Principles of Statistical Machine Learning
Basic Elements of Statistical Learning Theory
Vapnik-Chervonenkis Theory on Binary Classification

Ernest Fokoué
福尔特　教授

School of Mathematical Sciences
Rochester Institute of Technology
Rochester, New York, USA

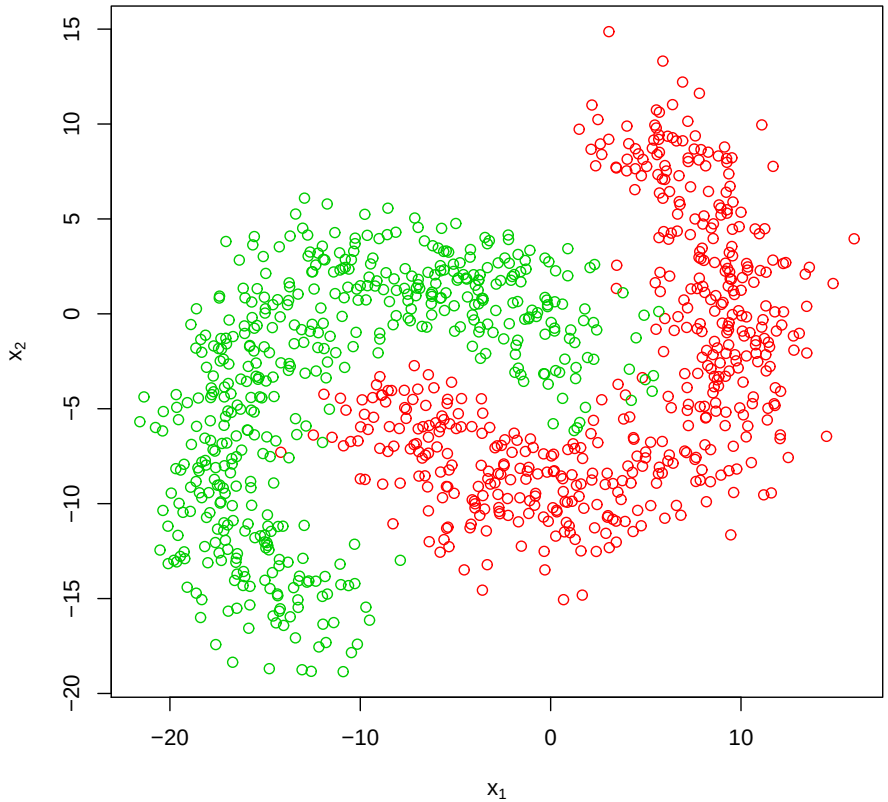
Statistical Machine Learning and Data Science
African Institute for Mathematical Sciences (AIMS)
Kigali (Rwanda)-January 2018

"To understand God's thoughts, one must study statistics ... the measure of His purpose."

Florence Nightingale

Binary Classification in the Plane (2D space)

Consider the following two dimensional binary classification task.



Question: Can the two classes ever be separated by a line? If not, what is the "best" classifier here?

Binary Classification in the Plane

For the binary classification problem introduced earlier:

- *A collection* $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ *of i.i.d. observations is* given
  - $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^2, i = 1, \cdots, n$. $\mathcal{X}$ *is the* input *space.*
  - $y_i \in \{-1, +1\}$. $\mathcal{Y} = \{-1, +1\}$ *is the* output *space.*
- *What is the probability law that governs the* $(\mathbf{x}_i, y_i)$*'s?*
- *What is the functional relationship between* $\mathbf{x}$ *and* y*?*
- What is the "best" approach to determining from the available observations, the relationship between $\mathbf{x}$ and y in such a way that, *given a new (unseen) observation* $\mathbf{x}^{\text{new}}$*, its class* y^{new} *can be predicted as accurately as possible.*

Basic Remarks on Classification

- *While some points clearly belong to one of the classes, there are other points that are either strangers in a foreign land, or are positioned in such a way that no automatic classification rule can clearly determine their class membership.*
- *One can construct a classification rule that puts all the points in their corresponding classes. Such a rule would prove disastrous in classifying new observations not present in the current collection of observations.*
- *Indeed, we have a collection of pairs* $(\mathbf{x}_i, y_i)$ *of observations coming from some unknown distribution* $\mathbb{P}(\mathbf{x}, y)$.

Basic Remarks on Classification

- *Finding an automatic classification rule that achieves the absolute very best on the present data is not enough since infinitely many more observations can be generated by $\mathbb{P}(\mathbf{x}, y)$ for which good classification will be required.*
- *Even the universally best classifier will make mistakes.*
- *Of all the functions in $\mathcal{Y}^{\mathcal{X}}$, it is reasonable to assume that there is a function f^* that maps any $\mathbf{x} \in \mathcal{X}$ to its corresponding $y \in \mathcal{Y}$, i.e.,*

$$\begin{array}{rccc} f^*: & \mathcal{X} & \to & \mathcal{Y} \\ & \mathbf{x} & \mapsto & f^*(\mathbf{x}), \end{array}$$

with the minimum number of mistakes.

Risk Minimization Revisited

- *Let f denote any generic function mapping an element $\mathbf{x}$ of $\mathcal{X}$ to its corresponding image $f(\mathbf{x})$ in $\mathcal{Y}$.*
- *Each time $\mathbf{x}$ is drawn from $\mathbb{P}(\mathbf{x})$, the disagreement between the image $f(\mathbf{x})$ and the true image y is called the loss, denoted by $\ell(y, f(\mathbf{x}))$.*
- *The expected value of this loss function with respect to the distribution $\mathbb{P}(\mathbf{x}, y)$ is called the risk functional of f. Generically, we shall denote the risk functional of f by $R(f)$, so that*

$$R(f) = \mathbb{E}[\ell(Y, f(X))] = \int \ell(y, f(\mathbf{x}))d\mathbb{P}(\mathbf{x}, y).$$

- *The best function f^* over the space $\mathcal{Y}^{\mathcal{X}}$ of all measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ is therefore*

$$f^* = \arg\inf_f R(f),$$

so that

$$R(f^*) = R^* = \inf_f R(f).$$

On the need to reduce the search space

- *Unfortunately, f^* can only be found if $\mathbb{P}(\mathbf{x}, y)$ is known. Therefore, since we do not know $\mathbb{P}(\mathbf{x}, y)$ in practice, it is hopeless to determine f^*.*
- *Besides, trying to find f^* without the knowledge of $\mathbb{P}(\mathbf{x}, y)$ implies having to search the infinite dimensional function space $\mathcal{Y}^{\mathcal{X}}$ of all mappings from $\mathcal{X}$ to $\mathcal{Y}$, which is an ill-posed and computationally nasty problem.*
- *Throughout this lecture, we will seek to solve the more reasonable problem of choosing from a function space $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, the one function $f^+ \in \mathcal{F}$ that best estimates the dependencies between $\mathbf{x}$ and y.*
- *It is therefore important to define what is meant by best estimates. For that, the concepts of loss function and risk functional need to be define.*

Loss and Risk in Pattern Recognition

For this classification/pattern recognition, the so-called 0-1 *loss function defined below is used. More specifically,*

$$\ell(y, f(\mathbf{x})) = \mathbf{1}_{\{Y \neq f(X)\}} = \begin{cases} 0 & \text{if} \quad y = f(\mathbf{x}), \\ 1 & \text{if} \quad y \neq f(\mathbf{x}). \end{cases} \tag{1}$$

The corresponding risk functional is

$$R(f) = \int \ell(y, f(\mathbf{x})) d\mathbb{P}(\mathbf{x}, y) = \mathbb{E}\left[\mathbf{1}_{\{Y \neq f(X)\}}\right] = \underset{(X,Y)\sim\mathbb{P}}{\Pr}[Y \neq f(X)].$$

The minimizer of the 0-1 *risk functional over all possible classifiers is the so-called* Bayes classifier *which we shall denote here by* f^* *given by*

$$f^* = \arg\inf_f \left\{ \underset{(X,Y)\sim\mathbb{P}}{\Pr}[Y \neq f(X)] \right\}.$$

Specifically, the Bayes' classifier f^* *is given by the posterior probability of class membership, namely*

$$f^*(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \left\{ \Pr[Y = y | \mathbf{x}] \right\}.$$

Function Class in Pattern Recognition

As stated earlier, trying to find f^* is hopeless. One needs to select a function space $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, and then choose the best estimator f^+ from $\mathcal{F}$, i.e.,

$$f^+ = \arg \inf_{f \in \mathcal{F}} R(f),$$

so that

$$R(f^+) = R^+ = \inf_{f \in \mathcal{F}} R(f).$$

For the binary pattern recognition problem, one may consider finding the best linear separating hyperplane, i.e.

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \{-1, +1\} | \ \exists \alpha_0 \in \mathbb{R}, (\alpha_1, \cdots, \alpha_p)^\top = \boldsymbol{\alpha} \in \mathbb{R}^p | \right.$$

$$\left. f(\mathbf{x}) = \text{sign}\left(\boldsymbol{\alpha}^\top \mathbf{x} + \alpha_0\right), \ \forall \mathbf{x} \in \mathcal{X} \right\}$$

Empirical Risk Minimization

- *Let* $\mathcal{D} = \Big\{(X_1, Y_1), \cdots, (X_n, Y_n)\Big\}$ *be an iid sample from* $\mathbb{P}(\mathbf{x}, \mathrm{y})$.
- *The empirical version of the risk functional is*

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq f(X_i)\}}$$

- *We therefore seek the best by empirical standard,*

$$\widehat{f} = \underset{f \in \mathcal{F}}{\mathtt{arg\,min}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq f(X_i)\}} \right\}$$

Since it is impossible to search all possible functions, it is usually crucial to choose the "right" function space $\mathcal{F}$.

