# An Application of Stability to Regularization in Hilbert Space

Udie Subre Abdul Razaq
Basic Education College
University of Babylon

Kawther Fawzi Hamza
Education College
University of Babylon

## Abstract

In this paper ,some definition and concepts of stability are given, an application of stability to regularization in Hilbert space are performed with modification. some illustration examples and conclusion, are listed.

## 1.Introduction

It has long been known that when trying to estimate an unknown function from data, one needs to find a tradeoff between bias and variance. Indeed, on one hand, it is natural to use the largest model in order to be able to approximate any function, while on the other hand, if the model is too large, then the estimation of the best function in the model will be harder given a restricted amount of data. Several ideas have been proposed to fight against this phenomenon. One of them is to perform estimation in several models of increasing size and then to choose the best estimator based on a complexity penalty (e.g. Structural Risk Minimization). One such technique is the bagging approach of Breiman (1996)[5] which consists in averaging several estimators built from random sub samples of the data. In the early nineties, concentration inequalities became popular in the probabilistic analysis of algorithms, due to the work of McDiarmid (1989) and started to be used as tools to derive generalization bounds for learning algorithms by Devroye (1991). Building on this technique, Lugosi and Pawlak (1994) obtained new bounds for the k-NN, kernel rules and histogram rules.

A key issue in the design of efficient machine learning systems is the estimation of the accuracy of learning algorithms. among the several approaches that have been proposed to this problem, one of the most prominent is based on the theory of uniform convergence of empirical quantities to their mean .this theory provides ways to estimate the risk (or generalization error) of a learning system based on an empirical measurement of its accuracy and measure of its complexity, such as the Vapnik-Chervonenkis(VC)dimension or the fat-shattering dimension. We explore here a different approach which is based on sensitivity analysis .sensitivity analysis aims at determining how much the variation of the input can influence the output of a system .it has been applied to many areas such as statistics and mathematical programming .in the latter domain ,it is often referred to as perturbation analys.

Uniform stability may appear as a strict condition . actually we will observe that many existing learning methods exhibit a uniform stability which is controlled by the regularization parameter and thus be very small. many algorithms such as Support Vector Machines(SVM) or classical regularization networks introduced by Poggio and Girosi (1990)[9] perform the minimization of a regularized objective function where the regularizer is a norm in a reproducing kernel Hilbert space (RKHS):

refers to the kernel. $k$ , where $N(f)=//f//_k^2$

**Some Definitions and Concepts**

**Def.(1)** ***Learning Algorithm*** [ 4]:- a learning algorithm is a function A from $Z^m$ into $F\subset Y^x$ which maps a learning set S onto a function $A_S$ from X to Y ,such that $X,Y\subset R$ are an input and output space respectively , and S a training set $S=\{z_1=(x_1y_1),\dots z_m=(x_m,y_m)\}$, of size $m$ in $Z=X\times Y$ drawn i.i.d. from an unknown distribution D.

**Def.(2)** ***Stability*** [3]:- consider the system

$$x'(t) = f[x(t),u(t)], x \in X \subset R^n, u \in U \subset R^m,$$

A stable $\bar{x}$ is an equilibrium state if there exists $\bar{u}$ such that $f(\bar{x},\bar{u}) = 0.$

**Def.(3)** ***Hilbert Space*** [2 ]:- an inner-product space which is infinite dimensional and complete is called a Hilbert space (real or complex ,according to whether the scalar filed is real or complex).

**Def.(4)** ***Hypothesis Stability*** [ 8] :-An algorithm A has hypothesis stability with respect to the loss function $\ell$ if the following holds

$$\forall i \in \{1,\ldots,m\}, E_{S,Z}[/\ell(A_S,z)/] \le \beta.$$

Note that this is the $L_1$ norm with respect to D, so that we can rewrite the above as

$$E_S[\| \ \ell(A_S,.) - \ell(A_{S^{\setminus i}},.) \|] \le \beta.$$

**Def.(5)** ***Pointwise Hypothesis Stability***[8]:- An algorithm A has point wise hypothesis stability $\beta$ with respect to the loss function $\ell$ if the following holds

$$\forall i \in \{1,...,m\}, E_S[/\ell(A_S,z_i) - \ell(A_{S\setminus i},z_i)/] \le \beta.$$

Another, weaker notion of stability was introduced by Kearns and Ron. It consists of measuring the change in the expected error of the algorithm instead of the average pointwise change.

