A Neural Model Approach for Regularization in the Mean Estimation Case

Sergio Decherchi, *Student Member, IEEE,* Mauro Parodi, *Member, IEEE*, and Sandro Ridella, *Member, IEEE*.

***Abstract*— Neural Networks are powerful tools for function approximation problems. A possible peculiar application of neural networks is that proposed here: estimating the univariate mean of a distribution from a finite sample. This problem characterizes a huge number of applicative and scientific problems. The Gaussian distribution case is analyzed, however the proposed analysis is of general validity and can be easily extended to other distributions. In particular the estimation problem is approached as a regularization problem and a solution to the selection of the regularization parameter is obtained via the employment of neural models. The paper, after introducing some theoretical results, presents two neural models, namely a MLP and a Circular Back Propagation Network, for the mean prediction. Experimental results show that neural networks can estimate the mean, in expectation, better than the usual sample mean formula.**

I. Introduction

Neural networks are widely used tools for a large variety of problems concerning function approximations [1]. Their universal approximation properties [1-3] allow one to cope with variegate inference problems. This work addresses the problem of the mean estimation and proposes two suitable neural models to that aim.

The mean estimation problem is typical not only in the statistical literature but it is pervasive in a vast number of applications (here for mean estimation is meant estimation of the mean value from finite samples and not time series prediction of an average value of a phenomenon as for instance in [11]). The seminal work of James and Stein [4] has showed that the apparently easy mean estimation problem is far from a trivial one.

The usual sample mean estimator response can be improved through a procedure which is a form of regularization. Finding the value of the regularization parameter for this procedure is not a minor problem. It will be shown how some traditional techniques, such as Generalized Cross Validation [5] or the Bayesian framework of Maximal Evidence for Neural Networks [6] have some important limitations. These limits can be overcome by using neural models such as Multilayer Perceptrons (MLP) and Circular Back Propagation networks (CBP) [7].

The studies on regularization problems, and function approximations, are of fundamental importance in learning theory [1-3]. Their importance originates from the observation that kernel methods such as Support Vector Machines, Kernel Logistic Regression, and Regularized Least Squares are all instances of regularization problems. To this regard the problem of obtaining an estimate of the 'optimal' regularization parameter can be considered a still open problem, particularly in small sample problems where cross validation can be impractical. The regularized mean problem [8] is a simple and controlled environment where new techniques of estimation of the regularization parameter can be experimented; in particular here two neural models are proposed.

The rationale behind this model choice is that the back propagation algorithm, together with a Montecarlo approach, allows to build effective 'neural' predictors for the mean value, where closed form formulas for the regularization parameters are not effective. The actual results of the learning stage is a 'table' of weights that can be used together with its associated neural model to predict the mean value. In this way the involved neural networks act as predictors not only for the mean value, but also, implicitly, for the regularization parameter. This is the reason why neural networks induce a 'neural' regularizer with respect to the regularized mean problem.

Some remarkable works of the machine learning community have recently been devoted to analyze the relations between regularization, kernel methods and estimation of statistical quantities [9],[10],[8]. In particular, [9] studies the possibility of embedding probability distributions in a Kernel Hilbert Space and then estimating Maximal Mean Discrepancy in that space; in [10] and [8] some oracle properties of regularization methods, the connection between shrinking methods (e.g. James-Stein estimation) and regularization, and the limits of a regularization based approach for mean estimation, are presented and studied. This work belongs to the line of research of these previous cited papers.

In the following the one-dimensional case of the mean is analyzed. Two neural models are built and compared: the first is a usual two layer MLP network, the second is a circular back propagation network [7]. The effectiveness of the models in predicting the mean value is showed via experimental results.

This work was partially supported by the University of Genoa under Grant "Academic Research Projects Area 09, 2008"

The authors are with University of Genoa, Italy, Department of Biophysical and Electronic Engineering, Via Opera Pia 11/A. (e-mail {sergio.decherchi, mauro.parodi, sandro.ridella} @unige.it)

II. The Regularized Mean Problem

A. Regularized Mean

Consider m samples $x_i \sim N(\mu,\sigma^2)$ as the elements of a set X. Based on these samples define the functional:

$$\Im(\xi;X,\alpha) \equiv \sum_{i=1}^{m}(x_i-\xi)^2 + \alpha\xi^2 \qquad (1)$$

where $\alpha \geq 0$ is a regularization parameter. When $\alpha = 0$ the minimum of $\Im(\xi;X,0)$ with respect to ξ leads to the usual sample mean estimation. Conversely , when $\alpha > 0$ a regularized mean value is obtained.

