| \left(\frac{\sqrt{\lambda_j}}{a_j} \right)_j \right\|_{l_2}$. From the necessary/sufficient propositions above, it follows that all those operators $\hat{A}$ for which $R_{\hat{A}} < \infty$ will satisfy the criterion $\hat{\mathrm{A}}^{-1}[\Phi(\mathcal{X})] \subset U_{l_2}$. We call such scaling/inverse operators *admissible*.

Unit Bounding Operator Entropy Numbers

1. Computing the Entropy Numbers of $\hat{A}$: Here we bound the entropy number for the operator $\hat{A}$, 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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_j \geq \cdots \geq 0$ be a non-increasing sequence of non-negative numbers, and let $\mathcal{D}\vec{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots)$ for $\vec{x} = (x_1, x_2, \ldots, x_j, \ldots) \in l_p$ be the diagonal operator from l_p into itself, generated by the sequence $(\sigma_j)_j$, where $1 \leq p \leq \infty$. The, for all $n \in \mathbb{N}$, $\sup_{j \in \mathbb{N}} \left(\frac{\sigma_1 \sigma_2 \ldots \sigma_j}{n}\right)^{\frac{1}{j}} \leq \epsilon_n(\mathcal{D}) \leq 6 \sup_{j \in \mathbb{N}} \left(\frac{\sigma_1 \sigma_2 \ldots \sigma_j}{n}\right)^{\frac{1}{j}}$.
3. Minimizing the Entropy Numbers by Choosing $\hat{A}$: We can exploit the freedom in choosing $\hat{A}$ 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.
4. Scaling Operators Lemma: Let $k : \mathcal{X} \times \mathcal{X}$ be a Mercer kernel with eigenvalues $(\lambda_s)_s$. Choose $a_j > 0$ for $j \in \mathbb{N}$ such that $\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2$, and define $A: (x_j)_j \mapsto (R_A \cdot a_j \cdot x_j)_j$ with $R_A \doteqdot C_k \left\| \left(\frac{\sqrt{\lambda_j}}{a_j}\right)_j \right\|_{l_2}$. Then $\epsilon_n(A: l_2 \to l_2) \leq \sup_{j \in \mathbb{N}} 6 C_k \left\| \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \right\|_{l_2} \left(\frac{\sigma_1 \sigma_2 \ldots \sigma_j}{n}\right)^{\frac{1}{j}}$.
5. Use of the Scaling Operator Lemma: The scaling operator lemma follows immediately by identifying $\mathcal{D}$ in the diagonal bounds statement with $\hat{A}$ in the scaling operator lemma. We can optimize $\left\| \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \right\|_{l_2}$ over all possible choices of A to obtain the optimal entropy number for $\hat{A}$. It turns out that the optimum for $\left\| \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \right\|_{l_2}$ (i.e., the infimum over $(a_s)_s$) is attainable when k 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.
6. Optimal Entropy Number for $\hat{A}$: There exists an $\hat{A}$ defined by $\hat{A}: (x_j)_j \mapsto (R_{\hat{A}} \cdot a_j \cdot x_j)_j$ with $R_{\hat{A}}, a_j \in \mathbb{R}^+$ that satisfies $\epsilon_n(\hat{A}) \leq \inf_{(a_s)_s: \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2} \sup_{j \in \mathbb{N}} 6 C_k \left\| \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \right\|_{l_2} \left(\frac{\sigma_1 \sigma_2 \ldots \sigma_j}{n}\right)^{\frac{1}{j}}$.

The SVM Operator

1. The SVM Feature Space: To recap, the hypothesis that an SVM generates may be expressed as $\langle \vec{w}, \overrightarrow{x^{\sim}} \rangle + b$, where both $\vec{w}$ and $\overrightarrow{x^{\sim}}$ are defined in the feature space $S = Span[\Phi(\mathcal{X})]$ and $b \in \mathbb{R}$.
2. 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 b for now, we consider the class $\mathcal{F}_\Lambda \doteqdot \{f_{\vec{w}} : \vec{x} \mapsto \langle \vec{w}, \vec{x} \rangle : \vec{x} \in \mathcal{S}, \|\vec{w}\| \leq \Lambda\} \subseteq \mathbb{R}^{\mathcal{S}}$. Note that $\mathcal{F}_\Lambda$ depends implicitly on k since $\mathcal{S}$ does.
3. l_∞^m Covering Numbers for $\mathcal{F}_\Lambda$: We seek the l_∞^m covering numbers for $\mathcal{F}_\Lambda$ induced by the kernel in terms of the parameter Λ which is the inverse of the size of the margin in the feature space as defined by the dot-product in $\mathcal{S}$ (see Vapnik and Chervonenkis (1974) and Vapnik (1995) for details).
4. 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 $\mathcal{T}$ as $\mathcal{T} = \mathcal{S}_{\overrightarrow{X_m}^{\sim}} \Lambda$ where $\Lambda \in \mathbb{R}^+$ and the operator $\mathcal{S}_{\overrightarrow{X_m}^{\sim}}$ is defined by $\mathcal{S}_{\overrightarrow{X_m}^{\sim}} : l_2 \to l_\infty^m \Longrightarrow \mathcal{S}_{\overrightarrow{X_m}^{\sim}} : \vec{w} \mapsto (\langle \overrightarrow{x_1}, \vec{w} \rangle, \ldots, \langle \overrightarrow{x_m}, \vec{w} \rangle)$ with $\overrightarrow{x_j}^{\sim} \in \Phi(\mathcal{X})$ for all j. 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 $\mathcal{T}$ and $\hat{A}$. 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 $\mathcal{S} \in \mathfrak{L}(\mathcal{H}, l_\infty^m)$ where $\mathcal{H}$ is a Hilbert Space. Then there exists a constant $c > 0$ such that for all $n, m \in \mathbb{N}$, $e_n(\mathcal{S}) \leq c\|\mathcal{S}\| \sqrt{\frac{\log\left(1+\frac{m}{n}\right)}{n}}$.
3. n vs. Input Dimensionality: Carl and Stephani (1990) state an additional condition, namely, that $n \leq m$. It turns out that for $n > m$, 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 ϵ-covering number for an extremely small ϵ for which $\log \mathcal{N}_m(\epsilon, \mathcal{F}) > m$.