**Def.(6)** ***Error Stability***[ 3]:- An algorithm A has error stability . with respect to the loss function $\ell$ if the following holds

$$\forall S \in Z^m, \forall i \in \{1,...,m\}, \left| E_z[\ell(A_S,z)] - Ez[\ell(A_{S\setminus i},z)] \ \right| \le \beta,$$

Which can also be written

$$\forall S \in Z^m, \forall i \in \{1,..., m\}, | \ R(S) - R^{\setminus i}(S) \ | \le \beta.$$

**Def.(7)** ***Uniform Stability***[9 ]:- An algorithm A has uniform stability $\beta$ with respect to the loss function $\ell$ if the following holds

$$\forall S \in Z^m, \forall i \in \{1,..., m\}, | \ R(S) - R^{\setminus i}(S) \ | \le \beta.$$

**Def.(8)** ***convex set*** [2 ]:- A set S in vector space X on the field F, is called convex if

$$\lambda x + (1-\lambda)y \in S,$$

$$\forall x, y \in S \quad , \forall \ 0 \le \lambda \le 1$$

*i.e*

$$\lambda S + (1-\lambda)S \subseteq S$$

**Def.(9)** ***A loss function***[6 ] :-A loss function $\ell$ defined on F×Y is $\sigma$-admissible with respect to F if the associated cost function c is convex with respect to its first argument and the following condition holds

$$\forall y_1, y_2 \in D, \forall y' \in Y, \left| c(y_1,y') - c(y_2,y') \right| \le \sigma \left| y_1 - y_2 \right|,$$

Where $D=\{y:\exists f\in F,\exists x\in X, f(x)=y\}$ is the domain of the first argument of c. thus in the case of the quadratic loss for example ,this condition is verified if Y is bounded and F is totally bounded, that is there exists $M\le\infty$ such that $\forall f\in F,\|f\|_\infty\le M$ and $\forall y\in Y,|y|\le M$, such that F is a convex subset of a linear space.

**Def.(10)** ***Classification Stability***[5 ] :- a real-valued classification algorithm A has classification stability $\beta$ if the following holds.

$$\forall S\in Z^m,\forall i\in\{1,\dots,m\},\left\|A_S(.)-A_{S^{\backslash i}}(.)\right\|_\infty\le\beta.$$

## 2. Survey of Online Kernel Methods[6]

The perception algorithm (Rosenblatt, 1958) is arguably one of the simplest online learning algorithms.

Given a set of labeled instances $\{(x_1y_1),(x_2,y_2)\dots(x_m,y_m)\}\subset X\times Y$ where $X\subseteq R^d$ and $y_i$ ∈ {±1} the algorithm starts with an initial weight vector $\theta=0$. It then predicts the label of a new instance x to be $\hat{y}=\sin g(\langle\theta,x\rangle)$.. If $\hat{y}$ differs from the true label $y$ then the vector $\theta=0$ is updated as $\theta=\theta+yx$. This is repeated until all points are well classified. The following result bounds the number of mistakes made by the perceptron algorithm (Freund and Schapire, 1999, Theorem 3) this generalizes the original result for the case when the points are strictly separable , i.e, when there exists a $\theta$ such that $\|\theta\|=1$ and $y_i\langle\theta,x_i\rangle\ge\gamma$ for all $(x_i,y_i)$.the so-called kernel trick has recently popularity in machine learning . as long as all operations of an algorithm can be expressed with inner product ,the kernel trick can used to lift the algorithm to a higher-dimensional feature space: the inner product in the feature space produced by the mapping $\phi:X\to H$ is represented by a kernel $k(x,x')=\langle\phi(x),\phi(x')\rangle_H$. We can now drop the condition $X\subseteq R^d$ but instead require that H be a reproducing kernel Hilbert space (RKHS).

## 3. Some Theorems About an Application of Stability to Regularization and Illustration Examples

Many algorithms such as Support Vector Machines (SVM) or classical regularization networks introduced by Poggio and Girosi (1990) perform the minimization of a regularized objective function where the regularizer is a norm in a reproducing kernel Hilbert space(RKHS): $N(f)=//f//_k^2$,

where k refers to the kernel .the fundamental property of a RKHS F is the so-called reproducing property which writes

$$\dots\dots(2)\ \forall f\in F,\forall x\in X, f(x)=\langle f,k(x,.)\rangle.$$

In particular this gives by Cauchy-Schwarz inequality

$$\forall f\in F,\forall x\in X,|f(x)|\le\|f\|_k\sqrt{k(x,x)} \qquad \dots\dots(3)$$

