A Tikhonov Approach to Calculate Regularisation Matrices

Cecilio Angulo, Andreu Català
Politechnical University of Catalonia
EUPVG, Av. Victor Balaguer, 08800 Vilanova i la Geltrú. Barcelona
cangulo@fabrique.upc.es, andreu@esaii.upc.es

Abstract

Regularisation is a popular method to overcome the ill-conditioning learning problems in neural networks. This is achieved by penalizing the performance criteria adding some prior distribution on the weights, usually a quadratic weight function, i.e. $E^d = w^t K w$. Find out the regularisation matrix ***K*** included in the regularisation solution of learning problems, is a difficult and computationally expensive task [2]. This paper provide an efficient and easy way for ***K*** matrix calculation and the development of this method for Zero and Second Order Regularisation.

1. Introduction

Let $S=\left\{ \left(\mathbf{x}_i, y_i\right) \in \Re^n \times \Re \mid i=1,\ldots,N \right\}$ be a set of data we want to approximate by a function *f*. The Tikhonov's regularisation approach [6] consists in looking for the function *f* that minimizes the functional [1], [4], [5], $H[f]=\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i)\right)^2 + \lambda\|Pf\|^2$ where *P* is a constraint operator, stabilizer or regulariser (usually a differential operator), $\|\cdot\|$ is a norm of the space functions (usually the L^2 norm), $\|Pf\|^2 = \int |Pf|^2 d\mathbf{x} =$ $= \int f(\mathbf{x}) \hat{P} P f(\mathbf{x}) d\mathbf{x}$, and λ *(regularisation parameter)* is a positive real number, controlling the compromise between the degree of smoothness of the solution and its closeness to the data.

From the point of view of training Radial Basis Functions Systems and Neurofuzzy models, regularisation is a popular method to constrain weight optimization [2]. We are looking for a function, $f(x,w)=G(x_i,x)$, where $G(x_i,x)$ is a combination of Radial Basis Functions, that minimizes the cost function $J=MSE+\lambda E^d$ where *MSE* is the Mean Square output Error and E^d is a penalizing error that may take different forms, but if it is represented by a quadratic weight function $E^d = w^t K w$, the solution to the cost function becomes $w = (R+K)^{-1} p$ where $R=G^T G$ is the autocorrelation matrix and $p=G^T y$ is the cross-correlation matrix.

In the next section we present the main results obtained by Poggio and Girosi [4], [5] to be applied in our approach to calculate the regularisation matrix. In the sections 3 and 4, we establish the relation of the Tikhonov's approach with the standard statistics approach and we present how the regularisation matrix is obtained when the Radial Basis Functions are fixed.

2. Tikhonov's regularisation solution recovers the RBF method

Minimization of the functional H leads to the associated Euler-Lagrange equations. For a functional $H[f]$ that can be defined

$$H[f] = \int_{\Re^r} F\left(f, f_x, f_y, \ldots, f_{xx}, f_{xy}, f_{yy}, \ldots, f_{yy\ldots y}\right) \mathrm{dxdy}$$

the Euler-Lagrange equations of extremity condition are

$$\mathrm{F}_f - \frac{\partial}{\partial x}\mathrm{F}_{f_x} - \frac{\partial}{\partial y}\mathrm{F}_{f_y} + \frac{\partial^2}{\partial x^2}\mathrm{F}_{f_{xx}} + \frac{\partial^2}{\partial y^2}\mathrm{F}_{f_{yy}} + \frac{\partial^2}{\partial x \partial y}\mathrm{F}_{f_{xy}} + \cdots + (-1)^n \frac{\partial^n}{\partial y^n}\mathrm{F}_{f_{yy\ldots y}} = 0 \cdot$$

If we define $\frac{\partial f(\mathbf{x})}{\partial f(\mathbf{x}')} = \delta(\mathbf{x} - \mathbf{x}')$ (Dirac's delta), for a functional that is only function of f we obtain the partial differential equation $\hat{P}Pf(\mathbf{x}) = \frac{1}{\lambda}\sum_{i=1}^{N}\left(y_i - f(\mathbf{x}_i)\right)\delta(\mathbf{x} - \mathbf{x}_i)$ where $\hat{P}$ is the adjoin of the differential operator P. Its solution can be written as the integral transformation of its right side with a kernel given by the Green's function of the differential operator $\hat{P}P$, the function G satisfying the following distributional differential equation $\hat{P}PG(\mathbf{x};\xi) = \delta(\mathbf{x} - \xi)$ *(Green's relation)*

Finally we obtain the solution of the regularisation problem

$$f(\mathbf{x}) = \sum_{i=1}^{N}\frac{y_i - f(\mathbf{x}_i)}{\lambda}G(\mathbf{x};\mathbf{x}_i) \overset{\mathrm{def}}{=} \sum_{i=1}^{N}\mathbf{w}_i\, G(\mathbf{x};\mathbf{x}_i) = \mathbf{G}(\mathbf{x}_i, \mathbf{x}) \cdot \mathbf{w}$$

where a polynomial term should be in the right side if the null space of the stabilizer is different of the null function. To obtain in a direct form the unknown coefficients $w_i=\lambda^{-1}(y_i-f(x_i))$, we use for the coefficients the linear system $(G+\lambda I)w=y$ where $(y)_i=y_i$, $(w)_i=w_i$, $(G)_{ij}=G(x_i;x_j)$.