4. Estimate for the Constant c: Williamson, Smola, and Scholkopf (2000) provide proof of a fairly tight bound for the constant of $c \leq 103$; however, there is reason to believe that c should be 1.86, the constant obtained for identity maps from l_2^m to l_∞^m.
5. Maurey's Theorem for Entropy Numbers (instead of Dyadic Entropy): The re-statement of Maurey's theorem in terms of $\epsilon_{2^{n-1}}(\mathcal{S}) = e_n(\mathcal{S})$, under the assumptions above, becomes $\epsilon_n(\mathcal{S}) \leq c\|\mathcal{S}\| \sqrt{\frac{\log\left(1+\frac{m}{1+\log n}\right)}{1+\log n}}$.
6. Carl and Stephani's Lemma (Carl and Stephani (1990)): Let E, F, and G be Banach spaces, with $\mathcal{R} \in \mathfrak{L}(F, G)$ and $\mathcal{S} \in \mathfrak{L}(E, F)$. Then for all $n, t \in \mathbb{N}$:
    a. $\epsilon_{nt}(\mathcal{RS}) \leq \epsilon_n(\mathcal{R}) \cdot \epsilon_t(\mathcal{S})$
    b. $\epsilon_n(\mathcal{RS}) \leq \epsilon_n(\mathcal{R}) \cdot \|\mathcal{S}\|$
    c. $\epsilon_n(\mathcal{RS}) \leq \epsilon_n(\mathcal{S}) \cdot \|\mathcal{R}\|$
    d. Note that b) and c) follow directly from a), and the fact that $\epsilon_1(\mathcal{R}) = \|\mathcal{R}\|$ for all $\mathcal{R} \in \mathfrak{L}(F, G)$, and $\epsilon_1(\mathcal{S}) = \|\mathcal{S}\|$ for all $\mathcal{S} \in \mathfrak{L}(E, F)$.

Bounds for SV Classes

1. Setup: Let k be a Mercer kernel, let Φ be induced from $\Phi: \mathcal{X} \to l_2 \Rightarrow \Phi: \vec{x} \mapsto \left(\phi_j(\vec{x})\right)_j \doteq \left(\sqrt{\lambda_j}\psi_j(\vec{x})\right)_j$, and let $\mathcal{T} = \mathcal{S}_{\widetilde{X_m}}\Lambda$, where $\mathcal{S}_{\widetilde{X_m}}$ is given from $\mathcal{S}_{\widetilde{X_m}}: l_2 \to l_\infty^m \Rightarrow \mathcal{S}_{\widetilde{X_m}}: \vec{w} \mapsto (\langle \overrightarrow{x_1}, \vec{w} \rangle, \dots, \langle \overrightarrow{x_m}, \vec{w} \rangle)$, and $\Lambda \in \mathbb{R}^+$. Let A be defined from $\epsilon_n(A) \leq \inf_{(a_s)_s: \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2} \sup_{j \in \mathbb{N}} 6C_k \left\|\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s\right\|_{l_2} \left(\frac{\sigma_1\sigma_2\dots\sigma_j}{n}\right)^{\frac{1}{j}}$, and suppose $\widetilde{\overrightarrow{x_j}} \mapsto \Phi\left(\overrightarrow{x_j}\right)$ for $j = 1, \dots, m$.
2. The SV Bounds: Then, the entropy numbers of $\mathcal{T}$ satisfy the following inequalities (c here is the same as specified earlier):

a. $\epsilon_n(\mathcal{T}) \leq c\|A\|\Lambda\sqrt{\frac{\log\left(1+\frac{m}{\log n}\right)}{\log n}}$

b. $\epsilon_n(\mathcal{T}) \leq 6\Lambda\epsilon_n(A)$

c. $\epsilon_{nt}(\mathcal{T}) \leq 6c\Lambda\sqrt{\frac{\log\left(1+\frac{m}{\log n}\right)}{\log n}}\epsilon_t(A)$

3. Optimal Bound for $\epsilon_n(\mathcal{T})$: The above set of bound provide several options for bounding $\epsilon_n(\mathcal{T})$. The optimal inequality bound to use depends on the rate of decay of the eigenvalues of k. The result gives effective bounds on $\mathcal{N}_m(\epsilon, \mathcal{F}_\Lambda)$ since $\epsilon_n(\mathcal{T}: l_2 \to l_\infty^m) \leq \epsilon_0 \Longrightarrow \mathcal{N}_m(\epsilon, \mathcal{F}_\Lambda) \leq n$.
4. Factorization of $\mathcal{T}$ to the Upper Bound $\epsilon_n(\mathcal{T})$: We will us the following factorization of $\mathcal{T}$ to upper-bound $\epsilon_n(\mathcal{T})$.