For given X and α define the regularized mean value $\bar{x}$ as:

$$\bar{x}(\alpha) \equiv \arg\min_{\xi} \Im(\xi;X,\alpha) \qquad (2)$$

Whose explicit expression is:

$$\bar{x}(\alpha) = \frac{\sum_{i=1}^{m} x_i}{m+\alpha} \qquad (3)$$

Expression (3) can be rewritten as:

$$\bar{x} = \lambda \bar{x}_0 \qquad (4.1)$$

where:

$$\lambda \equiv \frac{m}{m+\alpha} \qquad (4.2)$$

The quantity λ is useful in view of establishing an analogy with the work of James-Stein [4][8].

Following [4], a functional defining the quality of the estimator $\bar{x}$ with respect to μ is:

$$L(\lambda;X,\mu) \equiv \mathrm{E}_X\left\{(\bar{x}-\mu)^2\right\} \qquad (5)$$

The value of λ minimizing (5) is the oracular value:

$$\lambda^{orac} = \arg\min_{\lambda} L(\lambda;X,\mu) = \frac{\mu^2}{\mu^2 + \frac{\sigma^2}{m}} \qquad (6)$$

The term "oracular" derives from the fact that μ itself should be known.

Figure 1 shows an example of the advantage gained estimating μ as $\bar{x} = \lambda^{orac}\bar{x}_0$ instead of the non regularized value $\bar{x} = \bar{x}_0$. On the horizontal axis, μ ranges in the interval [-1,+1]; on the vertical axis, the gain $y = \mathrm{E}_X\left\{(\bar{x}_0-\mu)^2\right\} - \mathrm{E}_X\left\{(\lambda^{orac}\bar{x}_0-\mu)^2\right\}$ obtainable by oracular regularization, is reported using $\sigma^2 = 1$, $m = 10$.

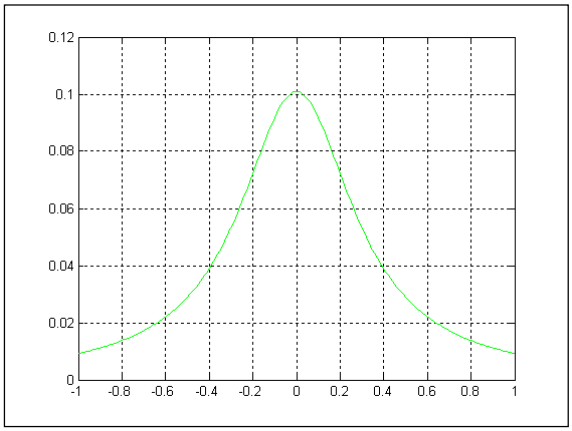


Figure 1. Gain of oracular regularized solutions against non regularized: x axis is μ range, y axis is the quality metric eq.(15)

B. Links with James-Stein Theory

Stein's theory [4] deals with the so called d-means problem that can be outlined as follows:

- define d unknown parameters, the means vector $\boldsymbol{\mu}$: correspondingly define d Gaussian probability density functions that share the same variance σ^2
- pick one sample for each p.d.f. and denote by X this samples vector.

Then the goal is to estimate the $\boldsymbol{\mu}$ vector as:

$$\bar{X} = \arg\min_{\xi} \mathrm{E}_X\left\{\| \boldsymbol{\xi} - \boldsymbol{\mu} \|_2^2\right\} \qquad (7)$$

In [4] it is shown that for $d \geq 3$ the estimator $\bar{X} = X$ is not the best possible estimator of $\boldsymbol{\mu}$ for the loss in (7). In [4] it is also proved that for $d \geq 3$ the best estimator for the vector $\boldsymbol{\mu}$ is:

$$\bar{X} = \left(1 - \frac{(d-2)\sigma^2}{\| X \|^2}\right) X \qquad (8)$$

This counterintuitive result can be interpreted as a regularized solution in which the coefficient $\left(1 - \frac{(d-2)\sigma^2}{\| X \|^2}\right)$ plays the role of λ defined in (4.2).

