[3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp, *Sparsity oracle inequalities for the Lasso*, Electronic J. Statist., **1** (2007), 169–194.

[4] A. Dalalyan, and A.B. Tsybakov, *Aggregation by exponential weighting and sharp oracle inequalities*, Proc. 20th Annual Conf. Learning Theory (COLT-2007), Lecture Notes in Artificial Intelligence, **4539** (2007), 97–111, Springer, Berlin-Heidelberg.

[5] A. Dalalyan, and A.B. Tsybakov. *Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity*, Machine Learning, **72** (2008), 39–61.

[6] K. Lounici. *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*, Electronic J. Statist., **2** (2008), 90–102.

Analysis of Elastic-Net Regularization

Lorenzo Rosasco

(joint work with Christine De Mol, Ernesto De Vito)

In many learning problems, a major goal besides prediction is that of *selecting the variables* that are *relevant to achieve good predictions.* In the problem of variable selection we are given a set $(\varphi_\gamma)_{\gamma\in\Gamma}$ of functions from the input space $\mathcal{X}$ into the output space $\mathcal{Y}$ and we aim at selecting those functions which are needed to find a good representation of the regression function f^* on the basis of n input-output samples. In last decade many different algorithms have been introduced to solve such problem, such as forward stepwise regression, Lasso and greedy algorithms. However these procedures have drawbacks if there are highly correlated features. To overcome this problem, Zou and Hastie suggest a new method, called the elastic-net regularization [3]. In our work we study several properties of this estimation procedure with the setting of statistical learning (see [2] for details). In particular, we prove consistency for prediction and variable selection under some adaptive and non-adaptive choices for the regularization parameter. As an extension of the setting originally proposed in [3], our setting is random-design regression where we allow the response variable to be vector-valued and we consider prediction functions which are linear combination of elements (*features*) in an infinite-dimensional dictionary. The elastic-net scheme is defined by the minimization of the empirical risk penalized with a (weighted) elastic-net penalty, that is, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of i.i.d random pairs in $(\mathcal{X}, \mathcal{Y})$, the estimator vector β_n^λ is

$$\beta_n^\lambda = \operatorname*{argmin}_{\beta\in\ell_2} \frac{1}{n}\sum_{i=1}^{n} |Y_i - f_\beta(X_i)|^2 + \lambda \sum_{\gamma\in\Gamma}(w_\gamma|\beta_\gamma| + \varepsilon\beta_\gamma^2)$$

$$f_\beta = \sum_{\gamma\in\Gamma} \beta_\gamma \varphi_\gamma,$$

where $(w_\gamma)_{\gamma\in\Gamma}$ is a family of positive weights enforcing more or less sparsity, λ is a regularization parameter controlling the trade-off between the empirical error and the penalty, and ε is a tuning positive parameter that controls the trade-off between the ℓ_1-penalty (pure Lasso) and the ℓ_2-penalty (regularized least-squares regression). The ℓ_1-penalty has selection capabilities since it enforces sparsity of the solution, whereas the ℓ_2-penalty induces a linear shrinkage on the coefficients leading to stable solutions.

Under the assumption that the features satisfy $\sup_{x\in\mathcal{X}}\sum_{\gamma\in\Gamma}\|\varphi_{\gamma(x)}\|_{\mathcal{Y}}^2<\infty$ and the noise $Y_i - f^*(X_i)$ has exponential tails, that is,

$$\mathbb{E}\left[\exp\left(\frac{\|Y_i - f^*(X_i)\|_{\mathcal{Y}}}{L}\right) - \frac{\|Y_i - f^*(X_i)\|_{\mathcal{Y}}}{L} - 1\Big|X_i\right] \qquad \leq \frac{\sigma^2}{2L^2},$$

we prove that, if the regularization parameter $\lambda = \lambda_n$ satisfies $\lim_{n\to\infty}\lambda_n = 0$ and $\lim_{n\to\infty}(\lambda_n\sqrt{n} - 2\log n) = +\infty$, then

$$\lim_{n\to\infty}\|\beta_n^{\lambda_n} - \beta^\varepsilon\|_2 = 0 \qquad \text{with probability one,}$$

where the vector β^ε, which we call the *elastic-net representation* of f^*, is the minimizer of

$$\min_{\beta\in\ell_2}\left(\sum_{\gamma\in\Gamma} w_\gamma|\beta_\gamma| + \varepsilon\sum_{\gamma\in\Gamma}|\beta_\gamma|^2\right) \qquad \text{subject to} \qquad \sum_{\gamma\in\Gamma}\beta_\gamma\varphi_\gamma = f^*.$$

The vector β^ε exists and is unique provided that the regression function f^* admits a *sparse representation on the dictionary*, i.e. $f^* = \sum_{\gamma\in\Gamma}\beta^*_\gamma\varphi_\gamma$ for at least a vector $\beta^*\in\ell_2$ such that $\sum_{\gamma\in\Gamma} w_\gamma|\beta^*_\gamma|$ is finite. Notice that, when the features are linearly dependent, there is a problem of identifiability since there are many vectors β such that $f^* = \sum_{\gamma\in\Gamma}\beta_\gamma\varphi_\gamma$. The elastic-net regularization scheme forces $\beta_n^{\lambda_n}$ to converge to β^ε. As a consequence of the above convergence result, one easily deduces the consistency of the corresponding prediction function $f_n := \sum_{\gamma\in\Gamma}(\beta_n^{\lambda_n})_\gamma\varphi_\gamma$, that is, $\lim_{n\to\infty}\mathbb{E}[|f_n - f^*|^2] = 0$ with probability one. When the regression function does not admit a sparse representation, we can still prove the previous consistency result for f_n provided that the regression function is bounded and the linear span of the features is dense in $L^2(\mathcal{X}, Q, \mathcal{Y})$, where Q is the marginal distribution of X. Both the above convergence results are based on the fact that β_n^λ is the fixed point of the following contractive map

