A NEW TRAINING SET-BASED REGULARIZATION FOR REGRESSION TECHNIQUES

Youness Naji, Laurent Le Brusquet and Gilles Fleury

Supélec, Department of Measurement
3 rue Joliot Curie, Plateau de Moulon, 91192 Gif-sur-Yvette, France
phone: +33 (0)1 69 85 14 23, fax: +33 (0)1 69 85 14 29, firsname.lastname@supelec.fr
web: www.supelec.fr/ecole/mesures/Bienvenue.html

ABSTRACT

The paper gives a new regularization criterion for the regression techniques where the overfitting problem may occur. The proposed criterion is not a penalization term calibrated from prior information but a penalization term calculated from the training set. It appears as an extension of the classic Tikhonov regularization constraint. It is shown that the statistical characterization of this penalization is possible. This characterization leads to an optimization criterion which does not depend on any hyperparameter.
The method is applied to a parametric regression technique (polynomial regression) and to a nonparametric regression technique (kernel approximation). For the first technique, overfitting is avoided. For the second one, the method gives an estimation of the kernel spread close to the optimal value.

1. PROBLEM STATEMENT

Consider an unknown process $f^* : [a,b] \to \mathbb{R}$ which is estimated from a training set $D_n = \{(y_i, z_i)_{i=1\cdots n}\}$ resulting from f^* [1, 2]:

$$z_i = f^*(y_i) + \varepsilon_i, \quad i = 1 \cdots n \tag{1}$$

where ε_i is a Gaussian additive noise independent with f^*.
The goal is to find an estimation $\hat{f}$ of f^* which limits the risk of overfitting [3]. In machine learning techniques, this risk may be evaluated by the integrated loss:

$$IL\left(\hat{f}\right) = \int_a^b \left(f^*(y) - \hat{f}(y)\right)^2 dy \tag{2}$$

Various techniques have been proposed for coping with the overfitting problem. Most of them fall into one of two categories: model selection [4] and regularization [5, 6]. In this paper, we introduce a new regularization criterion based on variability that automatically chooses the complexity of models: greater is the variability of the training set, greater is the allowed model complexity (see author's previous work in [7]).
The new criterion may be applied to parametric and nonparametric regression techniques. In this paper, it is illustrated on two particular techniques:

- a polynomial regression technique: f^* is modeled by one function in the class $\{P_\theta\}_{\theta\in\mathbb{R}^{d+1}}$ of polynomials of degree d:

$$P_\theta(y) = \sum_{k=0}^{d} \theta_k y^k$$

- a kernel technique: f^* is approximated by the nonparametric estimator of Nadaraya-Watson [8, 9]:

$$g_h^{NW(D_n)}(y) = \frac{\sum_{i=1}^{n} z_i K\left(\frac{y-y_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right)}$$

  $K(y)$ is the Gaussian kernel: $K(y) = \exp\left(-\frac{y^2}{2}\right)$.
  The spread h carries out a tradeoff between a smooth estimator (large h) and an unbiased one (small h).

Coefficients θ and h have to be estimated for both techniques. Section 2 gives the principle of the proposed approach to estimate them. It is applied in section 3 to polynomial regression and kernel approximation with training sets obtained by uniform and random sampling.

2. PROPOSED APPROACH

The suggested regularization consists in considering only models with a regularity equivalent to the regularity of the training set: models P_θ or $g_h^{NW(D_n)}$, indifferently noted f_β with $\beta = \theta$ or h.
The regularity of the training set (resp. model) is viewed as a measure of its variability: a training set (resp. model) is regular if it has a low variability. The variability criterion of a training set must only depend on the variations of f^* and thus not be very sensitive to the noise.
The variability of a function $g : [a,b] \to \mathbb{R}$ was defined by a classic criterion related to its fluctuations [10]:

$$V_{fc}(g) = \int_a^b \left(\frac{dg(y)}{dy}\right)^2 dy \tag{3}$$

The variability criterion proposed for the training set is an approximation of the integral in equation (3):

$$V_{TS}^0(D_n) = \sum_{i=1}^{n-1} \frac{(z_{i+1} - z_i)^2}{y_{i+1} - y_i}$$

We have assumed here, without loss of generality, that:

$$y_1 < y_2 < \cdots < y_n$$

The quadratic behavior of the estimator V_{TS} confers the advantage of having a bias and a variance independent of f^*.

