Non Positive SVM

Gaëlle Loosli
Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand
CNRS, UMR 6158, LIMOS, F-63173 AUBIERE
gaelle@loosli.fr

Stéphane Canu
LITIS EA 4108, INSA-Université de Rouen, Saint-Etienne-du-Rouvray, 76801, France
scanu@insa-rouen.fr

Abstract

Learning SVM with non positive kernels is is a problem that has been addressed in the last years but it is not really solved : indeed, either the kernel is *corrected* (as a pre-treatment or via a modified learning scheme), either it is used with some well-chosen parameters that lead to almost positive-definite kernels. In this work, we aim at solving the actual problem induced by non positive kernels, *i.e.* solving the stabilization system in the Kreĭn space associated with the non-positive kernel. We first describe this stabilization system, then we expose a simple algorithm based on the eigen-decomposition of the kernel matrix. While providing satisfying solutions, the proposed algorithm shows limitations in terms of memory storage and computational effort. The direct resolution is still an open question.

1 Kreĭn Space and SVM

From the first stages of SVM [10] , non positive kernels are proposed and used, in particular the $tanh$ kernel. In many application fields, some huge efforts are made to produce *true Mercer kernels* when the natural kernels turn out to be indefinite [4, 3]. Some author even study some kernels that are definite positive with high probablity [1]. However, until now, there is no adequate solver available. In [7, 11, 2], the authors propose to solve SVM with indefinite kernel considering that the indefinite kernel is a perturbation of a true Mercer kernel. In [5], the author states that learning with indefinite symmetric kernels is actually consisting in finding a stationary point, which is not unique but each of those performs correct separation. It has been shown [8] that learning with non positive kernel is actually solving the learning problem in a Kreĭn space instead of a Hilbert space. It has also been shown that in this situation, the learning problem is not a minimization anymore but a stabilization problem. This means that the solution is a saddle point of the cost function. In the remaining of the section, we briefly introduce the Reproducing Kernel Kreĭn Space (RKKS) and propose the stabilization system to be solved to train SVM in Kreĭn space.

Reproducing Kernel Kreĭn Space Kreĭn spaces are indefinite inner product spaces endowed with a Hilbertian topology. We recall here definitions from [8]

Definition 1.1 *Inner Product Let $\mathcal{K}$ be a vector space on the scalar field. An inner product $\langle ., . \rangle_{\mathcal{K}}$ on $\mathcal{K}$ is a bilinear form where for all $f, g, h \in \mathcal{K}, \alpha \in \mathbb{R}$:*

$$\begin{array}{l} \langle f, g \rangle_{\mathcal{K}} = \langle g, f \rangle_{\mathcal{K}} \\ \langle \alpha f + g, h \rangle_{\mathcal{K}} = \alpha \langle f, h \rangle_{\mathcal{K}} + \langle g, h \rangle_{\mathcal{K}} \\ \langle f, g \rangle_{\mathcal{K}} = 0, \quad \forall g \in \mathcal{K} \implies f = 0 \end{array}$$

Definition 1.2 *Kreĭn space An inner product space $(\mathcal{K}, \langle .,. \rangle_{\mathcal{K}})$ is a Kreĭn space if there exists two Hilbert spaces $\mathcal{H}_+, \mathcal{H}_-$ spanning $\mathcal{K}$ such that*

$$\begin{array}{l} \forall f \in \mathcal{K}, f = f_+ + f_-, \textit{ where } f_+ \in \mathcal{H}_+ \textit{ and } f_- \in \mathcal{H}_- \\ \forall f, g \in \mathcal{K}, \langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-} \end{array}$$

In [8, Proposition 6], the reproducing property is shown : in $\mathcal{K}$ a RKKS, there is a unique symmetric $k(x, x')$ with $k(x, .) \in \mathcal{K}$ such that for all $f \in \mathcal{K}, \langle f, k(x, .) \rangle_{\mathcal{K}} = f(x)$ and $k = k_+ - k_-$.

SVM in RKKS Applying this to SVM requires to interpret the stabilization setting. Following [6], we start from the fact that a (unconstrained) quadratic program in a Kreĭn space has a unique solution (if the involved matrix is non singular) which is in general a stationary point. In the case of SVM, we have to apply some box constraints that may exclude this unique solution. Moreover, the optimal constrained solution is not necessarily unique anymore. Let $x_i \in \mathcal{X}^d, i \in [1..\ell]$ be ℓ training points in dimension d, along with their label $y_i \in [-1, 1]$ representing the class each point belongs to in a classification problem. Let $\mathcal{K}$ be a Kreĭn space. Let $f \in \mathcal{K}$ be our objective function.