Bias-Variance Trade-Off

In traditional statistical estimation, one needs to address at the very least issues like: (a) the Bias *of the estimator; (b) the* Variance *of the estimator; (c) The* consistency *of the estimator; Recall from elementary point estimation that, if θ is the true value of the parameter to be estimated, and $\widehat{\theta}$ is a point estimator of θ, then one can decompose the total error as follows:*

$$\widehat{\theta} - \theta = \underbrace{\widehat{\theta} - \mathbb{E}[\widehat{\theta}]}_{\text{Estimation error}} + \underbrace{\mathbb{E}[\widehat{\theta}] - \theta}_{\text{Bias}} \tag{2}$$

Under the squared error loss, one seeks $\widehat{\theta}$ that minimizes the mean squared error,

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}[(\widehat{\theta} - \theta)^2] = \arg\min_{\theta \in \Theta} \text{MSE}(\widehat{\theta}),$$

rather than trying to find the minimum variance unbiased estimator (MVUE).

Bias-Variance Trade-off

Clearly, the traditional so-called bias-variance decomposition of the MSE *reveals the need for bias-variance trade-off. Indeed,*

$$\begin{aligned}\text{MSE}(\widehat{\theta}) = \mathbb{E}[(\widehat{\theta} - \theta)^2] &= \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2] + \mathbb{E}[(\mathbb{E}[\widehat{\theta}] - \theta)^2] \\ &= \text{variance} + \text{bias}^2\end{aligned}$$

If the estimator $\widehat{\theta}$ were to be sought from all possible value of θ, then it might make sense to hope for the MVUE. Unfortunately - an especially in function estimation as we clearly argued earlier - there will be some bias, so that the error one gets has a bias component along with the variance component in the squared error loss case. If the bias is too small, then an estimator with a larger variance is obtained. Similarly, a small variance will tend to come from estimators with a relatively large bias. The best compromise is then to trade-off bias and variance. Which is in functional terms translates into trade-off between approximation error and estimation error.

"When you have two competing theories that make exactly the same predictions, the simpler one is the better."

William of Ockham

Bias-Variance Trade-off

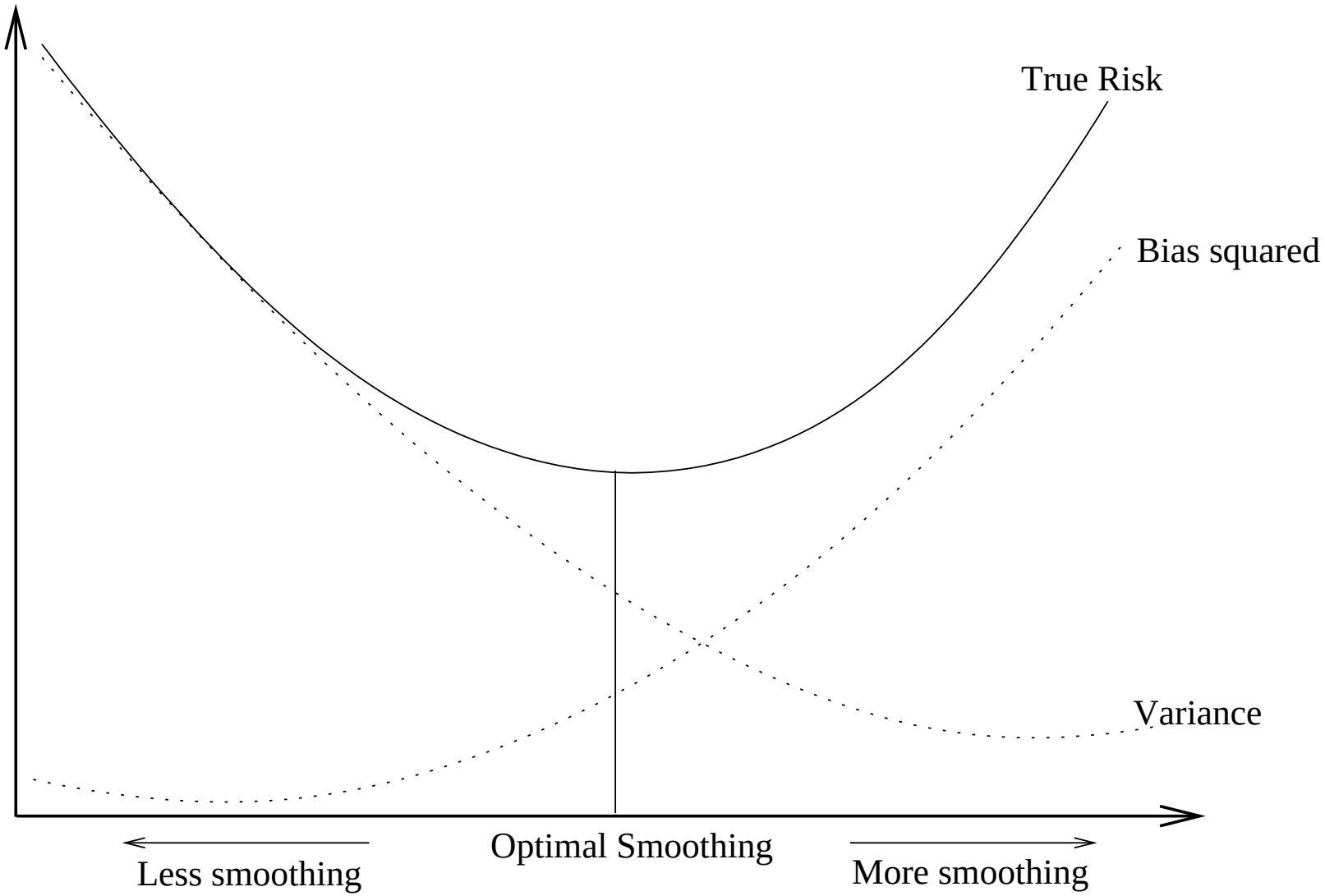


Figure: *Illustration of the qualitative behavior of the dependence of bias versus variance on a tradeoff parameter such as* λ *or* h*. For small values the variability is too high; for large values the bias gets large.*

Structural risk minimization principle

Since making the estimator of the function arbitrarily complex causes the problems mentioned earlier, the intuition for a trade-off reveals that instead of minimizing the empirical risk $\widehat{R}_n(f)$ *one should do the following:*

- *Choose a collection of function spaces* $\{\mathcal{F}_k : \ k = 1, 2, \cdots\}$*, maybe a collection of nested spaces (increasing in size)*
- *Minimize the empirical risk in each class*
- *Minimize the penalized empirical risk*

$$\min_k \min_{f \in \mathcal{F}_k} \widehat{R}_n(f) + \text{penalty}(k, n)$$

where $\text{penalty}(k, n)$ *gives preference to models with small estimation error. It is important to note that* $\text{penalty}(k, n)$ *measures the capacity of the function class* $\mathcal{F}_k$*. The widely used technique of regularization for solving ill-posed problem is a particular instance of structural risk minimization.*

Regularization for Complexity Control

- Tikhonov's Variation Approach to Regularization[Tikhonov, 1963]
  Find f that minimizes the functional

$$\widehat{R}_n^{(\mathtt{reg})}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(\mathbf{x}_i)) + \lambda\Omega(f)$$

  where $\lambda > 0$ is some predefined constant.

- Ivanov's Quasi-solution Approach to Regularization[Ivanov, 1962]
  Find f that minimizes the functional

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(\mathbf{x}_i))$$

  subject to the constraint

$$\Omega(f) \leq C$$

  where $C > 0$ is some predefined constant.

Regularization for Complexity Control

- Philips' Residual Approach to Regularization[Philips, 1962]
  Find f that minimizes the functional

$$\Omega(f)$$

  subject to the constraint

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f(\mathbf{x}_i)) \leq \mu$$

  where $\mu > 0$ is some predefined constant.

In all the above, the functional $\Omega(f)$ is called the regularization functional. $\Omega(f)$ is defined in such a way that it controls the complexity of the function f.

$$\Omega(f) = \|f\|^2 = \int_a^b (f''(t))^2 dt.$$

is a regularization functional used in spline smoothing.

Statistical Consistency

- **Definition:** *Let $\widehat{\theta}_n$ be an estimator of some scalar quantity θ based on an i.i.d. sample $X_1, X_2, \cdots, X_n$ from the distribution with parameter θ. Then, $\widehat{\theta}_n$ is said to be a* consistent *estimator of θ, if $\widehat{\theta}_n$ converges in probability to θ, i.e.,*

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{\mathrm{P}} \theta.$$

In other words, $\widehat{\theta}_n$ is a consistent *estimator of θ if, $\forall \epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{|\widehat{\theta}_n - \theta| > \epsilon\right\} = 0.$$

- *It turns out that for unbiased estimators $\widehat{\theta}_n$, consistency is straightforward as direct consequence of a basic probabilistic inequality like Chebyshev's inequality. However, for unbiased estimators, one has to be more careful.*

A Basic Important Inequality

定理

(Biename-Chebyshev's inequality) *Let X be a random variable with finite mean $\mu_X = \mathbb{E}[X]$ i.e. $|\mathbb{E}[X]| < +\infty$ and finite variance $\sigma_X^2 = \mathbb{V}(X)$, i.e., $|\mathbb{V}(X)| < +\infty$. Then, $\forall \epsilon > 0$,*

$$\Pr[|X - \mathbb{E}[X]| > \epsilon] \leq \frac{\mathbb{V}(X)}{\epsilon^2}.$$

It is therefore easy to see here that, with unbiased $\widehat{\theta}_n$, one has $\mathbb{E}[\widehat{\theta}_n] = \theta$, and the result is immediate. For the sake of clarity, let's recall here the elementary weak law of large numbers.