**Theorem** (1):[8 ]

let F be a reproducing kernel Hilbert space with kernel k such that $\forall x\in X,k(x,x)\le\kappa^2<\infty$.let $\ell$ be $\sigma$-admissible with respect to F .the learning algorithm A defined by

$$A_S = \arg\min_{g\in F} \frac{1}{m}\sum_{i=1}^{m} \ell(g, z_i) + \lambda\|g\|_k^2, \qquad .(4)$$

Has uniform stability $\beta$ with respect to $\ell$ with

$$\beta \leq \frac{\sigma^2\kappa^2}{2\lambda m}.$$

**Proof**

We have $\quad d_N(g, g') = \|g - g'\|_k^2.$

Thus, by the *(Let $\ell$ be admissible with respect to F, and N a functional defined on F such that for all training sets S, $R_r$ and $R_r^{\setminus i}$ have a minimum (not necessarily unique ) in F .let f denote a minimizer in F of $R_r$ , and for i=1,..,m ,let f denote a minimizer in F of $R_r^{\setminus i}$ .we have for any t=[0,1] ,*

$N(f) - N(f + t\Delta f) + N(f^{\setminus i}) - N(f^{\setminus i} - t\Delta f) \leq \dfrac{t\sigma}{\lambda m}|\Delta f(x_i)|$, *where* $\Delta f = (f^{\setminus i} - f).$ *)* ,gives

$$2\|\Delta f\|_k^2 \leq \frac{\sigma}{\lambda m}|\Delta f(x_i)|.$$

Using (2),we get

$$|\Delta f(x_i)| \leq \|\Delta f\|_k \sqrt{k(x_i, x_i} \leq k\|\Delta f\|_k,$$

So that

$$\|\Delta f\|_k \leq \frac{\sigma\kappa}{2\lambda m}.$$

Now we have ,by the $\sigma$-admissibility of $\ell$

$$|\ell(f, z)| - \ell(f^{\setminus i}, z) \leq \sigma|f(x) - f^{\setminus i}(x)| = \sigma|\Delta f(x)|, \qquad .......(4)$$

Which ,using (2)again , gives the results.

**Theorem** (2):[8 ]

Let A be the algorithm of theorem(1) where $\ell$ is a loss function associated to a convex cost function c(. , .) .we denote by B(.) a positive non-decreasing real-valued function such that for all $y \in D$ .

$$\forall y' \in Y, c(y, y') \leq B(y)$$

For any training set S, we have

$$\|f\|_k^2 \leq \frac{B(0)}{\lambda}, \qquad ...... (5)$$

And also

$$\forall z \in Z, 0 \leq \ell(A_S, z) \leq B(\kappa\sqrt{\frac{B(0)}{\lambda}})$$

Moreover , $\ell$ is $\sigma$-admissible where $\sigma$ can be taken as

$$\sigma = \sup_{y\ \in Y} \sup_{|y| \leq B(\kappa\sqrt{\frac{B(0)}{\lambda}}} \left|\frac{\partial c}{\partial y}(y, y')\ \right|$$

Proof We have for $\quad f = A_s,$

$$R_r(f) \leq R_r(\vec{0}) = \frac{1}{m}\sum_{i=1}^{m}\ell(\vec{0}, z_i) \leq B(0),$$

And also $R_r(f) \ge \lambda \| f \|_k^2$ which gives the first inequality .the second inequality follows from(3) .the last one is a consequence of the definition of $\sigma$ - admissibility.

**Theorem (3):[1 ]**

Let $\{(x_1 y_1), (x_2, y_2) \dots (x_m, y_m)\}$ be a sequence of labeled examples with ||xi|| ≤ R. Let $\theta$ be any vector with $\|\theta\| = 1$ and let $\gamma > 0$. Define the deviation of each example as

$d_i$ = max $(0, \gamma - y_i \langle \theta, x_i \rangle)$, and let D = $\sqrt{\sum_i d_i^2}$. Then the number of mistakes of the perception algorithm on this sequence is bounded by $(\frac{R+D}{\gamma})^2$.