Indeed, in the case of training set (y_i, z_i) regularly spaced (y_i resulting from a uniform sampling from $[a,b]$) and corrupted by a white noise of variance σ_ε^2 (cf. equation (1)), a calculation of mathematical expectation leads to:

$$\begin{aligned} \mathrm{E}_\varepsilon\left\{V_{TS}^0\right\} &= \sum_{i=1}^{n-1} \frac{(f^*(y_{i+1}) - f^*(y_i))^2}{y_{i+1} - y_i} + 2(n-1)^2\sigma_\varepsilon^2 \\ &\approx V_{fc}(f^*) + 2(n-1)\sigma_\varepsilon^2 \quad [n \gg 1] \end{aligned}$$

The expression is given for $a = 0$ and $b = 1$ but it may be obtained with any values of a and b.
It is then possible to correct the bias of the estimator:

$$V_{TS}(D_n) = \sum_{i=1}^{n-1} \frac{(z_{i+1} - z_i)^2}{y_{i+1} - y_i} - 2(n-1)^2\sigma_\varepsilon^2$$

The calculation of variance may also be achieved:

$$\mathrm{var}_\varepsilon\left\{V_{TS}\right\} \approx \sigma_\varepsilon^4(12n^3 - 4n^2)$$

Similar calculations may be performed in the case of training sets obtained by irregular sampling of the interval $[a,b]$. They lead to an unbiased estimation of $V_{fc}(f^*)$:

$$V_{TS}(D_n) = \sum_{i=1}^{n-1} \frac{(z_{i+1} - z_i)^2}{y_{i+1} - y_i} - 2(n-1)\sigma_\varepsilon^2\bar{\omega}$$

whose variance is:

$$\begin{aligned} \mathrm{var}_\varepsilon\left\{V_{TS}\right\} &\approx 4\sigma_\varepsilon^4 \left(3\bar{\omega}^2 - \bar{\omega}^2 + \sum_{i=2}^{n-1} \omega_i\omega_{i-1} \right. \\ &\quad \left. + \sum_{|i-j|\geq 2} \omega_i\omega_j \right) \end{aligned}$$

where $\omega_i = \dfrac{1}{y_{i+1} - y_i}$ and $\bar{\omega} = \dfrac{1}{n-1}\sum_{i=1}^{n-1} \omega_i$.
As $V_{TS}(D_n)$ is the sum of $(n-1)$ products of Gaussian variables, it could be proved that the $V_{TS}(D_n)$ distribution is asymptotically a Gaussian. In our case, n is sufficiently large so that the Gaussian assumption holds (it has been experimentally checked, see figure 1).

$$V_{TS} - V_{fc}(f^*) \sim \mathcal{N}\left(0, \mathrm{var}_\varepsilon\left\{V_{TS}\right\}\right)$$

Thanks to the statistical characterization of estimator V_{TS}, the conditional joint likelihood of the training set and the estimated $V_{TS}(D_n)$ may be calculated:

$$\begin{aligned} \mathscr{L}(D_n, V_{TS}(D_n)/\beta) \propto \prod_{i=1}^{n} \exp\left(-\frac{(z_i - f_\beta(y_i))^2}{2\sigma_\varepsilon^2} \right) \\ \exp\left(-\frac{\left(V_{fc}(f_\beta)) - V_{TS}(D_n)\right)^2}{2\mathrm{var}_\varepsilon\left\{V_{TS}\right\}} \right) \end{aligned}$$

The maximization of the log-likelihood leads to an estimation criterion for the required coefficients β:

$$\begin{aligned} \hat{\beta}(D_n) &= \arg\min_\beta \left[\sum_{i=1}^{n} (z_i - f_\beta(y_i))^2 \right. \\ &\quad \left. + \frac{\sigma_\varepsilon^2}{\mathrm{var}_\varepsilon\left\{V_{TS}\right\}} \left(V_{fc}(f_\beta)) - V_{TS}(D_n)\right)^2 \right] \end{aligned} \tag{4}$$

It is classically composed of a quadratic cost between the measured z_i and their estimated values (this is the empirical loss) and a penalization term. However, it may be noticed that contrary to classic regularization criteria, the penalization is calculated from the training set and that it does not depend on any hyperparameter.