$$\beta = \frac{1}{\tau+\varepsilon\lambda}\mathbf{S}_\lambda\left(\tau I - \Phi_n^*\Phi_n)\beta + \Phi_n^* Y\right) \tag{1}$$

where τ is a suitable relaxation constant, $\Phi_n^*\Phi_n$ is the matrix with entries $(\Phi_n^*\Phi_n)_{\gamma,\gamma'} = \frac{1}{n}\sum_{i=1}^n < \varphi_\gamma(X_i), \varphi_{\gamma'}(X_i) >_{\mathcal{Y}}$, Φ_n^*Y is the vector $(\Phi_n^*Y)_\gamma = \frac{1}{n}\sum_{i=1}^n < \varphi_\gamma(X_i), Y_i >_{\mathcal{Y}}$. Moreover, $\mathbf{S}_\lambda(\beta)$ is the soft-thresholding operator acting componentwise as follows

$$[\mathbf{S}_\lambda(\beta)]_\gamma = \begin{cases} \beta_\gamma - \frac{\lambda w_\gamma}{2} & \text{if} \quad \beta_\gamma > \frac{\lambda w_\gamma}{2} \\ 0 & \text{if} \quad |\beta_\gamma| \leq \frac{\lambda w_\gamma}{2} \\ \beta_\gamma + \frac{\lambda w_\gamma}{2} & \text{if} \quad \beta_\gamma < -\frac{\lambda w_\gamma}{2} \end{cases}.$$

As a by-product of (1), β_n^λ has only a finite number of non-zero components, corresponding to the features whose weight satisfies $w_\gamma < \frac{C_n}{\lambda}$, where C_n is a known constant. Moreover β_n^λ can be computed by means of an iterative algorithm. This procedure is completely different from the modification of the LARS algorithm used in [3] and is akin instead to the algorithm developed in [1].

Finally, we use a data-driven choice for the regularization parameter, based on the so-called balancing principle, to obtain non-asymptotic bounds which are adaptive to the unknown regularity of the regression function. More precisely, letting $\lambda_k = \lambda_0 q^k$ be a geometric sequence with $q > 1$, we define

$$\lambda_n^+ = \max\{\lambda_k | \|\beta_n^{\lambda_j} - \beta_n^{\lambda_{j-1}}\|_2 \leq \frac{4D}{\sqrt{n}\varepsilon\lambda_{j-1}} \text{ for all } j = 0, \ldots, k\},$$

where D is a suitable constant. If β^ε is such that for some unknown $a \in (0,1)$ it satisfies the a-priori bound

$$\|\beta^\lambda - \beta^\varepsilon\|_2 = O(\lambda^a) \qquad \text{where}$$
$$\beta^\lambda = \operatorname*{argmin}_{\beta\in\ell_2} \mathbb{E}[\|Y - f_\beta(X)\|_{\mathcal{Y}}^2] + \lambda \sum_{\gamma\in\Gamma} (w_\gamma|\beta_\gamma| + \varepsilon\beta_\gamma^2),$$

then we prove that $\|\beta^{\lambda_n^+} - \beta^\varepsilon\|_2 = O(n^{-\frac{a}{2(a+1)}})$.

References

[1] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, 57(11):1413–1457, 2004.

[2] C. De Mol, E. De Vito, and L. Rosasco. Elastic-Net Regularization in Learning Theory. preprint arXiv:0807.3423 (July 2008)

[3] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67(2):301–320, 2005.

Spectral Regularization for Multi-task Learning

Massimiliano Pontil

(joint work with Andreas Argyriou, Charles Micchelli, Yiming Ying)

We are interested in the problem of learning multiple regression or classification functions (tasks) simultaneously. We present a method for learning a set of features which are shared across the tasks [1]. The method is based on a non-convex regularizer which encourages the number of such features to be small. We highlight the observation that the method is equivalent to solving a convex optimization problem, for which there is an iterative algorithm. The algorithm has a simple interpretation and converges to an optimal solution.

1. **Notation.** We begin by introducing our notation. We let $\mathbb{R}$ be the set of real numbers and $\mathbb{R}_+$ the subset of nonnegative ones. If $w, u \in \mathbb{R}^d$, we define $\langle w, u\rangle := \sum_{i=1}^d w_i u_i$ and $\|w\|_2 = \sqrt{\langle w, w\rangle}$. If A is a $d \times T$ matrix we denote by $a^i \in \mathbb{R}^T$ and $a_t \in \mathbb{R}^d$ the i-th row and the t-th column of A respectively. We denote by $\mathbf{S}_{++}^d$ the set of symmetric and positive definite matrices. If D is a $d \times d$ matrix, we define $\operatorname{trace}(D) := \sum_{i=1}^d D_{ii}$. If $w \in \mathbb{R}^d$, we denote by $\operatorname{Diag}(w)$ or $\operatorname{Diag}(w_i)_{i=1}^d$