Another important consequence of the results in [4] is that, for $n=1$, the non regularized estimator $\bar{X} = X$ is the best for every value of $\boldsymbol{\mu}$ with respect to (7).

Summarizing the discussion and results obtained in [8], on the basis of Stein theory, one has that:

- The regularized mean problem can be linked to Stein theory.
- When μ ranges over $(-\infty,+\infty)$, a $\bar{x}_0$-based estimator cannot estimate μ better than $\bar{x}_0$ itself.
- For a limited range of μ one can address the problem to obtain, from a set of m samples, an estimator for μ better than $\bar{x}_0$. In particular the situations of major interest occur when both the

number of samples *m* and the μ^2/σ^2 are small.

C. Limits of Regularization and the Need of Viable Alternatives

This section develops some results that set the conceptual basis for the rest of the work: in particular, theorem 1 is the theoretical premise for the neural approach definition. An important aspect concerns the study of effectiveness of regularization, on the entire space of X and μ, when one has no prior knowledge on the distribution of μ. Such a problem leads to analyze the asymptotic behavior:

$$\mathrm{E}_\mu \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} \tag{9}$$

where the dependence of λ on the sample X has been evidenced. In particular one has to inquiry on the relationship between (9) and the non regularized cost $\mathrm{E}_\mu \mathrm{E}_X \left\{ (\bar{x}_0 - \mu)^2 \right\} = \sigma^2/m$; this analysis aims to show the limits of the classical approach and the need of an alternative model.

A possible way to study (9) is to assume a uniform prior $p(\mu)$ over a range $[-\gamma, \gamma]$ and then taking the limit $\gamma \to \infty$ (i.e. an improper prior): the convention E_μ^γ will be used to indicate an expectation integral computed in the interval $[-\gamma, \gamma]$. Formally one has to study the following multiple integral:

$$\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} =$$
$$= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \int_{-\gamma}^{+\gamma} \left\{ [\lambda(X)\bar{x}_0 - \mu]^2 \frac{1}{2\gamma} \frac{1}{(\sigma\sqrt{2\pi})^m} \right. \tag{10}$$
$$\left. \prod_{i=1}^{m} \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2} \right] \right\} d\mu \, dx_1 \ldots dx_m$$

The following result [8] clarifies a fundamental aspect on (10)

Theorem 1. (No Free Lunch for the Regularized Mean) *For any* $\lambda(X) < 1$, *and assuming the improper prior* $p(\gamma) = 1/(2\gamma)$, *the term* $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\}$ *fulfils the following properties*:

a) *if* $\gamma \to \infty$ *then* $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m \to 0$ *from positive values.*

b) *if* $\gamma \to 0$ *then* $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m < 0$.

c) *It exists at least a value* $\hat{\gamma}$ *such that*

1. *For* $\gamma < \hat{\gamma}$, $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m < 0$
2. *In* $\gamma = \hat{\gamma}$, $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m = 0$
3. *For* $\gamma > \hat{\gamma}$, $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m > 0$

The above theorem, among the various results, explicitly states that, for any sample based regularization strategy $\lambda(X) < 1$, and for $\gamma \to \infty$ the term $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda(X)\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m \to 0$ which amounts to say that regularization, or shrinking, is not effective. Similarly for $\lambda = \lambda^{orac}$ one finds the following result [8]:

Lemma 1. *For any* γ, *the inequality* $\mathrm{E}_\mu^\gamma \mathrm{E}_X \left\{ (\lambda^{orac}\bar{x}_0 - \mu)^2 \right\} \le \sigma^2/m$ *always holds true. For* $\gamma \to \infty$ *the term* $\mathrm{E}_\mu \mathrm{E}_X \left\{ (\lambda^{orac}\bar{x}_0 - \mu)^2 \right\} - \sigma^2/m$ *converges to 0 from negative values*.

This result confirms that the oracular regularizer (that is not sample based) is always useful when γ is finite.

In order to obtain a regularization strategy feasible for a finite γ, in [8] different approaches are studied: Leave one out (or generalized cross validation) [5], maximal evidence [6], and various attempts apt to approximate the oracular regularizer. Table 1 shows the closed form formulas obtained in [8]; there S^2 stands for $\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x}_0)^2$ and $\beta \equiv \max(|x_1|, |x_2|, \ldots, |x_m|)$. A sketch of proof of Maximal Evidence and Leave One Out is recorded in Appendix.

