

# An Introduction to Statistical Learning Theory

John Shawe-Taylor

Centre for Computational Statistics  
and Machine Learning

Department of Computer Science

UCL Engineering

University College London

`jst@cs.ucl.ac.uk`

June, 2011

Lammhult Summer School, June 2011

# STRUCTURE

## PART A

1. General Statistical Considerations
2. Basic PAC Ideas and proofs
3. Real-valued Function Classes and the Margin

## PART B

1. Rademacher complexity and Main Theory
2. Applications to classification
3. Conclusions

## Aim:

- Some thoughts on link between theory and practice
- Basic Techniques with some deference to history
- Insights into proof techniques and statistical learning approaches
- Concentration inequalities and relation to Rademacher approach

## What won't be included:

- The most general results
- Complete History
- Analysis of Bayesian inference
- Most recent developments, eg PAC-Bayes, local Rademacher complexity, etc.

# PART A

# Statistical learning Theory

- Statistical Learning Theory aims to analyse key quantities of interest to the practitioner:
  - Bounds of performance on new data
  - Bounds that will hold for the particular training set, not an average over possible training sets
  - Bounds that apply to the very flexible classes of functions used in practice
  - Bounds that can drive improved learning algorithms
  - Bounds that take into account all aspects of the data analysis

# Theories of learning

- Aim of any theory is to model real/ artificial phenomena so that we can better understand/ predict/ exploit them.
- SLT is just one approach to understanding/ predicting/ exploiting learning systems, others include Bayesian inference, inductive inference, statistical physics, traditional statistical analysis.

## Theories of learning cont.

- Each theory makes assumptions about the phenomenon of learning and based on these derives predictions of behaviour
  - every predictive theory automatically implies a possible algorithm
  - simply optimise the quantities that improve the predictions
- Each theory has strengths and weaknesses
  - generally speaking more assumptions, potentially more accurate predictions
  - BUT depends on capturing the right aspects of the phenomenon



# General statistical considerations

- Statistical models (not including Bayesian) begin with an assumption that the data is generated by an underlying distribution  $P$  typically not given explicitly to the learner.
- If we are trying to classify cancerous tissue from healthy tissue, there are two distributions, one for cancerous cells and one for healthy ones.

## General statistical considerations cont.

- Usually the distribution subsumes the processes of the natural/artificial world that we are studying.
- Rather than accessing the distribution directly, statistical learning typically assumes that we are given a ‘training sample’ or ‘training set’

$$S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$$

generated identically and independently (i.i.d.) according to the distribution  $P$ .

## Generalisation of a learner

- Assume that we have a learning algorithm  $\mathcal{A}$  that chooses a function  $\mathcal{A}_{\mathcal{F}}(S)$  from a function space  $\mathcal{F}$  in response to the training set  $S$ .
- From a statistical point of view the quantity of interest is the random variable:

$$\epsilon(S, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{(\mathbf{x}, y)} [\ell(\mathcal{A}_{\mathcal{F}}(S), \mathbf{x}, y)],$$

where  $\ell$  is a ‘loss’ function that measures the discrepancy between  $\mathcal{A}_{\mathcal{F}}(S)(\mathbf{x})$  and  $y$ .

## Generalisation of a learner

- For example, in the case of classification  $\ell$  is 1 if the two disagree and 0 otherwise, while for regression it could be the square of the difference between  $\mathcal{A}_{\mathcal{F}}(S)(\mathbf{x})$  and  $y$ .
- We refer to the random variable  $\epsilon(S, \mathcal{A}, \mathcal{F})$  as the generalisation of the learner.
- Note is random because of the dependence on the training set  $S$ .

## Example of Generalisation I

- We consider the Breast Cancer dataset from the UCI repository.
- Use the simple Parzen window classifier: weight vector is

$$\mathbf{w}^+ - \mathbf{w}^-$$

where  $\mathbf{w}^+$  is the average of the positive training examples and  $\mathbf{w}^-$  is average of negative training examples. Threshold is set so hyperplane bisects the line joining these two points.

## Example of Generalisation II

- Given a size  $m$  of the training set, by repeatedly drawing random training sets  $S$  we estimate the distribution of

$$\epsilon(S, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{(\mathbf{x}, y)} [\ell(\mathcal{A}_{\mathcal{F}}(S), \mathbf{x}, y)],$$

by using the test set error as a proxy for the true generalisation.

- We plot the histogram and the average of the distribution for various sizes of training set – initially the whole dataset gives a single value if we use training and test as the all the examples, but then we plot for training set sizes:

342, 273, 205, 137, 68, 34, 27, 20, 14, 7.

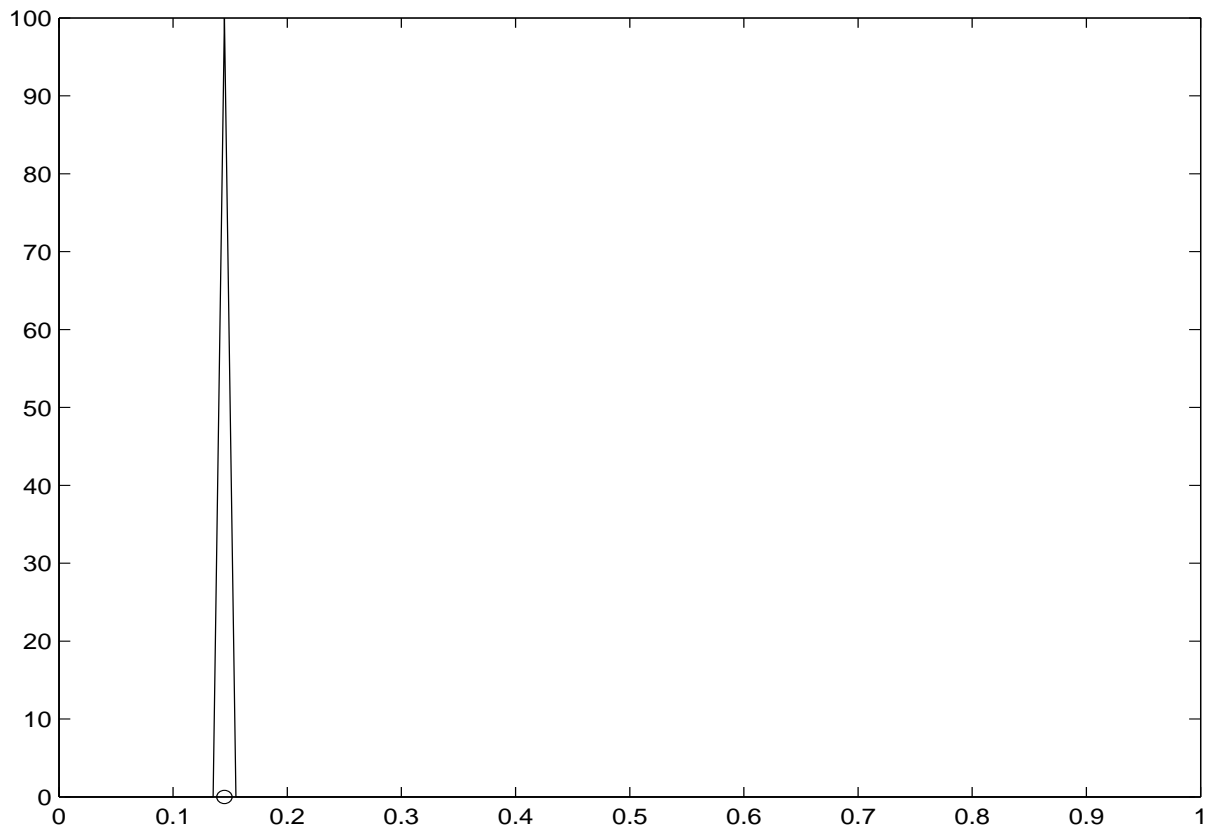
## Example of Generalisation III

- Since the expected classifier is in all cases the same:

$$\begin{aligned}\mathbb{E}[\mathcal{A}_{\mathcal{F}}(S)] &= \mathbb{E}_S[\mathbf{w}_S^+ - \mathbf{w}_S^-] \\ &= \mathbb{E}_S[\mathbf{w}_S^+] - \mathbb{E}_S[\mathbf{w}_S^-] \\ &= \mathbb{E}_{y=+1}[\mathbf{x}] - \mathbb{E}_{y=-1}[\mathbf{x}],\end{aligned}$$

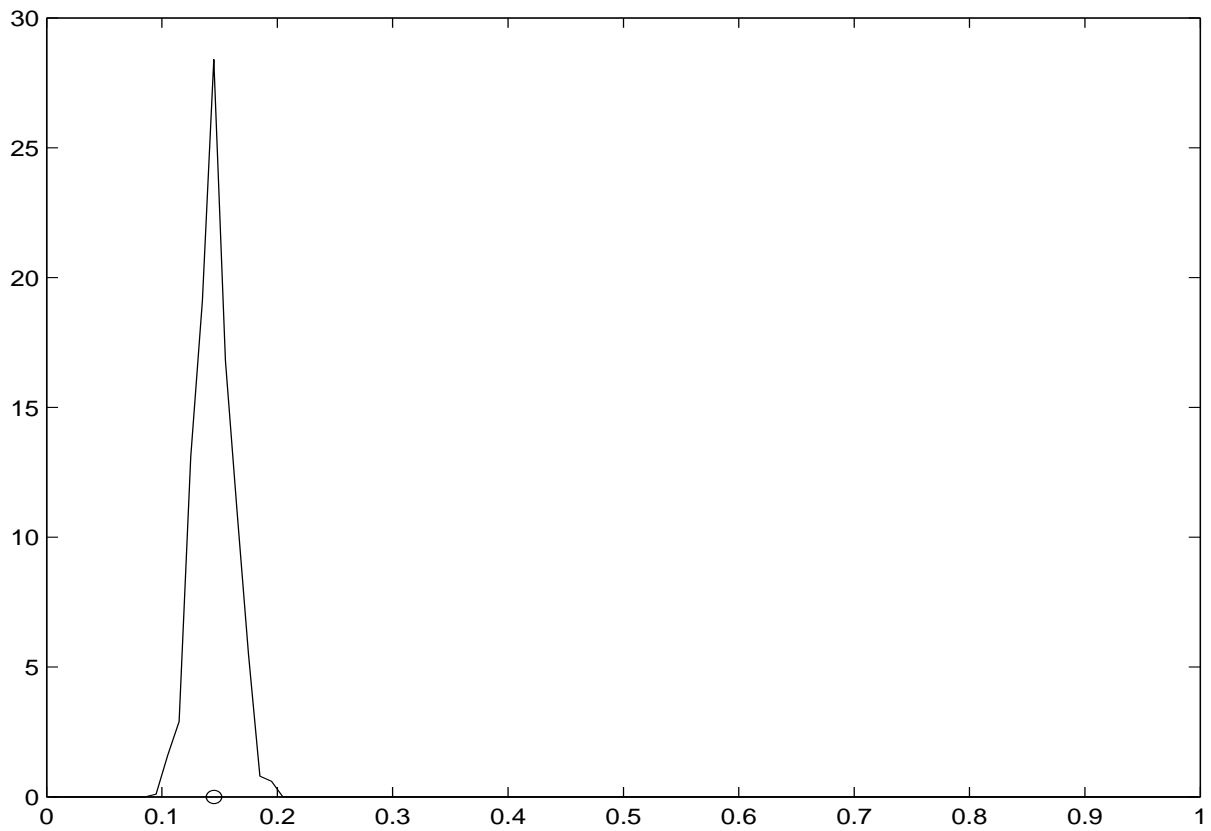
we do not expect large differences in the average of the distribution, though the non-linearity of the loss function means they won't be the same exactly.