If the operator P is translationally and rotationally invariant, $G = G(\|\mathbf{x} - \xi\|)$ will be a radial function and the method of Radial Basis Functions may be recovered.

2.1. Reduction of the computational complexity. Extension of the regularisation approach to moving centers: GRBF

The function that minimizes the functional H is specified by N coefficients, where N is the number of examples or training patterns; when N is very high the computation of the coefficients of the expansion can become a very time consuming operation.

To reduce the complexity of the problem, we concentrate the solution on a finite basis.

The approximated solution f^* to the regularisation problem implies a lower number of centers with a different distribution [3], $f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} w_\alpha \, G(\mathbf{x}; \mathbf{t}_\alpha)$ $n<<N$, where the coefficients w_α and the parameters t_α are unknown. If G is the Green's function of the operator $\hat{P}P$, the set of coefficients $\{w_\alpha, t_\alpha / \alpha=1,\ldots,n\}$ can be obtained in a simple way by minimizing $H[f^*]$.

The explicit form of the system equations to be satisfied for the coefficients depends on the specific restriction operator used.

This form of solution has the desirable property to be a universal approximation for continuos functions and to be the only choice that consistently recovered the correct solution in case $n = N$ and $\{t_i\}_i = \{x_i\}_i$.

2.2. A generalization of multidimensional splines.

We consider the stabilizer

$$\|P_1 f^*\|^2 = \int_{\Re^d} \sum_{m=0}^{\infty} a_m \left(P^m f^*(\mathbf{x})\right)^2 d\mathbf{x} = \lim_{M\to\infty} \sum_{m=0}^{M} a_m \|O^m f^*\|^2$$

with the functional associated to multidimensional splines

$$\|O^m f^*\| = \int_{\Re^d} \left(P^m f^*(\mathbf{x})\right)^2 d\mathbf{x} = \sum_{i_1\ldots i_m}^{d} \int_{\Re^d} \left(\frac{\partial^m}{\partial x_{i_1}\ldots\partial x_{i_m}} f^*(\mathbf{x})\right)^2 d\mathbf{x}$$

where $P^{2m}=\nabla^{2m}$, $P^{2m+1}=\nabla\nabla^{2m}$,and $\nabla^2=\Delta$ is the Laplacian. The coefficients a_m are positive real numbers.

The stabilizer is rotationally and translationally invariant and its Green's function satisfies the distributional differential equation:

$$\sum_{m=0}^{\infty}(-1)^m a_m \nabla^{2m} G(\mathbf{x}-\xi) = \delta(\mathbf{x}-\xi) \quad \textit{(Relation Green's Function - Stabilizer)}$$

By using the Green's formulas, the m-th term of P_1 can be written as

$$\int_{\Re^d} \left(P^m f^*(\mathbf{x})\right)^2 d\mathbf{x} = (-1)^m \int_{\Re^d} f^*(\mathbf{x}) \, P^{2m} f^*(\mathbf{x}) d\mathbf{x}$$

Because f^* is a expansion of Green's functions, each term containing G gives a delta function and the integral disappears, yielding

$$\|P_1 f^*\|^2 = \int_{\Re^d} \sum_{m=0}^{\infty} a_m \left(P^m f^*(\mathbf{x})\right)^2 d\mathbf{x} = (-1)^m \int_{\Re^d} \sum_{m=0}^{\infty} a_m \, f^*(\mathbf{x}) \, P^{2m} f^*(\mathbf{x}) d\mathbf{x} =$$

$$= \int_{\Re^d} \left(\sum_{\beta=1}^{n} w_\beta \, G(\mathbf{x}; \mathbf{t}_\beta)\right)\left(\sum_{\alpha=1}^{n} w_\alpha \delta(\mathbf{x}-\mathbf{t}_\alpha)\right) dx = \sum_{\alpha,\beta=1}^{n} w_\alpha w_\beta \, G(\mathbf{t}_\alpha; \mathbf{t}_\beta)$$

Defining a rectangular $N \times n$ - matrix G as $(G)_{i\alpha} = G(x_i; t_\alpha)$ and a symmetric $n \times n$

square matrix $(g)_{\alpha\beta} = G(t_\alpha ;t_\beta)$, the functional can be written as $H[f^*]=w^T(G^TG+\lambda g)w-2wG^Ty+yy$, a quadratic form in the coefficients w_{α}. For each fixed set of centers t_α, the coefficients optimal vector is then given by $w=(G^TG+\lambda g)^{-1} G^Ty$.