Weak Law of Large Numbers

Let X be a random variable with finite mean $\mu_X = \mathbb{E}[X]$ i.e. $|\mathbb{E}[X]| < +\infty$ and finite variance $\sigma_X^2 = \mathbb{V}(X)$, i.e., $|\mathbb{V}(X)| < +\infty$. Let $X_1, X_2, \cdots, X_n$ be a random sample of n observations drawn independently from the distribution of X, so that for $i = 1, \cdots, n$, we have $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$. Let $\bar{X}_n$ be the sample mean, i.e.,

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then, clearly, $\mathbb{E}[\bar{X}_n] = \mu$, and, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} \left\{ \Pr[|\bar{X}_n - \mu| > \epsilon] \right\} = 0. \tag{3}$$

This essentially expresses the fact that the empirical mean $\bar{X}_n$ converges in probability *to the theoretical mean μ in the limit of very large samples.*

Weak Law of Large Numbers

We therefore have

$$\bar{X}_n \xrightarrow[n\to\infty]{P} \mu.$$

With $\mu_{\bar{X}} = \mathbb{E}[\bar{X}_n] = \mu$ *and* $\sigma^2_{\bar{X}} = \sigma^2/n$, *one applyies Biename-Chebyshev's inequality and gets:* $\forall \epsilon > 0$,

$$\Pr[|\bar{X} - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}, \tag{4}$$

which, by inversion, is the same as

$$|\bar{X} - \mu| < \sqrt{\frac{1}{\delta}\frac{\sigma^2}{n}} \tag{5}$$

with probability at least $1 - \delta$.
Why is all the above of any interest to statistical learning theory?

Weak Law of Large Numbers

Why is all the above of any interest to statistical learning theory?

- *Equation* (3) *states the much needed consistency of* $\bar{X}$ *as an estimator of* μ.
- *Equation* (4), *by showing the dependence of on* n *and* ε *helps assess the rate at which* $\bar{X}$ *converges to* μ.
- *Equation* (5), *by showing a confidence interval helps compute bounds on the unknown true mean* μ *as a function of the empirical mean* $\bar{X}$ *and the confidence level* $1 - \delta$.
- *Finally, how does go about constructing estimators with all the above properties.*

Components of Statistical Machine Learning

Interestingly, all those 4 components of classical estimation theory, will be encountered again in statistical learning theory. Essentially, the 4 components of statistical learning theory consist of finding the answers to the following questions:

- *(a)* What are the necessary and sufficient conditions for the consistency of a learning process based on the ERM principle? *This leads to the Theory of consistency of learning processes.*
- *(b)* How fast is the rate of convergence of the learning process? *This leads to the Nonasymptotic theory of the rate of convergence of learning processes;*
- *(c)* How can one control the rate of convergence (the generalization ability) of the learning process?. *This leads to the Theory of controlling the generalization ability of learning processes;*
- *(d)* How can one construct algorithms that can control the generalization ability of the learning process?. *This leads to Theory of constructing learning algorithms.*

Error Decomposition revisited

A reasoning on error decomposition and consistency of estimators along with rates, bounds and algorithms applies to function spaces: indeed, the difference between the true risk $R(\widehat{f}_n)$ associated with $\widehat{f}_n$ and the overall minimum risk R^* can be decomposed to explore in greater details the source of error in the function estimation process:

$$R(\widehat{f}_n) - R^* = \underbrace{R(\widehat{f}_n) - R(f^+)}_{\text{Estimation error}} + \underbrace{R(f^+) - R^*}_{\text{Approximation error}} \tag{6}$$

A reasoning similar to bias-variance trade-off and consistency can be made, with the added complication brought be the need to distinguish between the true risk functional and the empirical risk functional, and also to the added to assess both pointwise behaviors and uniform behaviors. In a sense, one needs to generalize the decomposition and the law of large numbers to function spaces.

Approximation-Estimation Trade-Off

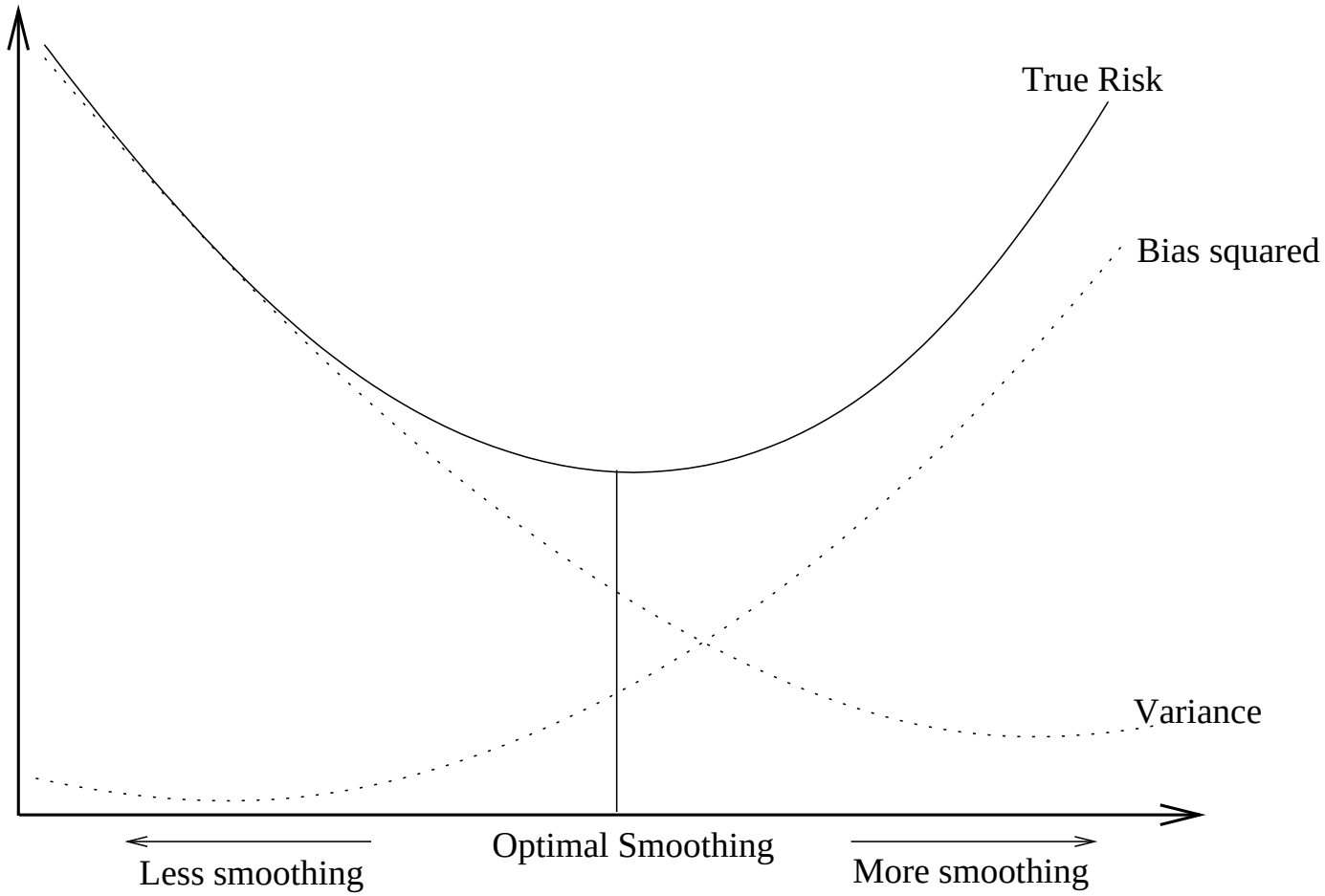


Figure: *Illustration of the qualitative behavior of the dependence of bias versus variance on a tradeoff parameter such as λ or h. For small values the variability is too high; for large values the bias gets large.*

Consistency of the Empirical Risk Minimization principle

- *The ERM principle is* consistent *if it provides a sequence of functions* $\hat{f}_n,\ n = 1, 2, \cdots$ *for which both the expected risk* $R(\widehat{f}_n)$ *and the empirical risk* $\widehat{R}_n(\widehat{f}_n)$ *converge to the minimal possible value of the risk* $R(f^+)$ *in the function class under consideration, i.e.,*

$$R(\hat{f}_n) \xrightarrow[n\to\infty]{P} \inf_{f\in\mathcal{F}} R(f) = R(f^+)$$

and

$$\widehat{R}_n(\hat{f}_n) \xrightarrow[n\to\infty]{P} \inf_{f\in\mathcal{F}} R(f) = R(f^+)$$

- *Vapnik discusses the details of this theorem at length, and extends the exploration to include the difference between what he calls* trivial *consistency and* non-trivial *consistency.*

Consistency of the Empirical Risk Minimization principle

- *To better understand consistency in function spaces, consider the sequence of random variables*

$$\xi^n = \sup_{f \in \mathcal{F}} \left| R(f) - \widehat{R}_n(f) \right|, \tag{7}$$

and consider studying

$$\lim_{n \to \infty} P \left\{ \sup_{f \in \mathcal{F}} \left| R(f) - \widehat{R}_n(f) \right| > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0.$$

- *Vapnik shows that the sequence of the means of the random variable ξ^n converges to zero as the number n of observations increases.*
- *He also remarks that the sequence of random variables ξ^n converges in probability to zero if the set of functions $\mathcal{F}$, contains a finite number m of elements. We will show that later in the case of pattern recognition.*

Consistency of the Empirical Risk Minimization principle

- *It remains then to describe the properties of the set of functions $\mathcal{F}$, and probability measure $\mathbb{P}(\mathbf{x}, y)$ under which the sequence of random variables ξ^n converges in probability to zero.*

$$\lim_{n\to\infty} P\left\{\left[\sup_{f\in\mathcal{F}}[R(f)-\widehat{R}_n(f)]>\varepsilon\right] \text{ or } \left[\sup_{f\in\mathcal{F}}[\widehat{R}_n(f)-R(f)]>\varepsilon\right]\right\}=0.$$

- *Recall that $\widehat{R}_n(f)$ is the realized disagreement between classifier f and the truth about the label y of $\mathbf{x}$ based on information contained in the sample $\mathcal{D}$.*
- *It is easy to see that, for a given (fixed) function (classifier) f,*

$$\mathbb{E}[\widehat{R}_n(f)] = R(f). \tag{8}$$

Note that while this pointwise unbiasedness of the empirical risk is a good bottomline property to have, it is not enough. More is needed as the comparison is against $R(f^+)$ or event better yet $R(f^*)$.

