A New Investigation about the Artificial Intelligence in Reproducing Kernel Hilbert Spaces: An Analytical Study

Nahid Khanlari †, Shima Asghari ††
Department of Applied Mathematics, Islamic Azad University, Hamedan Branch, Hamedan 65138, Iran

Abstract—The current paper aims to compute the artificial intelligence for regression in bounded subspaces of Reproducing Kernel Hilbert Spaces (RKHS) for the Support Vector Machine (SVM) regression. Both $\varepsilon$-insensitive loss function and general $L^p$ loss functions are studied. It is shown that the artificial intelligence is finiteness. This, in turn, confirms that the probability for regression machines in RKHS subspaces using the $L_\varepsilon$ or general $L^p$ loss functions is uniformly converged. Further, the results are verified in a new fashion in the case of introducing a bias to the functions in the RKHS.

Index Terms—Artificial Intelligence, Applied Mathematics, Reproducing Kernel Hilbert Spaces, Regression, Support Vector Machines, Regularization Networks

1 Introduction

The artificial intelligence of real-valued functions
\[ L(y - f(x)) = |y - f(x)|^p \]
and
\[ L(y - f(x)) = |y - f(x)|_f \]
with $f$ in a bounded sphere in a Reproducing Kernel Hilbert Space (RKHS), is computed in the current paper. Considering these loss functions, it is shown that the artificial intelligence is finite and then, an upper bound is computed for the dimension. The problem is solved by two solutions. A discussion on a simple argument, leading to a loose upper bound on the artificial intelligence, is introduced.

It is previously confirmed that when $L$ is used as loss function in a regression learning problem, a necessary and sufficient condition for uniform convergence in probability is finiteness of the artificial intelligence for all $\gamma > 0$ [20]. Accordingly, the results of the current paper confirm the uniform convergence of both RN and SVM regressions [21, 23, 24, and 27].

The problem of pattern recognition, with $L$ works as an indicator function, was considered in...
previous related works [21-39]. However, the fat-
shattering dimension [40] was introduced as the
substitute of the artificial intelligence [41-45]. By
presenting entropy numbers of operators as cover
of number arguments, a different approach is
followed to confirm uniform convergence for RN
and SVM [46-62]. However, regression as well as
the case of non-zero bias $b$ was not
comprehensively considered in both of them [63-
73].

According to the framework of statistical
learning theory, the problem of learning from
examples is taken into account in this study [74].
By randomly sampling from a space $XY \times X Y$ \includegraphics[width=1\textwidth]{XY.png}
with $X \subseteq R^d$, $Y \subseteq R$ a set of $\ell$
examples $(x_i, y_i), ..., (x_\ell, y_\ell)$ is generated
[75]. It is based on an unknown probability
distribution $P(x, y)$ [76]. It is assumed that $X$
and $Y$ are bounded [77]. The problem of learning is
defined as finding a function $f : X \rightarrow Y$, in
accordance to a set of examples, through it, the
value $y$, corresponded to new point $x \in X$, can
be predicted [78].

It is well-known that the problem of learning
from examples is ill-posed [79, 80]. Performing
Empirical Risk Minimization (ERM), using a
specified loss functions, and with limiting the
solution to the problem to be in a “small”
hypothesis space [80] is the tradition solution of
the problem. The goal of the solution is minimizing
the empirical risk

$$I_{emp}[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i)) \text{ with } f \in H,$$

where $L$ is the loss function measuring the error
as the difference of predicted, $f(x)$ and
actual, $y$, values and $H$ is a given hypothesis
space [81].

The hypothesis spaces of functions considered
in the current study are hyperplanes in some
feature space:

$$f(x) = \sum_{n=1}^{\infty} \alpha_n \phi_n(x) \tag{1}$$

with:

$$\sum_{n=1}^{\infty} \frac{\alpha_n^2}{\lambda_n} < \infty \tag{2}$$

where $\phi_n(x)$ is a set of given, linearly
independent basis functions, $\lambda_n$ are given non-
negative constants such that $\sum_{n=1}^{\infty} \lambda_n^2 < \infty$. The form
of functions’ spaces represented in Eq. (1) is similar
to which is used in Reproducing Kernel Hilbert
Spaces (RKHS) [81, 82] with kernel $K$
given by:

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(y) \tag{3}$$

Eq. (2) gives the RKHS norm of $f$, $\|f\|_K$
where $f$ is defined according to Eq. (1). However, the
number $D$ of features $\phi_n$ (if $D$ is finite, all sums
above are also finite) is the dimensionality of the
RKHS [83].