Proposition 1.1 *The initial primal problem is :*

$$\left\{ \begin{array}{ccc} \textit{stab}_{f,b,\xi} & \frac{1}{2}\langle f, f \rangle_{\mathcal{K}} + C \sum_{i=1}^{\ell} \xi_i & \\ \textit{st} & y_i(f(x_i) + b) \geq 1 - \xi_i & \forall i \in [1..\ell] \\ & \xi_i \geq 0 & \forall i \in [1..\ell] \end{array} \right. \tag{1}$$

According to the Kreĭn space's properties, we can formulate system (1) into a min-max system

$$\left\{ \begin{array}{ccc} \min_{f_+,b,\xi} \max_{f_-} & \frac{1}{2}\langle f_+, f_+ \rangle_{\mathcal{H}_+} - \frac{1}{2}\langle f_-, f_- \rangle_{\mathcal{H}_-} + & C_i \sum_{i=1}^{\ell} \xi_i \\ \text{st} & y_i(f_+(x_i) + f_-(x_i) + b) \geq 1 - \xi_i & \forall i \in [1..\ell] \\ & \xi_i \geq 0 & \forall i \in [1..\ell] \end{array} \right. \tag{2}$$

Proposition 1.2 *We claim that system (2) can be solved with the following setting :*

$$\left\{ \begin{array}{ccc} \min_{f_+,f_-,b,\xi} & \frac{1}{2}\langle f_+, f_+ \rangle_{\mathcal{H}_+} + \frac{1}{2}\langle f_-, f_- \rangle_{\mathcal{H}_-} + & C_i \sum_{i=1}^{\ell} \xi_i \\ \textit{st} & y_i(f_+(x_i) + f_-(x_i) + b) \geq 1 - \xi_i & \forall i \in [1..\ell] \\ & \xi_i \geq 0 & \forall i \in [1..\ell] \end{array} \right. \tag{3}$$

Let's define $f_+(.) = \sum_i^{\ell} \eta_i k_+(x_i, .)$ and $f_-(.) = -\sum_i^{\ell} \eta_i k_-(x_i, .)$.

$$\left\{ \begin{array}{ccc} \min_{\eta,b,\xi} & \frac{1}{2}\eta^\top (K_+ + K_-)\eta + C_i \sum_{i=1}^{\ell} \xi_i & \\ \textit{st} & Y((K_+ - K_-)\eta + be) \geq e - \xi & \\ & \xi_i \geq 0 & \forall i \in [1..\ell] \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ccc} \min_{\eta,b,\xi} & \frac{1}{2}\eta^\top \tilde{K}\eta + C_i \sum_{i=1}^{\ell} \xi_i & \\ \textit{st} & Y(K\eta + be) \geq e - \xi & \\ & \xi_i \geq 0 & \forall i \in [1..\ell] \end{array} \right. \tag{4}$$

where K_+ (resp. K_-) is the kernel matrix provided by the k_+ (resp. k_-), Y is a diagonal matrix containing y and e a unit vector.

In the remaining of this position paper, we propose a way to find a decomposition of the non-positive matrix that can fit the previous problem and that produces an exact solution to the stabilization problem. First we show how to solve the stabilization system in the primal using a spectral decomposition. Then we use the same reasoning starting from a dual stabilization system. We show that both methods lead to the same final algorithm (EigNPSVM). Finally, we provide simple experiments illustrating the ability of EigNPSVM to solve exactly a non positive SVM and we point out several research direction that would confirm our claim and hopefully lead to a less-demanding algorithm (without spectral decomposition).

2 Decomposition of the primal system according to eigenvalues

Assume that $f(.) = \sum_{i=1}^{\ell} \beta_i k(x_i, .)$. Then system (1) can be written as

$$\left\{ \begin{array}{cc} \text{stab}_{\beta,b,\xi} & \frac{1}{2}\beta^\top K\beta + Ce^\top \xi \\ \text{st} & Y(K\beta + be) \geq e - \xi \\ & \xi_i \geq 0 \qquad \forall i \in [1..\ell] \end{array} \right. \tag{5}$$

where K is the kernel matrix, Y is a diagonal matrix containing y and e a unit vector.