Consistency of the Empirical Risk

- *Remember that the goal of statistical function estimation is to devise a technique (strategy) that chooses from the function class $\mathcal{F}$, the one function whose true risk is as close as possible to the lowest risk in class $\mathcal{F}$.*
- *The question arises:* since one cannot calculate the true error, how can one devise a learning strategy for choosing classifiers based on it? *Tentative answer: At least devise strategies that yield functions for which the upper bound on the theoretical risk is as tight as possible, so that one can make confidence statements of the form:*
- *With probability $1 - \delta$ over an i.i.d. draw of some sample according to the distribution $\mathbb{P}$, the expected future error rate of some classifier is bounded by a function $g(\delta, \text{error rate on sample})$ of δ and the error rate on sample.*

$$\Pr\left\{\texttt{TestError} \leq \texttt{TrainError} + \phi(n, \delta, \kappa(\mathcal{F}))\right\} \leq 1 - \delta$$

Foundation Result in Statistical Learning Theory

*Theorem:(***Vapnik and Chervonenkis, 1971***) Let $\mathcal{F}$ be a class of functions implementing so learning machines, and let $\zeta = VCdim(\mathcal{F})$ be the VC dimension of $\mathcal{F}$. Let the theoretical and the empirical risks be defined as earlier and consider any data distribution in the population of interest. Then $\forall f \in \mathcal{F}$, the prediction error (theoretical risk) is bounded by*

$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{\zeta\left(\log\frac{2n}{\zeta}+1\right) - \log\frac{\eta}{4}}{n}} \tag{9}$$

with probability of at least $1-\eta$. or

$$\Pr\left\{\texttt{TestError} \leq \texttt{TrainError} + \sqrt{\frac{\zeta\left(\log\frac{2n}{\zeta}+1\right) - \log\frac{\eta}{4}}{n}}\right\} \leq 1-\eta$$

Bounds on the Generalization Error

For instance, using Chebyshev's inequality and the fact that $\mathbb{E}[\widehat{R}_n(f)] = R(f)$*, it is easy to see that, for given classifier* f *and a sample* $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$*,*

$$\Pr[|\widehat{R}_n(f) - R(f)| > \epsilon] \leq \frac{R(f)(1 - R(f))}{n\epsilon^2}.$$

To estimate the true but unknown error $R(f)$ *with a probability of at least* $1 - \delta$*, it makes sense to use inversion, i.e., set*

$$\delta = \frac{R(f)(1 - R(f))}{n\epsilon^2}, \quad \text{so that} \quad \epsilon = \sqrt{\frac{R(f)(1 - R(f))}{n\delta}}.$$

Owing to the fact that $\max_{R(f) \in [0,1]} R(f)(1 - R(f)) = \frac{1}{4}$*, we have*

$$\sqrt{\frac{R(f)(1 - R(f))}{n\delta}} < \sqrt{\frac{1}{4n\delta}} = \left(\frac{1}{4n\delta}\right)^{1/2}.$$

Bounds on the Generalization Error

- *Based on Chebyshev's inequality, for a given classifier f, with a probability of at least $1-\delta$, the bound on the difference between the true risk $R(f)$ and the empirical risk $\widehat{R}_n(f)$ is given by*

$$|\widehat{R}_n(f) - R(f)| < \left(\frac{1}{4n\delta}\right)^{1/2}.$$

- *Recall that one of the goals of statistical learning theory is to assess the rate of convergence of the empirical risk to the true risk, which translates into assessing how tight the corresponding bounds on the true risk are.*
- *In fact, it turns out many bounds can be so loose as to become useless. It turns out that the above Chebyshev-based bound is not a good one, at least compared to bounds obtained using the so-called hoeffding's inequality.*

Bounds on the Generalization Error

- **Theorem:**(Hoeffding's inequality) *Let $Z_1, Z_2, \cdots, Z_n$ be a collection of i.i.d random variables with $Z_i \in [a, b]$. Then, $\forall \epsilon > 0$,*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| > \epsilon\right] \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

- **corollary:**(hoeffding's inequality for sample proportions) *Let $Z_1, Z_2, \cdots, Z_n$ be a collection of i.i.d random variables from a Bernoulli distribution with "success" probability p. Let $\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Clearly, $\widehat{p}_n \in [0, 1]$ and $\mathbb{E}[\widehat{p}_n] = p$.*
- *Therefore, as a direct consequence of the above theorem, we have, $\forall \epsilon > 0$,*

$$\Pr\left[|\widehat{p}_n - p| > \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

Bounds on the Generalization Error

- *So we have,* $\forall \epsilon > 0$,

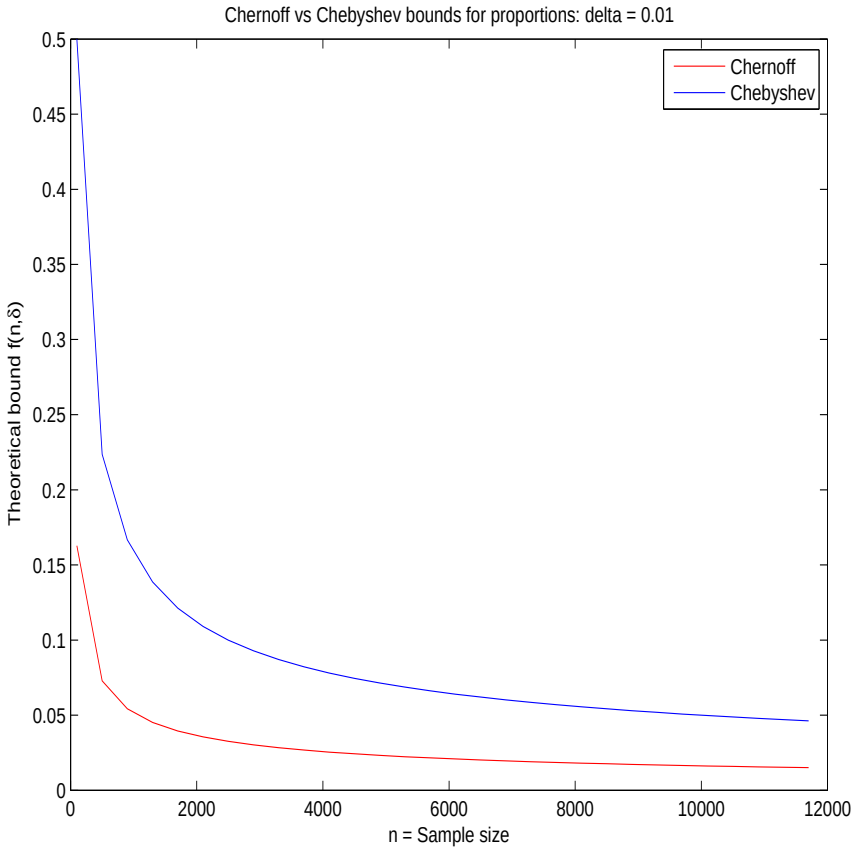
$$\Pr\left[|\widehat{p}_n - p| > \epsilon\right] \leq 2\exp\left(-2n\epsilon^2\right)$$

