Appearing in: *Lecture Notes in Computer Science : Artificial Neural Networks* - (ICANN'97.) vol. 1327, pp. 861-866.
Available at `http://www-poleia.lip6.fr/~zaragoza`

Multiple Multivariate Regression And Global Optimization in a Large Scale Thermodynamical Application.

H. Zaragoza, P. Gallinari
LIP6, *Université Pierre et Marie Curie*,
4, place Jussieu F-75252 PARIS cedex 05 (F).
`zaragoza@laforia.ibp.fr`

Abstract. We describe a large scale real-world application of neural networks for the modelization of heat radiation emitted by a source and observed through the atmosphere. For this problem, thousands of regressors need to be trained and incorporated into a single model of the process. On such large scale applications, standard techniques for the control of complexity are impossible to implement. We investigate the interest of i) integrating several regressors into a single neural network, and ii) refining the learned functions by optimizing simultaneously all regressors over a global function. The two approaches described offer a solution to these problems, and were crucial for the development of a fast and accurate model of radiation intensity.

1 Introduction

The motivation behind the research presented in this paper was the development of a system for the modelization of a complex physical process: the computation of *spectral intensity of radiation*. The problem is to provide fast and accurate computations of the intensity of a heat (radiation) source at different surrounding points in a turbulent and absorbing atmosphere. This thermodynamical problem is of primary importance for a number of applications in the aeronautical industry.

In Section 2 we give an overview of the physical model on which the application is based. We will see that the main problem that needs to be solved is fitting a very large number of functions (referred as *parameter functions*) which form a parametric base for the computation of the spectral intensity. The number of parameter functions that need to be determined, as well as the number of times these need to be evaluated, imposes practical constraints on the kind of techniques applicable to the regression. In a typical application such as ours, 35 parameter functions need to be fitted for each spectral point (given radiation frequency), and about 100 spectral points are needed to characterize the intensity of a single observation; this means that a total of about 3500 parameter functions need to be interpolated to construct the complete radiation model. Neural network (NN) applications of this size are very rare, because of the «manual work» usually necessary for good convergence of the models. To deal with such a great number of regressors, classical statistical methods (e.g. cross validation, bootstrap, etc.) cannot be used to determine hyper-parameters such as the learning rate, stopping criteria, or regularization constants. A final constraint is introduced by the necessity of fast computations: several millions of parameter function values need to be computed for a single radiation image. This imposes a practical constraint on the speed of the computation and, therefore on, the complexity of the regressors utilized.

In Section 3 we discuss how ordinary least square regression and NNs may be used to fit these functions independently. Straight-forward substitution of linear models by NNs brings a drastic improvement of accuracy, but greatly increases the model's computational cost. Instead, we use a single NN to approximate simultaneously all parameter functions. This leads to a reduction of the number of network parameters and improves generalization.

For many physical applications, approximated functions are (local) parameters of a much more complex (global) function of known physical significance. For these cases, we propose, in Section 4, a method that optimizes globally the physical function of interest, instead of optimizing individually each of its parameters.

2 The Physical Problem

Our goal is to compute the intensity (I) of a radiation source, at a point distant from the source, in an atmospheric media. At a given frequency, the *transmissivity* of a gas (τ) describes the behavior of radiation emitted through it. In an homogeneous (constant molecular composition) and isothermal (constant temperature and pressure) situation, the intensity can be computed in a straightforward manner. We are interested, however, in nonhomegenous and nonisothermal conditions, typical of aeroplane reactors. Computing the transmissivity of such media is an extremely difficult problem, because of the inherent nonlinearities of the process. To deal with such situations, the gas «column» (that is, the line going from the source to the observed point) is discretized into a sequence of N homogeneous and isothermal segments of different (unknown) spectral transmissivities (τ_s). Computing the intensity of radiation is then straightforward[1]:

$$I = \sum_{s = 1, N} (\tau_{s-1} - \tau_s) I_s^o + \text{const} \quad (1)$$

The difficult problem is the determination of τ_s for each segment, which is affected by local as well as global characteristics of the gas column.

There exists a very precise model to determine spectral transmissivity, named the Line-by-Line model (LBL), but this model is too computationally intensive for real-world applications and it is used only for validation purposes. A number of simplified models have been developed and are of common use. The correlated fictitious gases method (CKFG-N) is the most adequate of these for nonhomogeneous gases and has already been applied with success in a number of applications [1]. The CKFG-N model approximates the spectral transmissivity of a segment (τ_s) with a high dimensional parametric function[2]:

$$\tau_s = \prod_{j = 1, N} \left[\sum_{m = 1, M} A_m \exp\left(\sum_{i = 1, s} f_{jm}(t_i, p_i)\, l_i c_i \right) \right] \quad (2)$$

where f_{jm} are the *parameter functions*, which are two-dimensional unknown functions and are fitted from real (or LBL) data.

1.I_s^0 is a constant (the vacuum intensity of the segment *s*).
2.In equation (2), l_i and c_i, denote the length and molecular composition of segment *i* respectively, and A_m is a constant.