a. $U_{l_2} \subset l_2 \xrightarrow{\mathcal{T}} l_\infty^m$

b. $U_{l_2} \subset l_2 \xrightarrow{\Lambda} \Lambda U_{l_2} \subset l_2$

c. $\Lambda U_{l_2} \subset l_2 \xrightarrow{\mathcal{S}_{\Phi(\vec{X}_m)}} l_\infty^m$

d. $U_{l_2} \subset l_2 \xrightarrow{\mathrm{A}} \Lambda\mathcal{E} \subset l_2$

e. $\Lambda\mathcal{E} \subset l_2 \xrightarrow{\mathcal{S}_{\mathrm{A}^{-1}\Phi(\vec{X}_m)}} l_\infty^m$

5. Description of the Factorization Schematic: The statement a) follows from the definition of $\mathcal{T}$. The fact that b) – d) commutes, as in e), stems from the fact that since A is diagonal, it is self-adjoint, and so for any $\overrightarrow{x^\sim} \in \mathcal{S}$ $\langle \vec{w}, \overrightarrow{x^\sim} \rangle = \langle \vec{w}, A\mathrm{A}^{-1}\overrightarrow{x^\sim} \rangle = \langle A\vec{w}, \mathrm{A}^{-1}\overrightarrow{x^\sim} \rangle$.
6. Alternate Representation of $\mathcal{T}$: Instead of computing the entropy number of $\mathcal{T} = \mathcal{S}_{\overrightarrow{X_m^\sim}}\Lambda$ directly, which is difficult or wasteful, as the bound on $\mathcal{S}_{\overrightarrow{X_m^\sim}}$ does not take into account that $\overrightarrow{x^\sim} \in \mathcal{E}$, but simply assumes that $\overrightarrow{x^\sim} \in \rho U_{l_2}$ for some $\rho > 0$, we will represent $\mathcal{T}$ as $\mathcal{S}_{\mathrm{A}^{-1}\overrightarrow{X_m^\sim}}\mathrm{A}\Lambda$. This is more efficient as we constructed A such that $\mathrm{A}^{-1}\Phi(\mathcal{X}) \subseteq U_{l_2}$, filling a larger proportion of it than $\frac{1}{\rho}\Phi(\mathcal{X})$ does.
7. Proof of the SV Bounds: By construction of A, and due to the Cauchy-Schwartz inequality, we have $\left\|\mathcal{S}_{\mathrm{A}^{-1}\overrightarrow{X_m^\sim}}\right\| \leq 1$. Thus, applying the Carl and Stephani's lemma to the factorization of this version of $\mathcal{T}$, and using the Maurey's theorem, we get the SV bounds.

Asymptotic Rates of Decay for $\epsilon_n(\mathcal{T})$

1. Introduction: Eventually we seek asymptotic rates of decay for $\epsilon_n(A)$ – 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 $\epsilon_n(\mathcal{T})$ in terms of $\epsilon_n(A)$. By asymptotic here we mean asymptotic in n; this corresponds to asking how $\mathcal{N}_m(\epsilon, \mathcal{F})$ scales as $\epsilon \to 0$ for fixed m.
2. Rates Bounds on $\epsilon_n(\mathcal{T})$: Let k be a Mercer kernel and suppose that A is the scaling operator defined by for $j \in \mathbb{N}$ such that $\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2$, and define $A: (x_j)_j \mapsto (R_A \cdot a_j \cdot x_j)_j$ with $R_A \doteqdot C_k \left\| \left(\frac{\sqrt{\lambda_j}}{a_j}\right)_j \right\|_{l_2}$ such that for $j \in \mathbb{N}$ $a_j > 0$, and $\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2$.
    a. If $\epsilon_n(A) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\alpha}\right)$ for some $\alpha > 0$, then for fixed m $\epsilon_n(\mathcal{T}) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\alpha+\frac{1}{2}}\right)$.
    b. If $\log \epsilon_n(A) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\beta}\right)$ for some $\beta > 0$, then for fixed m $\epsilon_n(\mathcal{T}) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\beta}\right)$.
3. Proof Step #1 Splitting the Index n: From Maurey's theorem we know that $\epsilon_n(\mathcal{S}) = \mathcal{O}\left(\sqrt{\frac{1}{\log n}}\right)$. Using the 3rd SV Class bound, ignoring the constants and ignoring the constants and assuming m is fixed, we split the index n in the following way: $n = n^{\tau} n^{1-\tau}$ with $\tau \in (0, 1)$.
4. Proof Step #2 - Bounds for the First Case: For the first case, this yields $\epsilon_n(\mathcal{T}) \leq \epsilon_{n^{\tau}}(\mathcal{S}) \epsilon_{n^{1-\tau}}(A) = \mathcal{O}\left(\sqrt{\frac{1}{\log n^{\tau}}}\right) \cdot \mathcal{O}\left(\left[\frac{1}{\log n^{1-\tau}}\right]^{\alpha}\right) = \mathcal{O}\left(\frac{1}{\sqrt{\tau}(1-\tau)^{\alpha}} \left[\frac{1}{\log n}\right]^{\alpha+\frac{1}{2}}\right) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\alpha+\frac{1}{2}}\right)$.
5. Proof Step #3 - Bounds for the Second Case: For the first case, this yields $\epsilon_n(\mathcal{T}) \leq \epsilon_{n^{\tau}}(\mathcal{S}) \epsilon_{n^{1-\tau}}(A) = \mathcal{O}\left(\sqrt{\frac{1}{\tau \log n}}\right) \cdot \mathcal{O}\left(\left[\frac{1}{(1-\tau)\log n}\right]^{\beta}\right) = \mathcal{O}\left(\left[\frac{1}{\log n}\right]^{\beta}\right)$.
6. Impact of the Rate Bounds on $\epsilon_n(\mathcal{T})$: These bounds show that, in the first case, Maurey's theorem allows an exponent in the entropy number of $\mathcal{T}$, 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 n.