3. APPLICATIONS

The variability criterion proposed was used for the approximation of the function $f^* : [0,1] \to \mathbb{R}$ defined by:

$$z = f^*(y) = \frac{1}{2}\sin(4\pi y) + \frac{1}{2}\sin\left(\frac{20\pi y}{3}\right)$$

The training set D_n was generated according to (1) with a value of σ_ε leading to a SNR of 10dB.
In order to evaluate the proposed approach, an experimental comparison between the constrained regularization (CR) given by (4) and other regression methods is performed: 1000 realizations of non-uniform training sets are tested. Thus, the risk of overfitting may be detected if outliers in the integrated losses (equation (2)) exist.
The CR method can be applied as long as the Gaussian assumption for the $V_{TS}(D_n)$ distribution is valid. Figure 1 shows the frequency histogram of the 1000 estimations of $V_{TS}(D_n)$ and the corresponding χ^2 test. It proves than the Gaussian assumption holds.

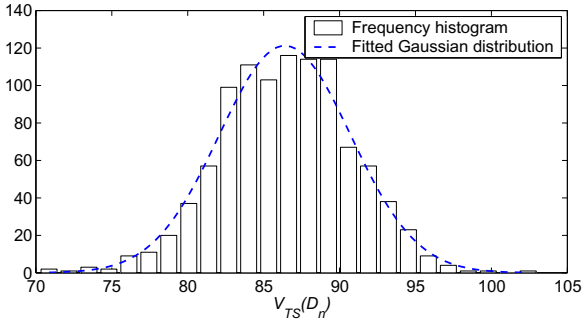

Figure 1: Frequency histogram of $V_{TS}(D_n)$ (calculated with the 1000 training sets). The dashed line indicates the fitted Gaussian used for the χ^2 test. The obtained P-value is 0.54.

3.1 Parametric Regression

The polynomial obtained by (4) was compared with the polynomial $P_{\theta'}$ estimated by a classic Tikhonov regularization method [10]:

$$\theta' = \arg\min_\theta \left[\sum_{i=1}^{n} (z_i - P_\theta(x_i))^2 + \lambda V_{fc}(P_\theta) \right]$$

where λ is chosen with the L-curve method [11]. It consists in plotting the empirical loss $\sum_{i=1}^{n} (z_i - P_\theta(x_i))^2$ versus the penalization term. The plot has an 'L-shape' when plotted on a *loglog* scale (see Figure 2). The location of the point of maximum curvature corresponds to the value of λ which gives the best tradeoff between both criteria.

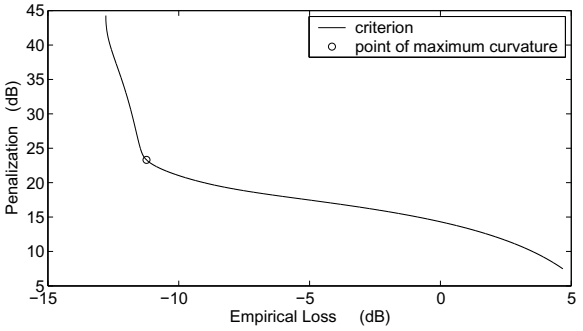

Figure 2: L-curve calculated with the training set of Fig 5.

Figure 3 and table 1 give the integrated losses for both methods. The proposed estimator outperforms the classic one since overfitting risk and mean error are smaller.

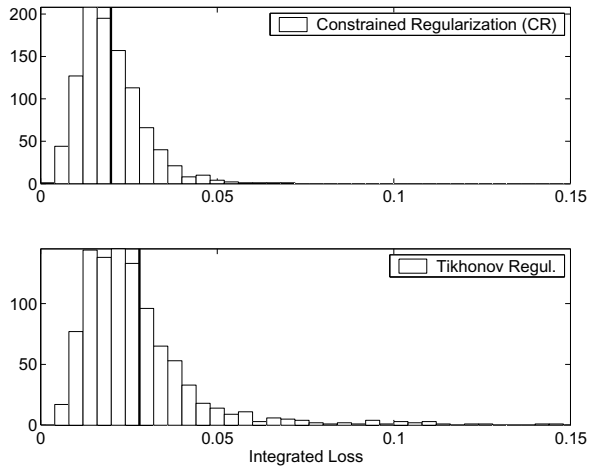

Figure 3: Polynomial regression: histograms of the integrated losses. The bold lines are the mean errors.

