universität
wien

Bachelor Thesis

Title
"Mathematical foundations of learning
and Clustering of high-dimensional time series"

Author
Boreiko Valentyn

Date

Supervisor: Philipp Grohs

Faculty of Mathematics

Vienna University

Contents

Contents

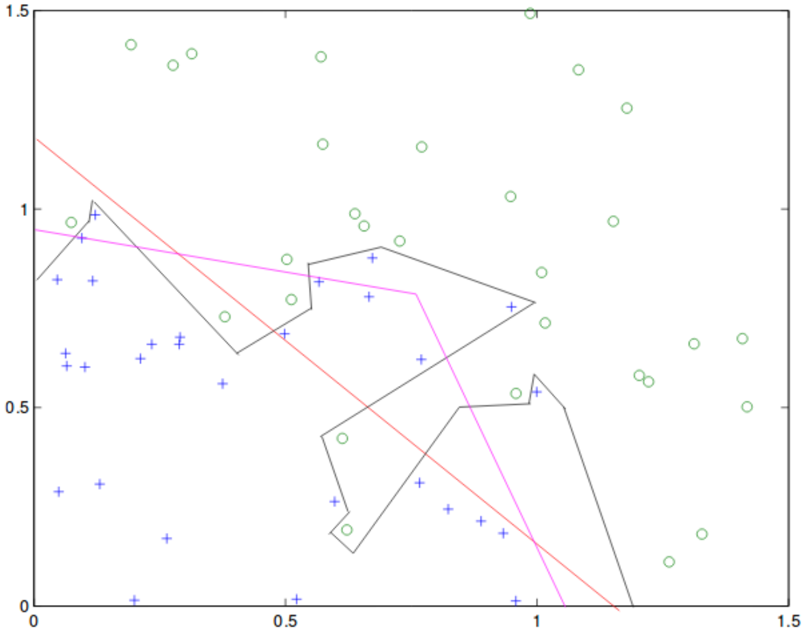
1 Introduction

Generally, when we talk about the learning, we have the following steps in our process:

- Observation of the phenomenon,
- Construction of the model of it,
- Making predictions on basis of this model.

The role of *Machine Learning* is to *automate* such process and the role of *Statistical Learning* is to *formalize* it.
The practical goal of such a learning would be to find such a function that *generalizes* well and is as simple as possible:
By *generalization* one means, that the optimal function found after training on the training data will work well on the test data. And by being as simple as possible one means controlling the *capacity* of the collection of functions that are considered as candidates for the optimal function. This is the one of the most important issues in deep learning in particular: why deep neural networks can generalize so good despite their great *capacity*.
The concept of generalization is so crucial, that I will introduce a more formal general definition in the introduction.

Figure 1: Example of fitting function for binary classification.



Definition 1 (Generalization gap). *Let $f_{\mathcal{A}(\boldsymbol{z})} : X \to Y$ be a model learned by an algorithm $\mathcal{A}$, with a training dataset $\boldsymbol{z} = ((x_1, y_1), ..., (x_m, y_m)) \in Z^m, (x_i)_{i=1}^m$*

being inputs and $(y_i)_{i=1}^m$ *- targets.*
Moreover, let L *be a loss function, that measures the difference between predicted label and a true label. It can be for example a squared loss* $L : X \times Y \to \mathbb{R}_+$, $L(x, y) = (x - y)^2$.
Additionally, let $\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim\rho}[L(f(x), y)]$, *where* ρ *is a probability measure on* $X \times Y$ *and* $\mathcal{R}_{\boldsymbol{z}}(f) := \frac{1}{m}\sum_{i=1}^m L(f(x_i), y_i)$.
Then the the generalization gap(also known as defect) is:

$$\mathcal{L}(f_{\mathcal{A}(\boldsymbol{z})}) := \mathcal{R}(f_{\mathcal{A}(\boldsymbol{z})}) - \mathcal{R}_{\mathbf{z}}(f_{\mathcal{A}(\boldsymbol{z})})$$

The difficulty in analysing the generalisation gap lies in the dependence of $f_{\mathcal{A}(\mathbf{z})}$ on the same sample $\mathbf{z}$ used in the definition of $\mathcal{R}_{\mathbf{z}}$. Different methods in statistical generalization theory have been developed to solve this issue. Some of them are *model complexity* and *stability* approaches:

- *Model complexity* method does so by taking a bound over a hypothesis space $\mathcal{H}$ of functions, decoupling $f_{\mathcal{A}(\mathbf{z})}$ from a particular sample $\mathbf{z}$:

  $$\mathcal{L}(f_{\mathcal{A}(\mathbf{z})}) \leq sup_{f\in\mathcal{H}}(\mathcal{R}(f) - \mathcal{R}_{\mathbf{z}}(f))$$

  And because the cardinality of $\mathcal{H}$ will be typically infinite, one needs to consider the ways to characterize it, e.g. by using capacity measures, that will be discussed later on.

- *Stability* method, by looking at how much the change of a data point in $\mathbf{z}$ can change $f_{\mathcal{A}(\mathbf{z})}$.

These concepts will be discussed in a more formal fashion and the way to find such an optimal function will be shown. After which, I will consider one case of learning, concretely - clustering or grouping of time series.

2 Mathematical Background

2.1 Basic concepts

Definition 2 (Mitchell, deeplearningbook)**.** *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E*

- The experience E typically consists of a dataset which consists of many examples. They can be labeled, then we speak of *supervised learning*, or unlabeled, in which case we speak of *unsupervised learning*.

- The task T can be:
  Regression, i.e. predicting a numerical value $f : \mathbb{R}^n \to \mathbb{R}$.
  Classification, i.e. mapping data $x \in \mathbb{R}^n$ to a category in $\{1, .., k\}$
  $f : \mathbb{R}^n \to \{1, .., k\}$ or to a histogram with respect to k categories $f : \mathbb{R}^n \to \mathbb{R}^k$.
  Density estimation, i.e. estimating a probability distribution on the space that the examples were drawn from $p : \mathbb{R} \to \mathbb{R}_+$.

- The performance measure P in classification tasks is typically the accuracy, i.e., the proportion of examples for which the model produces the correct output.
  Usually dataset is split into a *training*, *cross-validation* (the set to fine tune the parameteres, that cannot be done with training set) and a *test set*.

Regression - one example:

- The task: predict $f : \mathbb{R}^n \to \mathbb{R}$

- The experience: training data $(x_i^{train}, y_i^{train})_{i=1}^m$

- The performance measure: given test data $(x_i^{test}, y_i^{test})_{i=1}^m$ we evaluate the performance of an estimator $\hat{f} : \mathbb{R}^n \to \mathbb{R}$ as the *mean squared error*:

$$\frac{1}{n}\sum_{i=1}^{n} |\hat{f}(x_i^{test}) - x_i^{test})|^2$$

So the task would be to find the *empirical target function*:

$$f_{\mathcal{H},\mathbf{z}} := argmin_{f\in\mathcal{H}}\mathcal{E}_{\mathbf{z}}(f)$$

Where:

- $\mathcal{H} = span\{\phi_1, ..., \phi_l\} \subset C(\mathbb{R}^n)$ is the *hypothesis space.*
- $\mathbf{z} = (x_i, y_i)_{i=1}^m$ is training data.
- $\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2$ is the *empirical risk.*

Regression - estimate: Considering the following,

- Every $f \in \mathcal{H}$ can be written as $\sum_{i=1}^m w_i\phi_i$ and let $\mathbf{w} := (w_i)_{i=1}^l$.

- Let $\mathbf{A} = (\phi_j(x_i))_{i,j} \in \mathbb{R}^{m \times n}$.
- Let $\mathbf{y} := (y_i)_{i=1}^m$.
- With the new notation, we get:

  $$\mathcal{E}_{\mathbf{z}}(f) = ||\mathbf{A}\mathbf{w} - \mathbf{y}||^2.$$

  The minimizer is then $\mathbf{w} := \mathbf{A}^{\dagger}\mathbf{y}$, where $\mathbf{A}^{\dagger}$ is MoorePenrose pseudoinverse.
  And the estimate is:

  $$f_* := \sum_{i=1}^{l} (\mathbf{w}_*)_i \phi_i.$$

2.2 Remarks

Later on two approaches in learning theory will be discussed - one when the hypothesis space of functions $\mathcal{H}$ on X is much simpler and the development proceeds with a more combinatorial flavor. It is the one that uses VC dimension, Rademacher complecity and Growth function.
Another one with the function space $\mathcal{H} \subset \mathcal{C}(X)$, where $\mathcal{C}(X)-$ is the Banach space of continuous functions on X, with norm $\|f\|_\infty$ - leads to functional analysis. The VC dimension for example is then replaced by the radius R of a ball which defines the hypothesis space in RKHS.

2.3 Capacity Measures

A McDiarmid's concentration inequality will be of a great help to prove the useful bounds in this part. And in order to prove it, we would need Hoeffding's lemma first:

Lemma 1 (Hoeffding's lemma). *Let X be a r.v. with $E[X] = 0$ and $a \leq X \leq b, a < b$. Then:*

$$\forall t > 0 : E[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

Proof. Define $\phi(t) = \log E[e^{tX}]$, then:

$$\phi'(t) = \frac{E[Xe^{tX}]}{E[e^{tX}]}, \phi''(t) = \frac{E[X^2e^{tX}]}{E[e^{tX}]} - \left(\frac{E[Xe^{tX}]}{E[e^{tX}]}\right)^2.$$

$\phi''(t)$ is also a variance under the probability measure $\widetilde{\rho} = \frac{e^{tX}}{E[e^{tX}]}\rho$. But under any probability measure and considering that $X \in [a, b]$, we have that:

$$Var[X] = Var\left[X - \frac{a+b}{2}\right] \leq E\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4},$$

which together with the fundmental theorem of calculus and the fact that $\phi(0) = \phi'(0) = 0$ gives us:

$$\phi(t) = \int_0^t \int_0^s \phi''(x)\, dx\, ds \leq \frac{t^2(b-a)^2}{8}.$$

□

Now, becasue the notion of the conditional expectation will be widely used, we remind ourselves of the definition of the conditional expectation and prove the Radon-Nikodym Theorem. For a further use a notion of *absolute continuity* will be helpful:

Definition 3 (Absolute continuous measures)**.** *Let $M = (\Omega, \mathcal{B})$ be a measurable space, μ and λ - positive bounded measures on M.*
We say that λ is absolutely continuous with respect to μ, denoted by $\lambda << \mu$ if $\forall A \in \mathcal{B} : \mu(A) = 0 \implies \lambda(A) = 0$

The following theorem proves the existence and uniqueness of the conditional expectation, which will be defined later on.

Theorem 1 (Radon-Nikodym Theorem)**.** *Let $(\Omega, \mathcal{B}, P)$ be a probability space, ν - a positive bounded measure and $\nu << P$. Then there exists a unique r.v. $X \in L^1(\Omega, \mathcal{B}, P)$ s.t.:*

$$\nu(E) = \int_E X \, dP, \forall E \in \mathcal{B}.$$

We also denote then X by $\frac{d\nu}{dP}$

Proof. Define $Q(A) = \frac{\nu(A)}{\nu(\Omega)}$, which is a probability measure and $Q << P$. Now $\tilde{P} = \frac{P+Q}{2}$ is also a probability measure.
$H := L^2(\Omega, \mathcal{B}, \tilde{P})$ is in turn a Hilbert space. Let us define the functional on H:

$$L(Y) = \int_\Omega Y \, dQ.$$

L is a linear continuous functional. While linearity is clear, let us show the boundedness:

$$\begin{aligned}|L(Y)| &\leq \int |Y| \, dQ \leq \int |Y| \, dQ + \int |Y| \, dP = 2 \int |Y| \, d\tilde{P} \\ &\leq 2(\int |Y|^2 \, d\tilde{P})^{1/2}(*) \\ &= 2||Y||_2.\end{aligned}$$

$(*)$ can be proven for example by the use of Hölder's Inequality:

$$E[|ZY|] \leq (E|Z|^r)^{1/r}(E|Y|^s)^{1/s},$$

when one set for $\beta > \alpha > 0 : Z = |X|^\alpha, Y = 1$ and $r = \frac{\beta}{\alpha}, s = \frac{\beta}{\beta - \alpha}$:

$$\begin{aligned}E[|ZY|] = E|X|^\alpha \leq (E|X|^{r\alpha})^{1/r} &= (E|X|^\beta)^{\alpha/\beta} \\ \iff & ||X||_\alpha \leq ||X||_\beta.\end{aligned}$$

So now one has that L is continuous and linear on H and then by the Riesz representation theorem:

$$\begin{aligned}\forall Y \in H \quad \exists! Z \in H : L(Y) = (Y, Z) = \int YZ \, d\tilde{P} &= \int \frac{YZ}{2} \, dP + \int \frac{YZ}{2} \, dQ \\ &= \int Y \, dQ.(**)\end{aligned}$$

Therefore it follows that:

$$\forall Y \in H \quad \exists! Z \in H : \int \frac{YZ}{2}\, dP = \int Y(1-\frac{Z}{2})\, dQ.(***)$$

(**) gives us for $Y = \chi_A, A \in \mathcal{B}$ as well that:

$$Q(A) = \int Y\, dQ = \int_A Z\, d\tilde{P}.$$

Therefore:

$$0 \leq \frac{Q(A)}{\tilde{P}(A)} = \frac{\int_A (Z)\, d\tilde{P}}{\tilde{P}(A)} \leq \frac{\int_A Z\, d\tilde{P}}{Q(A)/2} = 2$$
$$\implies 0 \leq \int_A Z\, d\tilde{P} \leq \int_A 2\, d\tilde{P}.$$

It gives us:

$$0 \leq Z \leq 2 \quad \tilde{P} - a.s.$$

And now, if in $(***)$ we set $Y = \chi_{Z=2}$, then:

$$\int_{Z=2} (1 - Z/2)\, dQ = \int_{Z=2} Z/2\, dP \implies P[Z=2] = 0$$
$$\implies Q[Z=2] = 0 \quad (\textit{because } Q << P)$$
$$\implies \tilde{P}[Z=2] = 0$$
$$\implies 0 \leq Z < 2\tilde{P} - a.s.$$