7. 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 $\epsilon_n(\mathcal{T})$ on n, the effect of kernel is so strong that it completely dominates the $\frac{1}{\sqrt{n}}$ for arbitrary Hilbert spaces. An example of such a kernel is $k(x, y) = e^{-(x-y)^2}$.

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 $(\lambda_i)_i$ 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 $k(x, y) = e^{-(x-y)^2}$, which has non-compact support.
2. Integral Operator Induced Continuous Spectrum: As in the kernel above, the induced integral operator $[\mathcal{T}_k f](x) = \int_{-\infty}^{+\infty} k(x, y)\, f(x) dx$ has a continuous spectrum (i.e., a non-denumerable infinity of eigenvalues), and, thus, $\mathcal{T}_k$ 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.
3. 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 $\lambda_j = \Omega\left(e^{-j^2}\right)$ (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.
4. Kernels with Compact Support: We first consider the case where the support of $k \subseteq [-a, +a]$ for some $a < \infty$. We also further suppose that the data points $\vec{x}_j$ satisfy $\vec{x}_j \in [-b, +b]$.
5. Kernels with Compact Support applied to SV Hypothesis: If $k(\cdot,\cdot)$ is a convolution kernel (i.e., $k(x, y) = k(x - y, 0)$), then the SV hypothesis $h_k(\cdot)$ can be written as $h_k(x) = \sum_{j=1}^{m} \alpha_j k\left(x, \vec{x}_j\right) = \sum_{j=1}^{m} \alpha_j k_v\left(x, \vec{x}_j\right) \div h_{k_v}(x)$ for $v \geq 2(a + b)$ where $k_v(\cdot,\cdot)$ is the v-periodic extension of $k(\cdot,\cdot)$ (analogously, $k_v(x, y) = k_v(x - y, 0)$) $k_v(x, 0) \div \sum_{j=-\infty}^{j=+\infty} k(x - jv, 0)$.

Eigenvalues of $\mathcal{T}_{k_v}$

1. The d-dimensional Fourier Transform Operator: The d-dimensional Fourier transform is defined by $F: L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d) \Longrightarrow F_f(\vec{w}) \doteq \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d}^{\mathbb{R}^d} e^{-i\langle \vec{w}, \vec{x} \rangle} f(\vec{x}) d\vec{x}$. Then its inverse is defined by $F^{-1}: L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d) \Longrightarrow {F^{-1}}_f(\vec{x}) \doteq \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d}^{\mathbb{R}^d} e^{i\langle \vec{w}, \vec{x} \rangle} f(\vec{w}) d\vec{w}$. From the above, F is an isometry on $L_2(\mathbb{R}^d)$.
2. The Approach: We now relate the eigenvalues of $\mathcal{T}_{k_v}$ to the Fourier transform of $k(\cdot,\cdot)$. We do so for the case of $d = 1$, and state the general cases later.
3. Fourier Transform of the Symmetric Convolution Kernel - Statement: Let $k: \mathbb{R}^1 \to \mathbb{R}^1$ be a symmetric convolution kernel, let $K(\vec{w}) = F_f(\vec{w})$ denote the Fourier transform of $k(\cdot,\cdot)$, and k_v denote the v-periodical kernel derived from k (also assume that k_v exists). Then k_v has a representation as a Fourier series with $w_0 \doteq \frac{2\pi}{v}$ and $k_v(x-y) = \sum_{j=-\infty}^{j=+\infty} \frac{\sqrt{2\pi}}{v} K(jw_0) e^{ijw_0(x-y)} = \frac{\sqrt{2\pi}}{v} K(0) + \sum_{j=-\infty}^{j=+\infty} \frac{\sqrt{8\pi}}{v} K(jw_0) cos\big(jw_0(x-y)\big)$. Moreover, the eigenvalues λ_j of $\mathcal{T}_{k_v}$ satisfy $\lambda_j = \sqrt{2\pi} K(jw_0)$ for $j \in \mathbb{Z}$ and $C_k = \sqrt{\frac{2}{v}}$.
4. Proof Step #1 - Fourier Series Coefficients of k_v: Clearly, the Fourier series coefficients K_j of k_v exist (as k_v exists), with $K_j \doteq \frac{1}{\sqrt{v}} \int_{-\frac{v}{2}}^{\frac{v}{2}} e^{-ijw_0 x} k_v(x) dx$, and therefore, by the definition of k_v and the existence of $K(w)$, we conclude $K_j \doteq \frac{1}{\sqrt{v}} \int_{-\frac{v}{2}}^{\frac{v}{2}} \sum_{l=-\infty}^{l=+\infty} \big[e^{-ijw_0 x} k(x - lv)\big] dx = \frac{1}{\sqrt{v}} \sum_{l=-\infty}^{+\infty} \left[\int_{-\frac{v}{2}}^{\frac{v}{2}} e^{-ijw_0 x} k(x - lv) dx \right] = \sqrt{\frac{2\pi}{v}} K(jw_0)$.
5. Proof Step #2 - The Orthogonal L_2 Basis: This and the fact that $\left\{ x \mapsto \frac{1}{\sqrt{v}} e^{ijw_0 x} : j \in \mathbb{Z} \right\}$ form an orthogonal basis in $L_2\left(\left[-\frac{v}{2}, +\frac{v}{2}\right], \mathbb{C}\right)$ proves that $k_v(x-y) = \frac{\sqrt{2\pi}}{v} K(0) + \sum_{j=-\infty}^{j=+\infty} \frac{\sqrt{8\pi}}{v} K(jw_0) cos\big(jw_0(x-y)\big)$.