Equations (1) and (2) define a «computational chain». Our goal is the computation of the intensity (I). We first compute the segment transmissivity (τ_s) of each segment. For this, we will need the value of all parameter functions f_{jm} for the given segment's temperature and pressure (t_i,p_i). We have eliminated the dependance on the radiation frequency for simplicity of notation, but, basically, the whole process is repeated independently for each frequency of interest. This is why the number of parameter functions is so great. In (2) we see that NxM (here N = 5, M = 7) parameter functions are needed per frequency; since hundreds of frequencies are typically computed, several thousands of f_{jm} models are needed.

3 Parameter Regression

Physicists construct the parameter functions using ordinary least squares regression (OLS) independently for each f_{jm}. The nonlinearity of the dependance on temperature and pressure may be introduced in the regression directly providing nonlinear terms as input variables. Similarly, a standard neural network approach would be to build a small network to predict each f_{jm} separately.

In this application the number of points available for fitting the data was small. Instead of partitioning the data into a training and a validation set, we used four *columns* of known intensity (I) but unknown transmissivities, in order to test the generalization capabilities of our models. Therefore, the model is not tested with respect to the f_{jm} functions used for training, but rather with respect to the final value of I obtained after applying equations (1) and (2). The four columns are: the construction column (P-train), two columns dealing only with low or high pressure situations (P-low and P-high respectively) and a mixed pressure column (P-mix).

Model:	P train	P mix	P high	P low	mean	# par.
OLS	2.12	0.41	0.52	0.05	0.33	245
NN	0.01	0.10	0.13	0.01	0.08	1400
i-NN-35	0.01	0.09	0.06	0.05	0.07	1470
i-NN-10	0.04	0.03	0.07	0.05	0.05	630
i-NN-5	0.36	0.01	0.40	0.09	0.17	210

Table 1. Linear and NN Implementations (see text for notations).

The first two rows of Table 1 present the performances of the OLS and NN models. We see that independent NNs perform significantly better for the construction and the columns. This increase in performance must be contrasted with the enormous increase of computational cost when using an MLP of the type described. In the last two columns of Table 1 we show the mean validation error, as well as the number of parameters. This number is (at best) linearly related to the computational time needed to compute a solution. It seemed necessary at this point to reduce the model's complexity.

Several ideas lead naturally to the implementation of multiple function regressors for our application. The integration of regressors, when correlated, may improve generalization. In the case of linear regression it can be shown that a weighted sum of a set of regressors gives a better (or at worst equal) solution than the independent regressors. Appropriate weights may be estimated from data [3]. There do not exist to our knowledge similar results in the case of non-linear function approximation. However, a number of empirical investigations exist in the case of classification [2].

To test the interest of integrated function regressors, we built a single NN to estimate simultaneously the thirty-five functions needed for the computation of the intensity of radiation at a given frequency. Instead of using 35 MLPs with one output, we utilized a (fully connected, one hidden layer) MLP with thirty-five linear output units. We hope that function correlations will be captured at the hidden layer (which is now connected to all parameter functions) and used during the training phase to make up for the reduced number of parameters.

In Table 1 (last three rows) we present the performance of such an integrated architecture, for MLPs with 35, 10 and 5 hidden units. Results should be contrasted with the number of parameters and the mean test error of each architecture. The largest model, i-NN-35, has about the same number of parameters as the 35 independent NNs, but one fifth the number of hidden units. The accuracy obtained in the training column is similar to the one obtained by the simple NN scheme, and the mean error of generalization is slightly lower. The second integrated model, i-NN-10, has about one half the number of parameters than the previous two architectures, and it is not capable of learning as accurately the training functions (the error for the P-train column is higher), but reduces the mean test error by a factor of 1.7. We see here the double benefit of constraining the network while making use of the correlation of the parameter functions: we have almost doubled the accuracy of our network, using one half of the number of parameters. With respect to the linear network, i-NN-10 utilizes three times more parameters and reduces by seven the mean test error. Our last integrated model i-NN-5, shows that we may build systems with less parameters than the linear model, and still obtain better results. This contradicts the common intuition that we must «pay the price» of a large increase in parameters if we use nonlinear regression. While this is generally the case, it is not so when dealing with large families of nonlinearly correlated functions.

4 Global Optimization

Not all parameter functions (f_{jm}) are equally important for the accuracy of the intensity calculation (I). Depending on the wave-length, temperature and pressure as well as their inter-segment gradients, a few f_{jm} will dominate the computation of I. Our goal is not to obtain the best possible estimates of f_{jm} but, rather, the best possible estimate of I. Therefore, most resources should be put to estimate the most important f_{jm}.

The method described so far does not take this into account: by minimizing the mean square error of the f_{jm} functions we implicitly attribute equal importance (resources) to all functions and to all regions of the functional domain. This is particularly dangerous in the case of integrated regressors since shared weights and units may be dominated by functions which may be difficult to approximate but have negligeable effect on the ulterior intensity of radiation calculations.