# Error distribution: full dataset

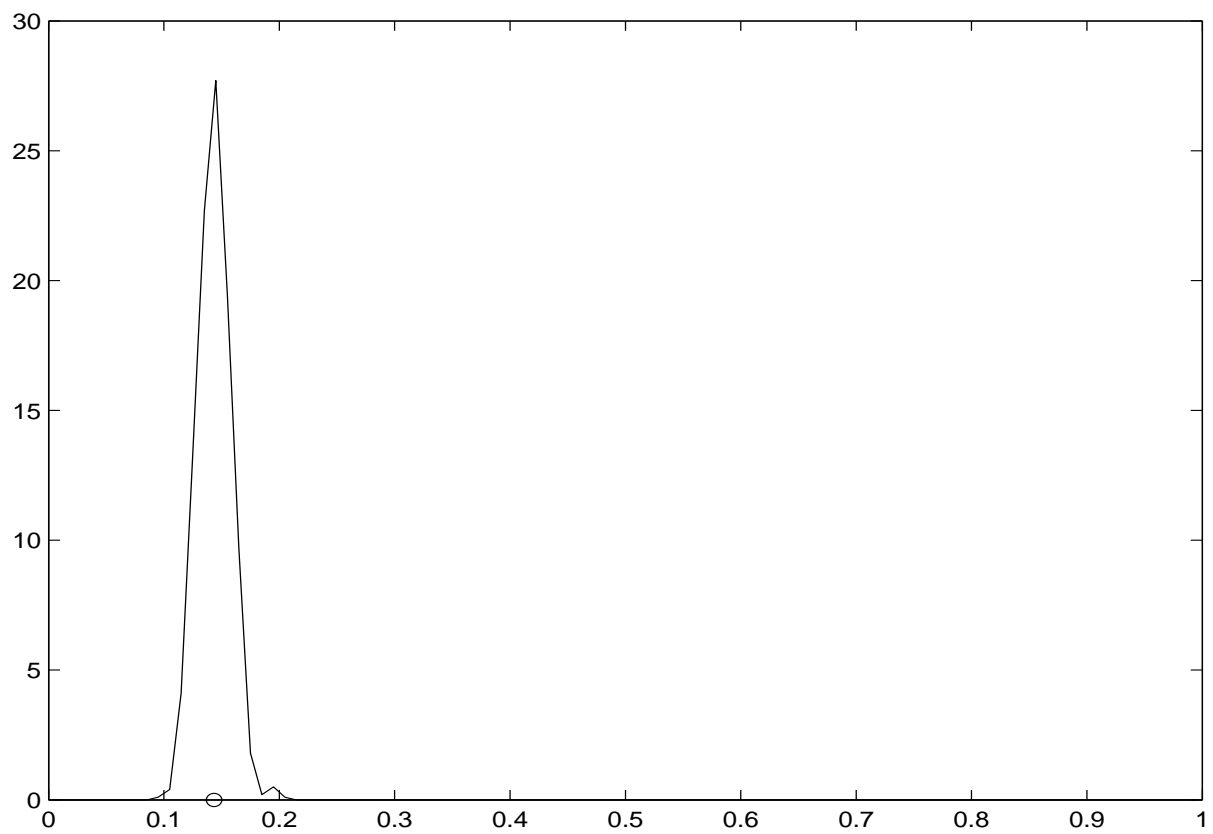




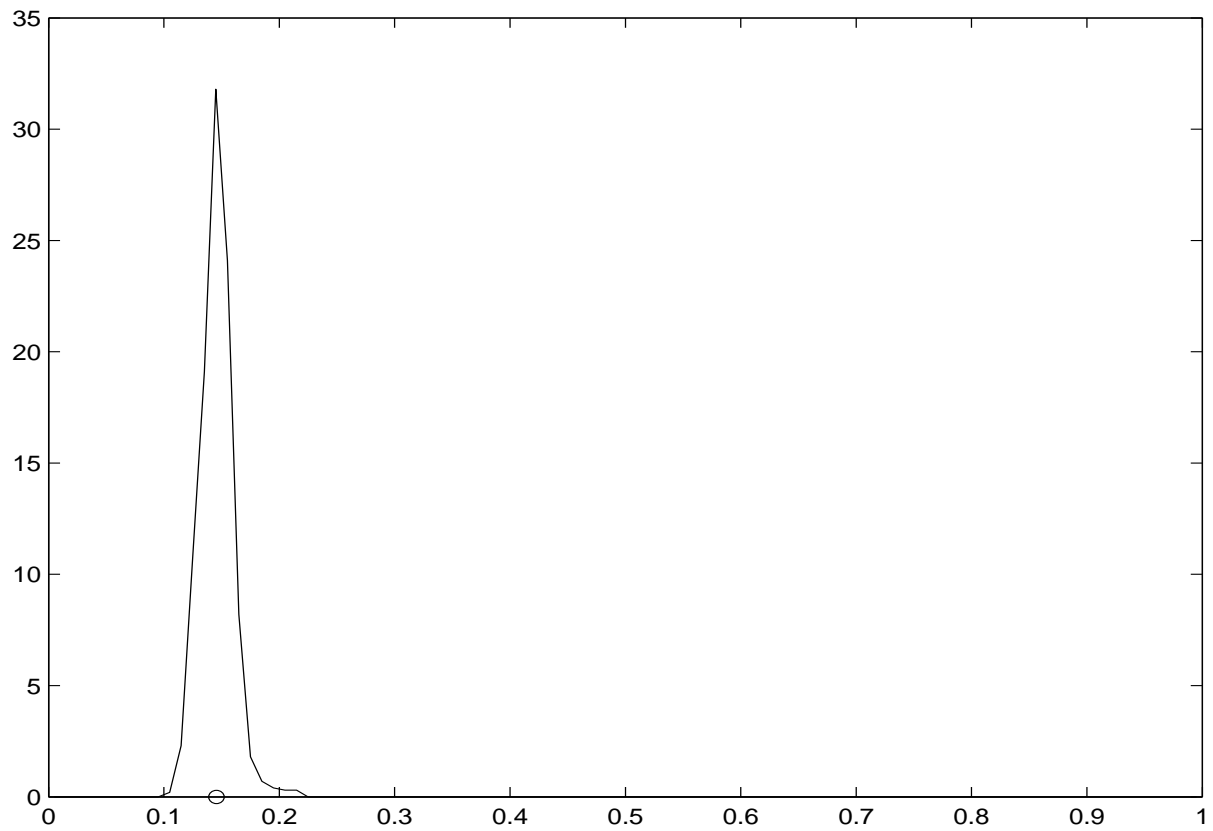
# Error distribution: dataset size: 342



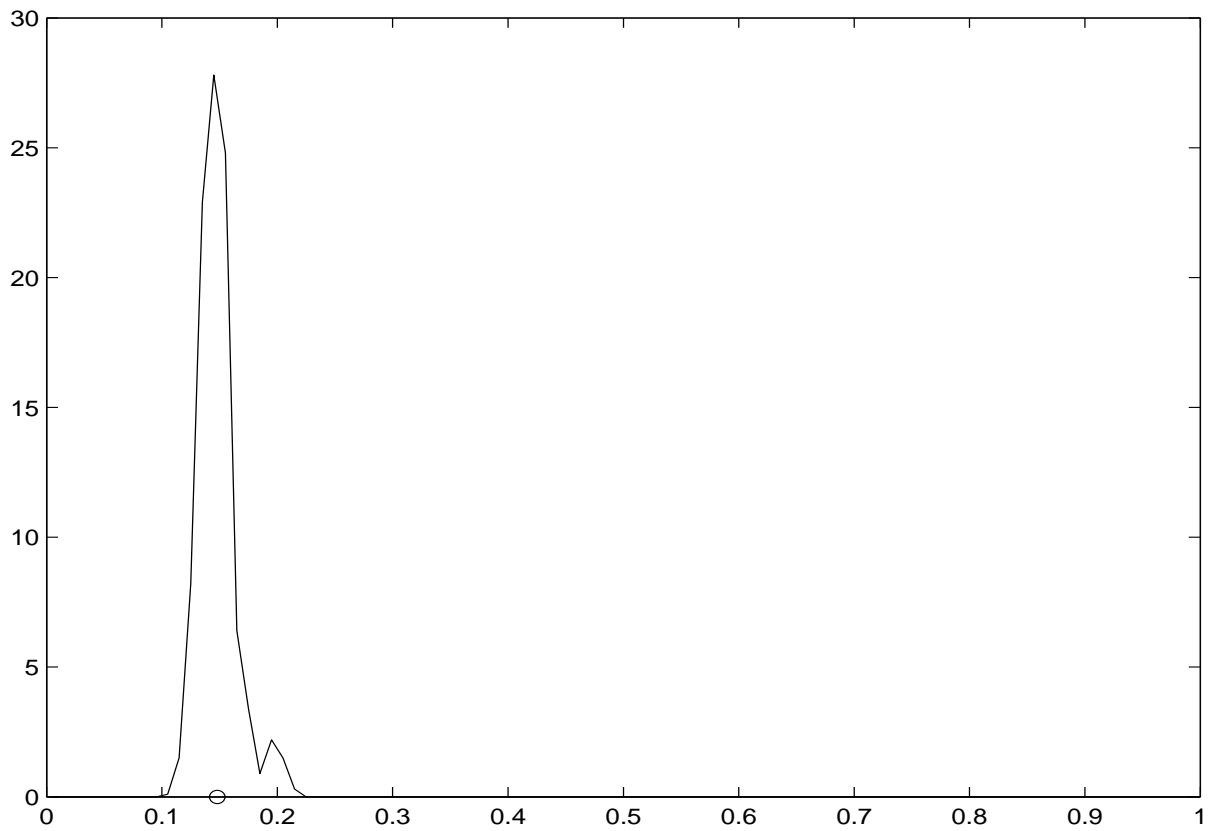
## Error distribution: dataset size: 273



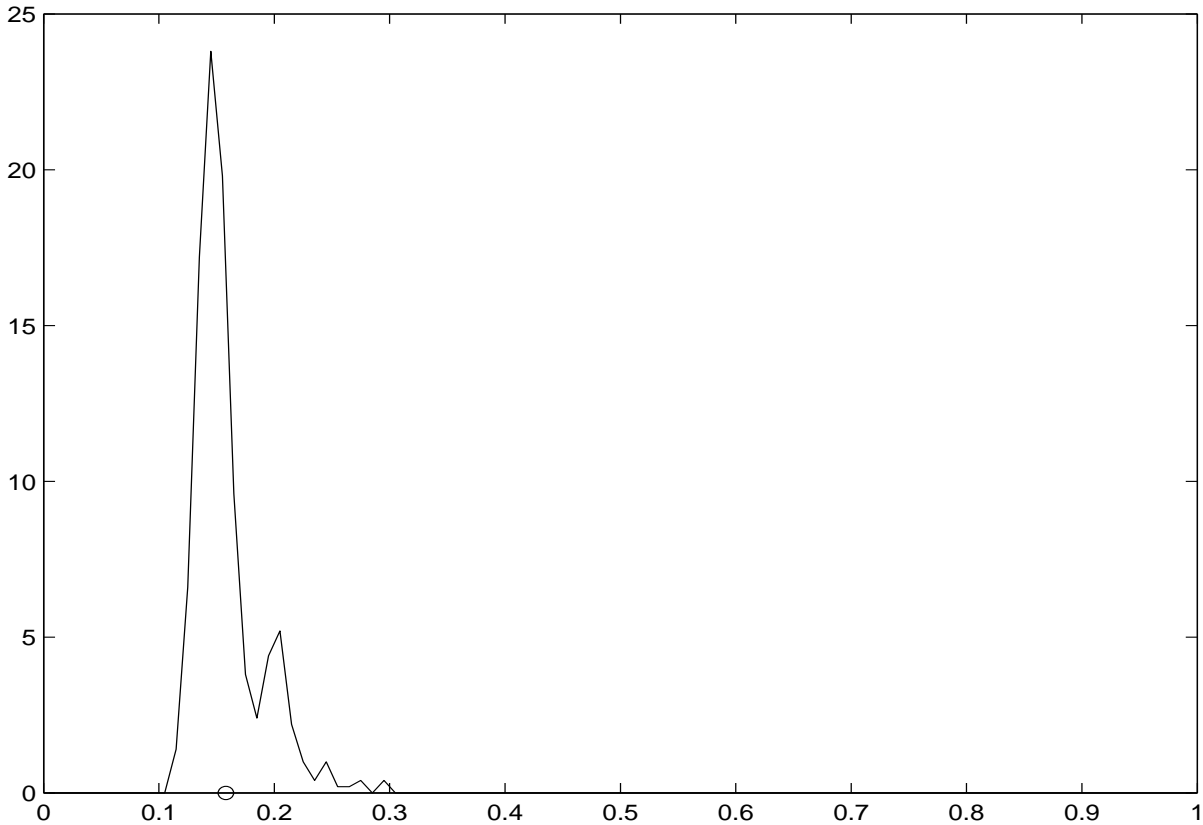
## Error distribution: dataset size: 205