By limiting the hypothesis space so that it
consists of functions in a RKHS with norm less
than a constant $A$, the general setting of above
mentioned learning becomes:

$$\text{Minimize : } \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i)) \text{ subject to } \|f\|_K \leq A^2 \tag{4}$$

The consistency of learning machines defined
by Eq. (4) is a critical issue. It was found that by
approaching the number of examples $(x_i, y_i)$
to infinity, the probable error of the solution should
converge in probability to the minimum expected
error in the hypothesis space [80, 83]. For learning
machines performing ERM in a hypothesis space
Eq. (4), it was shown that consistency is related to
uniform convergence in probability [84]. In
addition, depending on the artificial intelligence of
the considered hypothesis space, which indicates
the complexity of the space, necessary and
sufficient conditions for uniform convergence are
defined [84, 85].

The VC-dimension is typically used in statistical
learning theory as the measure of complexity.
However, when RKHS is dimensionally infinite,
the VC-dimension also is infinite both for $L_p$ and $L_\infty$, in the above learning setting. As a result, the VC-dimension is not applicable to investigate learning machines of the form Eq. (4). In this regard, another measure of complexity (e.g., the artificial intelligence) should be considered to demonstrate uniform convergence in infinite dimensional RKHS.

2 Results and Discussion

In the following, it is assumed that data $X$ are within a sphere of radius $R$ in the feature space defined by the kernel $K$ of the RKHS and $y$ is bounded between $-1$ and $1$. As a result of these assumptions, a theorem can be described as follows:

**Theorem.** The artificial intelligence $h$ for regression, considering $L_p$ or $L_\infty$ as loss functions, for hypothesis spaces

$$H_A = \left\{ f(x) = \sum_{n=1}^{\infty} \omega_n \phi_n(x) \left| \sum_{n=1}^{\infty} \omega_n^2 \lambda_n \leq A^2 \right. \right\}$$

and $y$ bounded, is finite for $\forall \gamma > 0$. If $D$ is the dimensionality of the RKHS,

$$h \leq O \left( \min \left( D, \frac{R^2 + 1}{\gamma^2} \left( A^2 + 1 \right) \right) \right)$$

then

**Proof.** Considering $L_1$ as loss function and $B$ as the upper bound of $L_1$, the rules for separating points can be decomposed according to the following:

**class 1** if $y_i - f_1(x_i) \geq s + \gamma$

or $f(x_i) - y_i \geq s + \gamma$

**class -1** if $y_i - f_1(x_i) \leq s - \gamma$

or $f(x_i) - y_i \leq s - \gamma$

(5)

for some $\gamma \leq s \leq B - \gamma$. It should be noted that, despite the number of $N$ points, the number of separations possible to get by rules Eq. (5) cannot more than the number of separations possible to get by the product of two indicator functions (of hyperplanes with margin):

$${\text{function (a): class 1}} \quad \text{if} \quad y_i - f_1(x_i) \geq s_1 + \gamma$$

$${\text{class -1}} \quad \text{if} \quad y_i - f_1(x_i) \leq s_1 - \gamma$$

$${\text{function (b): class 1}} \quad \text{if} \quad f_2(x_i) - y_i \geq s_2 + \gamma$$

$${\text{class -1}} \quad \text{if} \quad f_2(x_i) - y_i \leq s_2 - \gamma$$

(6)

where $f_1$ and $f_2$ are in $H_A$, $\gamma \leq s_1$, $s_2 \leq B - \gamma$. Recovering Eq. (5), for $s_1 = s_2 = s$ and for $f_1 = f_2 = f$, gives the results same as to what will be obtained if one follows Eq. (5). For example, if $y - f(x) \geq s + \gamma$ then indicator function (a) will result $-1$ and indicator function (b) will result $-1$. Hence, their product will result $+1$, same as Eq. (5). Therefore, it can be obviously seen that the number of separations for any considered set of points can be considerably increased if more freedom give to $f_1$, $f_2$, $s_1$, $s_2$ compared to when Eq. (5) is followed.

It is previously mentioned that the number of separations, for any $N$ points, is bounded by the growth function. However, it was shown that the growth function for products of indicator functions is enclosed by the product of the growth functions of the indicator functions. Moreover, the indicator functions in Eq. (6) are hyperplanes and its margin is in the $D + 1$ dimensional space of vectors $\{\phi_n(x), y\}$ where $R^2 + 1$ is the radius of the data, the norm of the hyperplane is bounded by $A^2 + 1$, (where 1 added in both cases due to $\gamma$), and the margin is at least $\frac{\gamma^2}{A^2 + 1}$. It was previously found that the artificial intelligence $h_\gamma$ of these hyperplanes is bounded

$$h_\gamma \leq \min \left( (D + 1) + 1, \frac{R^2 + 1}{\gamma^2} \left( A^2 + 1 \right) \right)$$