Physicists aware of these facts have traditionally biased by hand their regressors (modifying the training point distribution or weights, or with piece-wise linear regressors). Such an heuristic approach leads to very accurate systems, but necessitates of an expensive period of experimentation and calibration specific to each particular application. We sought to bias the regressors automatically. This can be done redefining the optimized cost function as the quadratic error of the intensity of radiation. Formally, the function minimized by the regressors presented in the previous sections is the mean quadratic error between the desired and obtained parameter functions, which may be written as:

$C = \frac{1}{2}\sum_{i,j,m}\left(f_{ijm} - \hat{f}_{ijm}\right)^2$, where f_{ijm} denotes the parameter function value $f_{jm}(t_i, p_i)$.

We propose to optimize directly I by minimizing instead a cost function of the form:

$$C' = \frac{1}{2}\sum_{\zeta}\left[I_\nu(\zeta) - \hat{I}_\nu(\zeta)\right]^2 , \qquad (3)$$

where I_ν denotes the radiation intensity at frequency ν and ζ denotes the set of sequences globally optimized. To solve for C' we need to compute its derivatives with respect to the approximated parameter functions which are, in fact, the output values of the regression models described in the previous sections. Derivatives may be obtained by the chain rule:

$$\frac{dC}{d\hat{f}_{ijm}} = \frac{dC}{d\hat{I}} \cdot \frac{d}{d\hat{f}_{ijm}}\hat{I} = -(I - \hat{I}) \cdot \frac{d}{d\hat{f}_{ijm}}\hat{I} ,$$

$$\frac{\partial I}{\partial f_{ijm}} = \left[\sum_{i' = i, N-1} \frac{\partial \tau_{i'}}{\partial f_{ijm}}(I_{i'+1} - I_{i'})\right] \frac{\partial \tau_N}{\partial f_{ijm}} I_N ,$$

$$\frac{\partial \tau_{i'}}{\partial f_{ijm}} = \left(\prod_{j' \neq j} \tau_{j'i'}\right) A_m \exp\left(\sum_{\iota = 1, i'} f_{\iota jm} l_\iota c_\iota\right) l_i c_i$$

The difficulties are twofold: the derivatives are recursive functions of the segment intensity, and partial derivatives need to be computed with respect to the thirty-five parameter functions *at every segment* (several thousand partial derivatives are therefore needed at each back-propagation pass of the complete model). We may use the same MLP architectures as before to optimize C' with the back-propagation implementation of gradient descent. This optimization presents however serious difficulties, since there is only a local guarantee of convergence and the computational cost of a single back-propagation pass is extraordinary. To ensure a faster and more subtle convergence of the direct optimization, we chose a two stage learning process in which parameter functions are first approximated to a reasonable degree (by the training method described in Section 4), and then global sequence optimization is carried out.

In Table 2 we present the results of this optimization scheme, on the previously discussed i-NN-10 model. The first stage consists in the optimization of parameter functions with a cost function similar to C, as described in Section 3 (the results obtained are the same as those of Table 1). We then retrain the MLP optimizing directly the intensity of radiation of a real gas column (P-Mix), with cost function of the form of C'. We present in the second column the values of the errors produced by this network on the optimized column (P-Mix), and on the two other validation columns (P-High and P-Low) as well as on the construction column (P-Train). We see that the error on the gas column optimized has been greatly reduced. More important, we see that the error on the other two validation columns has not been increased by this optimization but, on the contrary, has been slightly reduced. We can therefore expect very accurate predictions on columns similar to those for which it has been specialized, without deterioration of the accuracy on other type of columns. Finally, we note that the error on the construction column has increased as expected.

	local	+ global
P-Mix	***0.034***	***0.005***
P-High	0.066	0.057
P-Low	0.057	0.042
P-Train	0.024	0.056

Table 2. Intensity optimization (see text for notations).

5 Conclusions

We have presented an effective large-scale model of spectral radiation intensity based on neural networks. A straightforward substitution of linear models by neural networks lead to the development of several thousand neural networks. Instead, we developed networks that approximated simultaneously all functions within the same parametrical family. This model was much more accurate than the linear implementation, with only a moderate increase in the number of parameters. Furthermore, the choice on the number of hidden units of the model proved an effective parameter to weight the necessary compromise between accuracy and speed of computations. This type of model may be further extended to optimize a global error function instead of optimizing the sum of the squared output errors. This provides the advantage of selectively improving the accuracy of the functions that play an important role in the computation of the global solution.

References

[1] Rivière Ph., Soufia A. and Taine J. (1992). Correlated-k and fictitious gas methods for H2O near 2.7µm, *J. Quant. Spetrosc. Radiat. Transfer* 48(2), 187-203.

[2] Caruana R. (1994). Learning Many Related Tasks at the Same Time With Backpropagation, *Advances in Neural Information Systems* 7, 664-657.

[3] Breiman L. and Friedman, J.H. Predicting Multivariate Responses in Multiple Linear Regression. *Royal Statistical Society* (in press).

[4] Zaragoza H. (1997). Lessons from a large-scale neural network thermodynamical application: variable reduction, multiple-function approximation and global optimization. *Technical Report, LAFORIA-IBP (University of Paris 6).*