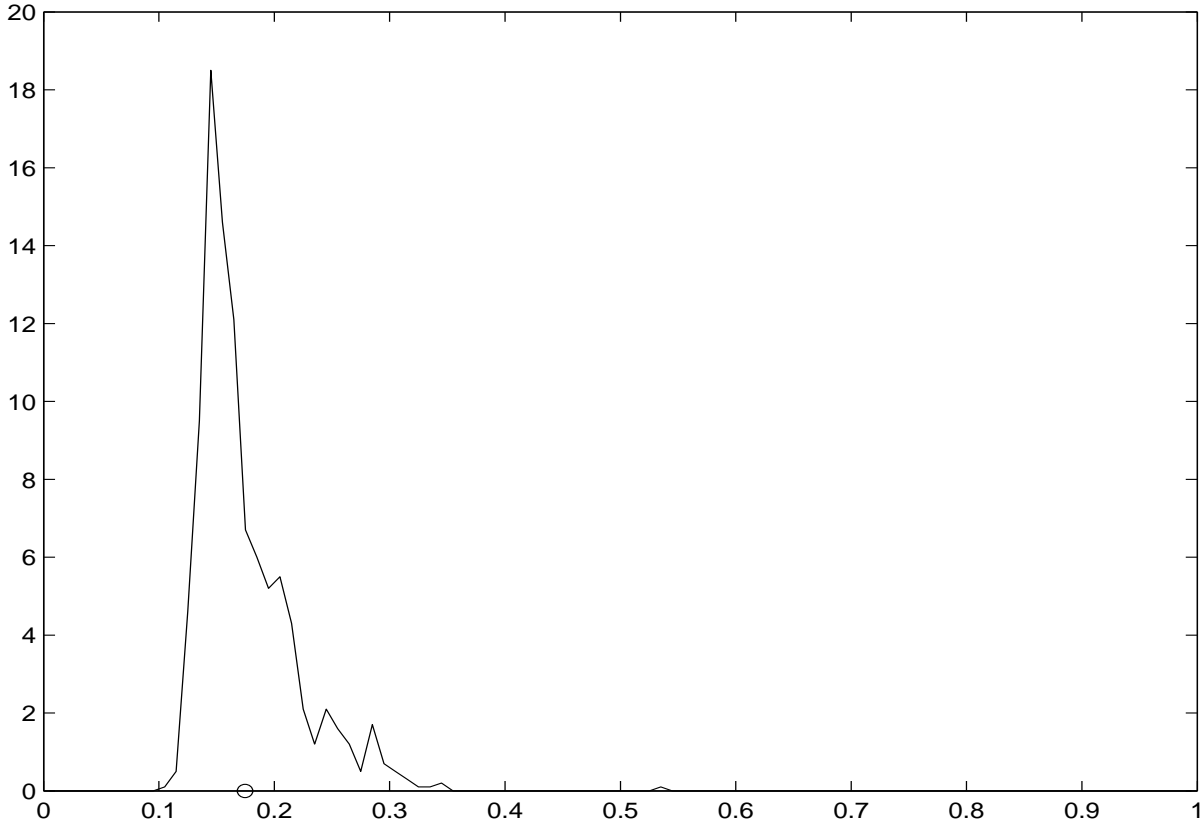
# Error distribution: dataset size: 137



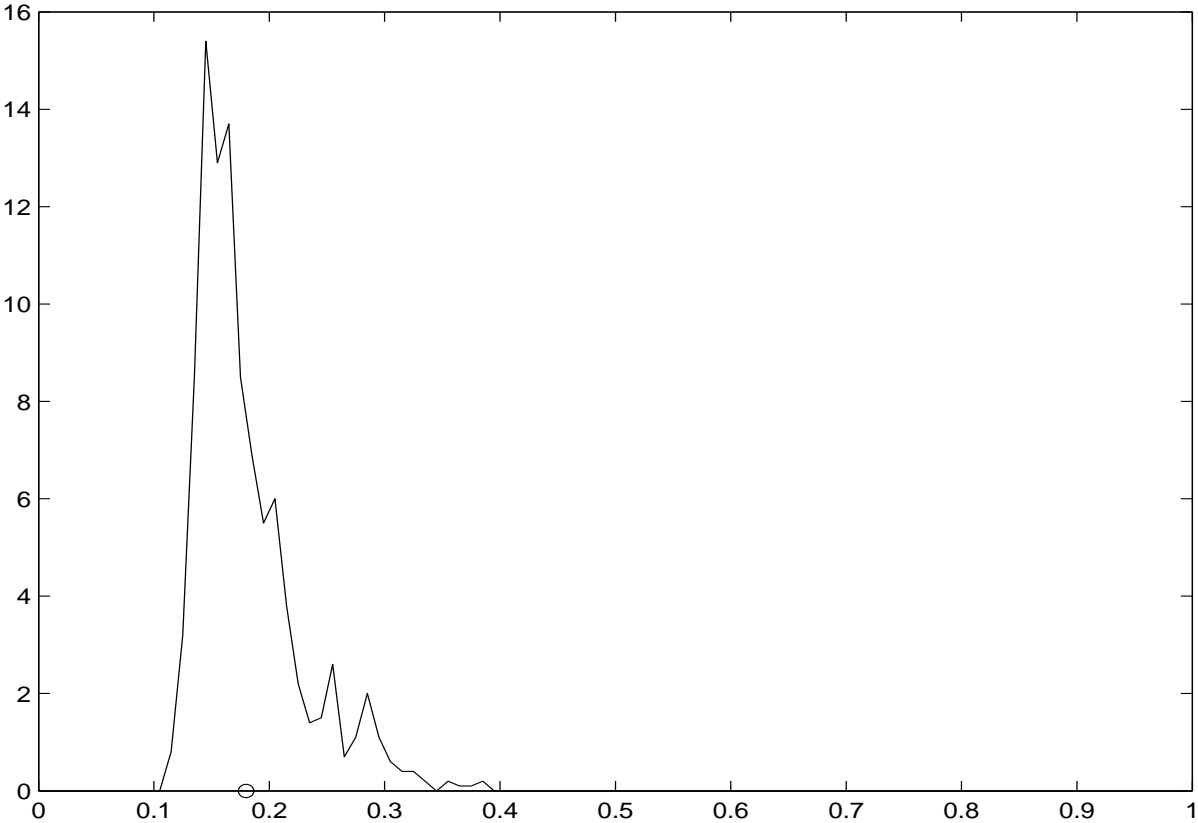
# Error distribution: dataset size: 68



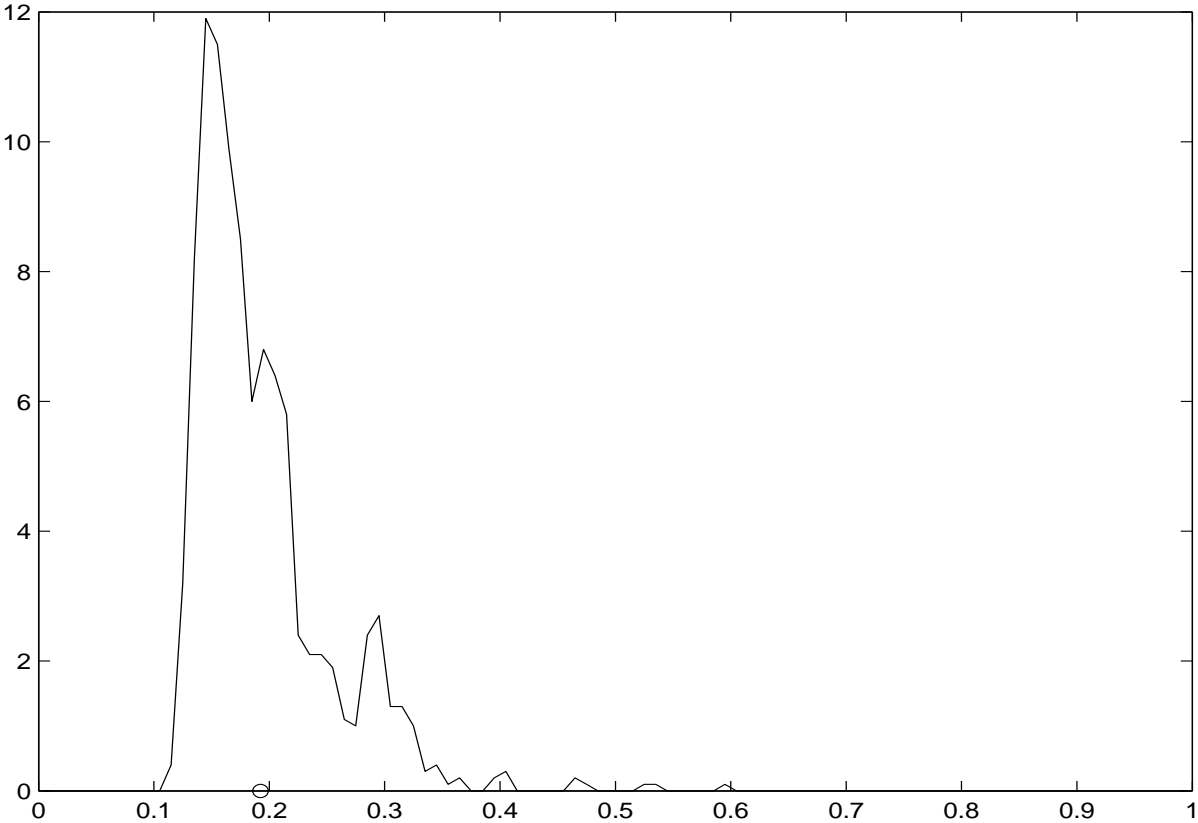
# Error distribution: dataset size: 34



# Error distribution: dataset size: 27

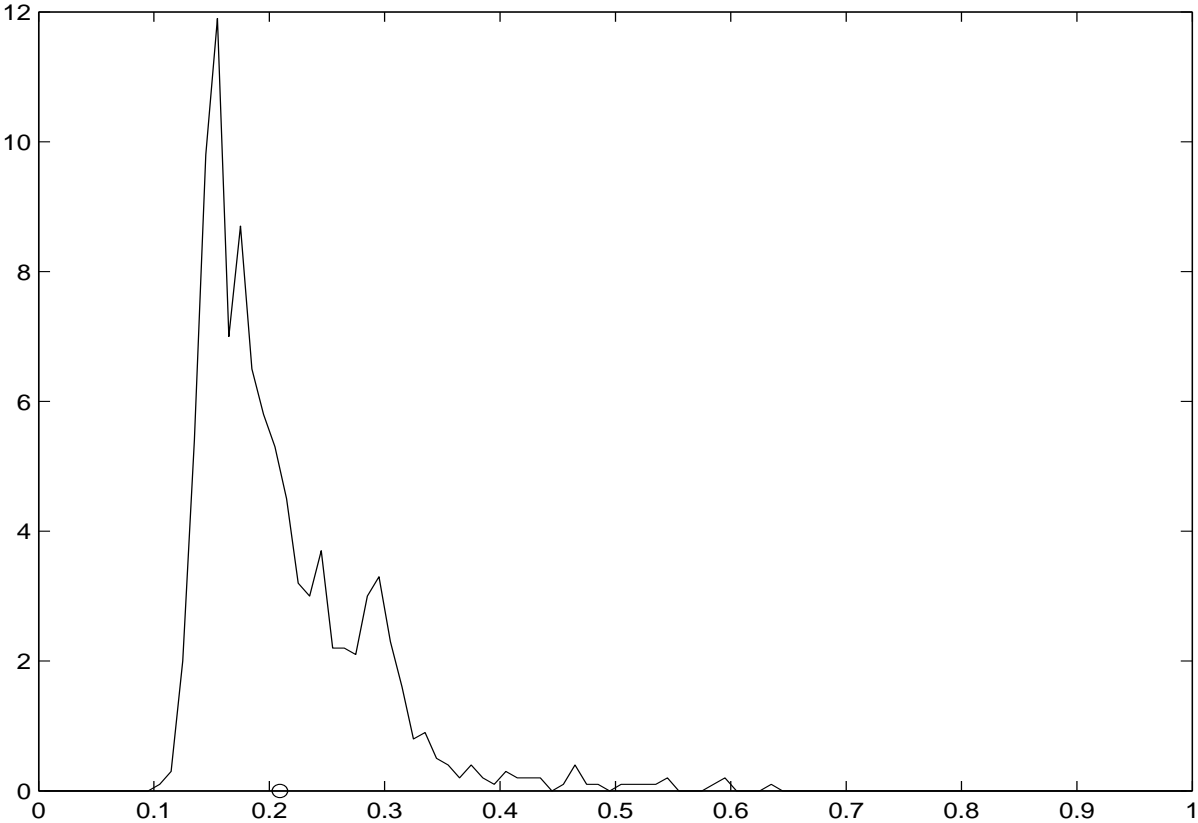


# Error distribution: dataset size: 20

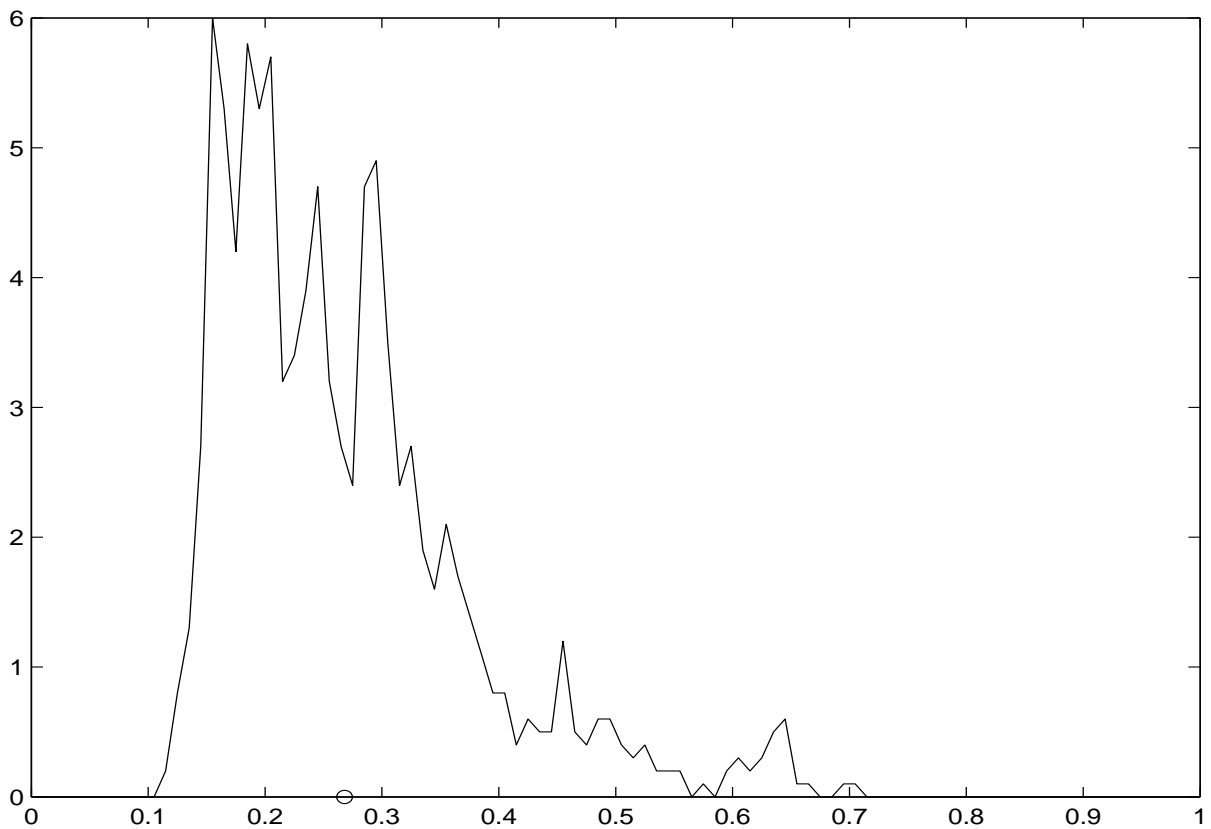




# Error distribution: dataset size: 14



## Error distribution: dataset size: 7



## Bayes risk and consistency

- Traditionally statistics has concentrated on analysing

$$\mathbb{E}_S [\epsilon(S, \mathcal{A}, \mathcal{F})].$$

- For example consistency of a classification algorithm  $\mathcal{A}$  and function class  $\mathcal{F}$  means

$$\lim_{m \rightarrow \infty} \mathbb{E}_S [\epsilon(S, \mathcal{A}, \mathcal{F})] = f_{\text{Bayes}},$$

where

$$f_{\text{Bayes}}(\mathbf{x}) = \begin{cases} 1 & \text{if } P(\mathbf{x}, 1) > P(\mathbf{x}, 0), \\ 0 & \text{otherwise.} \end{cases}$$

is the function with the lowest possible risk, often referred to as the Bayes risk.

## Expected versus confident bounds

- For a finite sample the generalisation  $\epsilon(S, \mathcal{A}, \mathcal{F})$  has a distribution depending on the algorithm, function class and sample size  $m$ .
- As indicated above, statistics has tended to concentrate on the mean of this distribution – but this quantity can be misleading, eg for low fold cross-validation.