Finally, we set again in $(***) Y = (\frac{Z}{2})^2 \chi_A, A \in \mathcal{B}$.
Then we have that $Y \in H, 0 \leq Y < 1 \quad (P, Q, \tilde{P}) - a.s.$ and:

$$\int_A \left(\frac{Z}{2}\right)^n \left(1 - \frac{Z}{2}\right) dQ = \int_A \left(\frac{Z}{2}\right)^{n+1} dP$$
$$\implies \int_A \left(1 - \left(\frac{Z}{2}\right)^{N+1}\right) dQ = \int_A \frac{Z}{2} \sum_{j=0}^{N} \left(\frac{Z}{2}\right)^j dP. \quad (\textit{sum over n=0 to N})$$

By the dominated convergence theorem and the fact that for $N \to \infty$ we have: $1 - \left(\frac{Z}{2}\right)^{N+1} \to 1 \quad \tilde{P} - a.s.$: left side converges to $\int_A dQ = Q(A)$, at the same time monotone convergenve theorem implies, that right side converges to $\int_A \frac{Z}{2-Z}\, dP$.
And if we set $X = \frac{Z}{2-Z}$, then:

$$\forall A \in \mathcal{B} : Q(A) = \int_A X\, dP.$$

From the Riesz representation theorem we have also, that Z and therefore X is unique. □

To continue with the definition of the conditional expectation, the next result will be of use:

Corollary 1. *Let Q and P are probability measures on $(\Omega, \mathcal{B})$ s.t. $Q << P$. Furthermore, let $\mathcal{G} \subset \mathcal{B}$ be a sub-σ-algebra. Then in $(\Omega, \mathcal{B})$:*

$$Q|_{\mathcal{G}} << P|_{\mathcal{G}} \text{ and } \frac{dQ|_{\mathcal{G}}}{dP|_{\mathcal{G}}} \text{ is } \mathcal{G}\text{-measureable}$$

Now we will show the definition of the conditional expectation:

Definition 4 (Conditional expectation)**.** *Suppose $X \in L^1(\Omega, \mathcal{B}, P)$ and let $\mathcal{G} \subset \mathcal{B}$ be a sub-σ-field. Then there exists a r.v. $E[X|\mathcal{G}]$, called the conditional expectation of X with respect to $\mathcal{G}$, such that:*

- *$E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable and integrable,*
- *$\forall G \in \mathcal{G}$ one has: $\int_G X\, dP = \int_G E[X|\mathcal{G}]\, dP$.*

Having introduced the conditional expectation, let us proceed to the McDiarmid's inequality:

Theorem 2 (McDiarmid's concentration inequality)**.** *Let $X_1, ..., X_m \in X^m$ - a set of independent r.v. and assume that there are $c_1, ..., c_m > 0$ such that $f : X^m \to \mathbb{R}$ satisfies for all $i \in 1, ..., m$:*

$$|f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i', ..., x_m)| \leq c_i,$$

then we have that generaly for $f(S) := f(X_1, ..., X_m)$:

$$P[f(S) - E[f(S)] \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}},$$
$$P[f(S) - E[f(S)] \leq -\epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}}.$$

Proof. First, we want to find a martingale difference, i.e. a sequence of random variables $V_1, V_2, ...$ with respect to $X_1, X_2, ...$ if for all $i > 0, V_i$ is a function of $X_1, ..., X_i$ and $E[V_{i+1}|X_1, ..., X_i] = 0$.
For this let's take:

$V = f(S) - E[f(S)]$, $V_k = E[V|X_1, ..., X_k] - E[V|X_1, ..., X_{k-1}] = E[f(S)|X_1, ..., X_k] - E[f(S)|X_1, ..., X_{k-1}], k > 1$, $V_1 = E[V|X_1] - E[V]$.

Due to the tower property of the conditional expectation one can see that $E[E[V|X_1, ..., X_k]|X_1, ...X_{k-1}] = E[V|X_1, ..., X_{k-1}]$, from which follows, that $E[V_k|X_1, ..., X_{k-1}] = 0$. Therefore, $(V_k)_{k=1}^m$ is a martingale sequence.
It also useful to note that:

$\sum_{k=1}^m V_k = E[f(X_1, ...X_m) - E[f(X_1, ...X_m)]|X_1, ..., X_m] - E[f(X_1, ..., X_m) - E[f(X_1, ..., X_m)]] = f(S) - E[f(S)] = V$.

Then let us define the upper and lower bounds for V_k by:

$$U_k = sup_x E[f(S)|X_1, ..., X_{k-1}, x] - E[f(S)|X_1, ..., X_{k-1}]$$
$$L_k = inf_x E[f(S)|X_1, ..., X_{k-1}, x] - E[f(S)|X_1, ..., X_{k-1}].$$

Then by assumption:

$$\forall k \in 1,..,m : U_k - L_k = sup_{x,x'} E[f(S)|X_1,...,X_{k-1},x] - E[f(S)|X_1,...,X_{k-1},x'] \leq c_k,$$

From which follows, that $L_k \leq V_k \leq L_k + c_k$.
Then, to simplify notation, we write $S_k = \sum_{i=1}^{k} V_i$. Then:

$$\begin{aligned}
\forall t > 0 : P[S_m \geq \epsilon] &= P[e^{tS_m} \geq e^{t\epsilon}] \\
&\leq \frac{E[e^{tS_m}]}{e^{t\epsilon}} \textit{(Markov's inequality)} \\
&= e^{-t\epsilon} E[E[e^{tS_m}]|X_1,...,X_{m-1}] \\
&= E[e^{tS_m - 1} E[e^{tV_m}|X_1,...,X_{m-1}]] \\
&\leq e^{-t\epsilon} E[e^{tS_{m-1}}] e^{\frac{t^2 c_m^2}{8}} (*) \\
&\leq e^{-t\epsilon} e^{t^2 \sum_{i=1}^{m} \frac{c_i^2}{8}} \quad (\textit{iterating the previous argument}) \\
&= e^{\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}} \quad (t = \frac{4\epsilon}{\sum_{i=1}^{m} c_i^2} \textit{ is chosen to minimize the upper bound}).
\end{aligned}$$

(*) is done by applying Hoeffding's lemma for conditional expectation, V_k taking values in $[L_k, L_k + c_k]$. □

Now having the McDiarmid's inequality at hand, we can proceed to an exploration of a series of useful upper bounds for the *generailsation gap.* Usually, we define the space of the functions that we look in for the solution as *hypothesis space*:

Definition 5 (Hypothesis space $\mathcal{H}$)**.** *Let $\mathcal{H}$ be a compact subset of the Banach space $\{f : X \to Y, continuous\}$ with norm $||f|| := max_{x \in X}|f(x)|$. We call it hypothesis space.*
The algorithm will work to find, as well as possible, the best approximation for f_ρ, and compactness is important to ensure the finite covering number (will be discussed later on).

As it was mentioned in the introduction, we would like to measure the capacity of such a space. There are following important capacity measures (not in a sense of the measure defined on a $\sigma - algebra$):

- *Rademacher complexity* - lets us derive learning guarantees with the simple proofs based on McDiarmid's inequality, but the computation can be NP-hard for some hypothesis sets.
- *Growth function*
- *VC-dimension*

First, the *Empirical Rademacher complexity* captures the degree to which a hyptothesis set can fit random noise. The formal definition is:

Definition 6 (Empirical Rademacher complexity)**.** *Let G be the family of functions $f : Z \to \mathbb{R}$, and $S = (z_1,...,z_m)$ a sample from Z. The empirical*

Rademacher complexity of G with respect to the sample S is then:

$$\widehat{\mathfrak{R}}_S(G) := E_\sigma[sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)] = E_\sigma[sup_{g\in G}\frac{\langle\sigma,\mathbf{g}_S\rangle}{m}]$$

Here we can see, as it was mentioned before, that the supremum of the inner product $sup_{g\in G}\frac{\langle\sigma,\mathbf{g}_S\rangle}{m}$ measures how well the class of functions G correlates with σ over the sample S.

The *Rademacher complexity* is then a measure of the expected noise-fitting ability of G:

Definition 7 (Rademacher complexity)**.** *For $m\in\mathbb{N}:\mathfrak{R}_m(G) := E_{S\sim\rho}[\widehat{\mathfrak{R}}_S(G)]$, where ρ is a probability measure on Z.*

So let us prove the generalization bound with the notion of the Rademacher complexity:

Theorem 3. *Let G be a space of functions $f : Z \to [0,1]$. Then:*

$$\forall\delta>0\quad\forall f\in G : P\left[E[f(z)] - \widehat{E}_S[f] \leq 2\mathfrak{R}_m(G) + \sqrt{\frac{log\frac{1}{\delta}}{2m}}\right] \geq 1-\frac{\delta}{2}, \quad (2.1)$$

and:

$$\forall\delta>0\quad\forall f\in G : P\left[E[f(z)] - \widehat{E}_S[f] \leq 2\widehat{\mathfrak{R}}_m(G) + 3\sqrt{\frac{log\frac{2}{\delta}}{2m}}\right] \geq 1-\frac{\delta}{2}, \quad (2.2)$$

where $\widehat{E}_S[f] = \frac{1}{m}\sum_{i=1}^{m} g(z_i)$.

Proof. In this proof we will apply McDiarmid's inequality to the function $\Phi(S) = sup_{f\in G}(E[f] - \widehat{E}_S[f])$.
If we take two samples S and $\tilde{S}$ such that w.l.o.g.: $S = (z_1, ..., z_m)$ and $\tilde{S} = (z_1, ..., \tilde{z}_m)$, then:

$$|\Phi(\tilde{S}) - \Phi(S)| \leq sup_{f\in G}|\widehat{E}_S[f] - \widehat{E}_{\tilde{S}}[f]| = sup_{f\in G}|\frac{f(z_m) - f(\tilde{z}_m)}{m}| \leq \frac{1}{m}.$$

Then, using McDiarmid's inequality:

$$\forall\delta>0 : P\left[\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{log\frac{2}{\delta}}{2m}}\right] \geq 1-\delta/2.$$

And now we want to bound $E_S[\Phi(S)]$:

$$\begin{aligned}
E_S[\Phi(S)] &= E_S[sup_{f\in G}E_{S'}[\widehat{E}_{S'}[f] - \widehat{E}_S[f]]] \\
&\leq E_{S,S'}[sup_{f\in G}(\widehat{E}_{S'}[f] - \widehat{E}_S[f])] = E_{S,S'}[sup_{f\in G}\frac{1}{m}\sum_{i=1}^{m}(f(z_i) - f(z_i'))] \quad \textit{(convexity of the supremum f} \\
&= E_{S,S',\sigma}[sup_{f\in G}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(z_i) - f(z_i')))] \\
&\leq E_{S,S',\sigma}[sup_{f\in G}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(z_i)))] - E_{S,S',\sigma}[sup_{f\in G}(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i'))] \\
&= 2E_{\sigma,S}[sup_{f\in G}\sum_{i=1}^{m}\sigma_i(f(z_i))] \\
&= 2\mathfrak{R}_m(G),
\end{aligned}$$

which gives us the bound (2.1), if we use δ instead of $\frac{\delta}{2}$,
And for the bound (2.2), we simply apply McDiarmid's inequality, considering that $|\mathfrak{R}_S(G) - \mathfrak{R}_{\tilde{S}}(G)| \leq \frac{1}{m}$ when S and $\tilde{S}$ differ only by one point and get that $P[\mathfrak{R}_m(G) \leq \widehat{\mathfrak{R}}_S(G) + \sqrt{\frac{log\frac{2}{\delta}}{2m}}] \geq 1 - \frac{\delta}{2}$.
Finally, by the subbaditivity of the measure we arrive at the bound (2.2) □

Now we want to reduce the case to the functions $f : \mathbb{R} \to \{-1, 1\}$ and for this let us introduce the notions of the *generalization error* and the *empirical error*, both of which are performance measures in case of the binary classification.

Definition 8 (Generalization error)**.** *Given labels mapping $f_l : \mathcal{X} \to \{-1, 1\}$ for the data with the probability measure ρ and the hypothesis $f_h : \mathcal{X} \to \{-1, 1\}$ that is being tested, the generalization error is:*

$$R(f_h) := P_{x\sim\rho}[f_h(x) \neq f_l(x)].$$

Because one doesn't have both probability measure ρ and generally all possible labels f_l are unknown, one uses *empirical error*:

Definition 9 (Empirical error)**.** *Let f_h be a hypothesis, f_l a labels, and a sample $S = (x_1, ..., x_m)$, the empirical error of f_h is defined as:*

$$\widehat{R}(f_h) := \frac{1}{m}\sum_{i=1}^{m}\chi_{\{x\in\mathcal{X}: f_h(x_i)\neq f_l(x_i)\}}.$$

Remark 1. *In the case of a binary loss we have:*
Let $\tilde{G}$ are the functions $f : \mathbb{R} \to \{-1, 1\}$ and $G = \{g : \mathbb{R} \to \{-1, 1\} : (x, y) \mapsto \chi_{\{x\in\mathcal{X}: f(x)\neq y\}}, f \in \tilde{G}\}$. Then for $S = ((x_1, y_1), ..., (x_m, y_m)) \subset \mathcal{X} \times \{-1, 1\}$ and $S_{\mathcal{X}}$, its projection: $S_{\mathcal{X}} = (x_1, ..., x_m)$ the following holds:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(G) &= E_\sigma\left[sup_{f\in\tilde{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\chi_{\{x\in\mathcal{X}: f(x)\neq y\}}\right] = E_\sigma\left[sup_{f\in\tilde{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{1 - y_i f(x_i)}{2}\right] \\
&= \frac{1}{2}E_\sigma\left[sup_{f\in\tilde{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S_{\mathcal{X}}}(\tilde{G})
\end{aligned}$$

From theorem 3 one has, that for function space $\mathcal{H}$, ρ - probability measure on $\mathcal{X}$, and a sample S_m drawn according to the measure ρ:

$$\forall\delta > 0 \forall f \in \mathcal{H} : P[R(f) \leq \widehat{R}(f) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{log\frac{1}{\delta}}{2m}}] \geq 1 - \delta$$

and

$$P[R(f) \leq \widehat{R}(f) + \widehat{\mathfrak{R}}_m(\mathcal{H}) + 3\sqrt{\frac{log\frac{2}{\delta}}{2m}}] \geq 1 - \delta.$$

One has that $\widehat{\mathfrak{R}_S}(H)$ can be hard to compute for some hypothesis spaces, therefore the more feasible combinatorial "measures" can be used instead. Now we will proceed with the *growth function* and derive a bound for the Rademacher complexity in terms of it.