All of table I regularizers are quite effective but in a narrow range of γ. This motivates the study of more efficient regularization strategies. Here, two different neural models are considered to predict the mean: the circular back propagation network [7], and a classical MLP.

III. Neural Models for the Mean

A. MLP and Circular Back Propagation Networks

Multi-layers perceptrons are among the most used and successful neural networks models. In particular one popular configuration is that with only one hidden layer.

Given an input vector **x** of dimension *d* and weights **w**, at the unit level, the MLP (before the sigmoid function) is governed by the following equation:

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i}^{d} w_i x_i \tag{11}$$

Expression (11) is a simple linear model of the input data. Representation ability of MLP derives from the non linear combination, due to sigmoids, and their combination in the subsequent and final layer. The conceptual point of MLP is that the intervention of non linear processing is concentrated on the second layer.

Circular back Propagation Networks improves on this model by augmenting the original space with one more input variable that already on the first layer endows a non linear

processing: this variable is $\sum_i^d x_i^2$. The practical results is that the unit function becomes:

$$f(\mathbf{x},\mathbf{w}) = w_0 + \sum_i^d w_i x_i + w_{d+1}\sum_i^d x_i^2 \qquad (12)$$

Where to the anti-symmetric MLP component is added a symmetric circular component.
In [7] it is shown that augmenting the space by this new added input gives the resulting CBP networks a greatly improved representation ability; at the meantime one has the crucial advantage to re-use back propagation algorithmic machinery.
In the following these two neural structures, namely a MLP and a CBP-like network, will be used as mean value predictors and their performances will be compared.
The problem of data over-fitting is not considered because, in the current case, this problem is absent: the entire population is presented to the network, thus weight decay or other regularization strategies for neural networks are not necessary.

B. Neural Setup

In order to define a neural model for the expected value μ, some assumptions and definitions are introduced:

- The value of σ and the range $\mu \in [-\mu_M, +\mu_M]$ where the solution lies are known.
- A single hidden layer neural network is used.
- The sample X is distributed as $N(\mu, \sigma^2)$.
- Denoting by g the NN output function, the adopted prediction model is:

$$g\mu_M + \bar{x}_0 = \bar{x} \qquad (13)$$

  Thus the NN has to predict the offset by which $\bar{x}_0$ should be modified to improve the estimation of μ
- Set as target values the normalized offset $\frac{\mu - \bar{x}_0}{\mu_M}$
- Set as instantaneous cost function

$$L(\bar{x}_0, \mu_M, \mu) = \left(\frac{\mu - \bar{x}_0}{\mu_M} - g\right)^2 \qquad (14)$$

The final goal is trying to minimize $\mathrm{E}_X\left\{(\mu - \bar{x})^2\right\}$ by using (14) in a predefined range of μ. This amounts to say that the following gap must be maximized:

$$y \equiv \mathrm{E}_X\left\{(\bar{x}_0 - \mu)^2\right\} - \mathrm{E}_X\left\{(\lambda\bar{x}_0 - \mu)^2\right\} \qquad (15)$$

Since the first non regularized term on the r.h.s. is constant, the minimization process is confined to the second regularized term. This aspect can be clarified considering

TABLE I. Closed form rational regularizers.

Regularization Strategy	Regularizer
Leave One Out (LOO) / Maximal Evidence	$\lambda^{loo} = \frac{\bar{x}_0^2 - \frac{S^2}{m}}{\bar{x}_0^2}$
Threshold LOO	$\lambda_{ht}^{loo} = \max\left(0, \frac{\bar{x}_0^2 - \frac{S^2}{m}}{\bar{x}_0^2}\right)$
Stein	$\lambda^{stein} = \frac{\bar{x}_0^2 + \frac{\sigma^2}{m}}{\bar{x}_0^2}$
Linear Oracular	$\lambda_{\bar{x}_0^2}^{orac} = \frac{\bar{x}_0^2}{\bar{x}_0^2 + \frac{\sigma^2}{m}}$
Non Linear LOO	$\lambda_\beta^{loo} = \frac{\beta^2 - \frac{S^2}{m}}{\beta^2}$
Non Linear Oracular	$\lambda_\beta^{orac} = \frac{\beta^2}{\beta^2 + \frac{\sigma^2}{m}}$
Non Linear Oracular S^2	$\hat{\lambda}_\beta^{orac} = \frac{\beta^2}{\beta^2 + \frac{S^2}{m}}$