## Expected versus confident bounds cont.

- Statistical learning theory has preferred to analyse the tail of the distribution, finding a bound which holds with high probability.
- This looks like a statistical test – significant at a 1% confidence means that the chances of the conclusion not being true for *my* training set with  $m$  examples are less than 1%.
- This is also the source of the acronym PAC: probably approximately correct, the ‘confidence’ parameter  $\delta$  is the probability that we have been misled by this particular training set.

## Probability of being misled in classification

- Aim to cover a number of key techniques of SLT. Basic approach is usually to bound the probability of being misled and set this equal to  $\delta$ .
- What is the chance of being misled by a single bad function  $f$ , i.e. training error  $\text{err}_S(f) = 0$ , while true error is bad  $\text{err}(f) > \epsilon$ ?

$$\begin{aligned} P_S \{ \text{err}_S(f) = 0, \text{err}(f) > \epsilon \} &= (1 - \text{err}(f))^m \\ &\leq (1 - \epsilon)^m \\ &\leq \exp(-\epsilon m). \end{aligned}$$

so that choosing  $\epsilon = \ln(1/t)/m$  ensures probability less than  $t$ .

## Finite or Countable function classes

If we now consider a function class

$$\mathcal{F} = \{f_1, f_2, \dots, f_n, \dots\}$$

and make the probability of being misled by  $f_n$  less than  $q_n \delta$  with

$$\sum_{n=1}^{\infty} q_n \leq 1,$$

then the probability of being misled by one of the functions is bounded by

$$P_S \left\{ \exists f_n: \text{err}_S(f_n) = 0, \text{err}(f_n) > \frac{1}{m} \ln \left( \frac{1}{q_n \delta} \right) \right\} \leq \delta.$$

This uses the so-called union bound – the probability of the union of a set of events is at most the sum of the individual probabilities.

## Finite or Countable function classes result

- The bound translates into a theorem: given  $\mathcal{F}$  and  $q$ , with probability at least  $1 - \delta$  over random  $m$  samples the generalisation error of a function  $f_n \in \mathcal{F}$  with zero training error is bounded by

$$\text{err}(f_n) \leq \frac{1}{m} \left( \ln \left( \frac{1}{q_n} \right) + \ln \left( \frac{1}{\delta} \right) \right)$$



## Some comments on the result

- We can think of the term  $\ln\left(\frac{1}{q_n}\right)$  as the complexity / description length of the function  $f_n$ .
- Note that we must put a prior weight on the functions. If the functions are drawn at random according to a distribution  $p_n$ , the expected generalisation will be minimal if we choose our prior  $q = p$ .
- Note that though  $q$  looks like a prior, the bound holds whether or not it is 'correct': this is the starting point of the PAC-Bayes analysis.

## What if uncountably many functions?

- We need a way to convert an infinite set to a finite one.
- Key idea is to replace measuring performance on a random test point with measuring on a second 'ghost' sample
- In this way the analysis is reduced to a finite set of examples and hence a finite set of classification functions.
- This step is often referred to as the 'double sample trick'

## Double sample trick

The result has the following form:

$$\begin{aligned} P^m \{ \mathbf{X} \in X^m : \exists h \in H : \text{err}_{\mathbf{X}}(h) = 0, \text{err}(h) \geq \epsilon \} \\ \leq 2P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : \\ \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \} \end{aligned}$$

If we think of the first probability as being over  $\mathbf{XY}$  the result concerns three events:

$$\begin{aligned} A(h) &:= \{ \text{err}_{\mathbf{X}}(h) = 0 \} \\ B(h) &:= \{ \text{err}(h) \geq \epsilon \} \\ C(h) &:= \{ \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \} \end{aligned}$$

## Double sample trick II

It is clear that

$$\begin{aligned} P^{2m}(C(h)|A(h)\&B(h)) &= P^{2m}(C(h)|B(h)) \\ &> 0.5 \end{aligned}$$

for reasonable  $m$  by a binomial tail bound.

## Double sample trick II

Hence, we have

$$\begin{aligned} P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : A(h) \& C(h) \} &\geq \\ P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : A(h) \& B(h) \& C(h) \} &= \\ P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : A(h) \& B(h) \} & \\ &P(C(h) | A(h) \& B(h)) \end{aligned}$$

It follows that

$$\begin{aligned} P^m \{ \mathbf{X} \in X^m : \exists h \in H : A(h) \& B(h) \} &\leq \\ &2P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : A(h) \& C(h) \} \end{aligned}$$

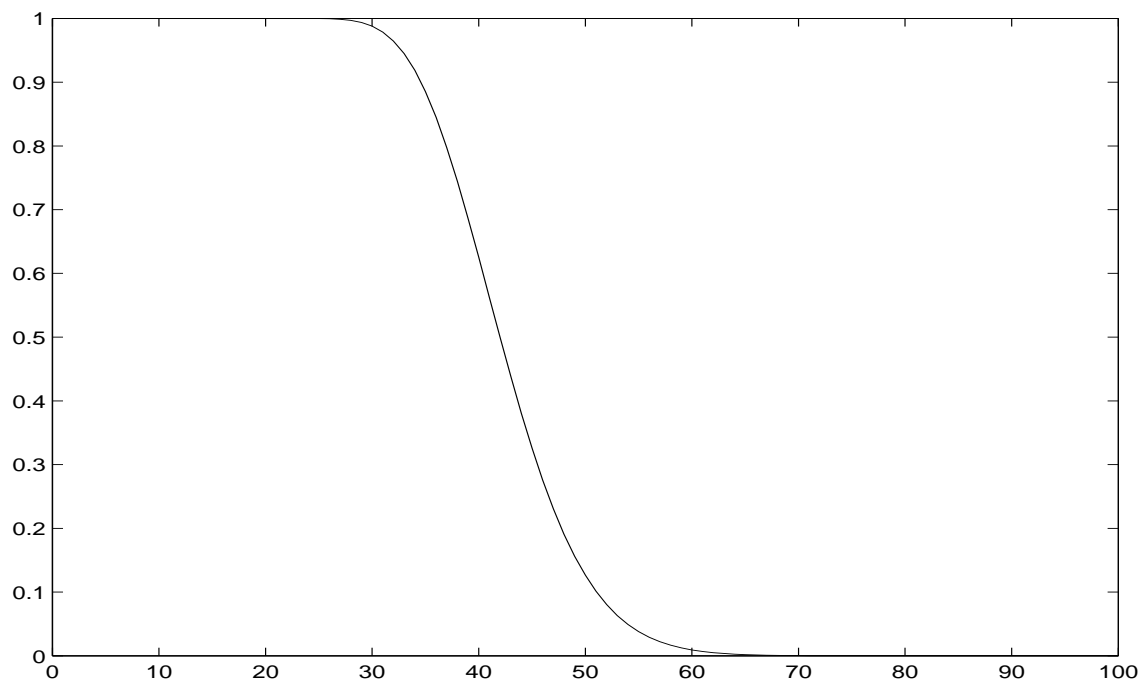
the required result.

## How many functions on a finite sample?

- Let  $H$  be a set of  $\{-1, 1\}$  valued functions.
- The growth function  $B_H(m)$  is the maximum cardinality of the set of functions  $H$  when restricted to  $m$  points – note that this cannot be larger than  $2^m$ , i.e.  $\log_2(B_H(m)) \leq m$
- For the statistics to work we want the number of functions to be much smaller than this as we will perform a union bound over this number.

## Examining the growth function

Consider a plot of the ratio of the growth function  $B_H(m)$  to  $2^m$  for linear functions in a 20 dimensional space:



# Vapnik Chervonenkis dimension

- The Vapnik-Chervonenkis dimension is the point at which the ratio stops being equal to 1:

$$\text{VCdim}(H) = \max\{m \quad : \quad \text{for some } \mathbf{x}_1, \dots, \mathbf{x}_m, \\ \text{for all } b \in \{-1, 1\}^m, \\ \exists h_b \in H, h_b(\mathbf{x}_i) = b_i\}$$

- For linear functions  $\mathcal{L}$  in  $\mathbb{R}^n$ ,  $\text{VCdim}(\mathcal{L}) = n + 1$ .



## Sauer's Lemma

- Sauer's Lemma (also due to VC and Shelah):

$$B_H(m) \leq \sum_{i=0}^d \binom{m}{i} \leq \left(\frac{em}{d}\right)^d,$$

where  $m \geq d = \text{VCdim}(H)$ .

- This is surprising as the rate of growth of  $B_H(m)$  up to  $m = d$  is exponential – but thereafter it is polynomial with exponent  $d$ .
- Furthermore, there is no assumption made about the functions being linear – this is a completely general result.

## Basic Theorem of SLT

We want to bound the probability that the training examples can mislead us about one of the functions we are considering using:

$$P^m \{ \mathbf{X} \in X^m : \exists h \in H : \text{err}_{\mathbf{X}}(h) = 0, \text{err}(h) \geq \epsilon \}$$

→ double sample trick →

$$\leq 2P^{2m} \{ \mathbf{XY} \in X^{2m} : \exists h \in H : \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \}$$

→ union bound →

$$\leq 2B_H(2m)P^{2m} \{ \mathbf{XY} \in X^{2m} : \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \}$$

Final ingredient is known as symmetrisation.

## Symmetrisation

- Consider generating a  $2m$  sample  $S$ . Since the points are generated independently the probability of generating the same set of points in a different order is the same.
- Consider a fixed set  $\Sigma$  of permutations and each time we generate a sample we randomly permute it with a uniformly chosen element of  $\Sigma$  – gives probability distribution  $P_{\Sigma}^{2m}$

## Symmetrisation cont.

- Any event has equal probability under  $P^{2m}$  and  $P_{\Sigma}^{2m}$ , so that

$$P^{2m}(A) = P_{\Sigma}^{2m}(A) = \mathbb{E}^{2m} [P_{\sigma \sim \Sigma}(A)]$$

- Consider particular choice of  $\Sigma$  the permutations that swap/leave unchanged corresponding elements of the two samples  $\mathbf{X}$  and  $\mathbf{Y}$  –  $2^m$  such permutations.

## Completion of the proof

$$\begin{aligned} P^{2m} \{ \mathbf{XY} \in X^{2m} : \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \} \\ \leq \mathbb{E}^{2m} [ P_{\sigma \sim \Sigma} \{ \text{err}_{\mathbf{X}}(h) = 0, \text{err}_{\mathbf{Y}}(h) \geq \epsilon/2 \text{ for } \sigma(\mathbf{XY}) \} ] \\ \leq \mathbb{E}^{2m} [ 2^{-\epsilon m/2} ] \\ = 2^{-\epsilon m/2} \end{aligned}$$

- Setting the right hand side equal to  $\delta/(2B_H(2m))$  and inverting gives the bound on  $\epsilon$ .

## Final result

- Assembling the ingredients gives the result: with probability at least  $1 - \delta$  of random  $m$  samples the generalisation error of a function  $h \in H$  chosen from a class  $H$  with VC dimension  $d$  with zero training error is bounded by

$$\epsilon = \epsilon(m, H, \delta) = \frac{2}{m} \left( d \log \frac{2em}{d} + \log \frac{2}{\delta} \right)$$

- Note that we can think of  $d$  as the complexity / capacity of the function class  $H$ .
- The bound does not distinguish between functions in  $H$ .

## Lower bounds

- VCdim *Characterises* Learnability in PAC setting: there exist distributions such that with probability at least  $\delta$  over  $m$  random examples, the error of  $h$  is at least

$$\max \left( \frac{d-1}{32m}, \frac{1}{m} \log \left( \frac{1}{\delta} \right) \right).$$

## Non-zero training error

- Very similar results can be obtained for non-zero training error.
- The main difference is the introduction of a square root to give a bound of the form

$$\epsilon(m, H, k, \delta) = k + O \left( \sqrt{\frac{d}{m} \log \frac{2em}{d}} + \sqrt{\frac{1}{m} \log \frac{2}{\delta}} \right)$$

for  $k$  training errors, which is significantly worse than in the zero training error case.

- PAC-Bayes bounds now interpolate between these two.



## Structural Risk Minimisation

- The way to differentiate between functions using the VC result is to create a hierarchy of classes:

$$H_1 \subseteq H_2 \subseteq \dots \subseteq H_d \subseteq \dots \subseteq H_K.$$

of increasing complexity/VC dimension.

- Can now find the function in each class with minimum empirical error  $k_d$  and choose between the classes by minimising over the choice of  $d$ :

$$\epsilon(m, d, \delta) = k_d + O \left( \sqrt{\frac{d}{m} \log \frac{2em}{d}} + \sqrt{\frac{1}{m} \log \frac{2K}{\delta}} \right)$$

which bounds the generalisation by a further application of the union bound over the  $K$  classes.

## Criticisms of PAC Theory

- The theory is certainly valid and the lower bounds indicate that it is not too far out – so can't criticise as stands
- Criticism is that it doesn't accord with experience of those applying learning.
- Mismatch between theory and practice.
- For example

# Support Vector Machines (SVM)

One example of PAC failure is in analysing SVMs: thresholded linear functions in very high dimensional feature spaces.

1. kernel trick means we can work in an infinite dimensional feature space ( $\Rightarrow$  infinite VC dimension) so that VC result does not apply:
2. and YET very impressive performance

## Support Vector Machines cont.

- SVM seeks linear function in a feature space defined by a mapping  $\phi$  specified implicitly via a kernel  $\kappa$ :

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

- For example the 1-norm SVM seeks  $\mathbf{w}$  to solve

$$\min_{\mathbf{w}, b, \gamma, \xi} \quad \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$

$$\text{subject to} \quad y_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \\ i = 1, \dots, m.$$

## Margin in SVMs

- Intuition behind SVMs is that maximising the margin makes it possible to obtain good generalisation despite the high VC dimension
- The lower bound implies that we must be taking advantage of a benign distribution, since we know that in the worst case generalisation will be bad.
- Structural Risk Minimisation cannot be applied to a hierarchy determined by the margin of different classifiers, since the margin is not known until we see the data, while the SRM hierarchy must be chosen in advance.