Definition 10 (Growth function)**.** *The growth function $\prod_{\mathcal{H}}(S) : \mathbb{N} \to \mathbb{N}$ is the maximum number of ways into which m points can be classified with functions in $\mathcal{H}$ for $S_m = (x_1, ..., x_m)$:*

$$\prod\nolimits_{\mathcal{H}}(S_m) := max_{\{(x_1,...,x_m)\subset\mathcal{X}\}}|\{(f(x_1), ..., f(x_m)) : f \in \mathcal{H}\}|$$

Further we will use the following lemma:

Theorem 4 (Massart's lemma)**.** *If $A \subset \mathbb{R}^m$ is a finite set, and $m = max_{\mathbf{x}\in A}\|\mathbf{x}\|_2$ then:*

$$E_\sigma[\frac{1}{m}sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i] \leq \frac{m\sqrt{2log|A|}}{m}$$

Proof. We have that, using convexity of exponential function:

$$\forall t > 0 : exp(tE_\sigma[sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i]) \leq E_\sigma[exp(tsup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i)]$$
$$\leq \sum_{x\in A}E_\sigma[exp(t\sum_{i=1}^{m}\sigma_i x_i)]$$
$$\leq \sum_{x\in A}\prod_{i=1}^{m}epx(\frac{t^2(2x_i)^2}{8})(From\ Hoeffding's\ inequality)$$
$$\leq \sum_{\mathbf{x}\in A}exp\left(\frac{t^2m^2}{2}\right) = |A|exp\left(\frac{(tm)^2}{2}\right).$$

It follows that:

$$E_\sigma[sup_{x\in A}\sum_{i=1}^{m}\sigma_i x_i] \leq \frac{log|A|}{t} + \frac{tm^2}{2}.$$

But then for $t = \frac{\sqrt{2log|A|}}{m}$, which minimizes the upper bound:

$$E_\sigma[sup_{\mathbf{x}\in A}\sum_{i=1}^{m}\sigma_i x_i] \leq m\sqrt{2log|A|}.$$

This gives the derivation of the statement. □

Now let us show the generalization bound for the case of the binary classification using growth function:

Theorem 5. *Let $\mathcal{H} = \{f : \mathbb{R} \to \{-1, 1\}\}$, then:*

$$\forall \delta > 0 : P\left[R(f) \leq \widehat{R}(f) + \sqrt{\frac{2log \prod_{\mathcal{H}}(S_m)}{m}} + \sqrt{\frac{log \frac{1}{\delta}}{2m}}\right] \geq 1 - \delta$$

Proof. Now, using Remark 1 and Massart's lemma for a sample $S_m = (x_1, ..., x_m)$ and $\mathcal{H}|_{S_m}$ and by the definition of the growth function:

$$\mathfrak{R}_m(\mathcal{H}) = E_{S_m,\sigma}[sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)] \leq E_{S_m}\left[\frac{\sqrt{2mlog|\mathcal{H}|_{S_m}|}}{m}\right] \leq \sqrt{\frac{2log \prod_{\mathcal{H}(S_m)}}{S_m}}.$$

This gives us the desired result. □

We will finish this introduction by defining VC-dimension and showing the upper bound for the generalisation gap.

Definition 11 (Vapnik-Chervonenkis dimension)**.** *It can be defined as the size of the largest set that can be fully shattered by hypothesis set $\mathcal{H}$ for $S_m = (x_1, ..., x_m)$:*

$$VC(\mathcal{H}) := max\{m : \prod_{\mathcal{H}}(S_m) = 2^m\}$$

For the proof of the bound one needs the following lemma and its corollary:

Lemma 2 (Sauer's lemma)**.** *For hypothesis set with $\mathcal{H}$ and $VC(\mathcal{H}) = d$. Then:*

$$\forall m \in \mathbb{N} : \prod_{\mathcal{H}}(S_m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Proof. By induction on $m + d$ we have:

- **The base:**
  Degenerate case for $n = 0 : \prod_{\mathcal{H}}(S_0) = 1 = \sum_{i=0}^{d} \binom{0}{i}$ Not a single point can be shattered for $d = 0 : \prod_{\mathcal{H}}(S_m) = \binom{m}{0}$.

- **Induction step:**
  Let us assume that the statement holds for any $\tilde{m} + \tilde{d} < m + d$, i.e. for S_m. Then for $\mathcal{H}_1 := \mathcal{H}|_{S_{m-1}}$ and $\mathcal{H}_2 := \mathcal{H} \backslash \mathcal{H}_1$, since $VC(\mathcal{H}_1) \leq VC(\mathcal{H})$ and $VC(\mathcal{H}_2) + 1 \leq VC(\mathcal{H})$:

$$\prod_{\mathcal{H}}(S_m) = |\mathcal{H}_1| + |\mathcal{H}_2| \leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i}$$
$$= \sum_{i=0}^{d} \binom{m-1}{i} + \binom{m-1}{i-1} = \sum_{i=0}^{d} \binom{m}{i}.$$

□

The generalization bound based on VC-dimension is then:

Corollary 2. *Let $\mathcal{H} = f : \mathbb{R} \to \{-1, 1\}$ and $VC(\mathcal{H}) = d$. Then for all $m \geq d$:*

$$\forall \delta > 0 \quad \forall f \in \mathcal{H} : R(f) \leq \widehat{R}(f) + \sqrt{\frac{2dlog\frac{em}{d}}{m}} + \sqrt{\frac{log\frac{1}{\delta}}{2m}}.$$

Proof. Using the Theorem 5 and the fact, that from the Sauer's lemma follows:

$$\begin{aligned}\prod_{\mathcal{H}}(S_m) &\leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} = \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \\ &= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m} \leq \left(\frac{m}{d}\right)^{d} e^{d}.\end{aligned}$$

□

2.4 Generalization in Deep Learning (Bengio et al.) TBD

2.5 General Regression

2.5.1 Estimating Sample Error

Let X be a compact and $Y = \mathbb{R}^k$ and a probability measure ρ on $Z := X \times Y$. A main concept is the *least squares error* of f defined by:

$$\mathcal{E}(f) := \int_Z (f(x) - y)^2 \, d\rho(x, y).$$

The learning problem is then to find function f, which minimize the the error $\mathcal{E}$. In general, such a function f will not exist for our problem. But we can look for a function f_ρ which minimizes the least squares error $\mathcal{E}$.
To do so, one can show, that $\mathcal{E}$ decomposes as a sum. Let us first introduce some definitions:

Definition 12 (Conditional Probability)**.** *Let p_ρ be the density of ρ. Then the measure $\rho(\cdot|x)$ on Y has a density:*

$$\frac{p_\rho(x, \cdot)}{\int_Y p_\rho(x, y) \, dy}.$$

Definition 13 (Regression Function)**.** *Regression function is defined as:*

$$f_\rho(x) := \int_Y y \, d\rho(y|x).$$

The regression function can be seen as the average of y coordinate of $\{y\} \times Y$.

Remark 2. *From $E[(y - f_\rho(x))] = 0$ it follows that*

$$\sigma^2(x) := \int_Y (y - f_\rho(x))^2 \, d\rho(y|x).$$

and by averaging over X:

$$\sigma_\rho^2 := \mathcal{E}(f_\rho).$$

Definition 14 (Marginal). *The marginal of ρ on X is defined as the measure:*

$$\rho_X(A) := \rho(A \times Y)$$

Proposition 1. *For every $f : X \to Y$:*

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 \, d\rho_X(x) + \sigma_\rho^2.$$

And since σ_ρ^2 is independent of $f_\rho(x)$, Proposition 1 implies that f_ρ has the smallest possible error among all functions $f : X \to Y$. Thus σ_ρ^2 represents a lower bound on the error $\mathcal{E}$. The goal is then to find a good approximation of f_ρ from random samples on Z.

Proof. Now, because $\forall x \in X : \int_Y (f_\rho(x) - y) = 0$, we have:

$$\begin{aligned}\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\ &= \int_X (f(x) - f_\rho(x))^2 + 2\int_X \int_Y (f(x) - f_\rho(x))(f_\rho(x) - y) + \int_X \int_Y (f_\rho(x) - y)^2 \\ &= \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.\end{aligned}$$

□

Definition 15 (Empirical error). *Let $\boldsymbol{z} = ((x_1, y_1), ..., (x_m, y_m)) \in Z^m$ be i.i.d. according to ρ.*
The empirical error of f w.r.t $\boldsymbol{z}$ is then defined as:

$$\mathcal{E}_{\boldsymbol{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

We cannot measure the theoretical error $\mathcal{E}(f)$, but we can do so with the empirical error $\mathcal{E}_{\mathbf{z}}(f)$, therefore it is useful to estimate the diference between them, which is defined as *defect function*:

Definition 16 (Defect function). *The defect of f is defined as:*

$$\mathcal{L}_{\boldsymbol{z} \in Z^m}(f) := \mathcal{E}(f) - \mathcal{E}_{\boldsymbol{z}}(f)$$

Definition 17 (Target function). *Target function in hypothesis space $\mathcal{H}$ is defined via*

$$f_{\mathcal{H}} := argmin_{f \in \mathcal{H}} \mathcal{E}(f)$$

And one can also find a bound for the defect function. Unlike the previous chapters, here we will use the complexity of the hypothesis space that is based on the covering numbers. Here we show uniform estimate of the defect.
But first we need to prepare the theorems and lemmas that will be useful in our proof.
The proof of the theorem follows from the following proposition, Hoeffding and Bernstein inequalities, and following lemma:

Proposition 2. *If for $j = 1, 2, |f_j(x) - y| \leq M$ then, for all $\boldsymbol{z} \in Z^m$:*

$$|\mathcal{L}_{\boldsymbol{z}}(f_1) - \mathcal{L}_{\boldsymbol{z}}(f_2)| \leq 4M||f_1 - f_2||_\infty$$

Proof. From the definition of the least squares error it follows that:

$$\begin{aligned}|\mathcal{E}(f_1) - \mathcal{E}(f_2)| &= |\int (f_1(x) - y)^2 - (f_2(x) - y)^2| \\ &= |\int (f_1(x) - f_2(x))^2(f_1(x) + f_2(x) - 2y)| \leq \|f_1 - f_2\|_\infty 2M,\end{aligned}$$

analogously:

$|\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| \leq \|f_1 - f_2\|_\infty 2M.$

Hence we have that:

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| = |(\mathcal{E}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1)) - (\mathcal{E}(f_2) - \mathcal{E}_{\mathbf{z}}(f_2))| \leq \|f_1 - f_2\|_\infty 4M.$$

□

Proposition 3 (Hoeffding's and Bernstein's inequalities)**.** *Let $\{\xi_i\}_{i=1}^m$ be independent random variables on a probability space Z with means μ_i, variances σ_i^2, $\Sigma^2 = \sum_{i=1}^m (\sigma_i^2)$ and satisfying $|\xi_i(z) - E(\xi_i)| \leq M$ for each i almost everywhere. Then for every $\epsilon > 0$:*

$$P(\sum_i^m (\xi_i - \mu_i) \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2mM^2}} \textit{ (Hoeffding's inequality)}$$

Hoeffdings inequality does not use any information about the random variables except the fact that they are bounded. If the variances of ξ_i are small, then we can get a sharper inequality from Bernsteins inequality:

$$P(\sum_i^m (\xi_i - \mu_i) \geq \epsilon) \leq e^{-\frac{\epsilon^2}{2(\Sigma^2 + \frac{1}{3}M\epsilon)}} \textit{ (Bernstein's inequality)}$$

Proof. To begin with, we will show that:

$$P[\sum_{i=1}^m (\xi_i - E[\xi_i]) > \epsilon] \leq exp\bigg(-\frac{\epsilon}{M}\bigg(\bigg(1 + \frac{\Sigma^2}{M\epsilon}\bigg) log\bigg(1 + \frac{M\epsilon}{\Sigma^2}\bigg) - 1\bigg)\bigg). \quad (2.3)$$

W.l.o.g.: $E[\xi_i] = 0, \sigma_i^2 = E[\xi_i^2]$. Then for $\forall c \in \mathbb{R} : c > 0$:

$$\begin{aligned}P[\sum_{i=1}^m (\xi_i - E[\xi_i]) > \epsilon] &= P[e^{\sum_{i=1}^m c\xi_i} > e^{c\epsilon}] \\ &\leq e^{-c\epsilon} E[e^{\sum_{i=1}^m c\xi_i}] = e^{-c\epsilon} \prod_{i=1}^m E[e^{c\xi_i}]\end{aligned}$$

Further, because $|\xi_i| \leq M \quad a.e.$ and $E[\xi_i] = 0$:

$$\begin{aligned}E[e^{c\xi_i}] &= 1 + \sum_{k=2}^{+\infty} \frac{c^k E(\xi_i^k)}{k!} \leq 1 + \sum_{k=2}^{\infty} \frac{c^k M^{k-2}\sigma_i^2}{k!} \\ &\leq exp\bigg(\sum_{k=2}^{\infty} \frac{c^k M^{k-2}\sigma_i^2}{k!}\bigg) \quad (1 + x \leq e^x) \\ &= exp\bigg(\frac{e^{cM} - 1 - cM}{M^2}\sigma_i^2\bigg).\end{aligned}$$

Thus:

$$P[\sum_{i=1}^{m}(\xi_i - E[\xi_i]) > \epsilon] \leq epx\bigg(\frac{e^{cM}-1-cM}{M^2}\Sigma_i^2 - c\epsilon\bigg).$$