that the expected neural cost is:

$$\begin{aligned}
&\mathrm{E}_X\{L(\bar{x}_0, \mu_M, \mu)\} = \mathrm{E}_X\left\{\left(\frac{\mu - \bar{x}_0}{\mu_M} - g\right)^2\right\} = \\
&= \frac{1}{\mu_M^2}\mathrm{E}_X\left\{(\mu - (g\mu_M + \bar{x}_0))^2\right\} = \\
&= \frac{1}{\mu_M^2}\mathrm{E}_X\left\{(\mu - \bar{x})^2\right\} =
\end{aligned} \qquad (16)$$

C. Network Configurations

The minimization process of (14) can be accomplished by a classical back propagation algorithm. Moreover, due to the number of involved patterns, an on-line version of back-propagation has been used.
The proposed network has two possible configurations: in

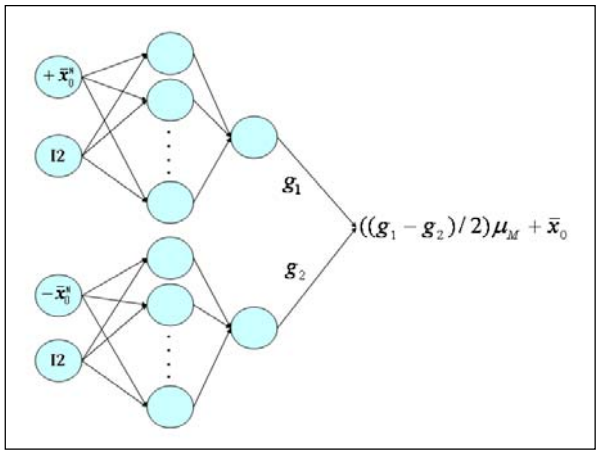


Figure 2. De-Biased Neural Predictor Scheme for the mean where: $\bar{x}_0^{\mathrm{N}} = \bar{x}_0 / \mu_M$ and I2 denotes the second input, whose value depends on the NN configuration (see text).

the first case the inputs are $(\bar{x}_0 / \mu_M, \beta^2 / \mu_M^2)$; in the circular case the inputs are $(\bar{x}_0 / \mu_M, \sum_{i=1}^{m} x_i^2 / \mu_M^2)$. In both cases it can happen that the expected symmetry of y as per eq.(15) around $\mu = 0$ is not exactly obtained. In order to avoid this phenomenon a balancing structure for the prediction has been employed (see figure 2). In this structure, two identical NNs are requested to predict on the predefined input pairs $(\bar{x}_0 / \mu_M, I2)$ and $(-\bar{x}_0 / \mu_M, I2)$. These two networks produce the respective predictions g_1 and g_2; the final $\bar{x}$ is computed as $((g_1 - g_2)/2)\mu_M + \bar{x}_0$. This expedient grants symmetric expectations over a wide range of μ.

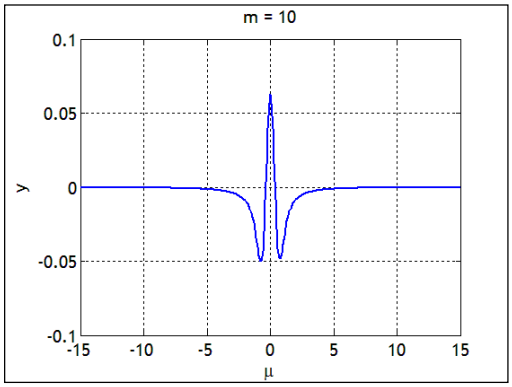
The value of the term *I2* is β^2 / μ_M^2 for both the inputs in fig.2 for the MLP (the first configuration) and $\sum_{i=1}^{m} x_i^2 / \mu_M^2$ for both the inputs in fig.2 for the CBP-like (the second configuration).