## Margin in SVMs cont

- Hence, we require a theory that can give bounds that are sensitive to serendipitous distributions – in particular we conjecture that the margin is an indication of such ‘luckiness’.
- The proof approach will rely on using real-valued function classes. The margin gives an indication of the accuracy with which we need to approximate the functions when applying the statistics.

# Covering Numbers

$\mathcal{F}$  a class of real functions defined on  $X$  and  $\|\cdot\|_d$  a norm on  $\mathcal{F}$ , then

$$\mathcal{N}(\gamma, \mathcal{F}, \|\cdot\|_d)$$

is the smallest size set  $U_\gamma$  such that for any  $f \in \mathcal{F}$  there is a  $u \in U_\gamma$  such that  $\|f - u\|_d < \gamma$ .

## Covering Numbers cont.

For generalization bounds we need the  $\gamma$ -growth function,

$$\mathcal{N}^m(\gamma, \mathcal{F}) := \sup_{\mathbf{X} \in X^m} \mathcal{N}(\gamma, \mathcal{F}, \ell_{\infty}^{\mathbf{X}}).$$

where  $\ell_{\infty}^{\mathbf{X}}$  gives the distance between two functions as the maximum difference between their outputs on the sample.



## Covering numbers for linear functions

- Consider the set of functions:

$$\left\{ \mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x}) \rangle \quad : \quad \mathbf{w} = \frac{1}{Z} \sum_{i=1}^m z_i \phi(\mathbf{x}_i), z_i \in \mathbb{Z}, \right.$$
$$\left. Z = \sum_{i=1}^m z_i \neq 0, \sum_{i=1}^m |z_i| \leq \frac{8R^2}{\gamma^2} \right\},$$

where  $R = \max_{1 \leq i \leq m} \|\phi(\mathbf{x}_i)\|$

- This is an explicit cover that approximates the output of any norm 1 linear function to within  $\gamma/2$  on the sample

$$\{\mathbf{x}_1, \dots, \mathbf{x}_m\}.$$

## Covering numbers for linear functions

- We convert the  $\gamma/2$  approximation on the sample problem into a classification problem, which is solvable with a margin of  $\gamma/2$ .
- It follows that if we use the perceptron algorithm to find a classifier, we will find a function satisfying the  $\gamma/2$  approximation with just  $8R^2/\gamma^2$  updates.
- This gives a sparse dual representation of the function. The covering is chosen as the set of functions with small sparse dual representations.
- Gives a bound on the size of the cover

$$\log_2 \mathcal{N}^{2m}(\gamma/2, \mathcal{F}) \leq k \log_2 \frac{e(2m + k - 1)}{k}, \text{ where } k = \frac{8R^2}{\gamma^2}.$$

## Second statistical result

- We want to bound the probability that the training examples can mislead us about one of the functions with margin bigger than fixed  $\gamma$ :

$$\begin{aligned} & P^m \{ \mathbf{X} \in X^m : \exists f \in \mathcal{F} : \text{err}_{\mathbf{X}}(f) = 0, m_{\mathbf{X}}(f) \geq \gamma, \text{err}_P(f) \geq \epsilon \} \\ & \leq 2P^{2m} \{ \mathbf{X}\mathbf{Y} \in X^{2m} : \exists f \in \mathcal{F} \text{ such that} \\ & \quad \text{err}_{\mathbf{X}}(f) = 0, m_{\mathbf{X}}(f) \geq \gamma, \text{err}_{\mathbf{Y}}(f) \geq \epsilon/2 \} \\ & \leq 2\mathcal{N}^{2m}(\gamma/2, \mathcal{F}) P^{2m} \{ \mathbf{X}\mathbf{Y} \in X^{2m} : \text{for fixed } f' \\ & \quad m_{\mathbf{X}}(f') > \gamma/2, m_{\mathbf{Y}(\epsilon m/2)}(f') < \gamma/2 \} \\ & \leq 2\mathcal{N}^{2m}(\gamma/2, \mathcal{F}) 2^{-\epsilon m/2} \leq \delta \end{aligned}$$

## Second statistical result cont.

- inverting gives

$$\epsilon = \epsilon(m, \mathcal{F}, \delta, \gamma) = \frac{2}{m} \left( \log_2 \mathcal{N}^{2m}(\gamma/2, \mathcal{F}) + \log_2 \frac{2}{\delta} \right)$$

i.e. with probability  $1 - \delta$  over  $m$  random examples a margin  $\gamma$  hypothesis has error less than  $\epsilon$ . Must apply for finite set of  $\gamma$  ('do SRM over  $\gamma$ ').

# Bounding the covering numbers

Have the following correspondences with the standard VC case (easy slogans):

- Growth function –  $\gamma$ -growth function
- Vapnik Chervonenkis dim – Fat shattering dim
- Sauer's Lemma – Alon *et al.*

## Generalization of SVMs

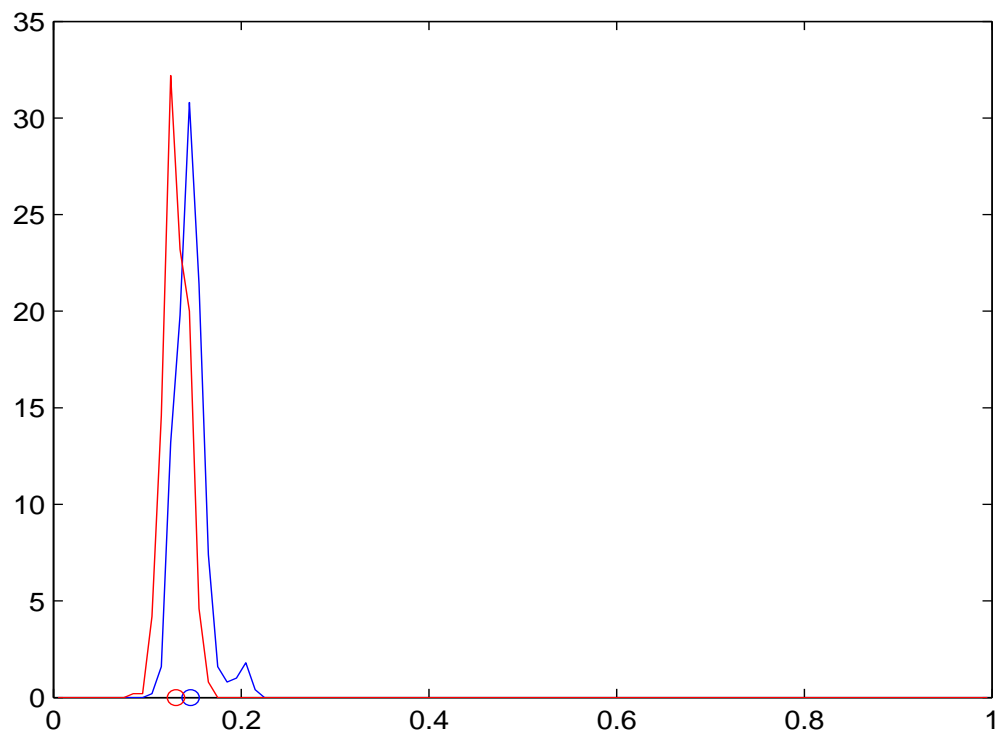
- Since SVMs are linear functions, can apply linear function bound.
- Hence for distribution with support in ball of radius  $R$ , (eg Gaussian Kernels  $R = 1$ ) and margin  $\gamma$ , have bound:

$$\epsilon(m, \mathcal{L}, \delta, \gamma) = \frac{2}{m} \left( k \log_2 \frac{e(2m + k - 1)}{k} + \log_2 \frac{m}{\delta} \right)$$

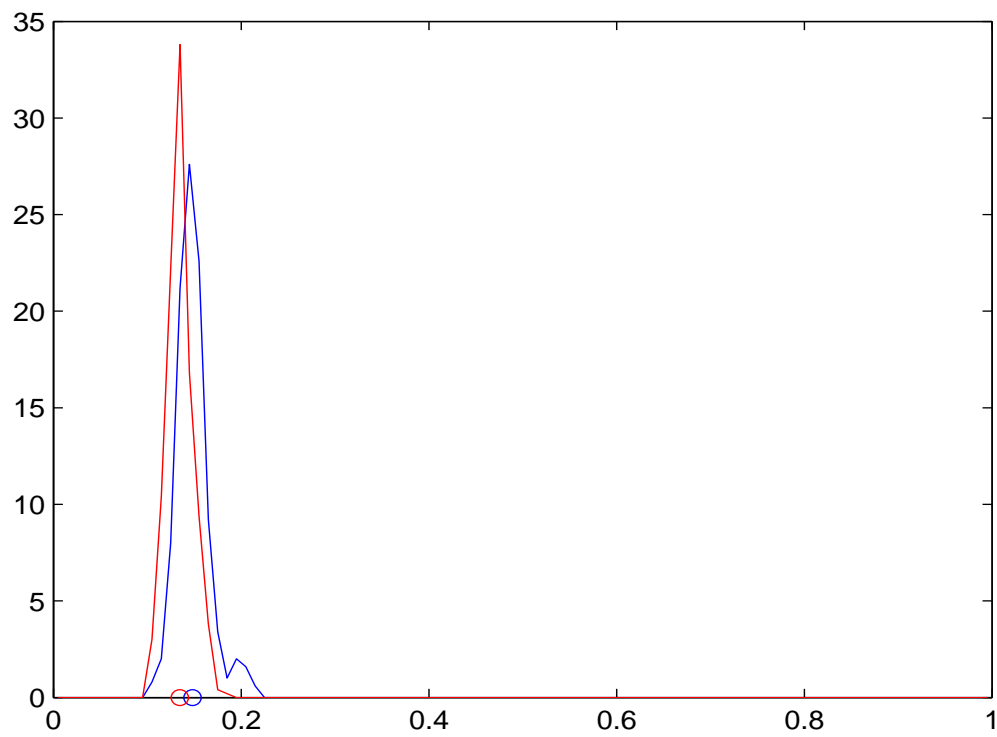
where  $k = \frac{8R^2}{\gamma^2}$ .

- Key property: no explicit dependence on dimension!