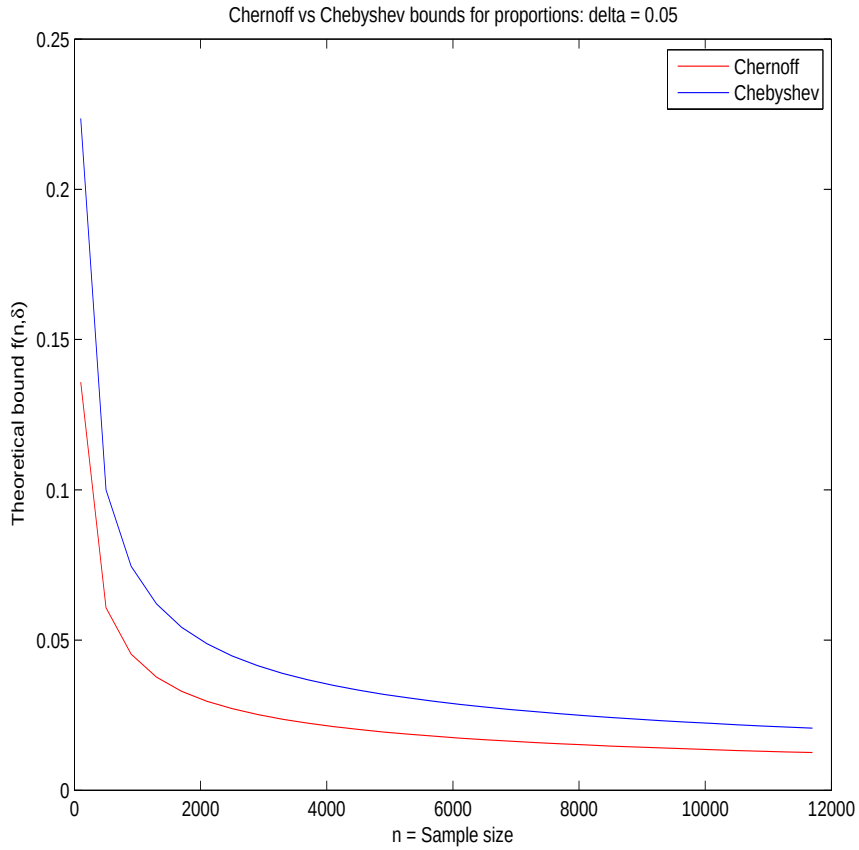
- *Now, setting* $\delta = 2\exp(-2\epsilon^2 n)$, *it is straightforward to see that the hoeffding-based* $1-\delta$ *level confidence bound on the difference between* $R(f)$ *and* $\widehat{R}_n(f)$ *for a fixed classifier* f *is given by*

$$|\widehat{R}_n(f) - R(f)| < \left(\frac{\ln\frac{2}{\delta}}{2n}\right)^{1/2}.$$

- *Which of the two bounds is tighter? Clearly, we need to find out which of* $\ln 2/\delta$ *or* $1/2\delta$ *is larger. This is the same as comparing* $\exp(1/2\delta)$ *and* $2/\delta$, *which in turns means comparing* $a^{(2/\delta)}$ *and* $2/\delta$ *where* $a = \exp(1/4)$. *With* $\delta > 0$, $a^{(2/\delta)} > 2/\delta$, *so that, we know that hoeffding's bounds are tighter. The graph also confirm this.*

Bounds on the Generalization Error





Beyond Chernov and Hoeffding

- *In all the above, we only addressed* pointwise convergence *of* $\widehat{R}_n(f)$ *to* $R(f)$, *i.e., for Fix a machine* $f \in \mathcal{F}$, *we studied the convergence of*

$$\widehat{R}_n(f) \quad \text{to} \quad R(f).$$

  Needless to mention that that pointwise convergence is of very little use here.

- *A more interesting issue to address is* uniform convergence. *That is, for all machines,* $f \in \mathcal{F}$, *determine the necessary and sufficient conditions for the convergence of*

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| > \epsilon \quad \text{to} \quad 0.$$

- *Clearly, such a study extends the Law of Large Numbers to function spaces, thereby providing tools for the construction of bounds on the theoretical errors of learning machines.*

Beyond Chernov and Hoeffding

- *Since uniform convergence requires the consideration of the entirety of the function space of interest, care needs to be taken regarding the dimensionality of the function space.*
- *Uniform convergence will prove substantially easier to handle for finite sample spaces than for infinite dimensional function spaces.*
- *Indeed, infinity dimensional spaces, one will need to introduce such concepts of the capacity of the function space, measured through devices such as the VC-dimension and covering numbers.*

Theorem: *If $\widehat{R}_n(f)$ and $R(f)$ are close for all $f \in \mathcal{F}$, i.e., $\forall \epsilon > 0$,*

$$\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)| \leq \epsilon,$$

then

$$R(\widehat{f}_n) - R(f^+) \leq 2\epsilon.$$

Proof: *Recall that we did define $\widehat{f}_n$ as the best function that is yielded by the empirical risk $\widehat{R}_n(f)$ in the function class $\mathcal{F}$. Recall also that $\widehat{R}_n(\widehat{f}_n)$ can be made as small as possible as we saw earlier. Therefore, with f^+ being the best true risk in class $\mathcal{F}$, we always have*

$$\widehat{R}_n(f^+) - \widehat{R}_n(\widehat{f}_n) \geq 0.$$

As a result,

$$\begin{aligned} R(\widehat{f}_n) &= R(\widehat{f}_n) - R(f^+) + R(f^+) \\ &= \widehat{R}_n(f^+) - \widehat{R}_n(\widehat{f}_n) + R(\widehat{f}_n) - R(f^+) + R(f^+) \\ &\leq 2 \sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| + R(f^+) \end{aligned}$$

Proof: *Recall that we did define $\widehat{f}_n$ as the best function that is yielded by the empirical risk $\widehat{R}_n(f)$ in the function class $\mathcal{F}$. Recall also that $\widehat{R}_n(\widehat{f}_n)$ can be made as small as possible as we saw earlier. Therefore, with f^+ being the best true risk in class $\mathcal{F}$, we always have*

$$\widehat{R}_n(f^+) - \widehat{R}_n(\widehat{f}_n) \geq 0.$$

As a result,

$$\begin{aligned} R(\widehat{f}_n) &= R(\widehat{f}_n) - R(f^+) + R(f^+) \\ &= \widehat{R}_n(f^+) - \widehat{R}_n(\widehat{f}_n) + R(\widehat{f}_n) - R(f^+) + R(f^+) \\ &\leq 2\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)| + R(f^+) \end{aligned}$$

Consequently,

$$R(\widehat{f}_n) - R(f^+) \leq 2\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)|$$

as required.

Corollary: *A direct consequence of the above theorem is the following:*

- *For a given machine* $f \in \mathcal{F}$,

$$R(f) \leq \widehat{R}_n(f) + \left(\frac{\ln \frac{2}{\delta}}{2n} \right)^{1/2}$$

with probability at least $1 - \delta$, $\forall \delta > 0$.

- *If the function class* $\mathcal{F}$ *is finite, ie*

$$\mathcal{F} = \{f_1, f_2, \cdots, f_m\}$$

where $m = |\mathcal{F}| = \#\mathcal{F} =$ Number of functions in the class $\mathcal{F}$ *then it can be shown that, for all* $f \in \mathcal{F}$,

$$R(f) \leq \widehat{R}_n(f) + \left(\frac{\ln m + \ln \frac{2}{\delta}}{2n} \right)^{1/2}$$

with probability at least $1 - \delta$, $\forall \delta > 0$.

Beyond Chernov and Hoeffding

- *It can also be shown that*

$$R(\hat{f}_n) \leq \widehat{R}_n(f^+) + 2\left(\frac{\ln m + \ln \frac{2}{\delta}}{2n}\right)^{1/2} \tag{10}$$

with probability at least $1-\delta$, $\forall \delta > 0$, *where as before*

$$f^+ = \arg\inf_{f\in\mathcal{F}} R(f) \quad \textit{and} \quad \hat{f}_n = \operatorname{argmin}_{f\in\mathcal{F}} \widehat{R}_n(f).$$

- *Equation* (10) *is of foundational importance, because it reveals clearly that the size of the function class controls the uniform bound on the crucial generalization error: Indeed, if the size* m *of the function class* $\mathcal{F}$ *increases, then* $R(f^+)$ *is caused to increase while* $R(\widehat{f}_n)$ *decreases, so that the trade-off between the two is controlled by the size* m *of the function class.*

Vapnik-Chervonenkis Dimension

- **Definition: (Shattering)** *Let $\mathcal{X} \neq \emptyset$ be any non empty domain. Let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be any non-empty class of functions having $\mathcal{X}$ as their domain. Let $S \subseteq \mathcal{X}$ be any finite subset of the domain $\mathcal{X}$. Then S is said to be shattered by $\mathcal{F}$ iff*

$$\{S \cap f \mid f \in \mathcal{F}\} = 2^S$$

In other words, $\mathcal{F}$ shatters S if any subset of S can be obtained by intersecting S with some set from $\mathcal{F}$.

- **Example:** *A class $\mathcal{F} \subseteq 2^{\mathcal{X}}$ of classifiers is said to shatter a set $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ of n points, if, for any possible configuration of labels $y_1, y_2, \cdots, y_n$, we can find a classifier $f \in \mathcal{F}$ that reproduces those labels.*

Vapnik-Chervonenkis Dimension

- **Definition(VC-dimension)** *Let $\mathcal{X} \neq \emptyset$ be any non empty learning domain. Let $\mathcal{F} \subseteq 2^{\mathcal{X}}$ be any non-empty class of functions having $\mathcal{X}$ as their domain. Let $S \subseteq \mathcal{X}$ be any finite subset of the domain $\mathcal{X}$. The VC dimension of $\mathcal{F}$ is the cardinality of the largest finite set $S \subseteq \mathcal{X}$ that is shattered by $\mathcal{F}$, ie*

$$VCdim(\mathcal{F}) := \max\Big\{|S| : S \text{ is shattered by } \mathcal{F}\Big\}$$

  Note: *If arbitrarily large finite sets are shattered by $\mathcal{F}$, then $VCdim(\mathcal{F}) = \infty$. In other words, if a small set of finite cardinality cannot be found that is shattered by $\mathcal{F}$, then $VCdim(\mathcal{F}) = \infty$.*

- **Example:** *The VC dimension of a class $\mathcal{F} \subseteq 2^{\mathcal{X}}$ of classifiers is the largest number of points that $\mathcal{F}$ can shatter.*