Table 1: Polynomial regression: statistics on the integrated losses (IL).

	Regular design		Random design	
values $\times 10^3$	CR	Tikhonov	CR	Tikhonov
$\langle IL \rangle$	16.8	19.3	19.9	28
$std(IL)$	6.7	8	8.9	22.1
IL_{max}	49.4	107.3	69.8	289.8

Figure 4 confirms this result: when comparing integrated losses for each simulation, the proposed method gives better results except in the cases where the Tikhonov regularization gives very small errors. Figure 5 gives an example when the L-curve method is not effective: even if the L-curve plot has a correct shape, the overfitting problem occurs. The problem is avoided with the Constrained Regularization.

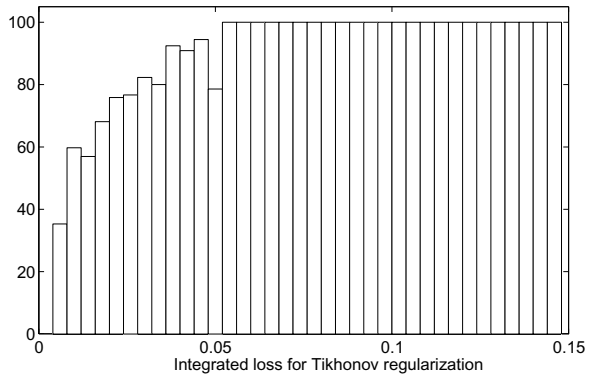

Figure 4: Empirical comparison between the L-curve method and the Constrained Regularization. The figure gives the number of cases (in percent) where the integrated loss of the proposed approach is smaller than the Tikhonov regularization integrated loss.

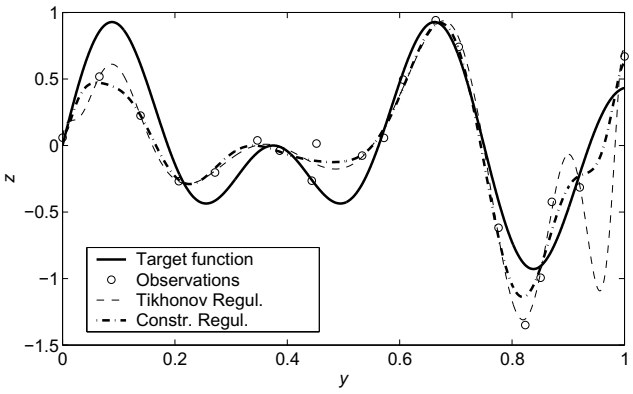

Figure 5: Example of polynomial regression (with $n = 20$, $d = 16$). For the Constrained Regularization: $IL = 0.0446$. For the Tikhonov regularization (with the L-curve method): $IL = 0.1047$.

3.2 Kernel Regression

The coefficient h of the Nadaraya-Watson estimator is chosen by the criterion (4). The result is compared with the estimator obtained with the optimal value of h:

$$h_{opt} = \arg\min_h \left[IL\left(g_h^{NW(D_n)} \right) \right]$$

Figure 6 and table 2 give the integrated losses obtained for the 1000 simulations (Figure 7 shows the simulation obtained with the training set of figure 5) with both estimated values of h. They prove that resolving (4) leads to a Nadaraya-Watson estimator close to the optimal one since error distributions are similar.

This result is confirmed by the relatives errors on h which are kept small (see Table 2):

$$\mathscr{E}_h(h_{est},\, h_{opt}) = \frac{|h_{est} - h_{opt}|}{h_{opt}}$$

Table 2: Nadaraya-Watson approximation: statistics on the integrated losses (IL) and the errors on h ($\mathscr{E}_h$).

	Regular design		Random design	
values $\times 10^3$	CR	optimal	CR	optimal
$\langle IL \rangle$	19	18.3	28.1	27
$std(IL)$	6.5	6.5	10.2	9.8
IL_{max}	49.8	45.8	84.2	74.4
$\langle \mathscr{E}_h \rangle$	9.5 %	×	12 %	×