By choosing $c = \frac{log(1+\frac{M\epsilon}{\Sigma^2})}{M}$ we can minimise the bound above and thus we get the desired result.
Now when we have the function $f : [0, \infty) \to \mathbb{R}$,

$$f(x) := (1+x)log(1+x) - x$$

and $\tilde{f} : [0, \infty) \to \mathbb{R}, \tilde{f}(x) := 2log(1+x) - 2x + xlog(1+x)$ with $\tilde{f}(0) = \tilde{f}'(0) = 0$, $\tilde{f}(x) = \frac{x}{(1+x)^2} \geq 0, x \geq 0$. This implies: $\tilde{f}(x) \geq 0 \iff log(1+x) - x \geq -\frac{xlog(1+x)}{2}, \forall x \geq 0$. Therefore:

$$\forall x \geq 0 : f(x) \geq \frac{xlog(1+x)}{2}.$$

For the function $g(x) = (6+2x)f(x) - 3x^2$ we have analogously: $g(x) = g'(x) = 0, g''(x) = \frac{4f(x)}{1+x} \geq 0$. Which means again that $\forall x \geq 0 : g(x) \geq 0 \iff f(x) \geq \frac{3x^2}{6+2x}$.
By using it together with the fact that (2.3) $\iff P[\sum_{i=1}^{m}(\xi_i - E[\xi_i]) > \epsilon] \leq exp(-\frac{\Sigma^2}{M^2}f(\frac{M\epsilon}{\Sigma^2}))$ one arrives at the Bernstein's inequality.
And to prove the Hoeffding's inequality we show, that due to convexity of the exponential function and assumption that $|\xi_i| \leq M \quad a.s.$ we have for $t = \frac{c\xi_i + cM}{2cM}$:

$$e^{c\xi_i} \leq te^{cM} + (1-t)e^{-cM} \quad a.e.$$

Then w.l.o.g $E[\xi_i] = 0$:

$$E[e^{c\xi_i}] \leq \frac{1}{2}(e^{-cM} + e^{cM}) = cosh(cM) = \sum_{j=0}^{\infty}\frac{(cM)^{2j}}{(2j)!} \leq \sum_{j=0}^{\infty}\frac{((cM)^2/2)^j}{j!}.$$

Therefore, together with Markov's inequality:

$$\begin{aligned} P[\sum_{i=1}^{m}(\xi_i - E[\xi_i]) > \epsilon] &\leq exp(\frac{m(cM)^2}{2} - c\epsilon) \\ &\leq e^{-\frac{\epsilon^2}{(2mM^2)}}.(\textit{For bound minimization } c = \frac{\epsilon}{mM^2}) \end{aligned}$$

□

Lemma 3. *Let* $\mathcal{H} = S_1 \cup ... \cup S_l$ *and* $\epsilon > 0$. *Then*

$$P_{\mathbf{z} \in Z^m}\{sup_{f \in \mathcal{H}}\mathcal{L}_{\mathbf{z}}(f) \geq \epsilon\} \leq \sum_{j=1}^{l} P_{\mathbf{z} \in Z^m}\{sup_{f \in S_j}\mathcal{L}_{\mathbf{z}}(f) \geq \epsilon\}.$$

Proof. The proofs follows from the subadditivity of the probability measure and from the fact that:

$$sup_{f \in \mathcal{H}}\mathcal{L}_{\mathbf{z}}(f) \geq \epsilon \iff \exists i \leq l : sup_{f \in S_i}\mathcal{L}_{\mathbf{z}}(f) \geq \epsilon.$$

□

Now we can approach the uniform bound that was mentioned before.

The notion of *covering number* will be used, so we will define it formally:

Definition 18. *Let S be a metric space and $s > 0$. The covering number $\mathcal{N}(S, s)$ is defined as the minimal $l \in \mathbb{N}$ such that there exists l disks in S with radius s covering S.*

Theorem 6. *Let $\mathcal{H} \subset C(X)$ be a hypothesis class. Assume that for all $f \in \mathcal{H}$ it holds that $|f(x) - y| < M$ a.e. and let $\sigma^2 := \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$. Then, for all $\epsilon > 0$:*

$$P_{\boldsymbol{z} \in Z^m}(sup_{f \in \mathcal{H}} |\mathcal{L}_{\boldsymbol{z}}(f)| \leq \epsilon) \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M})e^{-\frac{m\epsilon^2}{8M^4}}, (\textit{From Hoeffding's inequality})$$

$$P_{\boldsymbol{z} \in Z^m}(sup_{f \in \mathcal{H}} |\mathcal{L}_{\boldsymbol{z}}(f)| \leq \epsilon) \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M})e^{-\frac{\epsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2\epsilon)}}.(\textit{From Bernstein's inequality})$$

Proof. Let's take $s = \mathcal{N}(\mathcal{H}, \frac{\epsilon}{4M})$ and covering of $\mathcal{H}$ with the balls $B_j := B_{\frac{\epsilon}{4M}}(f_j)$. Now because $\mathcal{H}$ is compact each open cover has a finite subcover and therefore there is a finite n, s.t.: $\mathcal{H} = \bigcup_{i=1}^n B_i$.
By the assumptions and propositions 2, we have that:

$$\forall z \in Z^m \forall f \in \mathcal{H} : |\mathcal{L}_{\mathbf{z}}(f) - \mathcal{L}_{\mathbf{z}}(f_i)| \leq 4M \|f - f_j\|_\infty \leq 4M \frac{\epsilon}{4M} = \epsilon.$$

From this follows:

$$\begin{aligned} P[sup_{f \in \mathcal{H}} \mathcal{L}_{\mathbf{z}}(f) \geq 2\epsilon] &\leq P[\mathcal{L}_{\mathbf{z}}(f_j) \geq \epsilon] \\ &\leq e^{-\frac{m\epsilon^2}{2M^4}}.(*) \end{aligned}$$

Here (*) follows from the last theorem in case when $\xi_i^j = \frac{\xi^j(z_i)}{m}, \xi^j = -(f_j(x) - y), j = 1, ..., n$ and $|\xi_i^j - E[\xi_i^j]| \leq \frac{M}{m}, Var(\xi_i^j) = \frac{\sigma^2}{m^2}$, what gives us $\forall \epsilon > 0$:

$$P[\frac{1}{m}\sum_{i=1}^m \xi^j(z_i) - E[\xi^j] \geq \epsilon] \leq exp(-\frac{m\epsilon^2}{2(\sigma^2 + \frac{M\epsilon}{3})}), \quad (\textit{From Bernstein's inequality})$$

$$P[\frac{1}{m}\sum_{i=1}^m \xi^j(z_i) - E[\xi^j] \geq \epsilon] \leq exp(-\frac{m\epsilon^2}{2M^2}). \quad (\textit{From Hoeffding's inequality})$$

Now by using previous lemma and $\frac{\epsilon}{2}$ instead of ϵ we get the desired result. □

Here:

- $\mathcal{N}(S, s)$ is *the covering number* defined by:

Before we move on to the last estimation let us again introduce a definiton:

Definition 19. *For a given hypothesis space $\mathcal{H}$, the error in $\mathcal{H}$ of a function $f \in \mathcal{H}$ is defined as:*

$$\mathcal{E}_{\mathcal{H}}(f) := \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

This way we can also infer that:

Definition 20 (*Sample error* and *Approximation error*)**.**

$$\mathcal{E}(f_{\boldsymbol{z}}) = \mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},\boldsymbol{z}}) + \mathcal{E}(f_{\mathcal{H}}).$$

$\mathcal{E}(f_{\boldsymbol{z}})$ is called empirical error.

The second term in the sum depends on the choice of $\mathcal{H}$, but is independent of sampling. It is called approximation error. The first term, however, depends on sampling, and is called sample error.

Using this definition now one can talk about the *bias-variance problem.* Our goal is to make empirical error as small as possible. Assume now that sample size m is fixed. Then by enlarging hypothesis space $\mathcal{H}$, the approximation error will certainly decrease, but the sample will increase.

Definition 21 (Bias-variance problem)**.** *Assuming that sample size is fixed, the bias-variance problem consists of choosing the size of $\mathcal{H}$, so that the empirical error is minimized with high probability. One can say, that "bias" of the solution f coincides with the approximation error, and the "variance" with the sample error. Several parameters (radius of balls, dimensions, etc.) determine the "size" of $\mathcal{H}$, and one can fix all of them except one and minimize the error over this nonfixed parameter.*

High variance and low bias are known as *overfitting*, low variance and high bias, in turn, named as *underfitting.*

It is important to know, how well does $f_{\mathcal{H},\mathbf{z}}$ approximate $f_{\mathcal{H}}$. In other words, how small is the sample error expected to be. To provide the theorem, that gives the estimate, one need to use the following lemma:

Lemma 4. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Let $\epsilon > 0$ and $0 < \delta < 1$ such that:*

$$P_{\boldsymbol{z}\in Z^m}\{sup_{f\in\mathcal{H}}|\mathcal{L}_{\boldsymbol{z}}(f)| \leq \epsilon\} \geq 1-\delta,$$

then:

$$P_{\boldsymbol{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},\boldsymbol{z}}) \leq 2\epsilon\} \geq 1-\delta.$$

Which together with the respective Hoeffding's and Bernstein's inequalities gives us the following result:

Theorem 7. *Let $\mathcal{H} \subset C(X)$ be a compact subspace of $\mathcal{C}(X)$. Assume that for all $f \in \mathcal{H}$ it holds that $|f(x) - y| < M$ a.e. and let $\sigma^2 := \sup_{f\in\mathcal{H}} \sigma^2(f_Y^2)$. Then, for all $\epsilon > 0$:*

$$P_{\boldsymbol{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},\boldsymbol{z}}) \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})2e^{-\frac{m\epsilon^2}{32M^4}}.(\textit{From Hoeffding's inequality})$$

$$P_{\boldsymbol{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},\boldsymbol{z}}) \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})2e^{-\frac{m\epsilon^2}{8(4\sigma^2+\frac{1}{3}M^2\epsilon)}}.(\textit{From Bernstein's inequality})$$

Finally, this gives us estimate for m if we want to ensure that the probability, that the sample error is $\leq \epsilon$, is at least $1-\delta$:

$$m \geq \frac{32M^4}{\epsilon^2}\{\log(2\mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})) + \log(\frac{1}{\delta})\}.(\textit{From Hoeffding's inequality})$$

$$m \geq \frac{8(4\sigma^2 + \frac{1}{3}M^2\epsilon)}{\epsilon^2}\{\log(2\mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})) + \log(\frac{1}{\delta})\}.(\textit{From Bernstein's inequality})$$

2.5.2 Reproducing Kernel Hilbert Space

Motivation to use Reproducing Kernel Hilbert Space as the Hypothesis space is that every function in RKHS, that minimizes empirical risk function can be written as a linear combination of the kernel functions evaluated at the training points. This simplifies the empirical risk minimization problem from an infinite dimensional to a finite dimensional problem.

We will start with the definition of the *Mercer kernel*:

Definition 22. *Let X be a metric space. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric, when $K(x, y) = K(y, x)$ for all $x, y \in \mathcal{X}$ and is positive semidefinite when for all $\boldsymbol{x} = x_1, ..., x_k \subset \mathcal{X}$ the $k \times k$ Gramian matrix $K[\boldsymbol{x}]$ whose (i, j) entry is positive semidefinite. K is a Mercer kernel if it is continuous, symmetric, and positive semidefinite.*

And now we can introduce the RKHS $\mathcal{H}_K$ with the following theorem:

Theorem 8. *Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel.*
Then there exists a unique Hilbert space $(\mathcal{H}_K, \langle,\rangle_{\mathcal{H}_K})$ of
functions $K_x : x' \mapsto K(x', x)$ on X satisfying the following conditions:

- *for all $x \in X$, $K_x \in \mathcal{H}_K$,*
- *$\overline{\{K_x | x \in X\}} = \mathcal{H}_K$, and*
- *for all $f \in \mathcal{H}_K$ and $x \in X, f(x) = \langle K_x, f\rangle_{\mathcal{H}_K}$. It is called the reproducing property.*

Furthermore:
$\mathcal{H}_K$ consists of continuous functions, and for the inclusion $I_K : \mathcal{H}_K \to \mathcal{C}(X)$ we have: $\|I_K\| \leq C_K$, where $C_K := sup_{x \in X} \sqrt{K(x, x)}$.

Proof. When we denote by $\mathcal{H}_0 := span\{K_x | x \in X\}$ and for any $f, g \in \mathcal{H}_0, f = \sum_{i=1}^{m} \alpha_i K_{x_i}, g = \sum_{i=1}^{n} \beta_i K_{\tilde{x}_i}$ we define the inner product as:

$$\langle f, g\rangle := \sum_{(i,j) \in \{1..m\} \times \{1..n\}} \alpha_i \beta_j K(x_i, \tilde{x}_j)$$

One needs to check that it is indeed the inner product.
While symmetry, linearity in the first argument, and the implication $f = 0 \implies \langle f, f\rangle_{\mathcal{H}_0} = 0$ are clear and the fact that $\langle f, f\rangle_{\mathcal{H}_0} \geq 0$ follows from the positive semi-definiteness of the kernel, we need to show that $\langle f, f\rangle_{\mathcal{H}_0} = 0 \implies f = 0$. But from positive-semidefiniteness of K for the fix $x \in \mathcal{X}$ and $\tilde{\alpha}_i = c\alpha_i, i \in \{1..n\}, \tilde{\alpha}_{n+1} = f(x)$:

$$\begin{aligned} & c^2 \langle f, f\rangle_{\mathcal{H}_0} + 2c|f(x)|^2 + |f(x)|^2 K(x, x) \geq 0 \\ & \implies |f(x)|^4 \leq |f(x)|^2 \langle f, f\rangle_{\mathcal{H}_0} K(x, x) = 0, \quad when \langle f, f\rangle_{\mathcal{H}_0} = 0, \end{aligned}$$

which implies that $\forall x \in \mathcal{X} : f(x) = 0$. From the completion theorem, we have that there exists a unique completion of the metric space $(\mathcal{H}_0, d(f, g) =$

$\|f-g\|_{\mathcal{H}_0}$) up to isometries. As one sees, it satisfies 3 conditions stated in the theorem. And the additional statement can be proven for $f \in \mathcal{H}_K$ and $x \in X$:

$$\begin{aligned} |f(x)| &= |\langle K_x, f\rangle| \\ &\leq \|f\|_{\mathcal{H}_K}\|K_x\|_{\mathcal{H}_K} \quad (ByCauchy-Schwarzinequality) \\ &\leq \|f\|_{\mathcal{H}_K}\sqrt{K(x,x)}. \end{aligned}$$

Thus, $\|f\|_\infty \leq C_K\|f\|_{\mathcal{H}_K}$, which proves that $\|I_K\| \leq C_K$, and since f is the limit of elements $\mathcal{H}_0$ that are continuous, and convergence in $\|\|_{\mathcal{H}_K}$ implies convergence in $\|\|_\infty$. □

Let us give some examples of Mercer Kernels:

Example 1 (Dot product kernels)**.** *Let $X = \{x \in \mathbb{R}^n : ||x|| \leq R\}$. A dot product kernel is a function $K : X \times X \to \mathbb{R}$ given by*

$$K(x,y) = \sum_{d=0}^{\infty} a_d (x \cdot y)^d,$$

where $a_d \geq 0$ and $\sum_d a_d R^{2d} < \infty$.