by
Therefore, whenever $h \geq h_\gamma$, the growth function of the separating rules Eq. (5) is bounded by

$$g(\ell) \leq \left( \frac{e \ell}{h_\gamma} \right)^{h_\gamma}.$$  

Considering $h_\gamma^{reg}$ as the artificial intelligence, $h_\gamma^{reg}$ is limited to be smaller than the larger number $\ell$ for which

$$2^\ell \leq \left( \frac{e \ell}{h_\gamma} \right)^{h_\gamma}$$

inequality holds. In this regard, as $h \leq 5h_\gamma$, therefore

$$h_\gamma^{reg} \leq 5 \min \left( \frac{D + 2 \left( R^2 + 1 \right) \left( A^2 + 1 \right)}{\gamma^2} \right).$$

It is proved the theorem for the case of $L_1$ loss functions.

By rewriting Eq. (5) as follows, a same proof can be achieved for general $L_p$ loss functions:

**class 1** if $y_i - f(x_i) \geq (s + \gamma)^{\frac{1}{p}}$

or $f(x_i) - y_i \geq (s + \gamma)^{\frac{1}{p}}$

**class -1** if $y_i - f(x_i) \leq (s - \gamma)^{\frac{1}{p}}$

or $f(x_i) - y_i \leq (s - \gamma)^{\frac{1}{p}}$ (7)

Moreover, for $p > 1$,

$$\left( s + \gamma \right)^{\frac{1}{p}} \geq s^{\frac{1}{p}} + \frac{\gamma}{pB}$$

(\text{since } \gamma = \left( s + \gamma \right)^{\frac{1}{p}} - \left( s^{\frac{1}{p}} \right)^{\frac{1}{p}} \leq \left( s + \gamma \right)^{\frac{1}{p}} - s^{\frac{1}{p}} (pB))

$$\left( s - \gamma \right)^{\frac{1}{p}} \leq s^{\frac{1}{p}} - \frac{\gamma}{pB}$$  \text{(similarly). Similar to above argument, it can be found that the artificial intelligence is bounded by \[5 \min \left( \frac{D + 2 \left( pB \right)^2 \left( R^2 + 1 \right) \left( A^2 + 1 \right)}{\gamma^2} \right).\]

Finally, Eq. (5) can be rewritten as follows for the $L_\varepsilon$ loss function:

**class 1** if $y_i - f(x_i) \geq s + \gamma + \varepsilon$

or $f(x_i) - y_i \geq s + \gamma + \varepsilon$

**class -1** if $y_i - f(x_i) \leq s - \gamma + \varepsilon$

or $f(x_i) - y_i \leq s - \gamma + \varepsilon$ (8)

where calling $s' = s + \varepsilon$, the above mentioned proof can be used to find the upper bound on the artificial intelligence same as to that found for the $L_1$ loss function. (It should be noted that if the constraint $\gamma \leq s \leq B - \gamma$ is considered, it seems that it would have a little effect on the artificial intelligence for $L_\varepsilon$).

It can be concluded from these results that the artificial intelligence is still finite and is influenced only by

$$5 \min \left( \frac{D + 2 \left( R^2 + 1 \right) \left( A^2 + 1 \right)}{\gamma^2} \right)$$

when RKHS is dimensionally infinite.

3 Conclusions

A novel approach is introduced in the current paper to compute the artificial intelligence of RKHS when $L_p$ and $L_\varepsilon$ are considered as loss functions. It is found that better bounds can be achieved if $\varepsilon$ takes into account in the computations when $L_\varepsilon$ considered as loss function. As an instance, it is clearly proved that the artificial intelligence is bounded by

$$\frac{p^2 \left( B - \varepsilon \right)^2 \left( R^2 A^2 \right)}{\gamma^2},$$

when $\|f(x) - y\|_\varepsilon, p > 1$ considered as the loss function. However, it is found that $\varepsilon$ has a low influence (given that $\varepsilon \ll B$). Moreover, more
general loss functions can be introduced to the presented computations. Appearing the eigenvalues of the matrix $G$ in the computation of the artificial intelligence is very interested. By computing the number of separations for a given set of points, similar to that performed for the largest and smallest eigenvalues in the proofs, all the eigenvalues of $G$ can be considered. As a result of this computation, interesting relations can be found. In addition, for obtaining the bounds on the generalization performance of regression machines of the form Eq. (4), the bounds on the artificial intelligence can be effectively used.

Acknowledgment

The authors would like to thank Dr. Taher Lotfi for help and valuable discussions.

References

[65] E. Miña-Diaz, Asymptotics of polynomials orthogonal over the unit disk with respect to a polynomial weight without zeros on the unit circle, Journal of Approximation Theory, Volume 165, Issue 1, January 2013, Pages 41-59.