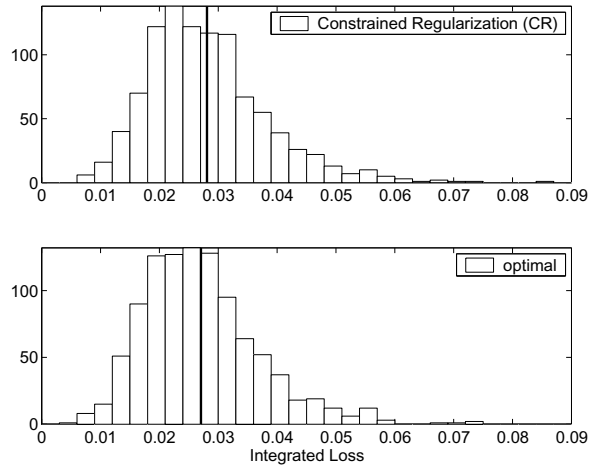

Figure 6: Nadaraya-Watson approximation: histograms of the integrated losses for $h = h_{est}$ and $h = h_{opt}$. The bold lines are the mean errors.

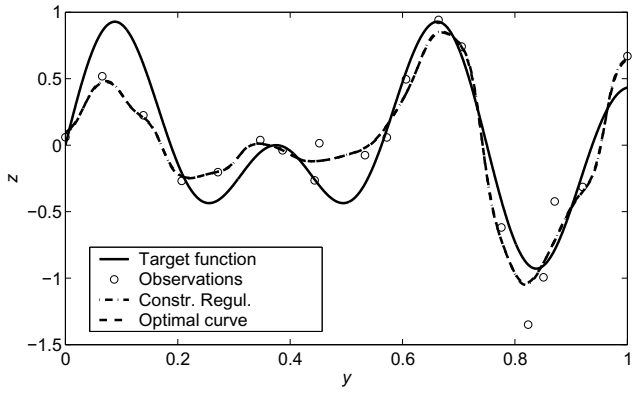

Figure 7: Nadaraya-Watson approximation with the training set of Fig. 5. For the Constrained Regularization: $h_{est} = 0.0287$, $IL = 0.0476$. For the optimal estimation: $h_{opt} = 0.0297, IL = 0.0475$.

4. CONCLUSION

The proposed approach is applied to regression techniques where a solution is sought within a training set. The main idea is to regularize the data fit process with prior information extracted from the training set data.
Thus, the variability of a training set (information linked to its regularity) is defined and used as a constraint on the searched solution. Indeed, forcing the model's variability to be close to the variability of the training set provides a solution that is compatible with the richness of the training set, thereby limiting the risk of overfitting.
Applying the proposed approach to scalar problems of parametric regression (polynomial regression) and nonparametric regression (kernel approximation) gives a series of robust solutions. The approach may be extended to other regression techniques, including non-scalar cases.
Works on new estimators of variability designed in the spectral field are underway. They aim at reducing the variance of the variability estimator.

REFERENCES

[1] V. Cherkassky and F. Mulier, *Learning from data*. Wiley, 1998.

[2] V. Vapnik, "An overview of statistical learning theory," *IEEE Transaction on Neural Networks*, Vol. 10, No. 5, Sept. 1999.

[3] C. Schaffer' "Overfitting Avoidance as Bias," *Machine Learning*, Vol. 10, No. 2, pp. 153–78, Feb 1993.

[4] M. Bekara, A. K. Seghouane and G. Fleury, "A small sample model selection criterion based on the kullback symetric divergence," in *IEEE Internationnal Conference on Acoustic Speech and Signal Processing ICASSP* , Vol. 6, pp. 145–148, 2003.

[5] A. E. Hoerl and R. Kennard,"Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 12, pp. 55-67, 1970.

[6] T. Poggio and F. Girosi, "Regularisation algorithms for learning that are equivalent to multilayer networks,"*Science*, No. 247, pp. 978–982, 1990.

[7] Y. Naji, L. Le Brusquet and G. Fleury, "Régression régularisée par contrainte de variabilité," in 19^{th} *Symposium on Image and Signal Processing GRETSI*, Vol. 2, pp. 88–91, Sept. 2003.

[8] E. A. Nadaraya, "On estimating regression," *Theory of probability and application*, Vol. 10, pp. 186–190, 1964.

[9] G. S. Watson, "Smooth regression analysis," *Sanklya Series*, A26, pp. 359–372, 1964.

[10] A. N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*. John Wiley, New York, 1977.

[11] P. C. Hansen, "Analysis of Discrete Ill-Posed Problems by means of the L-Curve," *SIAM review*, Vol. 34, pp. 561–580, 1992.