IV. Experimental Results

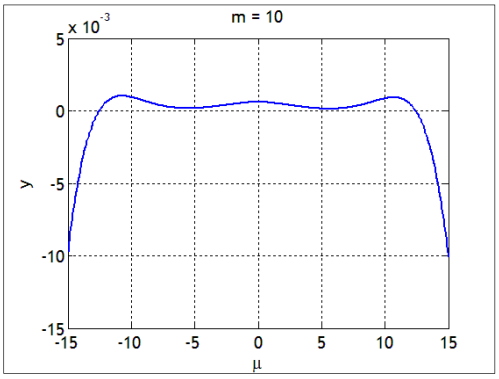
A. Experimental Setup

The values of the Gaussian parameters chosen for the experiments were a range for $\mu \in [-15,+15]$ by steps of 0.1 $\sigma^2 = 1$, and varying number of samples $m = [2,14]$. These values were used to evaluate the regularizers in Table I and the two neural models. The parameters of the neural networks were chosen as follows:

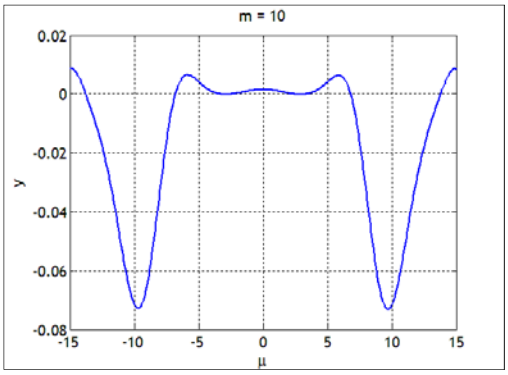
1. The number of patterns for each value of μ in the range $\mu \in [-15,+15]$ was 10^6 for training and 10^5 for test.



(a)



(b)



(c)

Figure 3. Results obtained by: (a) LOO/ME regularizer, (b) MLP, (c) CBP: *x* axis is μ, *y* axis is the gap eq.(15).

1. The learning rate was set to 10^{-4}.
2. The number of hidden neurons was 30 (this value follows from a preliminary model selection, not reported here for brevity).

Figure 3 (a,b,c) reports on the obtained results for $m = 10$ in the case of Maximal Evidence/ LOO, MLP, and CBP using (15) as quality metric; for the other m values the obtained results are qualitatively analogous.

A Matlab version of the software for the neural predictors,

which loads pre-computed weights and predicts the regularized mean value, has been implemented and is available under request.

B. Summarizing Comments

The obtained estimations of the mean value can be profitably compared with that obtained by rational regularizers [8].
In general, the rational regularizers in table I show a very regular behavior in terms of expectation [8]: in all the cases, a well defined range of μ exists where regularization is effective; this range is symmetric respect $\mu = 0$ and is quite narrow.
A similar qualitative behavior is obtained with the classical neural network proposed in this work: however, the range where regularization is effective is much larger than that obtained by rational regularizers [8].
The network with circular input, conversely, shows a peculiar and unexpected behavior: the mean is well predicted in the nearby of $\mu = 0$ and, quite surprisingly, also at the limits of the interval $[-\mu_M, +\mu_M]$. This behavior is emphasized by increasing values of samples *m*. Another interesting result concerns the gain *y* obtained when $\mu = 0$: the circular network seems more effective then the MLP at this point.
The absolute values of the regularized estimation gain *y* with respect to the unregularized one are minimal: this is due to the fact that these experiments, using a wide range of μ, were devoted to understand the widest range of μ for which regularization is effective. In order to build more effective regularizers in terms of the gain *y,* then, one should shrink the interval $\mu \in [-\mu_M, +\mu_M]$; in other words *effectiveness* depends on the a priori knowledge on the interval in which the true mean lies. Note also that in all this analysis the prior on μ is concentrated around $\mu = 0$; other choices are possible. These alternative possibilities translate in a biased regularization problem where results are analogous to that obtained here but centered around another μ value.