Defining $R= G^TG$; $p=G^Ty$; $K=g$, we obtain $w=(R+\lambda K)^{-1} p$. This expression for the coefficients is equivalent to that obtained for standard regularisation. So, our main work will be to show that functional cost are equivalent too.

2.3. Simplification to Multidimensional Splines

We consider the stabilizers $\|O^m f^*\|^2$, obtained in 2.2 when, $a_k=0, \forall k\neq m;\ a_m=1$ to profit all the last information. Green's function associated to this regulariser satisfies $(-1)^m \nabla^{2m} G(x-\xi)=\delta(x-\xi)$, where ∇^{2m} is the *m-th* Laplacian in *d* dimension, and the solution is

$$G(\mathbf{r}) = \begin{cases} \|\mathbf{r}\|^{2\mathrm{m}-\mathrm{d}} \cdot ln(\|\mathbf{r}\|) & (2\mathrm{m} > \mathrm{d}) \& (\mathrm{d} = \dot{2}) \\ \|\mathbf{r}\|^{2\mathrm{m}-\mathrm{d}} & \text{otherwise} \end{cases}$$

The approximated solution has then the following form, with $p_{m-1}(x)$ a polynomial of degree *m-1*.

$$\mathrm{f}^*(\mathbf{x}, w_i) = \sum_{i=1}^{n} \mathbf{w}_i \, \mathrm{G}(\mathbf{x};\mathbf{t}_i) + \mathrm{p}_{m-1}(\mathbf{x})$$

3. Multidimensional spline (m = 0) - Zero Order Regularisation

Green's function associated to this regulariser satisfies $G(x-\xi)=\delta(x-\xi)$, then the approximated solution has the following form

$$f^*(\mathbf{x}, \mathrm{w}_\mathrm{i}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_\mathrm{i} \delta(\mathbf{x} - \mathbf{t}_\mathrm{i})$$

The linear system associated to the coefficients is $\mathrm{w}=(\mathrm{G}^\mathrm{T}\mathrm{G}+\lambda \mathrm{g})^{-1}\ \mathrm{G}^\mathrm{T}\mathrm{y}$, where $(g)_{\alpha\beta} =G(t_\alpha -t_\beta)=Id$ or, applying definitions, $w=(R+\lambda K)^{-1} p$, where $K=Id$

The cost function or functional becomes

$$\mathrm{J} = \mathrm{MSE} + \lambda \|O^0 f^*\|^2 = \mathrm{MSE} + \lambda \int_{\Re^d} (\mathrm{f}^*(\mathbf{x}))^2 \,\mathrm{d}\mathbf{x} =$$

$$= \mathrm{MSE} + \lambda \int_{\Re^d} \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_\mathrm{i} \delta(\mathbf{x} - \mathbf{t}_\mathrm{i}) \right) \cdot \left(\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w}_\mathrm{j} \delta(\mathbf{x} - \mathbf{t}_\mathrm{j}) \right) \mathrm{d}\mathbf{x} = \mathrm{MSE} + \lambda \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w}_\mathrm{i}^2 =$$

$$= MSE + \lambda w^T w = MSE + \lambda w^T K w \text{ , where } K=Id$$

This is the commonest regularisation, known as ridge regression by statisticians. The penalty term $\mathbf{E}^\mathbf{d} = \int_{\Re^d} |f^*(\mathbf{x}, \mathrm{w}_\mathrm{i})| \mathrm{p}(\mathbf{x}) \mathrm{d}\mathbf{x}$ where $p(x)$ is the probability density function, represents the expected size of the output. This can be approximated by a quadratic penalty function given by, $E^d =w^Tw= =w^TKw$, where $K=Id$.

If we observe the linear system for the solution to this cost function in section 1, we conclude that Tikhonov's regularisation approach gives the same results that the RBF standard regularisation approach, and we obtain a specific form for the radial basis functions.