If we take X as above and $K(x,y) = 1 + x \cdot y$, then $\{1, x_1, ..., x_n\}$ consitutes an ONB of $\mathcal{H}_K$.

Example 2 (Translation invariant kernels)**.** *Let $f \in \mathcal{L}^2(\mathbb{R}^n)$ be continuous and even. Suppose the Fourier transform of f is nonnegative. Then the kernel*

$$K(x,y) = f(x-y)$$

is a Mercer kernel on $\mathbb{R}^n$.

Radial basis functions provide one interesting class of kernels:
They are of the form $K(x,y) = f(\|x-y\|^2)$, where f is univariate function on $[0,\infty)$. To show that the following kernles are Mercer kernels, we would need the definition of the *completely monotonic* function and the Bernstein's theorem.

Definition 23 (Completely monotonic function)**.** *A function $f : [0,\infty) \to \mathbb{R}$ is completely monotonic if it is:*

- $f \in \mathcal{C}[0,\infty)$
- $f \in \mathcal{C}^\infty(0,\infty)$
- $x \in (0,\infty), n \geq 0 : (-1)^n f^{(n)}(x) \geq 0$

Theorem 9 (Bernstein's theorem)**.** *A function $f : [0,\infty) \to \mathbb{R}$ is completely monotonic iff there is a finite positive Borel measure μ on $[0,\infty)$ s.t.:*

$$f(t) = \int_0^\infty e^{-xt} d\nu(x), t \in [0,\infty),$$

which is a Laplace transform of μ and we can write:

$$f(t) = \mathfrak{L}\nu(t).$$

Proof. For $\mathfrak{L}\nu(t)$ we have:

$$\begin{aligned}
&\frac{\mathfrak{L}\nu(t+h)-\mathfrak{L}\nu(t)}{h} = \int_{[0,\infty)} \frac{e^{-hx}-1}{h} e^{-tx}\, d\nu(x), \\
&\textit{and} \\
&|\frac{e^{-hx}-1}{he^{tx}}| \le \frac{|hx|e^{|hx|}}{|h|e^{tx}} \le \frac{c}{e^{\frac{t}{2}x}} \quad for|h| \le \frac{t}{4}
\end{aligned}$$

for some $c > 0$, and thus we can use the dominated convergence theorem:

$$(\mathfrak{L})'\nu(t) = lim_{h\to 0} \int_{[0,\infty)} \frac{e^{-hx}-1}{he^{tx}}\, d\nu(x) = -\int_{[0,\infty)} xe^{-tx}\, d\nu(x).$$

What by induction gives us that $f \in \mathcal{C}^{\infty}(0,\infty)$ and:

$$f^{(n)}(t) = \int_0^{\infty} (-x)^n e^{-tx}\, d\nu(x) \quad for \quad n \ge 0, t \in [0,\infty),$$

which together with $(-1)^n f^{(n)}(t) = \int_0^{\infty} x^n e^{-tx} \nu(\,dx) \ge 0$ let us prove that f is completely monotonic.
What concerns another direction we have that because $(-1)^n f^{(n)}$ is nonnegative and nonincreasing for $t > 0, n \ge 1$:

$$\begin{aligned}
&(-1)^n f^{(n)}(t) \le \frac{2}{t} \int_{\frac{t}{2}}^{t} (-1)^n f^n(x)dx = \frac{2}{t}(-1)^n (f^{(n-1)}(t) - f^{(n-1)}(\frac{t}{2})) \\
&\frac{2}{t} \le (-1)^n f^{(n-1)}(\frac{t}{2}).
\end{aligned}$$

Which by induction for any $n \ge 1$ gives:

$$\begin{aligned}
&(-1)^n f^{(n)}(t) \le (\prod_{i=1}^{n-1} \frac{2^i}{t}) f'(\frac{t}{2^{n-1}}) \le (\prod_{i=1}^{n-1} \frac{2^i}{t}) \frac{2^n}{t} (f(\frac{t}{2^{n-1}}) - f(\frac{t}{2^n})) \\
&= \frac{2^{\frac{n(n+1)}{2}}}{t^k} (f(\frac{t}{2^{n-1}}) - f(\frac{t}{2^n})),
\end{aligned}$$

and thus, considering as well that $f^{(n)}(t), n \ge 1$ for each n is continuous, we conclude:

$$\begin{aligned}
&f^{(n)}(t) = o(t^{-n}), t \to \infty, n \ge 1, \\
&(x-t)^n f^{(n)}(t) \to 0, x \to t, n \ge 1,
\end{aligned}$$

Therefore, for $n \geq 1, t \geq 0$:

$$\begin{aligned}
&lim_{b\to\infty}(f(t) - f(b)) = lim_{b\to\infty}(-\int_t^b f'(x)dx) \\
&= lim_{b\to\infty}(-(x-t)f'(x)|_{x=t}^b + \int_t^b (x-t)f''(x)dx) \\
&= lim_{b\to\infty}(\int_t^b (x-t)f''(x)dx) = ... = lim_{b\to\infty}(\frac{(-1)^{n+1}}{n!}\int_t^b (x-t)^n f^{(n+1)}(x)\,dx) \\
&= lim_{b\to\infty}(\frac{(-1)^{n+1}}{(n-1)!}\int_{\frac{t}{n}}^b (1-\frac{t}{nx})^n (nx)^n f^{(n+1)}(nx)\,dx \\
&= lim_{b\to\infty}(\frac{(-1)^{n+1}}{(n-1)!}\int_0^b (1-\frac{t}{nx})^n \chi_{[0,n]}(nx)^n f^{(n+1)}(nx)\,dx) \\
&= lim_{b\to\infty}(\frac{(-1)^{n+1}}{(n-1)!}\int_0^b (1-\frac{xt}{n})^n \chi_{[0,n]}(\frac{n}{x})^n f^{(n+1)}(\frac{n}{x})x^{-2}\,dx) \\
&= lim_{b\to\infty}(\frac{(-1)^{n+1}}{(n-1)!}\int_0^b (1-\frac{xt}{n})^n \chi_{[0,n]}\int_{\frac{1}{x}}^\infty (n\tilde{x})^n f^{(n+1)}(n\tilde{x})\,d\tilde{x}\,dx) \\
&= lim_{b\to\infty}(\int_0^b (1-\frac{xt}{n})^n \chi_{[0,n]}\,d\nu_n(x)).
\end{aligned}$$

The functions $\nu_n(t)$ are non-decreasing, continuous, thus right continuous, therefore we can define measure as $\nu_n(t_1) - \nu_n(t_2)$. Then we show that the total variation of ν_n is bounded and thus by Helly's First theorem there exists a convergent subsequence ν_{n_k} that converges pointwise to some nondecreasing ν on compact $[0, \hat{x}]$.
First of all the total variation of ν_n is $V_0^\infty \nu_n = \int_0^\infty |\nu_n'(u)|\,du = \int_0^\infty \nu_n'(u)\,du$ because $\nu_n'(x) = \frac{(-1)^{n+1}}{(n-1)!}(\frac{n}{x})^n f^{(n+1)}(\frac{n}{x})\frac{1}{x^2}$ exists and is Lebesgue integrable. Secondly, the total variation is bounded because :

$$\begin{aligned}
V_0^\infty \nu_n &= \int_0^\infty \frac{(-1)^{n+1}}{(n-1)!}(\frac{n}{x})^n f^{(n+1)}(\frac{n}{x})\frac{1}{t^2}\,dx \\
&= lim_{b\to\infty}(f(0) - f(b)).
\end{aligned}$$

Then because $(1-\frac{tx}{n})^n\chi_{[0,n]}$ is a sequence of non-decreasing measurable functions that converges uniformly to e^{-tx}, from the monoton convergence theorem we have:

$$f(t) = \lim_{b\to\infty} f(b) + \int_0^\infty e^{-tx}\,d\tilde{\nu}(x),$$

and thus letting $\nu = \tilde{\nu} + lim_{b\to\infty}f(b)\delta_0$, we have the statement of the theorem. □

Using this theorem, we show then the following example is the special case of translation invariant kernels. Before that we want to provide a proposition that verifies the positive semidefiniteness for the kernels given by *radial basis functions*.

Proposition 4. *Let $X \in \mathbb{R}^n, f : [0, \infty) \to \mathbb{R}$ and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $K(x, y) = f(||x - y||^2)$. If f is completely monotonic, then K is positive semidefinite.*

Proof. Bernstein's theorem gives us that there is a finite Borel measure ν on $[0, \infty)$, s.t.:

$$f(t) = \int_0^\infty e^{-tx}\, d\nu(x) \forall t \in [0, \infty).$$

Then, because for $x \in [0, \infty)$ we have $\widehat{e^{-x\|y\|^2}} = (\frac{\pi}{x})^{\frac{n}{2}} e^{-\frac{\|\xi\|^2}{4x}}$, and therefore $e^{-x\|y\|^2} = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \sqrt{\frac{\pi}{x}} e^{-\frac{\|\xi\|^2}{4x} + iy\xi}\, d\xi$. □

Thus:

$$\forall (x_1, ..., x_m) : \quad \sum_{i,j=1}^{m} K(x_i, x_j) = \int_0^\infty \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \widehat{f}(\xi) e^{i(x_i - x_j)\xi}\, d\xi\, d\nu(x)$$

$$= \int_0^\infty \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} (\frac{\pi}{x})^{\frac{n}{2}} e^{-\frac{\|\xi\|^2}{4x}} \left| \sum_{i=1}^{m} c_i e^{-ix_i\xi} \right|^2 d\xi d\nu(x) \geq 0.$$

Example 3 (Radial basis functions)**.** *Let $c > 0$. The following functions are Mercer kernels on any subset $X \subset \mathbb{R}^n$:*
$(1) K(x, t) = e^{\frac{-||x-t||^2}{c^2}}$ *(Gaussian)*
$(2) K(x, t) = (c^2 + ||x - t||^2)^{-\alpha}, \alpha > 0$ *(Inverse multiquadratics)*

2.5.3 Group invariant kernels

One interesting class of kernels are group invariant kernels. Encoding signals or building similarity kernels that are invariant to the action of a group is a key problem in unsupervised learning, as it reduces the complexity of the learning task.

First let us introduce the Haar measure and Haar integral and show that Haar-integration kernel is invariant to the group action.

For this it will be usefull to use the notions of the topological group and locally compact group.

Definition 24 (Topological group)**.** *A group G, together with a topology on the set G, s.t. group multiplication $G \times G \to G, (x, y) \mapsto xy$ and inversion $G \to G, x \mapsto x^{-1}$ are continuous maps is called topological.*

Example 4. *Some examples of the topological groups are:*

- *Any group with a discrete topology, because every subset is open. Such a group is called discrete group.*
- *Additive group $(\mathbb{R}, +)$ with usual topology, because $x_n \to x, y_n \to y \implies x_n - y_n \to x - y$.*
- *A group of real invertible matrices $GL_n(\mathbb{R})$.*

Definition 25 (Locally compact group)**.** *If every point of the topological space possesses a compact neighborhood, then it is called locally compact.*
A topological group is called locally compact when it is Hausdorff and locally compact.

Proposition 5. *Every topological group G that is T_1 is also T_2.*

Proof. A space X is T_2 iff the diagonal $\Delta_X \subset X \times X$ is closed. Now because G is a topological group, we have that the map $G \times G \to G, (x, y) \mapsto xy^{-1}$ is continuous, and diagonal is then closed as its inverse of the closed singleton $\{e\}$. It is closed, because G is T_1. □

The Haar measure is then based on the notions of the left-invariance and Radon measure. So let us introduce both of them.

Definition 26 (Radon measure)**.** *For a measure space $(X, \mathcal{B}, \mu)$, where $\mathcal{B}$ is a Borel σ-algebra μ is called:*

- *locally finite if for any $x \in X$ there is a neighborhood U_x, s.t. $\mu(U_x) < \infty$,*
- *inner regular if $\forall B \in \mathcal{B} : \mu(B) = sup\{\mu(K)|K \subset B\}$ for K compact,*
- *outer regular if $\forall B \in \mathcal{B} : \mu(B) = inf\{\mu(O)|B \subset O\}$ for O open.*
  Radon measure is then locally finite inner regular measure ν on $\mathcal{B}$.

Definition 27 (Left-invartiant measure)**.** *Let G be a locally compact group. A measure μ on the Borel σ-algebra $\mathcal{B}$ is called left-invariant if:*

$$\forall A \in \mathcal{B} \quad \forall g \in G : \quad \mu(gA) = \mu(A)$$

The following theorem which is at the same time a definition will be presented without the proof:

Theorem 10. *Let G be locally compact group. There exists a unique up to positive multipliers non-zero left-invariant Radon measure on G. It is called Haar measure. The corresponding integral is then Haar-integral.*

Then let us look at how it can be applied in kernel methods:

Definition 28 (Invariant Haar-Integration Kernels)**.** *We consider a subset $\mathcal{X}$ of the hypershpere in d dimensions $\mathcal{S}^{d-1}$. Let $\rho_{\mathcal{X}}$ be a measure on $\mathcal{X}$. Consider a kernel k_0 on $\mathcal{X}$, such as radial basis function kernel. Let G be a group that acts on $\mathcal{X}$ with a normalized Haar measure μ. G is assumed to be compact and unitary group. Define now an invariant kernel K between $x, z \in \mathcal{X}$ through Haar-integration as follows:*

$$K(x, z) = \int_G \int_G k_0(g_1 x, g_2 z) d\mu(g_1) d\mu(g_2).$$

As we are integrating over the entire group, it follows that $K(g_1 x, g_2 z) = K(x, z)$, $\forall g_1, g_2 \in G$, $\forall x, z \in \mathcal{X}$ which can be proven for the case of the characteristic, step functions and then by the monoton convergence theorem for the measurable functions.