# Error distribution: dataset size: 205

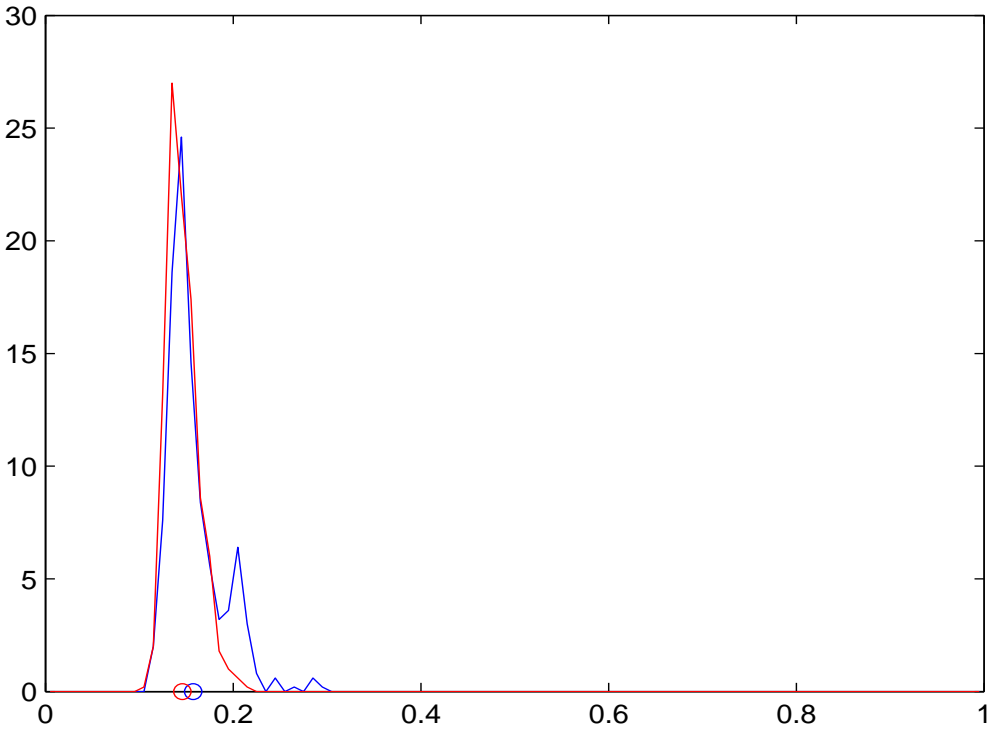


# Error distribution: dataset size: 137

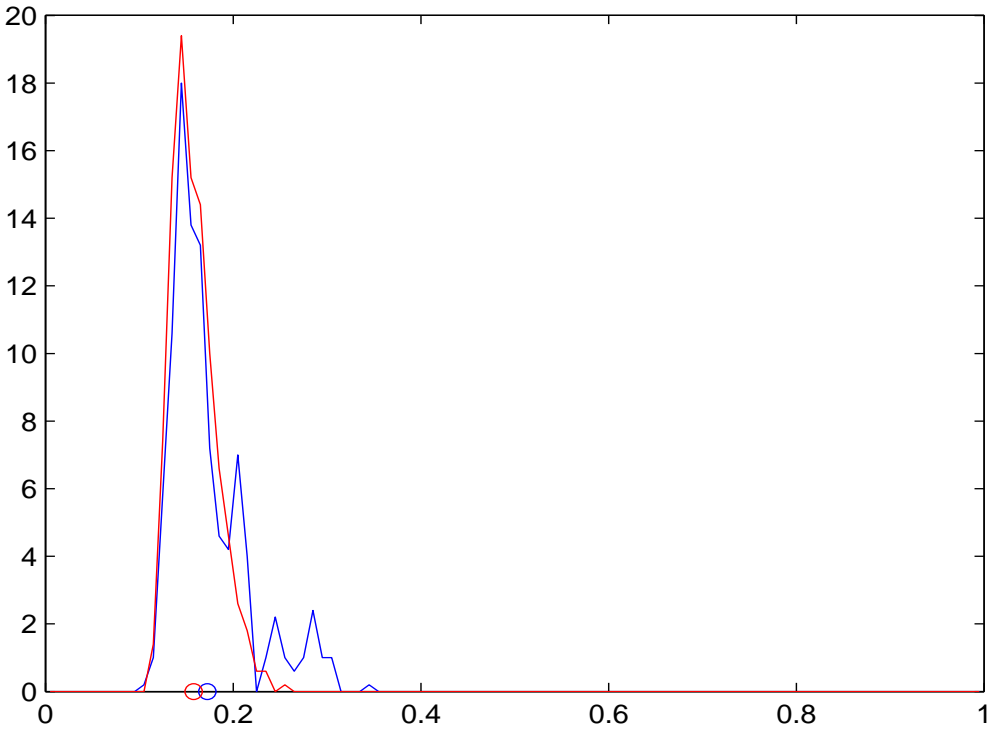




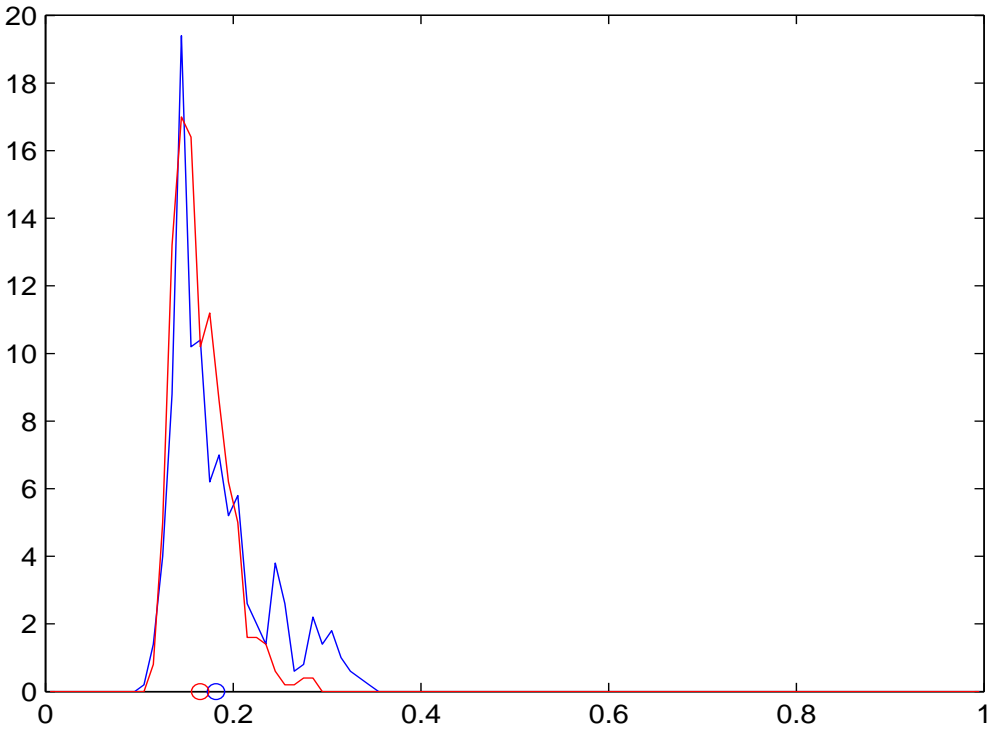
# Error distribution: dataset size: 68



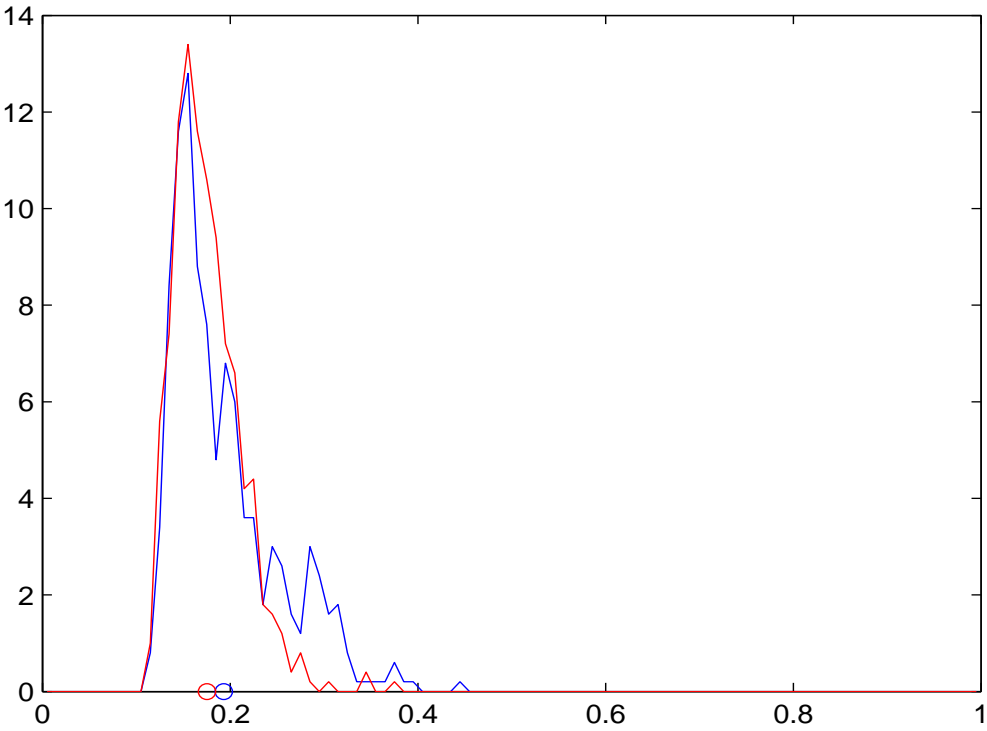
# Error distribution: dataset size: 34



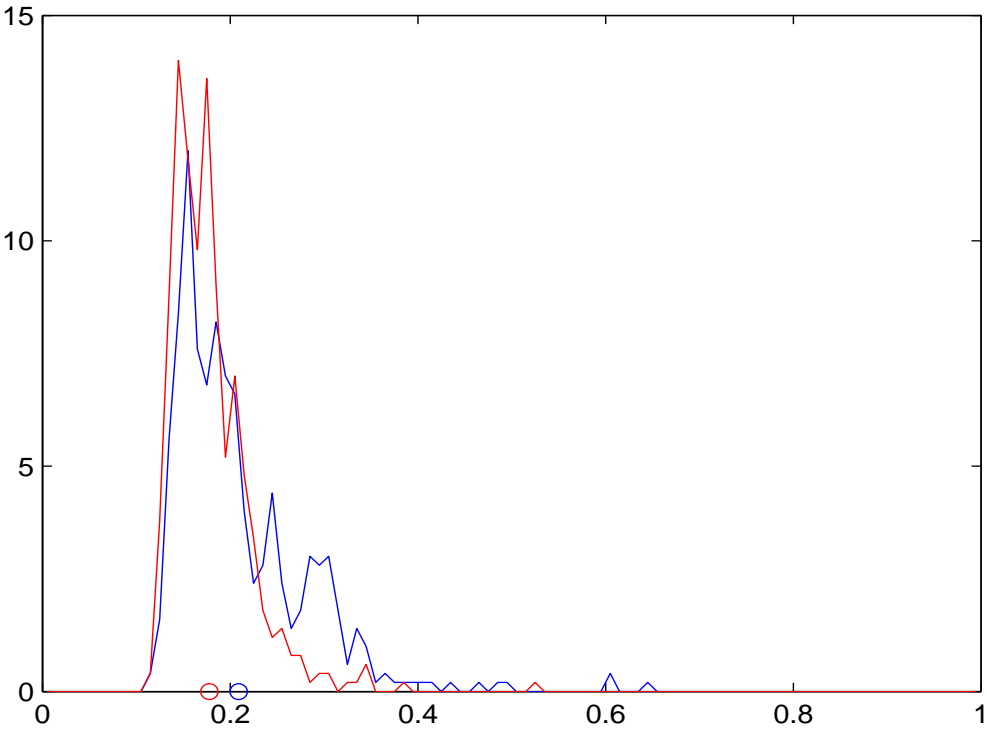
# Error distribution: dataset size: 27



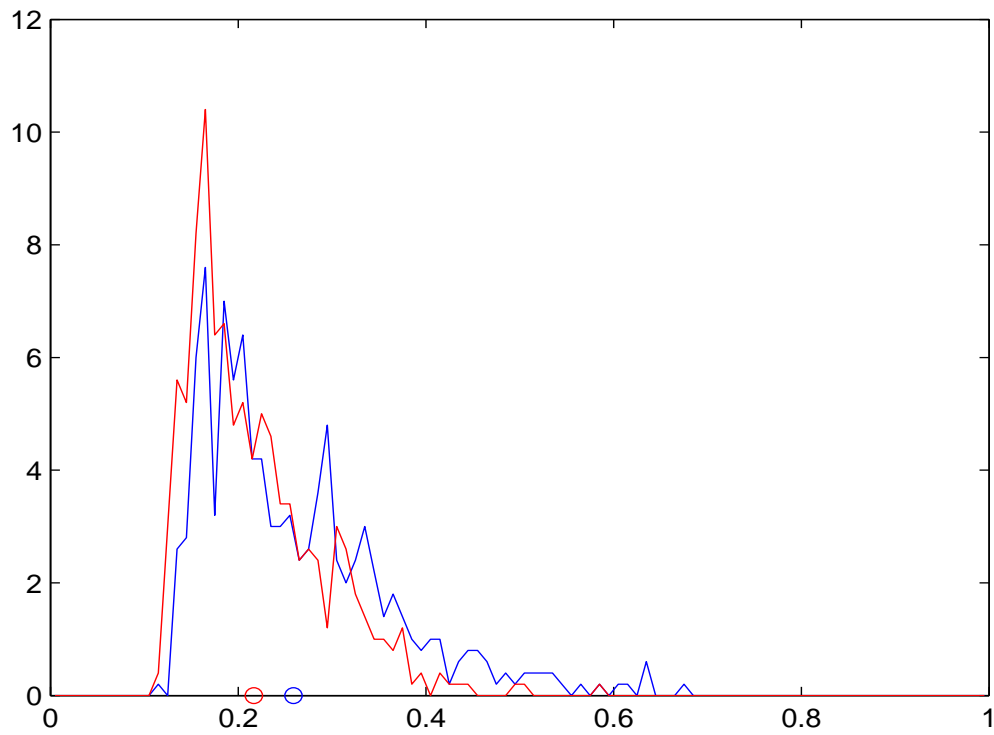
# Error distribution: dataset size: 20



# Error distribution: dataset size: 14



# Error distribution: dataset size: 7

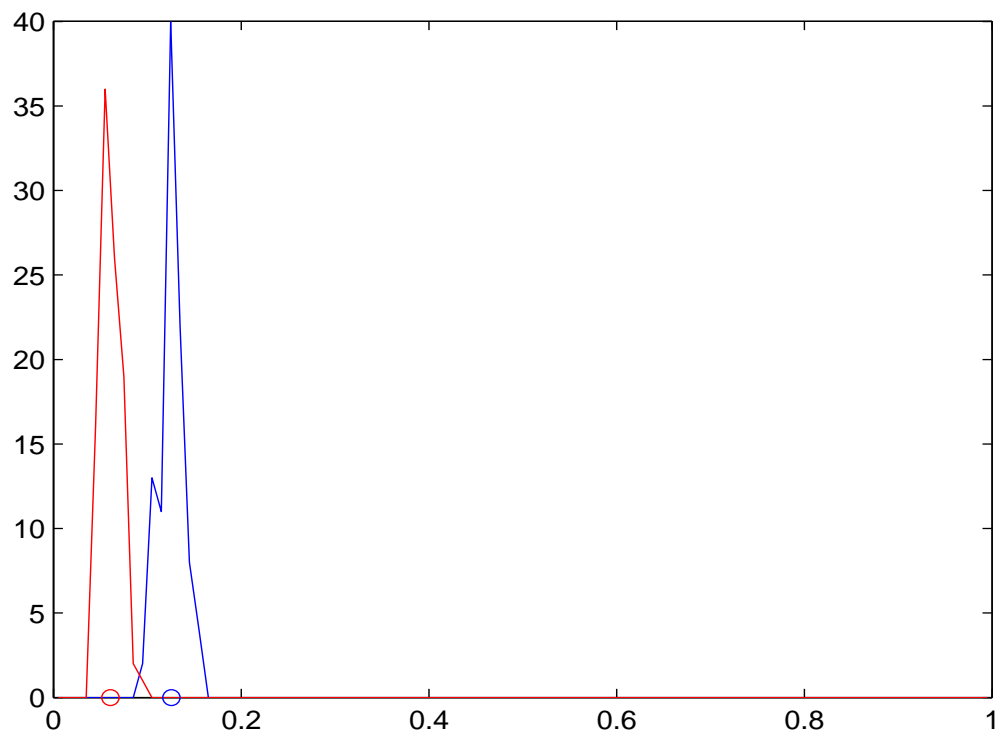


## Using a kernel

- Can consider much higher dimensional spaces using the kernel trick
- Can even work in infinite dimensional spaces, eg using the Gaussian kernel:

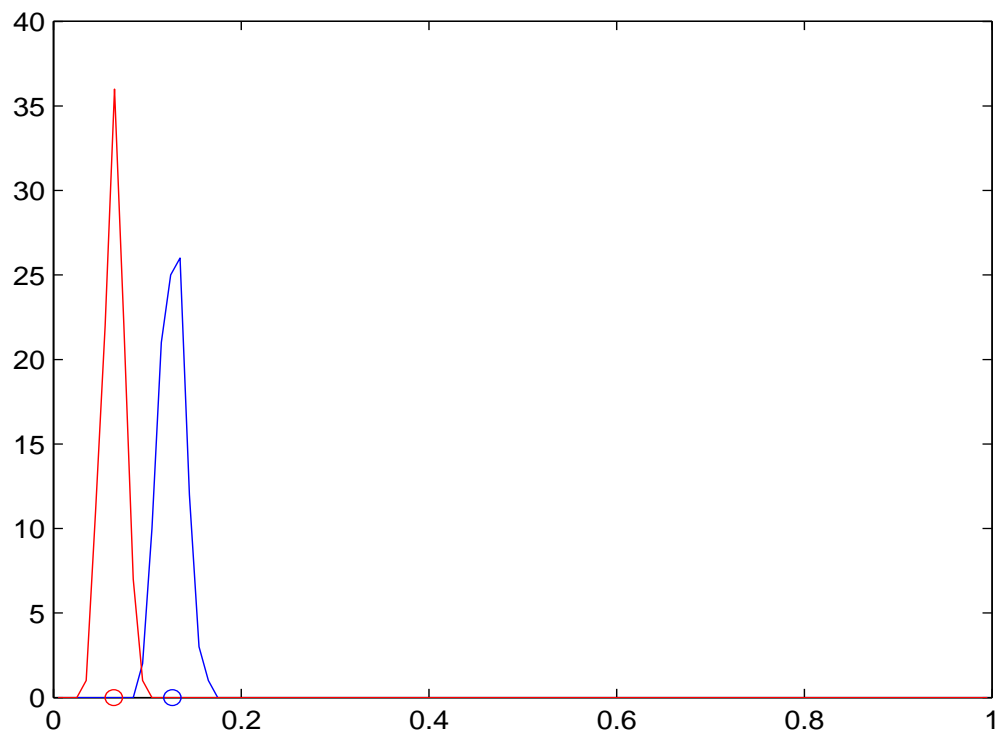
$$\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right)$$

# Error distribution: dataset size: 342





# Error distribution: dataset size: 273



# Conclusions

- Outline of philosophy and approach of SLT
- Central result of SLT
- Touched on covering number bounds for margin based analysis
- Application to analyse SVM learning

# STRUCTURE

## PART A

1. General Statistical Considerations
2. Basic PAC Ideas and proofs
3. Real-valued Function Classes and the Margin

## PART B

1. Rademacher complexity and Main Theory
2. Applications to classification
3. Conclusions

# PART B

## Concentration inequalities

- Statistical Learning is concerned with the reliability or stability of inferences made from a random sample.
- Random variables with this property have been a subject of ongoing interest to probabilists and statisticians.

## Concentration inequalities cont.

- As an example consider the mean of a sample of  $m$  1-dimensional random variables  $X_1, \dots, X_m$ :

$$S_m = \frac{1}{m} \sum_{i=1}^m X_i.$$

- Hoeffding's inequality states that if  $X_i \in [a_i, b_i]$

$$P\{|S_m - \mathbb{E}[S_m]| \geq \epsilon\} \leq 2 \exp\left(-\frac{2m^2\epsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}\right)$$

Note how the probability falls off exponentially with the distance from the mean and with the number of variables.

## Concentration for SLT

- We are now going to look at deriving SLT results from concentration inequalities.
- Perhaps the best known form is due to McDiarmid (although he was actually re-presenting previously derived results):

## McDiarmid's inequality

**Theorem 1.** Let  $X_1, \dots, X_n$  be independent random variables taking values in a set  $A$ , and assume that  $f : A^n \rightarrow \mathbb{R}$  satisfies

$$\sup_{x_1, \dots, x_n, \hat{x}_i \in A} |f(x_1, \dots, x_n) - f(x_1, \dots, \hat{x}_i, x_{i+1}, \dots, x_n)| \leq c_i,$$

for  $1 \leq i \leq n$ . Then for all  $\epsilon > 0$ ,

$$P\{f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

- Hoeffding is a special case when  $f(x_1, \dots, x_n) = S_n$



## Using McDiarmid

- By setting the right hand side equal to  $\delta$ , we can always invert McDiarmid to get a high confidence bound: with probability at least  $1 - \delta$

$$f(X_1, \dots, X_n) < \mathbb{E}f(X_1, \dots, X_n) + \sqrt{\frac{\sum_{i=1}^n c_i^2}{2} \log \frac{1}{\delta}}$$

- If  $c_i = c/n$  for each  $i$  this reduces to

$$f(X_1, \dots, X_n) < \mathbb{E}f(X_1, \dots, X_n) + \sqrt{\frac{c^2}{2n} \log \frac{1}{\delta}}$$

## Rademacher complexity

- Rademacher complexity is a new way of measuring the complexity of a function class. It arises naturally if we rerun the proof using the double sample trick and symmetrisation but look at what is actually needed to continue the proof:

## Rademacher proof beginnings

For a fixed  $f \in \mathcal{F}$  we have

$$\mathbb{E}[f(\mathbf{z})] \leq \hat{\mathbb{E}}[f(\mathbf{z})] + \sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right).$$

where  $\mathcal{F}$  is a class of functions mapping from  $Z$  to  $[0, 1]$  and  $\hat{\mathbb{E}}$  denotes the sample average.

We must bound the size of the second term. First apply McDiarmid's inequality to obtain ( $c_i = 1/m$  for all  $i$ ) with probability at least  $1 - \delta$ :

$$\sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \leq \mathbb{E}_S \left[ \sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

## Deriving double sample result

- We can now move to the ghost sample by simply observing that  $\mathbb{E}[h] = \mathbb{E}_{\tilde{S}} [\hat{\mathbb{E}}[h]]$ :

$$\mathbb{E}_S \left[ \sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] =$$
$$\mathbb{E}_S \left[ \sup_{h \in \mathcal{F}} \mathbb{E}_{\tilde{S}} \left[ \frac{1}{m} \sum_{i=1}^m h(\tilde{\mathbf{z}}_i) - \frac{1}{m} \sum_{i=1}^m h(\mathbf{z}_i) \middle| S \right] \right]$$

## Deriving double sample result cont.

Since the sup of an expectation is less than or equal to the expectation of the sup (we can make the choice to optimise for each  $\tilde{S}$ ) we have

$$\mathbb{E}_S \left[ \sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] \leq \mathbb{E}_S \mathbb{E}_{\tilde{S}} \left[ \sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i)) \right]$$

## Adding symmetrisation

Here symmetrisation is again just swapping corresponding elements – but we can write this as multiplication by a variable  $\sigma_i$  which takes values  $\pm 1$  with equal probability:

$$\begin{aligned} \mathbb{E}_S \left[ \sup_{h \in \mathcal{F}} \left( \mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] &\leq \\ &\leq \mathbb{E}_{\sigma S \tilde{S}} \left[ \sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i)) \right] \\ &\leq 2 \mathbb{E}_{S \sigma} \left[ \sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(\mathbf{z}_i) \right] \\ &= R_m(\mathcal{F}), \end{aligned}$$

## Rademacher complexity

where

$$R_m(\mathcal{F}) = \mathbb{E}_{S\sigma} \left[ \sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^m \sigma_i f(\mathbf{z}_i) \right].$$

is known as the Rademacher complexity of the function class  $\mathcal{F}$ .

- Rademacher complexity is the expected value of the maximal correlation with random noise – a very natural measure of capacity.
- Note that the Rademacher complexity is distribution dependent since it involves an expectation over the choice of sample – this might seem hard to compute.

## Main Rademacher theorem

Putting the pieces together gives the main theorem of Rademacher complexity: with probability at least  $1 - \delta$  over random samples  $S$  of size  $m$ , every  $f \in \mathcal{F}$  satisfies

$$\mathbb{E} [f(\mathbf{z})] \leq \hat{\mathbb{E}} [f(\mathbf{z})] + R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$



## Empirical Rademacher theorem

- Since the empirical Rademacher complexity

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^m \sigma_i f(\mathbf{z}_i) \mid \mathbf{z}_1, \dots, \mathbf{z}_m \right]$$

is concentrated, we can make a further application of McDiarmid to obtain with probability at least  $1 - \delta$

$$\mathbb{E}_{\mathcal{D}} [f(\mathbf{z})] \leq \hat{\mathbb{E}} [f(\mathbf{z})] + \hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

## Relation to VC theorem

- For  $H$  a class of  $\pm 1$  valued functions with VC dimension  $d$ , we can upper bound  $\hat{R}_m(H)$  using Hoeffding's inequality to upper bound

$$P \left\{ \sum_{i=1}^m \sigma_i f(\mathbf{x}_i) \geq \epsilon \right\} \leq \exp \left( -\frac{\epsilon^2}{2m} \right)$$

for a fixed function  $f$ , since the expected value of the sum is 0, and the maximum change by replacing a  $\sigma_i$  is 2.

## Relation to VC theorem cont.

- By Sauer's lemma there are at most  $(em/d)^d$  we can bound the probability that the sum is bounded by  $\epsilon$  for all functions by

$$\left(\frac{em}{d}\right)^d \exp\left(-\frac{\epsilon^2}{2m}\right) =: \delta.$$

- Taking  $\delta = m^{-1}$  and solving for  $\epsilon$  gives

$$\epsilon = \sqrt{2md \ln \frac{em}{d} + 2m \ln(m)}$$

## Rademacher bound for VC class

- Hence we can bound

$$\begin{aligned}\hat{R}_m(H) &\leq \frac{2}{m} (\delta m + \epsilon(1 - \delta)) \\ &\leq \frac{2}{m} + \sqrt{\frac{8(d \ln(em/d) + \ln(m))}{m}}\end{aligned}$$

- This is equivalent to the PAC bound with non-zero loss, except that we could have used the growth function or VC dimension measured on the sample rather than the sup over the whole input space.

# Application to large margin classification

- Rademacher complexity comes into its own for Boosting and SVMs.

## Application to Boosting

- We can view Boosting as seeking a function from the class ( $H$  is the set of weak learners)

$$\left\{ \sum_{h \in H} a_h h(\mathbf{x}) : a_h \geq 0, \sum_{h \in H} a_h \leq B \right\} = \text{conv}_B(H)$$

by minimising some function of the margin distribution.

- Adaboost corresponds to optimising an exponential function of the margin over this set of functions.
- We will see how to include the margin in the analysis later, but concentrate on computing the Rademacher complexity for now.

## Rademacher complexity of convex hulls

Rademacher complexity has a very nice property for convex hull classes:

$$\begin{aligned}\hat{R}_m(\text{conv}_B(H)) &= \frac{2}{m} \mathbb{E}_\sigma \left[ \sup_{h_j \in H, \sum_j a_j \leq B} \sum_{i=1}^m \sigma_i \sum_j a_j h_j(\mathbf{x}_i) \right] \\ &\leq \frac{2}{m} \mathbb{E}_\sigma \left[ \sup_{h_j \in H, \sum_j a_j \leq B} \sum_j a_j \sum_{i=1}^m \sigma_i h_j(\mathbf{x}_i) \right] \\ &\leq \frac{2}{m} \mathbb{E}_\sigma \left[ \sup_{h_j \in H} B \sum_{i=1}^m \sigma_i h_j(\mathbf{x}_i) \right] \\ &\leq B \hat{R}_m(H).\end{aligned}$$

## Rademacher complexity of convex hulls cont.

- Hence, we can move to the convex hull without incurring any complexity penalty for  $B = 1$ !



## Rademacher complexity for SVMs

- The Rademacher complexity of a class of linear functions with bounded 2-norm:

$$\left\{ \mathbf{x} \rightarrow \sum_{i=1}^m \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}) : \alpha' \mathbf{K} \alpha \leq B^2 \right\} \subseteq \\ \subseteq \{ \mathbf{x} \rightarrow \langle \mathbf{w}, \phi(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq B \} \\ = \mathcal{F}_B,$$

where we assume a kernel defined feature space with

$$\langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle = \kappa(\mathbf{x}, \mathbf{z}).$$

## Rademacher complexity of $\mathcal{F}_B$

$$\begin{aligned}\hat{R}_m(\mathcal{F}_B) &= \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}_B} \frac{2}{m} \sum_{i=1}^m \sigma_i f(\mathbf{x}_i) \right] \\ &= \mathbb{E}_\sigma \left[ \sup_{\|\mathbf{w}\| \leq B} \left\langle \mathbf{w}, \frac{2}{m} \sum_{i=1}^m \sigma_i \phi(\mathbf{x}_i) \right\rangle \right] \\ &\leq \frac{2B}{m} \mathbb{E}_\sigma \left[ \left\| \sum_{i=1}^m \sigma_i \phi(\mathbf{x}_i) \right\| \right] \\ &= \frac{2B}{m} \mathbb{E}_\sigma \left[ \left( \left\langle \sum_{i=1}^m \sigma_i \phi(\mathbf{x}_i), \sum_{j=1}^m \sigma_j \phi(\mathbf{x}_j) \right\rangle \right)^{1/2} \right] \\ &\leq \frac{2B}{m} \left( \mathbb{E}_\sigma \left[ \sum_{i,j=1}^m \sigma_i \sigma_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \right] \right)^{1/2} = \frac{2B}{m} \sqrt{\sum_{i=1}^m \kappa(\mathbf{x}_i, \mathbf{x}_i)}\end{aligned}$$



## Applying to 1-norm SVMs

We take the following formulation of the 1-norm SVM:

$$\begin{aligned} \min_{\mathbf{w}, b, \gamma, \xi} \quad & -\gamma + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i (\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b) \geq \gamma - \xi_i, \xi_i \geq 0, \\ & i = 1, \dots, m, \text{ and } \|\mathbf{w}\|^2 = 1. \end{aligned} \tag{1}$$

Note that

$$\xi_i = (\gamma - y_i g(\mathbf{x}_i))_+,$$

where  $g(\cdot) = \langle \mathbf{w}, \phi(\cdot) \rangle + b$ .

- The first step is to introduce a loss function which upper bounds the discrete loss

$$P(y \neq \text{sgn}(g(\mathbf{x}))) = \mathbb{E} [\mathcal{H}(-yg(\mathbf{x}))],$$

where  $\mathcal{H}$  is the Heaviside function.