6. Proof Step #3 - Choice of the Trigonometric Basis Function: Furthermore, we are interested in real-valued basis functions for $k(x-y)$. The functions $\psi_0(x) \doteqdot \frac{1}{\sqrt{v}}$, $\psi_j(x) \doteqdot \sqrt{\frac{2}{v}}\, cos(jw_0x)$, and $\psi_{-j}(x) \doteqdot \sqrt{\frac{2}{v}}\, sin(jw_0x)$ for all $j \in \mathbb{N}$ satisfy $\left\|\psi_j\right\|_{L_2} = 1$, $j \in \mathbb{Z}$, and form an eigen-system of the integral operator defined by k_v with the corresponding eigenvalues $\sqrt{2\pi}K(jw_0)$. Finally, one can see that $C_k = \sqrt{\frac{2}{v}}$ by computing the maximum over $j \in \mathbb{N}$ and $x \in \left[-\frac{v}{2}, +\frac{v}{2}\right]$.
7. Extension to non-compact $\mathcal{T}_k$ Support: Thus, even though $\mathcal{T}_k$ may not be compact, $\mathcal{T}_{k_v}$ can be (if $\left(K(jw_0)\right)_{j\in\mathbb{N}} \subset l_2$, for example). The above Fourier transform can be applied whenever we can form $k_v(\cdot)$ from $k(\cdot)$. Clearly, $k(x) = \mathcal{O}(x^{-1-\epsilon})$ for some $\epsilon > 0$ suffices to ensure that the sum $k_v(x,0) \doteqdot \sum_{j=-\infty}^{j=+\infty} k(x-jv,0)$ converges.
8. 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 v

1. Asymptotics of $K(w)$: From the Riemann-Lebesgue lemma, it is clear that for integrable $k(\cdot)$ of bounded variations (surely any kernel one would choose would satisfy that criterion), one has $K(w) = \mathcal{O}\left(\frac{1}{w}\right)$.
2. Tradeoff in the Choice of v: There is a tradeoff in choosing v in that, for large enough w, $K(w)$ is a decreasing function of w (at least as fast as $\frac{1}{w}$), and thus, from $\lambda_j = \sqrt{2\pi}K(jw_0)$, we get that $\lambda_j = \sqrt{2\pi}K\left(\frac{2\pi j}{v}\right)$ is an increasing function of v. This suggests that one should choose a small value of v. But a small v will lead to high empirical error (as the kernel "wraps around" and its localization properties are lost, i.e., its $VCDim$ becomes high) and large C_k.

3. An Approach for picking v: There are several approaches for picking a value for v. One obvious way is to *a priori* pick some $\overline{\epsilon} > 0$ and choose the smallest v such that $|k(x) - k_v(x)| \leq \overline{\epsilon}$ for all $x \in \left[-\frac{v}{2}, +\frac{v}{2}\right]$. Thus, one would obtain a hypothesis $h_{k_v}(x)$ uniformly within $C\overline{\epsilon}$ of $h_k(x)$ where $\sum_{j=1}^{m}|\alpha_j| \leq C$.

Extension to d-dimension

1. The Statement: Assume that $k(\vec{x})$ is v-periodic in each direction $[\vec{x} = (x_1, \ldots, x_d)]$, we get $\lambda_j = (2\pi)^{\frac{d}{2}}K(\vec{j}w_0) = (2\pi)^{\frac{d}{2}}K(w_0\|\vec{j}\|)$ for radially symmetric k, and finally for the eigen-functions we get $C_k = \left(\frac{2}{v}\right)^{\frac{d}{2}}$.
2. Kernel Bandwidth Choice: Here we examine how a difference choice for the bandwidth of the kernel, i.e., after letting $k_\sigma(\vec{x}) \doteqdot \sigma^d k(\sigma\vec{x})$, affects the eigen-spectrum of the corresponding operator. We have $K_\sigma(\vec{w}) = K\left(\frac{\vec{w}}{\sigma}\right)$, hence scaling a kernel by σ means more densely spaced eigenvalues in the spectrum of the integral operator $\mathcal{T}_{k_\sigma}$.