**Theorem (4): [8 ]**

Let A be an algorithm with uniform stability $\beta$ with respect to a loss function $\ell$ such that $0 \le \ell(A_S, z) \le M$, for all $z \in Z$ and all sets S .then ,for any $m \ge 1$, and any $\delta \in (0,1)$, , the following bounds hold with probability at least $1-\delta$ over the random draw of the sample S,

$$R \le R_{emp} + 2\beta + (4m\beta + M)\sqrt{\frac{\ln 1/\delta}{2m}}, \qquad \text{…….(6)}$$

And

$$R \le R_{loo} + \beta + (4m\beta + M)\sqrt{\frac{\ln 1/\delta}{2m}}. \qquad \text{………(7)}$$

**Example** (1);[8 ]

stability of bounded SVM regression

Assume k is a bounded kernel, that is $k(x,x) \le \kappa^2$ and Y=[0,B].

Consider the loss function

$$\ell(f,z) = |f(x) - y|_{\varepsilon} = \left\{ \begin{array}{ll} \text{o} & \text{if } |\text{f(x) - y}| \le \varepsilon \\ & \\ |\text{f(x) - y}| - \varepsilon & otherwise \end{array} \right\}$$

This function is 1-admissible and we can state B(y)=B .the SVM algorithm for regression with a kernel k can be defined as And we thus get the following stability bound

$$\beta \le \frac{\sigma^2 \kappa^2}{2\lambda m}.$$

$A_S = \arg\min_{g \in F} \frac{1}{m} \sum_{i=1}^{m} \ell(g, z_i) + \lambda \|g\|_k^2$, Moreover, by theorem (2) we have

$$\forall z \in Z, 0 \le \ell(A_S, z) \le \kappa \sqrt{\frac{B}{\lambda}}$$

Plugging the above into theorem (4) gives the following bound

$$R \le R_{emp} + \frac{\kappa^2}{\lambda m} + \left(\frac{2\kappa^2}{\lambda} + \kappa\sqrt{\frac{B}{\lambda}}\right)\sqrt{\frac{\ln 1/\delta}{2m}} \qquad \text{…….(8)}$$

Such that $R_{emp}$ is the simplest estimator (empirical error).

**Example**(2):[11]

Absolute stability of control system

Suppose system of differential equation

$$\dot{x} = Ax \ ,$$

………(9)

Is represent the free motion to body .such that $A = (a_{ij})$ matrix of order $n \times n$ ,constant and stable matrix and non-singular.

Let x is vector (column) in $R^n$ and let system of control arbitress by the following equation

$$\dot{x} = Ax - ub$$
$$\dot{u} = f(Z) \qquad \text{……… (10)}$$
$$Z = c^T x - pu$$

Such that c, b are constant vectors in $R^n$ and $c^T$ is transpose to vector c . p and u are elements in $R$ and f is element in admissible characteristic function ʒ, then system (10) is absolutely stable if for all solution ( x ,u) to system(10) is $(x(t), u(t)) \to (0,0)$ when t is converge to $\infty^+$.the variable of control of system (10) is u ,let us take transformation

$$\dot{x} = y$$
$$z = c^T x - pu$$

That trans the system (10) to the new formal following ;

$$\dot{y} = Ay - bf(z)$$
$$\dot{z} = x^T y - pf(z) \qquad \text{…… (11)}$$

Its keep on the estate stability ,we must that transformation non-singular ,that is

$$\begin{vmatrix} A & -b \\ c^T & -p \end{vmatrix} \neq 0 \ , \quad \text{then} \quad p = c^T A^{-1} b$$

So the original point is unique control point of system (11) .

**Example** (3):[10]

stability of regularized least squares regression

We will consider the bounded case Y=[0,B] .the regularized least squares regression algorithm is defined by

$$A_S = \arg\min_{g \in F} \frac{1}{m} \sum_{i=1}^{m} \ell(g, z_i) + \lambda \|g\|_k^2 ,$$

Where $\ell(f, z) = (f(x) - y)^2$.

We can state $B(y) = B^2$ so that is 2B-admissible by theorem (2) .also we have

$$\forall z \in Z, 0 \le \ell(A_S, z) \le \kappa \sqrt{\frac{B}{\lambda}} \ .$$

The stability bound for this algorithm is thus

$$\beta \le \frac{2B^2 \kappa^2}{\lambda m}$$

So that we have the generalization error bound

$$R \le R_{emp} + \frac{4B^2 \kappa^2}{\lambda m} + \ (\frac{8\kappa^2 B^2}{\lambda} + 2B)\sqrt{\frac{\ln 1/\delta}{2m}} . \qquad \text{……..(12)}$$

Remark

The function $f$ is said to be element in admissible characteristic function ʒ, if

أطلب العلم من المهد الى اللحد
كلية التربية الأساسية ـ جامعة بابل
١٩٩٤م
١٤١٣هـ