Vapnik-Chervonenkis Dimension

- **Remarks:** *If $VCdim(\mathcal{F}) = d$, then there exists a finite set $S \subseteq \mathcal{X}$ such that $|S| = d$ and S is shattered by $\mathcal{F}$. Importantly, every set $S \subseteq \mathcal{X}$ such that $|S| > d$ is not shattered by $\mathcal{F}$. Clearly, we do not expect to learn anything until we have at least d training points. Intuitively, this means that an infinite VC dimension is not desirable as it could imply the impossibility to learn the concept underlying any data from the population under consideration. However, a finite VC dimension does not guarantee the learnability of the concept underlying any data from the population under consideration either.*
- **Fact:** *Let $\mathcal{F}$ be any finite function (concept) class. Then, since it requires 2^d distinct concepts to shatter a set of cardinality d, no set of cardinality greater than $\log |\mathcal{F}|$ can be shattered. Therefore, $\log |\mathcal{F}|$ is always an upper bound for the VC dimension of finite concept classes.*

Vapnik-Chervonenkis Dimension

- *To gain insights into the central concept of VC dimension, we herein consider a few examples of practical interest for which the VC dimension can be found.*
- **VC dimension of the space of separating hyperplanes:** *Let $\mathcal{X} = \mathbb{R}^p$ be the domain for the binary $Y \in \{-1, +1\}$ classification task, and consider using hyperplanes to separate the points of $\mathcal{X}$. Let $\mathcal{F}$ denote the class of all such separating hyperplanes. Then,*

$$VCdim(\mathcal{F}) = p + 1$$

Intuitively, the following pictures for the case of $\mathcal{X} = \mathbb{R}^2$ help see why the VC dimension is $p + 1$.

Foundation Result in Statistical Learning Theory

Theorem: (**Vapnik and Chervonenkis, 1971**) *Let $\mathcal{F}$ be a class of functions implementing so learning machines, and let $\zeta = VCdim(\mathcal{F})$ be the VC dimension of $\mathcal{F}$. Let the theoretical and the empirical risks be defined as earlier and consider any data distribution in the population of interest. Then $\forall f \in \mathcal{F}$, the prediction error (theoretical risk) is bounded by*

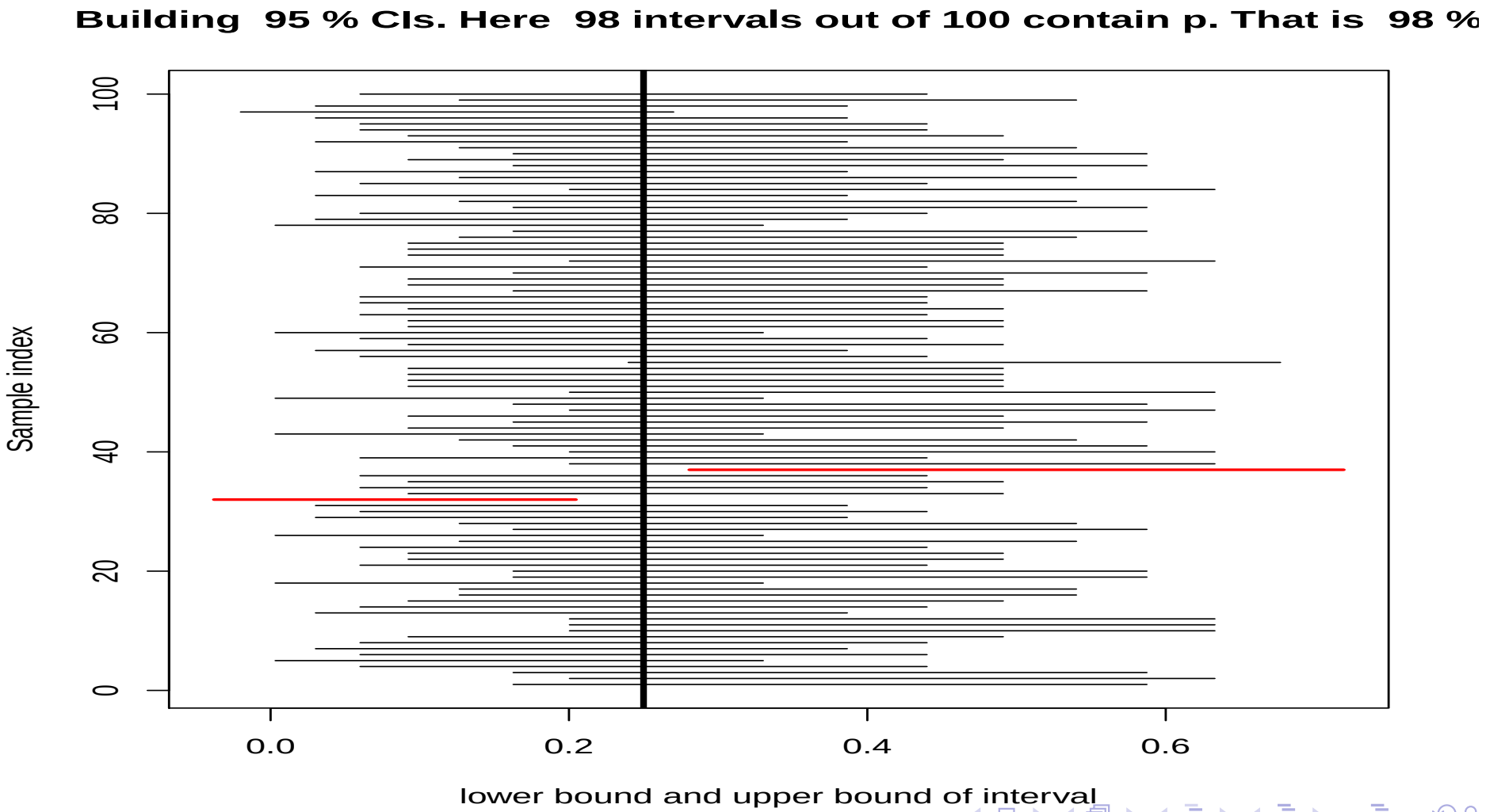
$$R(f) \leq \hat{R}_n(f) + \sqrt{\frac{\zeta\left(\log\frac{2n}{\zeta} + 1\right) - \log\frac{\eta}{4}}{n}} \tag{11}$$

with probability of at least $1 - \eta$. or

$$\Pr\left\{\texttt{TestError} \leq \texttt{TrainError} + \sqrt{\frac{\zeta\left(\log\frac{2n}{\zeta} + 1\right) - \log\frac{\eta}{4}}{n}}\right\} \leq 1 - \eta$$

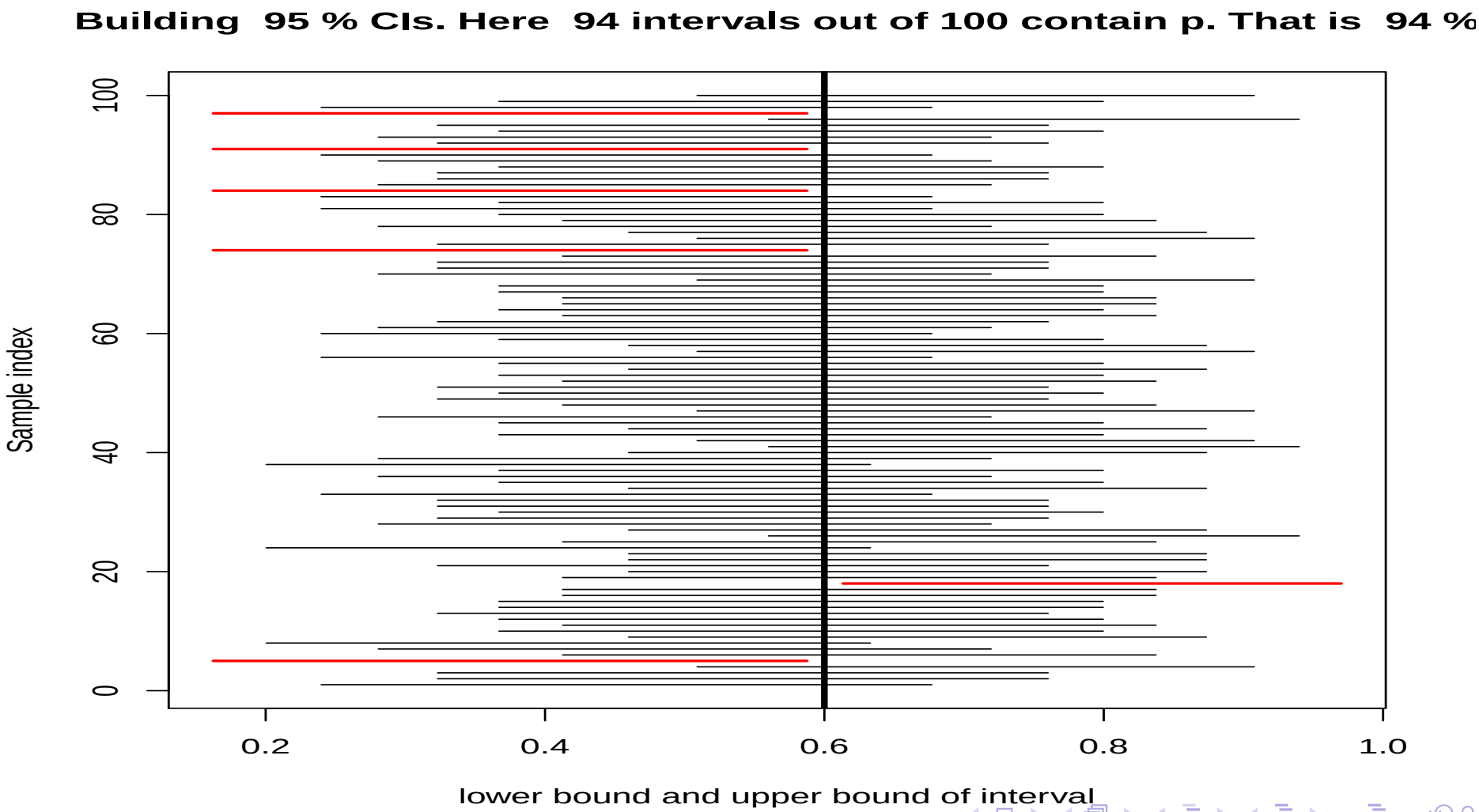
Confidence Interval for a proportion

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \text{ with } 100(1-\alpha)\% \text{ confidence}$$