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 Give Decay Rates

Asymptotic/Non-asymptotic Decay of Covering Numbers

1. Introduction: Here we show how the asymptotic behavior of $\epsilon_n(A: l_2 \to l_2)$, where A is the scaling operator introduced earlier, depends on the eigenvalues of $\mathcal{T}_k$.
2. 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 $L_2(\mathcal{X}, \mathbb{C})$. Furthermore, while the analysis here states the propositions as asymptotic results as Prosser (1966) does, the proofs provide non-asymptotic details with explicit constants.
3. 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 $\hat{A}$, i.e., $\epsilon_n(\hat{A}) \leq \inf_{(a_s)_s: \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2} \sup_{j \in \mathbb{N}} 6C_k \left\|\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s\right\|_{l_2} \left(\frac{\sigma_1 \sigma_2 \dots \sigma_j}{n}\right)^{\frac{1}{j}}$. If the eigenvalues were unsorted, one could obtain far too small numbers for the geometrical mean of $\lambda_1, \lambda_2, \dots, \lambda_j$.
4. 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., $k(x, y) = k(x - y)$, there is nothing in the formulations of the propositions themselves that requires this. When we consider the d-dimensional case, we shall see that with rotationally invariant kernels, degenerate systems of eigenvalues are generic.
5. Explicit Decay of $(\lambda_j)_j$: We shall consider the specific cases where $(\lambda_j)_j$ decays asymptotically in some polynomial or exponential degree. In this case, we choose the sequence A for which we can evaluate $\epsilon_n(A) \leq \inf_{(a_s)_s: \left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s \in l_2} \sup_{j \in \mathbb{N}} 6C_k \left\|\left(\frac{\sqrt{\lambda_s}}{a_s}\right)_s\right\|_{l_2} \left(\frac{\sigma_1 \sigma_2 \dots \sigma_j}{n}\right)^{\frac{1}{j}}$ explicitly. In what follows, by the eigenvalues of kernel k we refer to the sorted eigenvalues of the induced integral operator $\mathcal{T}_k$.

References

- Prosser, R. T. (1966): The ϵ-entropy and ϵ-capacity of certain Time-Varying Channels *Journal of Mathematical Analysis and Applications* **16** 553-573.

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.
6. 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 $\mathfrak{I}$ and a set of training points, we try to pick out the function $f \in \mathfrak{I}$ that minimizes the following expression: $\mathcal{L}(f) + \mathcal{L}(Training\ Data \mid f)$ where $\mathcal{L}(f)$ is the length of the code used to encode the function $f \in \mathfrak{I}$.
2. $\mathcal{L}(Training\ Data \mid f)$: The term $\mathcal{L}(Training\ Data \mid f)$ denotes the length of the code needed to express the given training data with the aid of the function f. The idea is simple; if f fits the training data well, this part of the code will be very small.
3. MDL vs. Classical SLT: $\mathcal{L}(Training\ Data \mid f)$ corresponds to the training error the function f makes on the data. The term $\mathcal{L}(f)$ measures the complexity of f. 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 $\mathfrak{I}$, whereas in MDL it is a function of f.
4. Building Universal Function Codes: One way to approach the above is to decompose $\mathfrak{I}$ into subsets $\mathfrak{I}_1 \subset \mathfrak{I}_2 \subset \cdots$. Then we encode the elements of each subset with a fixed length code that assigns each member of $\mathfrak{I}_i$ the same length. If $\mathfrak{I}_i$ is finite with N elements, it is possible to encode each member $f \in \mathfrak{I}_i$ using $\log N$ bits. If $\mathfrak{I}_i$ is infinite, one may use the VC dimension (or some complexity measure) to encode the function.

5. MDL Coding Complexity: The *coding complexity* of $\mathfrak{I}_i$ can be defined as the smallest code length one can achieve for encoding the functions $f \in \mathfrak{I}_i$. It is uniquely defined, and is *universal* in the sense that it does not rely on a specific scheme.
6. Function Coding Technique: Given a function $f \in \mathfrak{I}_i$, we encode our data in several steps; we first encode the index i of the function class, then we use a uniform code to encode which element of $\mathfrak{I}_i$ the function f is, and finally code the data with help of f. The code length then becomes $\mathcal{L}(i) + \mathcal{L}(f \mid f \in \mathfrak{I}_i) + \mathcal{L}(Training\ Data \mid f)$. The goal of this segmented approach is, of course, to choose the hypothesis f that minimizes this code.
    a. Function Class Code $\mathcal{L}(i)$ => Usually for $\mathcal{L}(i)$ one chooses a uniform code of integers, as long as we are dealing with finitely many classes $\mathfrak{I}_i$. Thus, this entity is typically constant across all classes.
    b. $\mathcal{L}(f \mid f \in \mathfrak{I}_i)$ => This term is identical for all $f \in \mathfrak{I}_i$. It does not distinguish between different functions within $\mathfrak{I}_i$, but only from with functions that come from different $\mathfrak{I}_i$'s.
    c. $\mathcal{L}(Training\ Data \mid f)$ => This term explains how well the given f can explain the training data.

MDL Analyses

1. MDL for a Fixed Function Class $\mathfrak{I}$: If we are given just one fixed function class $\mathfrak{I}$ (without splitting it down into sub-classes $\mathfrak{I}_i$), the code length essentially depends on some measure of complexity of $\mathfrak{I}$ plus an additional term that accounts for how well f 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 $\mathfrak{I}_i$ – 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 $\mathcal{P}$. This probability class is described by one/more parameters, i.e., it is of the form $\mathcal{P} = \{\mathcal{P}_\alpha | \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ is the set of parameters values, and α the corresponding distribution parameters (e.g., if it is normal, this may be the mean/variance).
2. 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.
    a. Maximum Likelihood Estimate => The frequentist approach de facto estimates the likelihood of the data given the parameter α realization, i.e., $\mathcal{P}(Data \mid \alpha)$. The MLE computes the parameter that maximizes this likelihood, i.e., $\hat{\alpha}(Data) = \arg\max_{\alpha \in \mathcal{A}} \mathcal{P}(Data \mid \alpha)$. 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 $\mathcal{P}_\alpha$ which, for each parameter α, 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:
    a. Cox (1961) introduces the fundamental axioms that allow expressing beliefs using probability calculus.