4. Multidimensional spline (m = 2) - Second Order Regularisation

The distributional differential equation associated is, in this case, $(\nabla^2)^2G(x-\xi)=\delta(x-\xi)$ and the solution is

$$G(\mathbf{r})=\begin{cases}\|\mathbf{r}\|^{4-d}\cdot \mathit{ln}(\|\mathbf{r}\|) & d=2\\ \|\mathbf{r}\|^{4-d} & d\neq 2\end{cases}$$

This solution is known as thin plate spline when d=2. The approximated function

$$f^*(\mathbf{x},\mathrm{w}_i)=\sum_{i=1}^{n}\mathrm{w}_i\,G(\mathbf{x};\mathbf{t}_i)+p_1(\mathbf{x})$$

has coefficients that accomplish the linear system associated as in the previous case but now *K=g*. The cost function or functional is given by

$$\mathrm{J}=\mathrm{MSE}+\lambda\|O^2f^*\|^2=\mathrm{MSE}+\lambda\int_{\Re^d}(P^2f^*(\mathbf{x}))^2\,d\mathbf{x}=\int_{\Re^d}f^*(\mathbf{x})(\nabla^2)^2f^*(\mathbf{x})d\mathbf{x}=$$

$$=\mathrm{MSE}+\lambda\left(\sum_{i=1}^{n}\mathrm{w}_i\,G(\mathbf{t}_j;\mathbf{t}_i)+p_1(\mathbf{t}_j)\right)\left(\sum_{j=1}^{n}\mathrm{w}_j\right)=\mathrm{MSE}+\lambda\mathbf{w}^{\mathrm{T}}\mathbf{g}\mathbf{w}+\sum_{j=1}^{n}\mathrm{w}_j\,p_1(\mathbf{t}_j)\overset{(1)}{=}$$

$$=MSE+\lambda w^Tgw=MSE+\lambda w^TKw\text{ , where K=g}$$

(1)If $n = N$ and $\{x_i\}_i=\{t_i\}_i$, the term polinomic term is 0 [4]*; in other cases this condition is a new linear system of equations to be accomplished for the moving centers and the coefficients.*

This form of regularisation makes the assumption that the function is smooth, and hence the expected curvature of the output is used to penalize the cost function,

$$\mathbf{E}^{\mathrm{d}}=\int_{\Re^d}\left|\frac{d^2f^*(\mathbf{x},\mathrm{w}_i)}{d\mathbf{x}^2}\right|p(\mathbf{x})d\mathbf{x}$$

where p(x) is the probability density function. E^d can be approximated by [2]

$$\mathbf{E}^{\mathrm{d}}=\sum_{i=1}^{d}\left(\frac{d^2f^*(\mathbf{x},\mathrm{w}_i)}{d\mathbf{x}^2}\right)^2=\sum_{i=1}^{d}[\mathbf{k}_i^{\mathrm{T}}\mathbf{w}]^2=\sum_{i=1}^{d}\mathbf{w}^{\mathrm{T}}\mathbf{k}_i\mathbf{k}_i^{\mathrm{T}}\mathbf{w}=\mathbf{w}^{\mathrm{T}}\mathbf{K}\mathbf{w}$$

To obtain the regularisation matrix *K* in the previous expression is computationally intensive, because is necessary to compute the 2nd differential of the multidimensional basis functions.

Therefore, the Tikhonov's regularisation approach identifies the *K* matrix and calculations are no difficult if we choose the specified radial basis functions from the Green's condition.

5. Conclusions and future research.

In this paper we have employed the Thikonov's regularisation approach to provide an efficient and easy expression for the regularisation matrix K, if the radial basis functions are indicated, and we develop this method for the Zero and Second Order Regularisation [2]. It could be interesting to express the approximated function expansion over other class of functions, calculate the differential equations satisfied and find out the new expressions for the regularisation matrix. On the other hand, higher order regularisation can be compared with our approach; we have only shown two examples where analogies are direct, but the results, $K=g$ and the cost function equivalence, are general equivalencies.

Many options exist to select radial basis functions, but it has been pointed out that the thin plate splines, that are associated with the second order regularisation of the Tikhonov's approach, works very well in terms of modeling accuracy [7]. This practical result could have a motivation in the best approximation property of this class of functions in the ideal case, when $n=N$ and centers are the original training patterns, and the fact that the approximated function is a reduced version of the total one.

REFERENCES.

[1] Bishop C. M. "Neural Networks for Pattern Recognition". Clarendon Press, Oxford, 1995.

[2] Bossley K.M. "Regularisation theory applied to neurofuzzy modelling". Technical Report ISIS-TR3, Dept. of Electronics and Computer Science, University of Southampton, 1997.

[3] Broomhead D.S. & Lowe D. "Multivariable functional interpolation and adaptive networks". Complex Systems, vol. 38(5), pp. 181-200, 1982

[4] Poggio T. & Girosi F. "A Theory of Networks for Approximation and Learning". A.I.Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[5] Poggio T. & Girosi F. "Networks for Approximation and Learning". Proceedings of the IEEE, vol. 78, No. 9, September 1990. pp 1481-1497.

[6] Tikhonov A.N. & Arsenin V.Y. "Solutions of Ill-posed Problems". W.H. Winston, Washington D.C., 1977.

[7] Warwick K. and others. "Dynamic Systems in Neural Networks". In 'Neural Network Engineering in Dynamic Control Systems', K.J. Hunt, G.R. Irwin and K. Warwick Eds, 1995.