Remark 3. *The kernel K is continuous, symmetric and if k_0 is a positive deifinite kernel, so is the kernel K, therefore it is a Mercer kernel and has the corresponding RKHS.*

Now for the binary classification problem one has that, given the labeled training set $S = ((x_1, y_1), ..., (x_m, y_m)) \in \mathcal{X} \times \{-1, +1\}$ the decision boundary f^*_m, the derivation of which in general for the RKHS will be shown in Proposition 11, is group-invariant as well:

$$\forall g \in G \forall x \in \mathcal{X} f^*_m(gx) = \sum_{i=1}^{m} \alpha_i K(gx, x_i) = \sum_{i=1}^{m} \alpha_i K(x, x_i) = f^*_m(x).$$

Moreover it leads to the reduced sample complexity.

2.5.4 Hypothesis spaces associated with an RKHS

As we stated before, it is important for the hypothesis space in order to have a finite covering number to be compact.

First of all let us introduce the ArzelàAscoli theorem, that is used to prove the needed Propositions.

Theorem 11 (ArzelàAscoli theorem). *Let X be compact and S be a subset of $\mathcal{C}(X)$. Then S is a compact subset of $\mathcal{C}(X)$ if and only if S is closed, bounded and equicontinuous.*

Where *equicontinuous* is defined as follows:

Definition 29 (Equicontinuous). *A subset S of $\mathcal{C}(X)$ is said to be equicontinuous at $x \in X$ if for every $\epsilon > 0$ there exists a neighborhood U_ϵ of x, such that for all $y \in U_\epsilon$ and $f \in S, |f(x) - f(y)| < \epsilon$. The set S is said to be equicontinous when it is so at every $x \in X$.*

Now we can show, how can a compact hypothesis space be found. To use the Arzelà-Ascoli we need first to have a bounded, closed, equicontinuous subset of $\mathcal{C}(X)$.

Proposition 6. *Let K be a Mercer kernel on a compact metric space X, and $\mathcal{H}_K$ be its RKHS. For all $R > 0$, the ball $B_R := \{f \in \mathcal{H}_K : ||f||_K \leq R\}$ is a closed subset of $\mathcal{C}(X)$.*

Proof. To show the closedness, let us take a convergent in $\mathcal{C}(X)$ sequence $(f_n)_{n \in \mathbb{N}}$ to some $f \in \mathcal{C}(X)$:

$$\forall x \in X : lim_{n \to \infty} f_n(x) = f(x).$$

Furthermore, using Riesz representation theorem and because B_R is weakly compact as a closed ball of a Hilbert space, we have that there exists a subsequence $(f_{n_k})_{k \in \mathbb{N}}$ and $\tilde{f} \in B_R$, s.t.:

$$\begin{aligned} \forall h \in \mathcal{H}_K : \quad & lim_{k \to \infty} \langle f_{n_k}, h \rangle = \langle \tilde{f}, h \rangle \\ \implies \forall x \in X : f(x) = \quad & lim_{k \to \infty} f_{n_k}(x) = lim_{k \to \infty} \langle f_{n_k}, K_x \rangle = \langle \tilde{f}, K_x \rangle = \tilde{f}(x) \\ \implies f = \tilde{f} \implies f \in B_R. & \end{aligned}$$

This proves the statement. □

Proposition 7. *Let K be a Mercer kernel on a compact metric space X, and $\mathcal{H}_K$ be its RKHS. For all $R > 0$, the set B_R as defined above is compact.*

Proof. Using Arzelà-Ascoli theorem we only need to show, that B_R is equicontinuous, i.e. when $\forall x \in X \quad \forall \epsilon > 0 \quad \exists V_x \quad \forall y \in V_x \quad \forall f \in B_R : |f(x) - f(y)| < \epsilon$. We have that $X \times X$ is compact, since X is compact. It also follows, that K being continuous on $X \times X$ is uniformly continuous on $X \times X$.
Therefore:

$$\begin{aligned}
&\forall x \in X \quad \forall \epsilon > 0 \quad \exists \delta > 0 \quad \forall y \in B_\delta(x) \quad \forall f \in B_R : \\
&|f(x) - f(y)| = |\langle f, K_x - K_y \rangle| \leq \|f\|_{\mathcal{H}_K} \|K_x - K_y\|_{\mathcal{H}_K} \\
&\leq R\sqrt{(K(x,x) - K(x,y) + K(y,y) - K(x,y))} \leq R\sqrt{2\epsilon}
\end{aligned}$$

This shows that B_R is also equicontinuous and thus compact. □

This way if we let X to be compact and $K : X \times X \to \mathbb{R}$ be a Mercer kernel, then by previous proposition, for all $R > 0$ we can consider B_R to be a hypothesis space.

2.5.5 Computation of empirical target function

What is interesting is the way we can find the empirical target function $f_{B_R,\mathbf{z}}$ in the infinite dimensional hypothesis space B_R.

Let K be a Mercer kernel, and $\mathcal{H}_K$ its RKHS, and $\mathbf{z} \in Z^m$. Let $\mathcal{H}_{K,\mathbf{z}} := span\{K_{x_1}, ..., K_{x_m}\}$ and $P : \mathcal{H}_K \to \mathcal{H}_{K,\mathbf{z}}$ be an orthogonal projection. Then since both f and $P(f)$ are in $\mathcal{H}_K$, for all $f \in B$ and $i = 1, ..., m$:

$$f(x_i) = \langle f, K_{x_i} \rangle = \langle P(f), K_{x_i} \rangle = (P(f))(x_i)$$

Therefore we have that $\mathcal{E}_{\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(P(f))$. It follows that if we let $B \subset \mathcal{H}_K$ be such that $P(B) \subset B$, then if $\mathcal{E}_{\mathbf{z}}$ can be minimized in B, then such a minimizer can be chosen in $P(B)$. One example follows from this:

Example 5. *In many situations - for example, when B is convex - the empirical target function $f_{\mathbf{z}}$ may be chosen in $\mathcal{H}_{K,\mathbf{z}}$. The norm restricted to $\mathcal{H}_{K,\mathbf{z}}$ is given by:*

$$||\sum_{i=1}^{m} c_i K_{x_i}||^2 = c^T K[\boldsymbol{x}]c.$$

And when $B = B_R$, we have that $f_{B_R,\boldsymbol{z}} = \sum_{i=1}^{m} c_i^* K_{x_i}$, where $c^* \in \mathbb{R}^m$ is a solution of the following problem:

$$min \frac{1}{m} \sum_{j=1}^{m} (\sum_{i=1}^{m} c_i K(x_i, x_j) - y_j)^2 \ s.t. \ c^T K[\boldsymbol{x}]c \leq R^2.$$

2.5.6 Mercer's theorem

We begin with the definition of the operator defined by a kernel

Definition 30. *Let $K : X \times X \to \mathbb{R}$ be a continuous function. Then the linear map $L_K : \mathcal{L}^2(X, \mu) \to \mathcal{C}(X)$, where μ is a finite Borel measure, is given by:*

$$(L_K f)(x) = \int K(x,t)f(t)d\mu(t), x \in X,$$

is well defined. Composition with inclusion $\mathcal{C}(X) \hookrightarrow \mathcal{L}^2(X, \mu)$ gives a linear operator, and the function K is said to be kernel of L_K

To find a representation of the Mercer kernel as a sum of convergent sequences of product functions, we need first to use the spectral theorem for an operator L_K, which is compact and self-adjoint and positive.

First of all, let us remind the spectral theorem:

Theorem 12. *If L is a compact self-adjoint linear operator on a Hilbert space H. Then there exists in H an orthonormal basis $\{\phi_1, ...\}$ consisting of eigenvectors of L, and each eigenvalue is real. If λ_k is the eigenvalue corresponding to ϕ_k, then either the set $\{\lambda_k\}$ is finite or $\lambda_k \to 0, k \to \infty$. In addition, $max_{k\geq 1}|\lambda_k| = \|L\|$. When L is also positive, then $\lambda_k \geq 0, k \geq 1$, and if L is strictly positive, then $\lambda_k > 0, k \geq 1$*

The fact that in case when K is a Mercer kernel L_K is positive, compact and self-adjoint follows from the following propositions:

Proposition 8. *If K is continuous, then $L_K : \mathcal{L}^2(X, \mu) \to \mathcal{C}(X)$ is well defined and compact. Additionally, $\|L_K\| \leq \sqrt{\mu(X)}C_K^2$, where C_K is defined as before when we have introduced the RKHS.*

Proposition 9. *If K is symmetric, then L_K is self-adjoint. If K is also positive semidefinite, then L_K is positive.*

And therefore all the results of the theorem for the positive, compact, self-adjoint operators apply to L_K, when K is a Mercer kernel.

The following theorem will be useful for the Mercer's theorem.

Theorem 13. *If μ is a Borel measure on X, and $K : X \times X \to \mathbb{R}$ is a Mercer kernel, and λ_k and ϕ_k are the kth eigenvalue and the corresponding orthogonal eigenfunction. Then $\{\sqrt{\lambda_k}\phi_k | \lambda_k > 0\}$ forms an orthonormal system in $\mathcal{H}_K$.*

Proof. By the reproducing property of RKHS and definition of L_K:

$$\begin{aligned}\langle \sqrt{\lambda_i}\phi_i, \sqrt{\lambda_j}\phi_j \rangle_{\mathcal{H}_K} &= \left\langle \frac{1}{\sqrt{\lambda_i}} \int_X K_x \phi_i(x)\, d\mu(x), \sqrt{\lambda_j}\phi_j \right\rangle_{\mathcal{H}_K} \\ &= \sqrt{\frac{\lambda_j}{\lambda_i}} \int_X \phi_i(x)\phi_j(x)\, d\mu(x) = \sqrt{\frac{\lambda_j}{\lambda_i}} \langle \phi_i, \phi_j \rangle_{\mathcal{L}^2(X,\mu)} = \delta_{ij},\end{aligned}$$

where δ_{ij} is the Kronecker delta.
This proves the statement. □

Theorem 14. *For a Mercer kernel $K : X \times X \to \mathbb{R}$ and λ_k, ϕ_k being kth positive eigenvalue and corresponding orthonormal eigenfunction we have:*

$$\forall x, y \in X : K(x,y) = \sum_{k\geq 1} \lambda_k \phi_k(x)\phi_k(y),$$

and the series converges absolutely (for each x, y) and uniformly on $X \times X$.

Proof. Using the Fourier series expansion of $K_x \in \mathcal{L}^2(X,\mu)$, orthonormal basis being $\{\phi_k\}_{k\geq 1}$ and that fact that $\forall f \in ker(L_K) : \langle K_x, f\rangle_{\mathcal{L}^2(X,\mu)} = \int_X K(x,y)f(y)\,d\mu(y) = 0$, we have that for K_x as a function in $\mathcal{L}^(X,\mu)$:

$$K_x = \sum_{k\geq 1}\langle K_x, \phi_k\rangle_{\mathcal{L}^2(X,\mu)}\phi_k = \sum_{k\geq 1} L_K(\phi_k)(x)\phi_k = \sum_{k\geq 1}\lambda_k\phi_k(x)\phi_k.$$

Now we need to show that this series converges absolutely and uniform on $X \times X$. Having the result that $\{\sqrt{\lambda_k}\phi_k\}_{k\geq 1}$ is an orthonormal system of $\mathcal{H}_K$ and using the Parseval's theorem we have for fix $x \in X$ and for the Fourier coefficients of the function $K_x \in \mathcal{H}_K$ with respect to this system:

$$\sum_{k\geq 1} |\langle\sqrt{\lambda_k}\phi_k, K_x\rangle_{\mathcal{H}_K}|^2 = \sum_{k\geq 1}\lambda_k|\phi_k(x)|^2 \leq \|K_x\|^2 \leq \|K_x\|^2_{\mathcal{H}_K} = K(x,x) \leq C_K^2.$$

Thus, we have that for a fix $x \in X$ and $\forall y \in X$:

$$|\sum_{k=m}^{m+l}\lambda_k\phi_k(x)\phi_k(y)| \leq \left(\sum_{k=m}^{m+l}\lambda_k|\phi_k(x)|^2\right)^{1/2}\left(\sum_{k=m}^{m+l}\lambda_k|\phi_k(y)|^2\right)^{1/2} \leq C_K\left(\sum_{k=m}^{m+l}\lambda_k|\phi_k(x)|^2\right)^{1/2}.$$

This tends to zero uniformly, for $y \in X$, thus $\sum_{k\geq 1}\lambda_k\phi_k(x)\phi_k(y)$ as a function of y converges absolutely and uniformly to a continuous function f_x.
Therefore, as functions in $\mathcal{L}^2(X,\mu) : K_x = f_x \quad \forall y \in X$. And also $\sum_{k\geq 1}\lambda_k\phi_k(x)\phi_k(y)$ converges to $K(x,y)$ uniformly on $X \times X$. □

Corollary 3. *It follows that:*

$$K(x,x) = \sum_{k\geq 1}\lambda_k\phi_k^2(x) = \sum_{k\geq 1}\langle K_x, \sqrt{\lambda_k\phi_k}\rangle^2_{\mathcal{H}_K} = \|K_x\|^2_{\mathcal{H}_K},$$

which due to the Parseval's theorem gives us that $\{\sqrt{\lambda_k}\phi_k|\lambda_k > 0\}$ is even an orthonormal basis of $\mathcal{H}_K$.