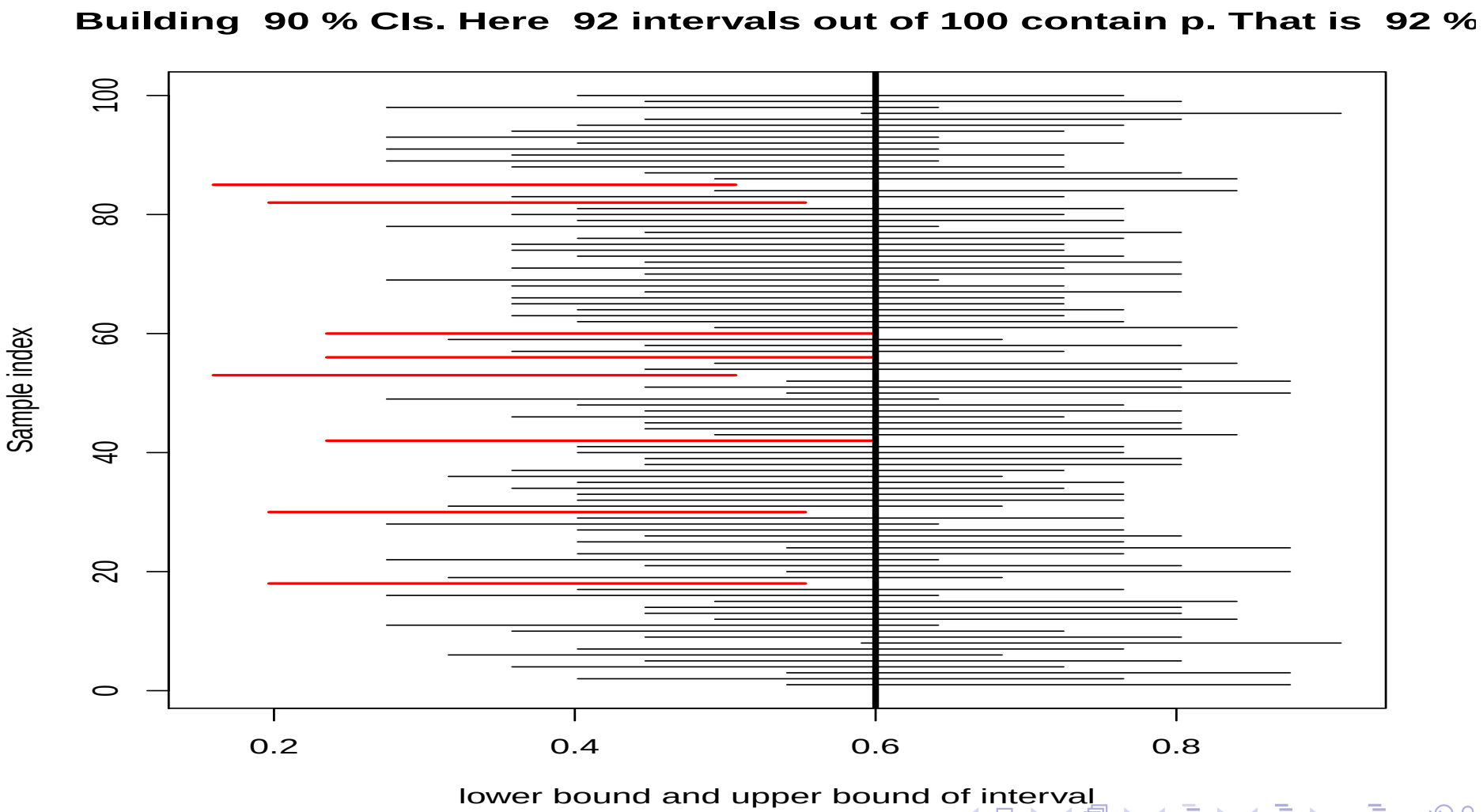
Confidence Interval for a proportion

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \text{ with } 100(1-\alpha)\% \text{ confidence}$$



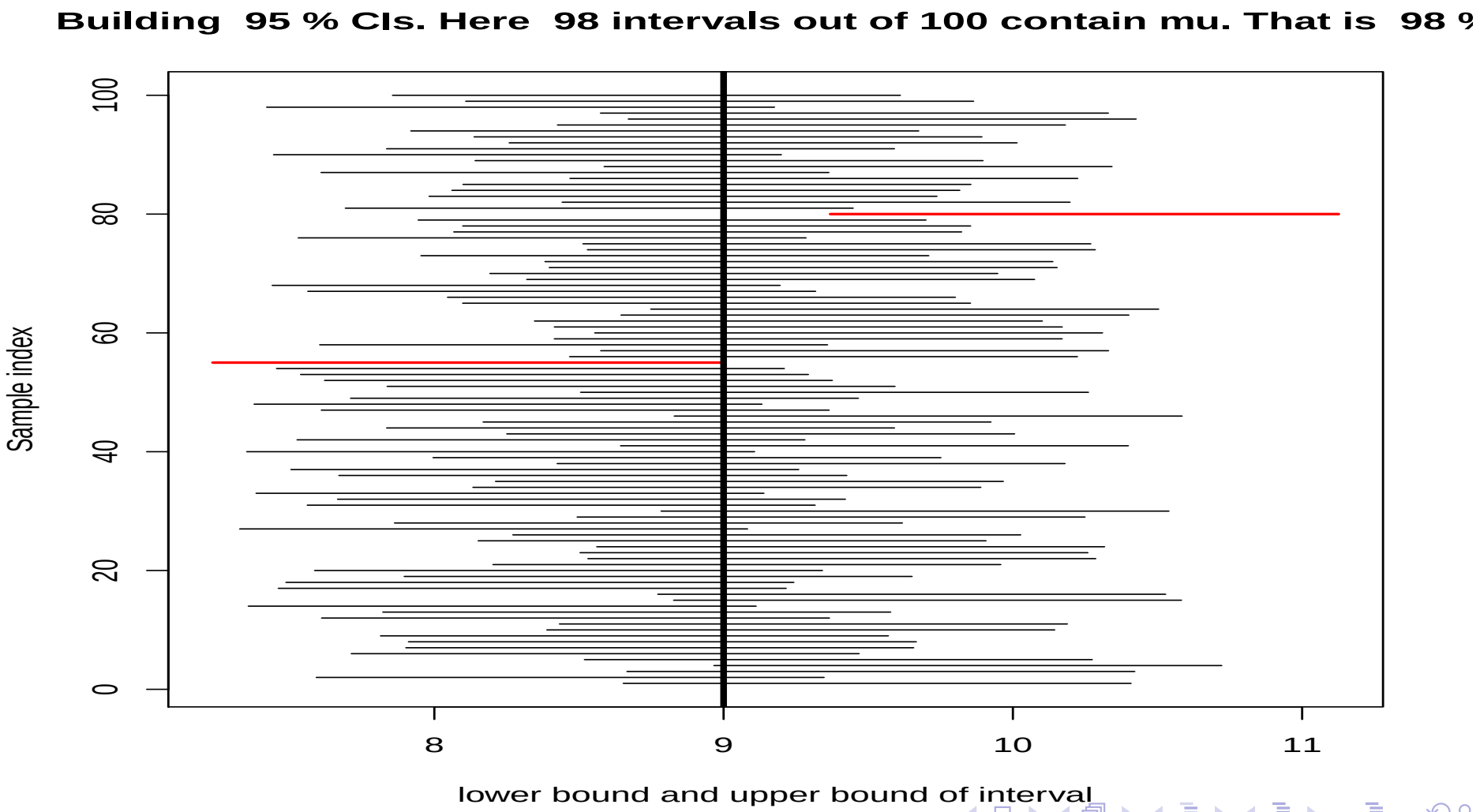
Confidence Interval for a proportion

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \text{ with } 100(1-\alpha)\% \text{ confidence}$$