V. Conclusions

The present research deals with an unconventional utilization of neural networks: the prediction of the mean value. Some key results have been developed: the main of them shows that regularization, and in general prediction, of the mean value cannot be improved over all the parameters range while effective predictors, that work better than the classical sample mean formulation, exist inside a finite interval of the true mean.
Neural networks offer a notable possibility in implementing these predictors efficiently. Their efficiency has been proved on various ranges of μ. The quality of the results depends on the neural architecture.
This work has addressed the problem of estimating the mean from a Gaussian distribution, however the approach followed here can be used for any distribution.
Future works will extend the results concerning the regularized mean to the more general setting of regression problems where still Stein-like results hold [12]. In that context analogous non linear approaches will be studied in order to predict the 'optimal' value of the regularization parameter.

Appendix

Sketch of Proof of Leave One Out

For the Leave One Out one has to compute the leave one out estimate:

$$\bar{\xi}_i(\alpha) = \frac{\sum_{i \neq k, k=1}^{m} x_k}{m + \alpha - 1}$$

Consequently the error to be minimized is:

$$L_{loo}(\alpha; X) \equiv \sum_{i=1}^{m} (x_i - \bar{\xi}_i)^2$$

After some manipulations the minimum of the previous equation is yielded when:

$$\lambda^{loo} = \frac{\bar{x}_0^2 - \frac{S^2}{m}}{\bar{x}_0^2}$$

where $S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x}_0)^2$.

Sketch of Proof of Maximal Evidence

As first step one decouple the regularization parameter α in two: τ, φ. Then the functional (1) becomes

$$\aleph(\xi, \tau, \varphi) \equiv \frac{\tau}{2} \sum_{i=1}^{m} (x_i - \xi)^2 + \frac{\varphi}{2} \xi^2 \equiv \tau E_D + \varphi E_W$$

Assuming a Gaussian prior over ξ and using Bayes theorem one can compute the Evidence $p(D)$. Then one maximizes $\ln(p(D))$. The obtained sufficient condition for the maximum is the solution of the second order equation

$$\alpha^2 \left[(m-1)\bar{x}_0^2 - s^2 \right] + \alpha \left\{ m \left[(m-1)\bar{x}_0^2 - 2s^2 \right] \right\} - m^2 s^2 = 0$$

Where s^2 is the biased estimator of the variance.
The final solution is

$$\lambda^{me} = \frac{\bar{x}_0^2 - \frac{S^2}{m}}{\bar{x}_0^2}$$

For the complete details on these proofs see [8]

REFERENCES

[1] T. Poggio, and F. Girosi. "Networks for Approximation and Learning", Proceedings of the IEEE (special issue: Neural Networks I: Theory and Modeling), Vol. 78, No. 9, 1481-1497, September 1990.

[2] P. C. Kainen, V. Kůrková, M. Sanguineti, "Complexity of Gaussian Radial-Basis Networks Approximating Smooth Functions'". Journal of Complexity, vol. 25, pp. 63-74, 2009.

[3] V. Kůrková, M. Sanguineti, "Geometric Upper Bounds on Rates of Variable-Basis Approximation'". IEEE Transactions on Information Theory, vol. 54, pp. 5681-5688, 2008.

[4] W James, C .Stein, "Estimation with quadratic loss", Proc. Fourth Berkeley Symp. Math. Statist. Prob., 1, pp. 361–379, 1961.

[5] G. Wahba, P. Craven, "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation". Numer. Math., 31, 377-403,1979.

[6] D.J.C. MacKay "A practical Bayesian framework for backpropagation networks" Neural Computation 4(3) pp.448-472.

[7] S, Ridella, S. Rovetta, R. Zunino "Circular backpropagation networks for classification", IEEE Transactions on Neural Networks, (8) pp. 84-97, 1997.

[8] S. Decherchi, M. Parodi, S. Ridella, "The Regularized Mean Problem", *submitted*, 2010.

[9] Borgwardt, K., A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf and A. Smola: "Integrating Structured Biological data by Kernel Maximum Mean Discrepancy". Bioinformatics 22**(4)**, e49-e57 08 2006.

[10] P. Liang, F. Bach, G. Bouchard, M. I. Jordan. "Asymptotically optimal regularization in smooth parametric models". Advances in Neural Information Processing Systems (NIPS), 2009.

[11] E. Sertel, H.K.Cigizoglu, D.U. Sanli, "Estimating daily mean sea level heights using artificial neural networks". Journal of Coastal Research, May 2008.

[12] J.B. Copas, "Regression, Prediction and Shrinkage", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 45, No. 3, pp. 311-354, 1983.