## Applying the Rademacher theorem

- Consider the loss function  $\mathcal{A} : \mathbb{R} \rightarrow [0, 1]$ , given by

$$\mathcal{A}(a) = \begin{cases} 1, & \text{if } a > 0; \\ 1 + a/\gamma, & \text{if } -\gamma \leq a \leq 0; \\ 0, & \text{otherwise.} \end{cases}$$

- By the Rademacher Theorem and since the loss function  $\mathcal{A} - 1$  dominates  $\mathcal{H} - 1$ , we have that

$$\begin{aligned} \mathbb{E} [\mathcal{H}(-yg(\mathbf{x})) - 1] &\leq \mathbb{E} [\mathcal{A}(-yg(\mathbf{x})) - 1] \\ &\leq \hat{\mathbb{E}} [\mathcal{A}(-yg(\mathbf{x})) - 1] + \\ &\quad \hat{R}_m((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \end{aligned}$$

## Empirical loss and slack variables

- But the function  $\mathcal{A}(-y_i g(\mathbf{x}_i)) \leq \xi_i/\gamma$ , for  $i = 1, \dots, \ell$ , and so

$$\mathbb{E} [\mathcal{H}(-y g(\mathbf{x}))] \leq \frac{1}{m\gamma} \sum_{i=1}^m \xi_i + \hat{R}_m((\mathcal{A} - 1) \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

- The final missing ingredient to complete the bound is to bound  $\hat{R}_m((\mathcal{A} - 1) \circ \mathcal{F})$  in terms of  $\hat{R}_m(\mathcal{F})$ .
- This will require a more detailed look at Rademacher complexity.

## Rademacher complexity bounds

- First simple observation:

$$\text{for } a \in \mathbb{R}, \quad \hat{R}_m(a\mathcal{F}) = |a| \hat{R}_m(\mathcal{F}),$$

since  $|a|f$  is the function achieving the sup for some  $\sigma$  for  $a\mathcal{F}$  iff  $f$  achieves the sup for  $\mathcal{F}$ .

- We are interested in bounding RC

$$\hat{R}_m(\mathcal{L} \circ \mathcal{F}) \leq L \hat{R}_m(\mathcal{F})$$

for class  $\mathcal{L} \circ \mathcal{F} = \{\mathcal{L} \circ f : f \in \mathcal{F}\}$ , where  $\mathcal{L}$  satisfies

$$|\mathcal{L}(a) - \mathcal{L}(b)| \leq L|a - b|,$$

i.e.  $\mathcal{L}$  is a Lipschitz function with constant  $L > 0$ .

## Rademacher complexity bounds cont.

- In our case  $\mathcal{L} = \mathcal{A} - 1$  and  $L = 1/\gamma$ .
- By above it is sufficient to prove for case  $L = 1$  only, since then

$$\hat{R}_m(\mathcal{L} \circ \mathcal{F}) = L\hat{R}_m((\mathcal{L}/L) \circ \mathcal{F}) \leq L\hat{R}_m(\mathcal{F})$$



## Contraction proof ( $L = 1$ )

- We show

$$\mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i \mathcal{L}(f(\mathbf{x}_i)) \right] \leq \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}} \sigma_1 f(\mathbf{x}_1) + \sum_{i=2}^m \sigma_i \mathcal{L}(f(\mathbf{x}_i)) \right]$$

and apply induction to obtain the full result.

## Contraction proof cont.

- We take the permutations in pairs:

$$(1, \sigma_2, \dots, \sigma_m) \text{ and } (-1, \sigma_2, \dots, \sigma_m)$$

The result will follow if we show that for all  $f, g \in \mathcal{F}$  we can find  $f', g' \in \mathcal{F}$  such that ( $f_i = f(\mathbf{x}_i)$  etc.)

$$\begin{aligned} \mathcal{L}(f_1) + \sum_{i=2}^m \sigma_i \mathcal{L}(f_i) - \mathcal{L}(g_1) + \sum_{i=2}^m \sigma_i \mathcal{L}(g_i) \\ \leq f'_1 + \sum_{i=2}^m \sigma_i \mathcal{L}(f'_i) - g'_1 + \sum_{i=2}^m \sigma_i \mathcal{L}(g'_i). \end{aligned}$$

## Contraction proof end

- If  $f_1 \geq g_1$  take  $f' = f$  and  $g' = g$  to reduce to showing

$$\mathcal{L}(f_1) - \mathcal{L}(g_1) \leq f_1 - g_1 = |f_1 - g_1|$$

which follows since  $|\mathcal{L}(f_1) - \mathcal{L}(g_1)| \leq |f_1 - g_1|$ .

- Otherwise  $f_1 < g_1$  and we take  $f' = g$  and  $g' = f$  to reduce to showing

$$\mathcal{L}(f_1) - \mathcal{L}(g_1) \leq g_1 - f_1 = |f_1 - g_1|$$

which again follows from

$$|\mathcal{L}(f_1) - \mathcal{L}(g_1)| \leq |f_1 - g_1|.$$

## Final SVM bound

- Assembling the result we obtain:

$$\begin{aligned} P(y \neq \text{sgn}(g(\mathbf{x}))) &= \mathbb{E}[\mathcal{H}(-yg(\mathbf{x}))] \\ &\leq \frac{1}{m\gamma} \sum_{i=1}^m \xi_i + \frac{2}{m\gamma} \sqrt{\sum_{i=1}^m \kappa(\mathbf{x}_i, \mathbf{x}_i)} + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \end{aligned}$$

- Note that for the Gaussian kernel this reduces to

$$P(y \neq \text{sgn}(g(\mathbf{x}))) \leq \frac{1}{m\gamma} \sum_{i=1}^m \xi_i + \frac{2}{\sqrt{m}\gamma} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

## Final Boosting bound

- Applying a similar strategy for Boosting with the 1-norm of the slack variables we arrive at Linear programming boosting that minimises

$$\sum_h a_h + C \sum_{i=1}^m \xi_i,$$

where  $\xi_i = (1 - y_i \sum_h a_h h(\mathbf{x}_i))_+$ .

- with corresponding bound:

$$\begin{aligned} P(y \neq \text{sgn}(g(\mathbf{x}))) &= \mathbb{E} [\mathcal{H}(-yg(\mathbf{x}))] \\ &\leq \frac{1}{m} \sum_{i=1}^m \xi_i + \hat{R}(H) \sum_h a_h + 3 \sqrt{\frac{\ln(2/\delta)}{2m}} \end{aligned}$$

## Linear programming machine

- Controversy of why boosting works and relation to bagging.
- The previous bound suggests an optimisation similar to that of SVMs.
- seeks linear function in a feature space defined explicitly.
- For example using the 1-norm it seeks  $\mathbf{w}$  to solve

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \|\mathbf{w}\|_1 + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, \\ & i = 1, \dots, m. \end{aligned}$$

## Linear programming boosting

- Very slight generalisation considers the features as a set  $\mathbf{H}_{ij}$  of ‘weak’ learners (and include the constant function as one weak learner and negative of each weak learner):

$$\begin{aligned} \min_{\mathbf{a}, \xi} \quad & \|\mathbf{a}\|_1 + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i \mathbf{H}_i \mathbf{a} \geq 1 - \xi_i, \xi_i \geq 0, a_i \geq 0 \\ & i = 1, \dots, m. \end{aligned}$$

## Alternative version

- Can explicitly optimise margin with 1-norm fixed:

$$\begin{aligned} \max_{\rho, \mathbf{a}, \xi} \quad & \rho - D \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y_i \mathbf{H}_i \mathbf{a} \geq \rho - \xi_i, \xi_i \geq 0, a_j \geq 0 \\ & \sum_{j=1}^N a_j = 1. \end{aligned}$$

- Dual has the following form:

$$\begin{aligned} \min_{\beta, \mathbf{u}} \quad & \beta \\ \text{subject to} \quad & \sum_{i=1}^m u_i y_i \mathbf{H}_{ij} \leq \beta, j = 1, \dots, N, \\ & \sum_{i=1}^m u_i = 1, 0 \leq u_i \leq D. \end{aligned}$$



## Column generation

Can solve the dual linear programme using an iterative method:

- 1 initialise  $u_i = 1/m, i = 1, \dots, m, \beta = \infty, J = \emptyset$
- 2 choose  $j^*$  that maximises  $f(j) = \sum_{i=1}^m u_i y_i \mathbf{H}_{ij}$
- 3 if  $f(j^*) \leq \beta$  solve primal restricted to  $J$  and exit
- 4  $J = J \cup \{j^*\}$
- 5 Solve dual restricted to set  $J$  to give  $u_i, \beta$
- 6 Go to 2

- Note that  $u_i$  is a distribution on the examples
- Each  $j$  added acts like an additional weak learner
- $f(j)$  is simply the weighted classification accuracy
- Hence gives 'boosting' algorithm - with previous weights updated satisfying error bound
- Guaranteed convergence and soft stopping criteria

## Multiple kernel learning

- MKL puts a 1-norm constraint on a linear combination of kernels:

$$\left\{ \kappa(\mathbf{x}, \mathbf{z}) = \sum_{t=1}^N z_t \kappa_t(\mathbf{x}, \mathbf{z}) : z_t \geq 0, \sum_{t=1}^N z_t = 1 \right\}$$

and trains an SVM while optimizing  $z_t$  – a convex problem, c.f. group Lasso.

- obtain corresponding bound:

$$P(y \neq \text{sgn}(g(\mathbf{x})))$$

$$\leq \frac{1}{m\gamma} \sum_{i=1}^m \xi_i + \frac{1}{\gamma} \hat{R}_m \left( \bigcup_{t=1}^N \mathcal{F}_t \right) + 3 \sqrt{\frac{\ln(2/\delta)}{2m}}$$

## Bounding MKL

- Need a bound on

$$\hat{R}_m \left( \mathcal{F} = \bigcup_{t=1}^N \mathcal{F}_t \right)$$

where  $\mathcal{F}_t = \{\mathbf{x} \rightarrow \langle \mathbf{w}, \phi_t(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq 1\}$ .

- First note further applications of McDiarmid gives with probability  $1 - \delta_0$  of a random selection of  $\sigma^*$ :

$$\hat{R}_m(\mathcal{F}) \leq \frac{2}{m} \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i^* f(\mathbf{x}_i) + 4 \sqrt{\frac{\ln(1/\delta_t)}{2m}}$$

and  $\frac{2}{m} \sup_{f \in \mathcal{F}_t} \sum_{i=1}^m \sigma_i^* f(\mathbf{x}_i) \leq \hat{R}_m(\mathcal{F}_t) + 4 \sqrt{\frac{\ln(1/\delta_t)}{2m}}$

with probability  $1 - \delta_t$

## Bounding MKL

- Hence taking  $\delta_t = \delta/2(N + 1)$  for  $t = 0, \dots, N$

$$\begin{aligned} \hat{R}_m \left( \mathcal{F} = \bigcup_{t=1}^N \mathcal{F}_t \right) &\leq \frac{2}{m} \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i^* f(\mathbf{x}_i) + 4 \sqrt{\frac{\ln(2(N + 1)/\delta)}{2m}} \\ &\leq \frac{2}{m} \max_{1 \leq t \leq N} \sup_{f \in \mathcal{F}_t} \sum_{i=1}^m \sigma_i^* f(\mathbf{x}_i) + 4 \sqrt{\frac{\ln(2(N + 1)/\delta)}{2m}} \\ &\leq \frac{2}{m} \max_{1 \leq t \leq N} \hat{R}_m(\mathcal{F}_t) + 8 \sqrt{\frac{\ln(2(N + 1)/\delta)}{2m}} \end{aligned}$$

with probability  $1 - \delta/2$ .

# Conclusions

- Outline of philosophy and approach of SLT
- Central result of SLT
- Touched on covering number analysis for margin based analysis
- Moved to consideration of Rademacher complexity.
- Case of RC for classification giving bounds for two of the most effective classification algorithms: SVMs and Boosting.
- Similar bounds can be obtained for regression and other tasks such as PCA.

## Where to find out more

**Web Sites:** `www.support-vector.net` (SV Machines)

`www.kernel-methods.net` (kernel methods)

`www.kernel-machines.net` (kernel Machines)

`www.neurocolt.com`

`www.pascal-network.org`

## References

- [1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive Dimensions, Uniform Convergence, and Learnability. *Journal of the ACM*, 44(4):615–631, 1997.
- [2] Amiran Ambroladze and John Shawe-Taylor. Complexity of pattern classes and lipschitz property. In *Fifteenth*

*International Conference on Algorithmic Learning Theory, ALT'04*, pages 181–193. Springer Verlag, 2004.

- [3] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
- [4] M. Anthony and N. Biggs. *Computational Learning Theory*, volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1992.
- [5] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.
- [6] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Math J.*, 19:357–367, 1967.
- [7] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 43–54, Cambridge, MA, 1999. MIT Press.
- [8] P. L. Bartlett. The sample complexity of pattern

classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

- [9] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
- [10] S. Boucheron, G. Lugosi, , and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, pages vol.16, pp.277–292, 2000.
- [11] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
- [12] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.
- [13] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to



- boosting. In *Computational Learning Theory: Eurocolt '95*, pages 23–37. Springer-Verlag, 1995.
- [14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Stat. Assoc.*, 58:13–30, 1963.
- [15] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.
- [16] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, pages 443 – 459, 2000.
- [17] J. Langford and J. Shawe-Taylor. PAC bayes and margins. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003. MIT Press.
- [18] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.
- [19] C. McDiarmid. On the method of bounded differences. In 141 London Mathematical Society Lecture Notes Series, editor, *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge, 1989.
- [20] R. Schapire, Y. Freund, P. Bartlett, and W. Sun

Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998. (To appear. An earlier version appeared in: D.H. Fisher, Jr. (ed.), Proceedings ICML97, Morgan Kaufmann.).

- [21] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
- [22] J. Shawe-Taylor and N. Cristianini. On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*, 48(10):2721–2735, 2002.
- [23] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.
- [24] J. Shawe-Taylor, C. Williams, N. Cristianini, and J. S. Kandola. On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *Proceedings of the 13th International Conference on Algorithmic Learning Theory (ALT2002)*, volume 2533,

pages 23–40, 2002.

- [25] M. Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126:505–563, 1996.
- [26] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.
- [27] V. Vapnik and A. Chervonenkis. Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:915 – 918, 1968.
- [28] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.
- [29] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.