2.5.7 Least squares regularization

Let's now add a penalization term in the error to avoid overfitting. Here the setting of a compact space is abandoned and the hypothesis space will be a RKHS $\mathcal{H}_K$.
This way the *regularized error* and the *regularized empirical error* are defined as follows:

Definition 31. *Let the whole RKHS $\mathcal{H}_K$ be a hypothesis space. Then the regularized error $\mathcal{E}_\gamma$ is defined by:*

$$\mathcal{E}_\gamma(f) = \int_Z (f(x)-y)^2 + \gamma||f||^2_{\mathcal{H}_K},$$

with f_γ being its minimizer over $\mathcal{H}_K$; and the regularized emprirical error $\mathcal{E}_{\boldsymbol{z},\gamma}$ is defined by:

$$\mathcal{E}_{\boldsymbol{z},\gamma}(f) = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2 + \gamma||f||^2_{\mathcal{H}_K},$$

with $f_{\boldsymbol{z},\gamma}$ minimizing it over $\mathcal{H}_K$.

Since the hypothesis space is not compact, the existence of f_γ and $f_{\mathbf{z},\gamma}$ is not obvious.

Using the following result on Lagrange multipliers, one can prove that $f_{\mathbf{z},\gamma}$ exists and is unique:

Proposition 10. *Let U be a Hilbert space and $F, H : U \to \mathbb{R}$ smooth. Let $c \in U$ be a solution of the problem*

$$minF(f) \ s.t. \ H(f) \leq 0.$$

Then there exist $\mu, \lambda \in \mathbb{R}$ with

$$\mu DF(c) + \lambda DH(c) = 0.$$

If $H(c) < 0$ then $\lambda = 0$ and $\mu \neq 0$. If $DH(c) \neq 0$ then we can take $\mu = 1$ and $\lambda \geq 0$

And so now the proposition follows:

Proposition 11. *Let $\boldsymbol{z} \in Z^m$ and $\gamma > 0$. The empirical target function can be expressed as*

$$f_{\boldsymbol{z}}(x) = \sum_{i=1}^{m} a_i K(x, x_i),$$

where $\boldsymbol{a} = (a_1, ..., a_m)$ is the unique solution of the linear equation:

$$(\gamma m Id + K[\boldsymbol{x}])\boldsymbol{a} = \boldsymbol{y}$$

Proof. Using Corollary 3 we have: $\forall f \in \mathcal{H}_K : f = \sum_{k\geq 1} \frac{c_k}{\sqrt{\lambda_k}} \sqrt{\lambda_k}\phi_k$, and $\|f\|^2_{\mathcal{H}_K} = \sum_{k\geq 1} \frac{c_k^2}{\lambda_k}$. Then $\mathcal{E}_{\mathbf{z},\gamma}$ is minimized by setting

$$c_k = \lambda_k \sum_{i=1}^{m} \alpha_i \phi_k(x_i),$$

where $a_i = \frac{y_i - f(x_i)}{\gamma m}$.
Then, using Mercer's theorem:

$$f(x) = \sum_{k\geq 1} \lambda_k \sum_{i=1}^{m} a_i \phi_k(x_i)\phi_k(x) = \sum_{i=1} a_i K(x, x_i).$$

Inserting now this in the definition of a_i and multiplying both sides by γm, we have that: $(\gamma m)Id + K[\mathbf{x}])\mathbf{a} = \mathbf{y}$, and there is a unique solution, because $K[\mathbf{x}]$ is positive semidefinite and addition of the identity to it will be positive definite. □

2.6 Coherence function

We consider the situation where we have a number of time series and wish to explore the relation between them. Here we look first at the cross-correlation and then at the cross-spectral density and coherence. We will mainly introduce new definitions.

Definition 32 (Cross-covariance and cross-correlation)**.** *Let* (x_t, y_t) *represent a pair of signals that are jointly wide-sense stationary, and* $\mu_x = \mathbf{E}[x_t], \mu_x = \mathbf{E}[x_t]$ *are the means of the corresponding time series, the cross-covariance is defined as:*

$$\sigma_{xy}(T) = E[(x_t - \mu_x)(y_{t+T} - \mu_y)],$$

and cross-correlation, given is in turn defined as:

$$r_{xy}(T) = \frac{\sigma_{xy}(T)}{\sqrt{\sigma_{xx}(0)\sigma_{yy}(0)}}.$$

Cross spectral density is a Fourier transform of the cross-covariance function:

Definition 33 (Cross-spectral density)**.** *It is defined by:*

$$S_{xy}(w) = \int_{-\infty}^{\infty} r_{xy}(t)e^{-jwt}.$$

Definition 34 (Coherence)**.** *The coherence between* x_t *and* y_t *is then represented as:*

$$C_{xy}(w) = \frac{|S_{xy}(w)|^2}{S_{xx}(w)S_{yy}(w)}.$$

3 Application

3.1 The data

- **Length(cm)** *Target length* in multibeam data is estimated using the maximum distance between any two above threshold samples in the target, measured by angle and range from the transducer.

- **Perimeter(cm)** *Target perimeter* is calculated as the sum of the target circumference contributions from the above-threshold samples in the target. The measurement uses the full sample space. For targets that have only one sample, the target length will be $0cm$ but the target perimeter will be the sum of the sides of the sample.

- **Area(cm^2)** *Target area* is defined as the sum of the areas of all samples within the boundary of the target. Where:

$$\textit{Area of sample} := (\frac{\textit{(Stop range of sample)}^2 * \textit{(Beam angle)}}{2} - \frac{\textit{(Start range of sample)}^2 * \textit{(Beam angle)}}{2}).$$

- **Thickness(cm)** *Target thickness* is the maximum range (interval) covered by the outline of samples per beam in the target.

- **Compactness** A common compactness measure is the *Circularity ratio*. It is the ratio of area of the shape to the area of a circle having the same perimeter. It is accepted that a circle has the most compact shape.

$$\textit{Circularity ratio} := \frac{4\pi * \textit{Area}}{(\textit{Perimeter})^2}.$$

  However in pattern recognition literature , the inverse of the *Circularity ratio* is often discussed. Hence, Echoviews target compactness is calculated as:

$$\textit{Circularity ratio} := \frac{(\textit{Perimeter})^2}{4\pi * \textit{Area}}.$$

- **Intensity variation** *Target intensity variation* is calculated as:

$$CV := \frac{\sigma}{\mu}.$$

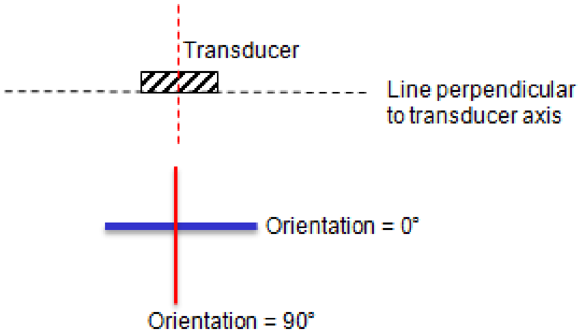
  Where:

  - CV := Coefficient of Variation of the intensity of the samples in the target.
  - $\sigma := \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$.
  - μ := mean intensity of samples in the target.
  - x_i := intensity value of sample i in the target.
  - N := number of samples in the target.

- **Length across beam(cm)** This property is faster to calculate than the *Target length* property and it can be more accurate for situations where fish are swimming perpendicular to the central beam. Target length across beams is determined by: 1. Finding for each beam in the target, the range of the midpoint between the shallowest sample and deepest sample. 2. Finding the length of the line that traverses the target when the midpoints are joined

- **Range extent (cm)** *Target range extent* is calculated as the range difference between the shallowest sample and the deepest sample in the target.

- **Orientation (degrees)** *Target orientation* property is the angle between the *target length* derived axis and the line perpendicular to the transducer axis. The *target length* axis is defined as the line joining the two most distant samples in the target. The available range of values is as follows, $0° < \textit{Target orientation} < 180°$.



3.2 The method

In order to cluster the fishes in two classes, weve conducted the following steps:

- **Selected features:**
  For the sake of having the features that uniqely characterize the movement of the fish we have selected the following features: *Area,Thickness, Compactness, Intensity variation, Length across beam, Perimeter,Orientation*. Features such as *Upstream, X angle, Range*, and *X distance* are not representing their movement, becasuse they depend more on the position of the sonar than on the motion of the objects that we are trying to cluster. The features such as *Length* and *Range extent* can be replaced by the *Length across beam* and *Thickness* correspondingly. Below is the code for the selection:

```
#add the time series wiht selected paramters to the new_samples_list
^^Ifor j in range(from_col,from_col+Y_batch[0,:].size):
^^I^^IY_batch_temp = data[glob_count-count:glob_count\
    +1*(glob_count == data[:,0].size-1),j] #get the selected parameters
```

- **Selected length of the time series:**
  Time order index is defined by the property *FrameEV*. It has been rescaled so that the time series for each particular fish is indexed from the time point 0. Consequently their size was lengthened for the sake of applying the linear interpolation later on:

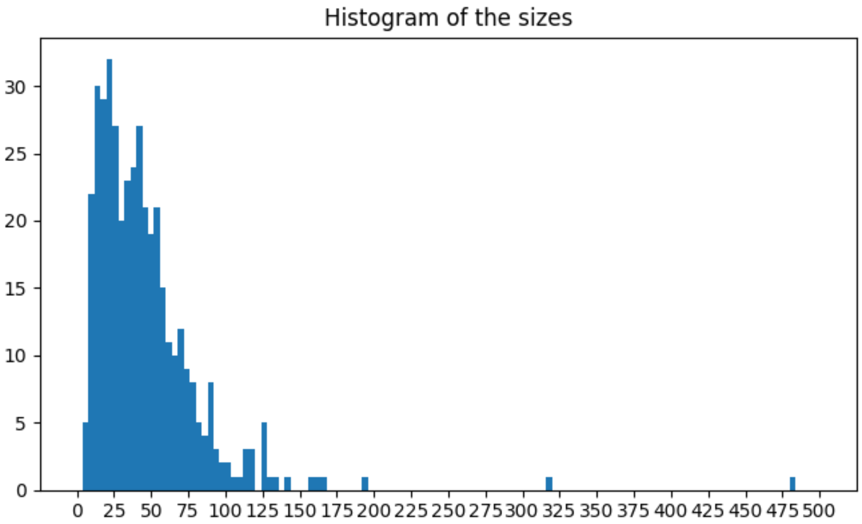
```
_X_temp = X_batch-X_batch[0] #rescale the index of the
#time order of the time series
X_temp = np.arange(_X_temp[-1]+1) #create new time intervals
#for the interpolated time series
```

  Later on the fishes that are represented by the time intervals of the length 8 were omitted and the number of such fishes is saved in the varibale *num of cuts*:

```
if len(X_temp)<=cut_num:
^^I#drop the times series if their interpolated lentgth
^^I# <= cut_num
^^Iif (j-from_col)==0:
^^I^^Inum_of_cuts+=1
```

  All other time series after the application of the linear interpolation are added to the list *new samples list*:

```
else:
^^I#add other time series to the new_samples_list
^^Iif (j-from_col)==0:
^^I^^Inew_samples_list.append([])
^^I^^Inew_samples_list[int(i)-2 + 1*(glob_count\
^^I^^I== data[:,0].size-1)-num_of_cuts].append([])
^^I^^Inew_samples_list[int(i)-2 + 1*(glob_count\
^^I^^I== data[:,0].size-1)-num_of_cuts][0] \
^^I^^I= int(i)-2 + 1*(glob_count == data[:,0].size-1)
^^Inew_samples_list[int(i)-2 + 1*(glob_count \
^^I== data[:,0].size-1)-num_of_cuts].append([])
^^Inew_samples_list[int(i)-2 + 1*(glob_count \
^^I== data[:,0].size-1)-num_of_cuts][j-from_col+1] = Y_temp
```



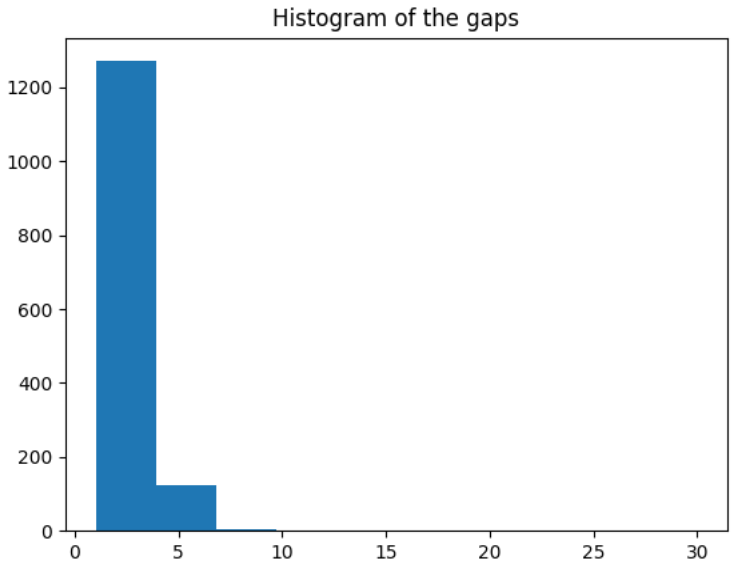
- **Time series with length $>$ 8 are linearly interpolated:**

```
Y_temp = np.interp(X_temp,_X_temp,Y_batch_temp) #interpolate
```

- **Gaps in the time intervals:**
  We have to make sure that the gaps in the indexing of the time series are not too big and apper not too often. Otherwise it would be difficult or impossible to realistically interpolate them.

```
holes_list.append([])
for nn,ii in enumerate(X_temp):
^^Itry:
^^I^^Iif (map(int,list(X_temp))[nn+1]\
^^I^^I-map(int,list(X_temp))[nn])!=1:
^^I^^I^^Iholes_list[int(i)-2+1*(glob_count\
^^I^^I^^I== data[:,0].size-1)]\
^^I^^I^^I.append(map(int,list(X_temp))[nn+1]\
^^I^^I^^I-map(int,list(X_temp))[nn]-1)
^^I^^I^^I_holes_list.append(map(int,list(X_temp))\
^^I^^I^^I[nn+1]-map(int,list(X_temp))[nn]-1)
```

Then the histogram of the number of gaps as well as 20 of the greatest of them are printed out:



- **The coherence function:**
  Afterwards we want to simplify our features (*Area, Thickness, Compactness, Intensity variation, Length across beam, Perimeter, Orientation*), which are repre- sented by the high-dimensional time series. To do so, we use the Coher- ence function ($C_{xy}(w) = \frac{|S_{xy}(w)|^2}{S_{xx}(w)S_{yy}(w)}$) pairwise between the features. And we do so for the each fish: we take two pairs of features with the smallest energy (that are the least similar) $(min_{(x,y)\in X\times X:x\neq y}(\sum_w C_{xy}(w))$, where X is the set of all features), assuming that the sampling rate is 7 f/s. The corresponding code is given below:

```
features = np.zeros((len(new_samples_list),6))  #new features
#vector for the fishes - corresponds to the 6 frequencies

for q in range(len(new_samples_list)):
    fish_n=q
    Cxy = [] #list of the coherence functions for all
^^I     #possible pairs of the parameters
```

```
             #(Area,Thickness,Compactness,Intensity_variation,
             #Length_across_beam Perimeter,Orientation) for each fish
    for enum,i in enumerate([z for z in combinations(range(len\
    (new_samples_list[fish_n])-1),2)]):
        Cxy.append([])
        f, Cxy[enum] = signal.coherence(new_samples_list[fish_n]\
        [1+i[0]], new_samples_list[fish_n][1+i[1]],sr,nperseg\
        =len(new_samples_list[fish_n][1+i[0]])/1.85)
        #calculation of the coherence function
        #with the window size of length (1/1.85) of the original
        #time series' length

    _sum_small = heapq.nsmallest(2,[sum(k) for k in Cxy if not\
    np.isnan(k).any()])
    #selection of two pairs of parameters with the least energy
    #of the coherence function

    _sum_small_ind = [i for i,x in enumerate([sum(k) for k in \
    Cxy if not np.isnan(k).any()]) if x in _sum_small]
```

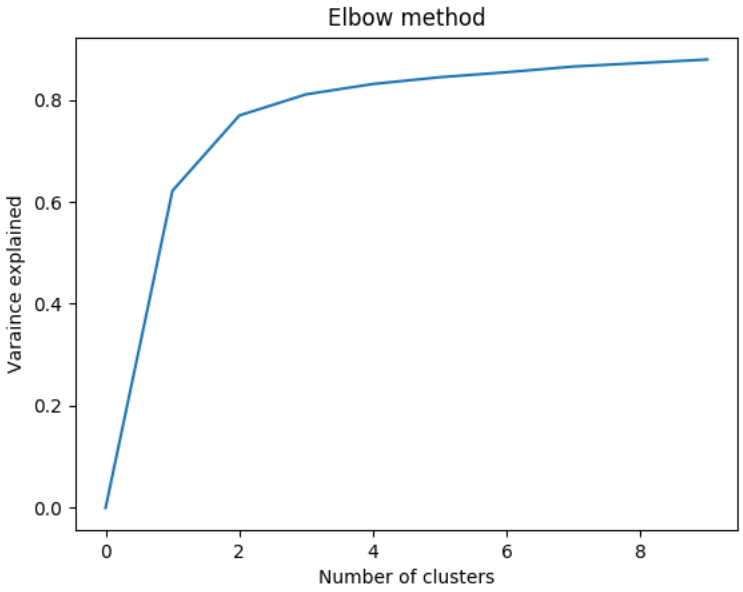
- **The transformation of our features:**
  Then we take 2 sets of 3 frequencies each (lists *biggest w1, biggest w2* in our program) that correspond to the 3 greatest frequencies in the two selected Coherence functions (for each of the P two sets of frequencies should hold: $w_1, w_2, w_3 = argmax_w(min_{\{(x,y)\in X\times X: x\neq y\}}(\sum_w C_{xy}(w))$ and then write them to the corresponding element of the array features:

```
    biggest_w1 = heapq.nlargest(3,Cxy[_sum_small_ind[0]])
    #selection of the three frequencies with the highest energy
    #in the first pair
    biggest_w2 = heapq.nlargest(3,Cxy[_sum_small_ind[1]])
    #selection of the three frequencies with the highest energy
    #in the second pair
    w = biggest_w1+biggest_w2 #list of the 6 selected frequencies
    features[q] = w
```

- **So each fish is now represented by 6 features.**

- **Clustering:**
  Consequently we apply the K-Means clustering algorithm for the newly selected attributes. On top of that we evaluate the variance explained by our clusters using Elbow method:

```
#k-elbow method with K-means clustering algorithm
def elbow(XX, n):
    kMeansVar = [KMeans(n_clusters=k).fit(XX) for k in range(1, n)]
    centroids = [X.cluster_centers_ for X in kMeansVar]
    k_euclid = [cdist(XX, cent) for cent in centroids]
    dist = [np.min(ke, axis=1) for ke in k_euclid]
    wcss = [sum(d**2) for d in dist]
    tss = sum(pdist(XX)**2)/XX.shape[0]
    bss = (tss-wcss)/tss
    plt.plot(bss)
    plt.title('Elbow method')
    plt.xlabel('Number of clusters')
    plt.ylabel('Varaince explained')
    plt.show()
    kmeans = KMeans(n_clusters=2).fit(XX)
    labels = kmeans.labels_
    return labels
```



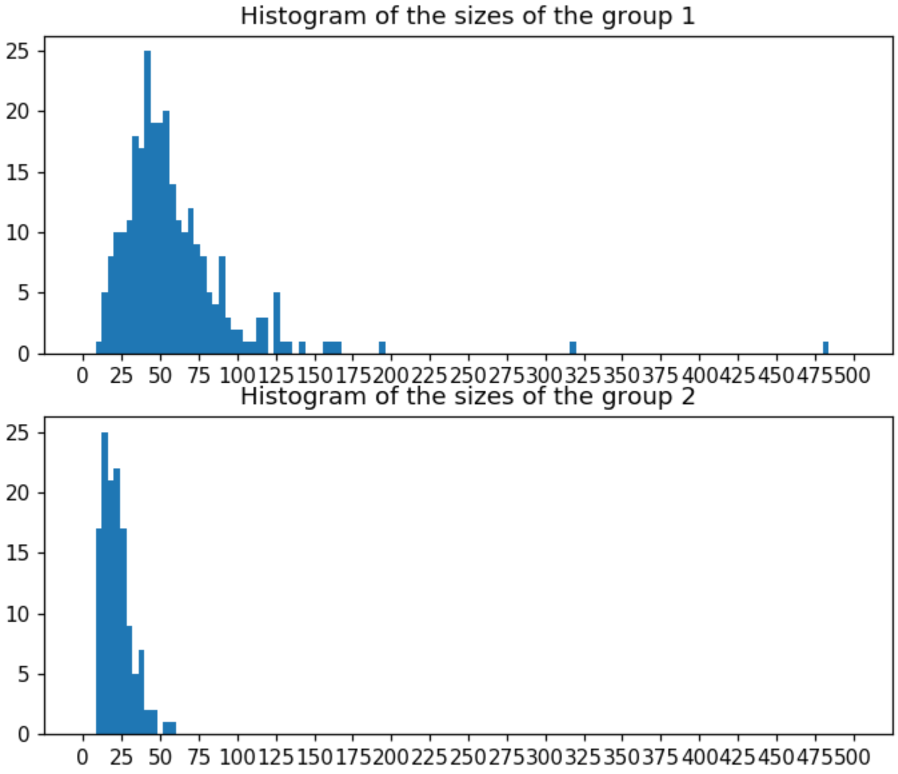
Using Elbow method one could see, that two clusters explain the most variance. We calculate the quality of the clustering as:
$R^2 = \frac{\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i)^2 - \sum_{i=1}^{K}\sum_{x_j \in C_i}(x_j - c_i)^2}{\sum_{i=1}^{N}(x_i - \frac{1}{N}\sum_{i=1}^{N} x_i)^2}$,
where N is the number of fishes that are represented by the time series of the length $> 8, K$ is the number of the clusters that were found, and c_i is the centroid of the corresponding cluster.

- **The clusters found:**
  Having said that, we want now to identify the groups of the fishes that were found. Considering that our function *elbow* returns the labels for the two clusters, we implement it this way:

```
labels = elbow(features,11) #Cluster the fishes with new
#features and plot the variance explained
```

```
#Get the clusters and the quantity of fishes that are in that clusters
group1 = [i for (i,k) in enumerate(labels) if k==0]
group2 = [i for (i,k) in enumerate(labels) if k==1]
numbers_of_fishes = []
for i in new_samples_list:
^^Inumbers_of_fishes.append(i[0])
_temp = np.array(labels)*(1+np.array(numbers_of_fishes))
group_1 = _temp[_temp!=0]
_temp = (np.array(labels)==0)*(1+np.array(numbers_of_fishes))
group_2 = _temp[_temp!=0]
print "Group N1: ",group_1
print "Group N2: ",group_2
print "Len Group N1: ", len(group_1)
print "Len Group N2: ", len(group_2)
```

We see that the groups found this way are of the sizes 129 and 273, which indicates that they can be considered significant in their size.

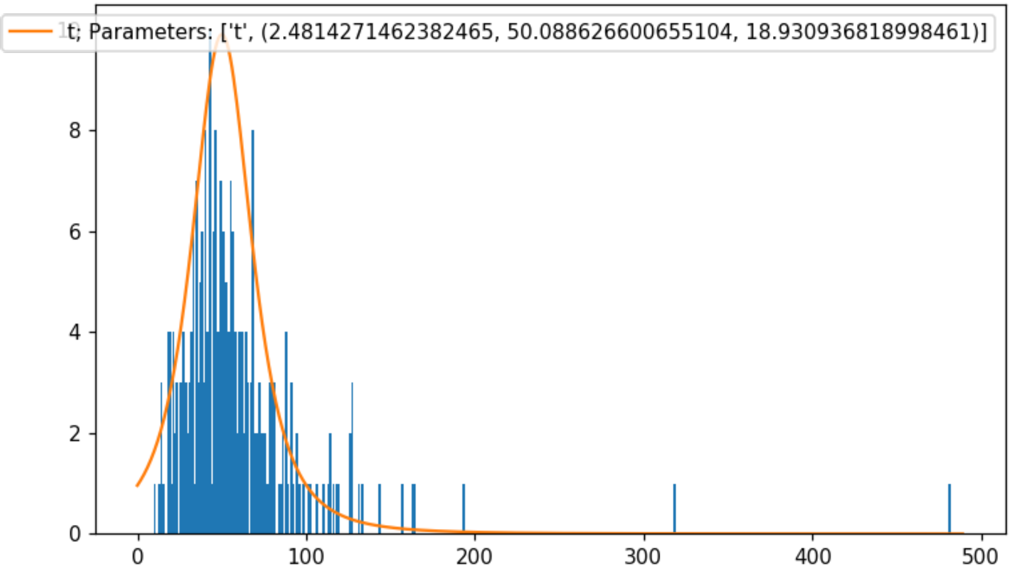


- **Kolmogorov-Smirnov test and conclusions:**
  In order to be sure that the the found clusters correspond to the different groups of fishes we will: 1) Implement Kolmogorov-Smirnov test and

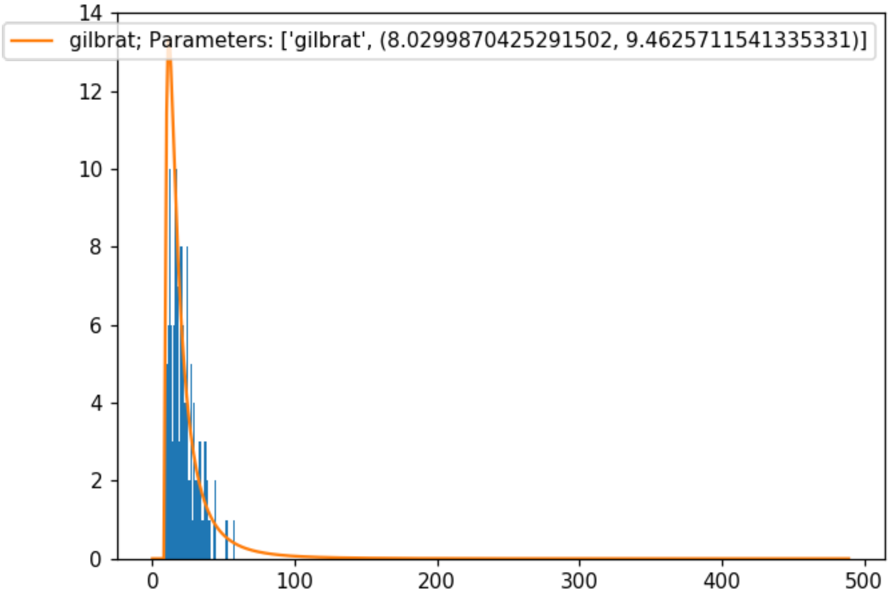
check, whether the null hypothesis is rejected at the given level α (e.g. $\alpha = 0.001$). The Kolmogorov- Smirnov statistics is defined as follows: $D_{n,m} = sup_x |F_{1,n}(x) - F_{2,m}(x)|$, where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively.
The null hypothesis is rejected at level α if: $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}}$.
Here n and m are the sizes of first and second sample respectively. The value of $c(\alpha)$ is given in the table below for the most common levels of α

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

2) Fit different distributions to the lengths of the groups found and visually find that these distributions are unlike. Note: only the first part is necessary to statistically reject or accept the hypothesis tested. The second part will give us more intuitive and visuall representation of the comparison. The results of the first and second parts: 1) Kolmogorov-Smirnov test shows us that the distributions differ drasti- cally and the null hypothesis is rejected even at level $0.001((0.919) > (0.208)$, here 0.919 is the K-S-statistic and 0.208 is the corresponding value). 2) One can also find the distributions that fit out sample data the best. The difference is seen visually. For the first group the optimally fitted distribution is t-distribution with parameters: $degress\ of\ freedom = 2.487$, $location = 49.696$, $scale = 18.885$.



For the second group the optimally fitted distribution is genaralized Pareto distribution with parameters: $shape\ paremeter = 0.241$, $location = 8.999$, $scale = 10.723$.


gilbrat; Parameters: ['gilbrat', (8.0299870425291502, 9.4625711541335331)]
14
12
10
8
6
4
2
0
0
100
200
300
400
500