b. Jaynes (2003) addresses philosophical/practical issues with these axioms.
c. O'Hagan (1994) provides a very general treatment.
d. 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 $\mathcal{P}(Data \mid \alpha)$, the likelihood term. In addition, combining $\mathcal{P}(\alpha)$ with the prior, we can also compute the posterior distribution $\mathcal{P}(\alpha \mid Data)$ as $\mathcal{P}(\alpha \mid Data) \propto \mathcal{P}(\alpha)\,\mathcal{P}(Data \mid \alpha)$.
4. MAP Estimator: The MAP approach involves choosing the $\widehat{\alpha_{MAP}}$ that maximizes the posterior probability $\widehat{\alpha_{MAP}} = \underset{\alpha \in \mathcal{A}}{\arg max}\, \mathcal{P}(\alpha \mid Data)$.
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: $-\log \mathcal{P}(\alpha \mid Data) = \underset{\alpha \in \mathcal{A}}{\arg max}\,[-\log \mathcal{P}(\alpha) - \log \mathcal{P}(Data \mid \alpha)]$. Like SLT, $\log \mathcal{P}(Data \mid \alpha)$ is the term that describes the model fit to the data.
7. $-\log \mathcal{P}(\alpha)$ 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 $\mathcal{P}(\alpha)$, thus increasing $-\log \mathcal{P}(\alpha)$. From coding theory, this indicates that, as the model complexity increases, greater is $-\log \mathcal{P}(\alpha)$. 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 $f \in \mathfrak{I}$. However, this *hits* our assumption of distribution agnosticity (we are OK with that in this section, given that we try to cater to NFT).
4. Encoding Prior Knowledge via the Loss Function l: 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").
5. 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 $f \in \mathfrak{I}$: Here the assumptions of usefulness of a classifier are encoded by the appropriate choice of $f \in \mathfrak{I}$. Generally, this is not very useful, as it still leaves significant wiggle room in the choice – the exception where the encoding via f becomes important is in Bayesian approaches.
2. Shortcomings of the Capacity Measures: While both $f \in \mathfrak{I}$ and the data points go together into the construction of the $\mathfrak{I}$ capacity/complexity measure, they do not typically weight the

contribution of a specific f (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 $\Im$, either via knowledge a *priori* or *a posteriori*.

3. Classifier Function Specific Prior Knowledge: The Bayesian approach is one way to incorporate the classifier function specific prior knowledge. A prior distribution $\pi(f)$ expresses our belief on how to place emphasis on the classifier function space. However, $\pi(f)$ should be chosen well before the statistical inference process is started.
4. 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 $1-\delta$, $R(f) \leq R_{emp}(f) + \sqrt{\frac{\log\frac{1}{\pi(f)}+\log\frac{1}{\delta}}{2n}}$. This is the simplest among the PAC bounds (Boucheron, Bousquet, and Lugosi (2005)), and does not involve the capacity term for $\Im$; instead, it penalizes the individual functions according to $\pi(f)$.
    a. 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.
5. 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.

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 $f_{Bayes} \in \mathfrak{I}$ for some known, small function space $\mathfrak{I}$, the approximation error becomes zero, and the Bayes' consistency criterion then reduces to consistency with respect to $\mathfrak{I}$. If the above is not the case, then the only other known method is to ensure that the nested function spaces contain f_{Bayes} "*in the limit*".

Nested Function Spaces

1. The Concept: As the sample size n increases, so does the complexity of the function space $\mathfrak{I}_n$. The standard construction is to choose the spaces $\mathfrak{I}_n$ such that they constitute an expanding sequence of nested function spaces, i.e., $\mathfrak{I}_1 \subset \mathfrak{I}_2 \subset \mathfrak{I}_3$ .... Two conditions are needed to ensure this classifier is Bayes consistent.
2. Zero Estimation Error: $Estimation\ Error \to 0$ as $n \to \infty$. 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., $\mathfrak{I}_n$ does not grow too fast as n increases.
3. Zero Approximation Error: $Approximation\ Error \to 0$ as $n \to \infty$. This is possible only if, for some large n, either f_{Bayes} is contained in $\mathfrak{I}_n$, or it can be approximated by it closely enough.
4. Necessary and Sufficient Condition for Bayes' Consistency: This comes from Devroye, Gyorfi, and Lugosi (1996). Let $\mathfrak{I}_1 \subset \mathfrak{I}_2 \subset \mathfrak{I}_3$ ... be a sequence of nested function spaces,

and consider the classifier set described by $f_n = \underset{f \in \mathfrak{I}}{\arg min} R_{emp}(f)$. Assume that for a distribution $\mathbb{P}$ the following conditions are satisfied:

a. The VC Dimension of the spaces in $\mathfrak{I}_n$ satisfies $\mathcal{VC}(\mathfrak{I}_n).\frac{\log n}{n} \to 0$ as $n \to \infty$, and

b. $R\left(f_{\mathfrak{I}_n}\right) \to R\left(f_{Bayes}\right)$ as $n \to \infty$.

Then the sequence of classifiers $\mathfrak{I}_n$ is Bayes' consistent.

5. Bayes' Consistency – Sample: Say we choose the function space such that $\mathcal{VC}(\mathfrak{I}_n) \approx n^{\alpha}$ for some $\alpha \in]0,1[$. Then the first condition is satisfied, as $\mathcal{VC}(\mathfrak{I}_n).\frac{\log n}{n} \approx \frac{\log n}{n^{1-\alpha}} \to 0$ as $n \to \infty$. However, if $\alpha = 1$, then $\mathcal{VC}(\mathfrak{I}_n).\frac{\log n}{n} \to \infty$ as $n \to \infty$.