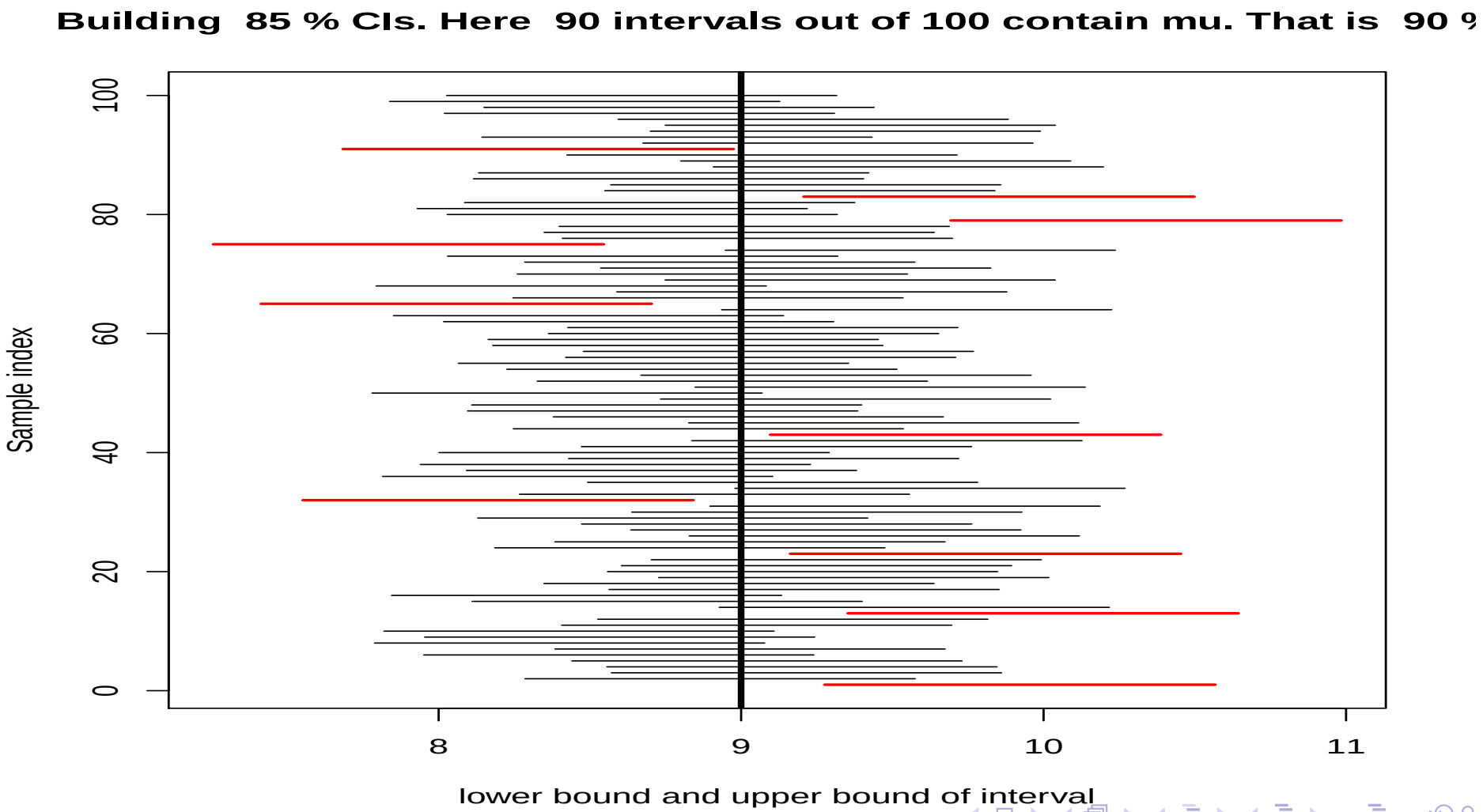
Confidence Interval for a population mean

$$\mu \in \left[\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2)}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] \text{ with } 100 \times (1-\alpha)\% \text{ confidence}$$

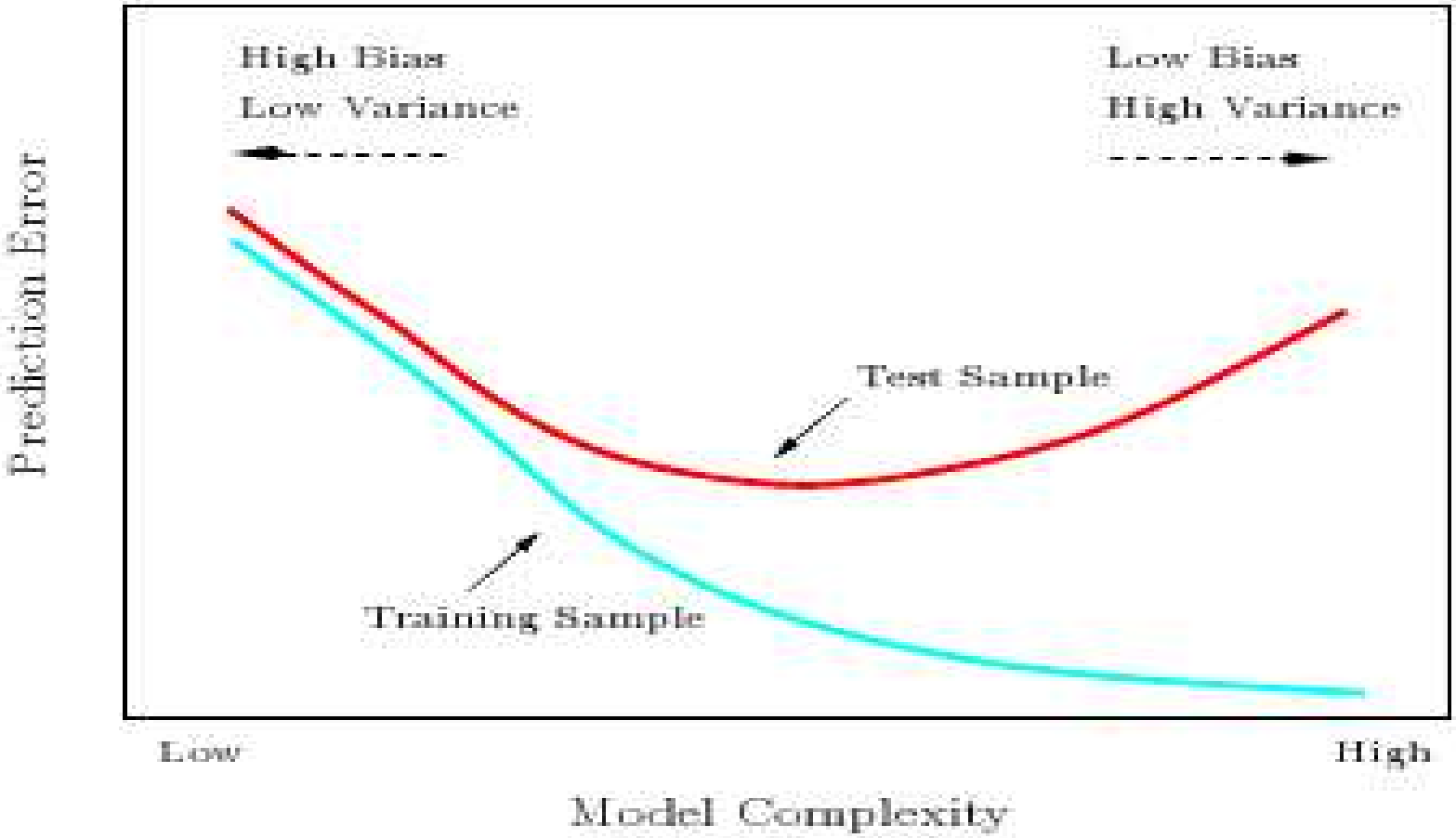


Confidence Interval for a population mean

$$\mu \in \left[\bar{x} - z_{\alpha/2}\sqrt{\frac{\sigma^2)}{n}}, \bar{x} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right] \text{ with } 100 \times (1-\alpha)\% \text{ confidence}$$



Effect of Bias-Variance Dilemma of Prediction



- *Optimal Prediction achieved at the point of bias-variance trade-off.*

Appeal of the VC Bound

- **Note:** *One of the greatest appeals of the VC bound is that, though applicable to function classes of infinite dimension, it preserves the same intuitive form as the bound derived for finite dimensional $\mathcal{F}$.*
- *Essentially, using the VC dimension concept, the number L of possible labeling configurations obtainable from $\mathcal{F}$ with $VCdim(\mathcal{F}) = \zeta$ over $2n$ points verifies*

$$L \leq \left[\frac{en}{\zeta}\right]^{\zeta}. \tag{12}$$

- *The VC bound is simply obtained by replacing $\log|\mathcal{F}|$ with L in the expression of the risk bound for finite dimensional $\mathcal{F}$.*
- *The most important part of the above theorem is the fact that the generalization ability of a learning machine depends on both the empirical risk and the complexity of the class of functions used, which is measured here by the VC dimension of (***Vapnik and Chervonenkis, 1971***).*

Appeal of the VC Bound

- *Also, the bounds offered here are distribution-free, since no assumption is made about the distribution of the population.*
- *The details of this important result will be discussed again in chapter 6 and 7, where we will present other measures of the capacity of a class of functions.*
- **Remark:** *From the expression of the VC Bound, it is clear that an intuitively appealing way to improve the predictive performance (reduce prediction error) of a class of machines is to achieve a trade-off (compromise) between small VC dimension and minimization of the empirical risk.*
- *At first, it may seen as if the VC bound is acting in a way similar to the number of parameters, since it serves as a measure of the complexity of $\mathcal{F}$. In this spirit, the following is a possible guiding principle.*

Appeal of the VC Bound

- *At first, it may seen as if the VC bound is acting in a way similar to the number of parameters, since it serves as a measure of the complexity of $\mathcal{F}$. In this spirit, the following is a possible guiding principle.*
- **Intuition:** *One should seek to construct a classifier that achieves the best trade-off (balance, compromise) between complexity of function class - measured by VC dimension- and fit to the training data -measured by empirical risk.*
- *Now equipped with this sound theoretical foundation one can then go on to the implementation of varioous learning machines. We shall use R to discover some of the most commonly learning machines.*

Machine Learning CRAN Task View in R

Let's visit the website where most of the R community goes

http://www.r-project.org

Let's install some packages and get started

```
install.packages('ctv')
library(ctv)

install.views('MachineLearning')
install.views('HighPerformanceComputing')
install.views('Bayesian')
install.views('Robust')
```

Let's load a couple of packages and explore

```
library(e1071)
library(MASS)
library(kernlab)
```