Regularization

1. Penalizing Regularizer: An implicit way of working with nested function spaces is the principle of regularization. Instead of minimizing the empirical risk $R_{emp}(f)$ and the expressing the generalization ability of the resulting classifier f_n using some capacity measure of the underlying function class $\mathfrak{I}$, one can directly minimize the regularized risk expressed as $R_{Reg}(f) = R_{emp}(f) + \lambda\Omega(f)$.
2. Regularizer Behavior: $\Omega(f)$ is called the regularizer, and it is designed to punish complexity. For instance, to punish functions with large fluctuations, one can choose a $\Omega(f)$ that is small for functions that are smooth/vary smoothly, and large for functions that fluctuate a lot. Thus, for linear classifiers one chooses $\Omega(f)$ as the inverse of the margin of a function.
3. Regularizer Trade-off Constant: λ weights $R_{Emp}(f)$ vs. $\Omega(f)$. High λ implies greater penalty, and one ends up choosing functions with low empirical risk. Low λ implies more empirical risk.
4. Regularization Principle: This principle chooses the f_n that minimizes the regularization risk $R_{Reg}(f)$. 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:

a. The nested function space f_n from $\Im$ contain the Bayes' classifier
b. The impact of λ should be such that, as $n \to \infty$, the ERM hypothesis space contribution dominates the penalty contribution
c. The regularizing penalty should still have an impact at $n < \infty$, i.e., as $n \to \infty$, λ should not go to 0 too fast (Steinwart (2005)).

6. Function Class Complexity vs Function Complexity: While ERM worries about function class complexity, regularization, de facto, is concerned with the *complexity* as measured by $\Omega(f)$ on an individual function.

Achieving Zero Approximation Error

1. Approximation Theory to achieve Zero Approximation Error: Essentially we need to ensure that $f_{Bayes} \in \Im$ or that f_{Bayes} is approximatable from $\Im$ for large enough n. 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.
2. Approximation Error for Binary Classifiers: If f approximates g (where f and g are binary classifiers), and their L_1 risk is less than δ, so is their $0-1$ risk, i.e., $\mathbb{P}[f(x) \neq sgn\{g(x)\}] < \delta$. Thus, to prove that the approximation error of a function space $\Im$ is smaller than δ, we have to show that every $f \in \Im_{all}$ can be approximated up to δ in the norm by functions from $\Im$.
3. Approximation Theory Application - Real Numbers: Results of the above type are abundant in the literature. For example, if $\mathcal{X}$ 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 $\Im_n$ as the spaces of polynomials with at most a degree of d_n where d_n grows slowly with n. 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 n should be so as to ensure that the error is smaller than a certain quantity (i.e., for estimation error ≤ 0.1, we need $n = 100{,}000$ samples).
2. 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 $\mathbb{P}(x, y)$, the convergence of the approximation error is dependent on it.
3. Nested Function Space: Unless one makes additional assumptions on the true distribution, for any fixed sequence $\{\Im_n\}$ of nested function spaces, the rate of convergence of the approximation error can be arbitrarily low.
4. 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 X_i's are not i.i.d. anymore.
5. Near i.i.d. Sampling: While for nearly independent sampling such as stationary α-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)).
6. 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)).
7. 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 $\mathbb{P}(x, y)$ for which a given classifier performs better than any other, we can always construct another $\mathbb{P}'(x, y)$ 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* $\mathbb{P}(x, y)$ (uniform $\mathbb{P}(x, y)$ 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 $n \to \infty$ sample sizes independent of f (remember f translates into $\mathbb{P}(x, y)$). However, at finite sizes, some f_n may be more effective than the other.
5. 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 $\Im$ - more complexity, less limiting assumption on $\mathbb{P}(x, y)$.
6. 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 χ 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 $\mathbb{P}$ on χ, one needs a σ-algebra on the space, which is typically the Borel σ -algebra. So the hidden assumption is that $\mathbb{P}$ is a Borel measure.
3. Topological Consistency: The fact that $\mathbb{P}$ is a Borel measure means that the topology of $\mathbb{R}$ 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 $\varepsilon > 0$. For every $n \in \mathbb{N}$, and for every classifier f_n, there exists a distribution $\mathbb{P}$ with Bayes' risk 0 such that the expected risk of f_n is greater than $\frac{1}{2} - \varepsilon$.
2. Statement on Algorithmic Consistency: An algorithm is consistent if for any probability measure $\mathbb{P}$, $\lim_{n\to\infty} R(f_n) = R^*$ almost surely.

3. No Free Lunch At All: For any algorithm, and any sequence (a_n) that converges to 0, there exists a probability distribution $\mathbb{P}$ such that $R^* = 0$ and $R(f_n) \leq a_n$.
4. VC Consistency of ERM: The ERM algorithm is consistent if for any probability measure $\mathbb{P}$, $R(f_n) = R(f^*)$ in probability, and $R_n(f_n) = R(f^*)$ in probability.
5. VC non-trivial Consistency of ERM: The ERM algorithm is non-trivially consistent for the set $\Im$ and the probability distribution $\mathbb{P}$ if for any $c \in \mathbb{R}$, $\inf_{f \in \Im: Pf > c} P_n(f) \rightarrow \inf_{f \in \Im: Pf > c} P(f)$ 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*.