Eigen decomposition The stabilization task means that we want to minimize according to the positive components and maximize according the negative components. We use spectral decomposition to identify those components. Let V be the column matrix of eigenvectors and Λ be the diagonal matrix of corresponding eigenvalues. We have $K = V\Lambda V^\top$. We sort V and Λ according the sign of eigenvalues, such that $V = [V_+, V_-]$ and $\Lambda = [\Lambda_+, 0; 0, \Lambda_-]$. Let us define $\gamma = V^\top \beta$. The stabilization process can be separated between a minimization of the positive part on the one hand and the maximization of the negative part on the other hand:

$$\left\{ \begin{array}{cc} \text{stab}_{\gamma,b,\xi} & \frac{1}{2}\gamma^\top \Lambda\gamma + Ce^\top \xi \\ \text{st} & Y(V\Lambda\gamma + be) \geq e - \xi \\ \forall i \in [1..\ell] & \xi_i \geq 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{cc} \min\limits_{\gamma_+,b,\xi} \max\limits_{\gamma_-} & \frac{1}{2}\gamma_+^\top \Lambda_+\gamma_+ + \frac{1}{2}\gamma_-^\top \Lambda_-\gamma_- + Ce^\top \xi \\ \text{st} & Y(V_+\Lambda_+\gamma_+ + V_-\Lambda_-\gamma_- + be) \geq e - \xi \\ \forall i \in [1..\ell] & \xi_i \geq 0 \end{array} \right. \tag{6}$$

Two separated problems Here we show how to transform the min-max system into a full minimization. If γ_- is already optimal and fixed:

$$\left\{ \begin{array}{cc} \min\limits_{\gamma_+,b,\xi} & \frac{1}{2}\gamma_+^\top \Lambda_+\gamma_+ + Ce^\top \xi \\ \text{st} & Y(V_+\Lambda_+\gamma_+ + V_-\Lambda_-\gamma_- + be) \geq e - \xi \\ \forall i \in [1..\ell] & \xi_i \geq 0 \end{array} \right. \Rightarrow \quad KKT \left\{ \begin{array}{c} \gamma_+^\top \Lambda_+ - \alpha^\top Y V_+\Lambda_+ = 0 \\ \alpha^\top Ye = 0 \\ Ce^\top - \alpha^\top - \eta^\top = 0 \\ \alpha_i \geq 0 \quad \forall i \in [1..\ell] \\ \eta_i \geq 0 \quad \forall i \in [1..\ell] \end{array} \right. \tag{7}$$

The same can be applied to the maximization part, admitting that γ_+, b, ξ are already optimal and fixed. In the same time, we transform max into min by changing the sign of the objective function.

$$\left\{ \begin{array}{cc} \min\limits_{\gamma_-} & -\frac{1}{2}\gamma_-^\top \Lambda_-\gamma_- \\ \text{st} & Y(V_+\Lambda_+\gamma_+ + V_-\Lambda_-\gamma_- + be) \geq e - \xi \end{array} \right. \Rightarrow \quad KKT \left\{ \begin{array}{c} \gamma_-^\top \Lambda_- + \alpha^\top Y V_-\Lambda_- = 0 \\ \alpha_i \geq 0 \quad \forall i \in [1..\ell] \end{array} \right. \tag{8}$$

Note that α are the same Lagrange multipliers as in system (7) since they apply to the same constraint of the original system (6).

Global minimization primal system System (7) and (8) are reassembled to produce a full minimization system, equivalent to the stabilization one (4).

$$\left\{ \begin{array}{cc} \min\limits_{\gamma,b,\xi} & \frac{1}{2}\gamma^\top \tilde{\Lambda}\gamma + Ce^\top \xi \\ \text{st} & Y(V\Lambda\gamma + be) \geq e - \xi \\ & \xi_i \geq 0 \qquad \forall i \in [1..\ell] \end{array} \right. \tag{9}$$

where $\tilde{\Lambda} = [\Lambda_+, 0; 0, -\Lambda_-]$. This system has the same shape as system (4). This let us think that the spectral decomposition is a good candidate to solve the stabilization problem in Kreĭn space.

Resolution via the dual From the primal problem (9), using KKT optimality conditions (7) and (8), the dual comes quite easily as

$$\left\{ \begin{array}{cc} \max\limits_{\alpha} & -\frac{1}{2}\alpha^\top \tilde{G}\alpha + \alpha^\top e \\ \text{st} & \alpha^\top y = 0 \\ & 0 \leq \alpha_i \leq C \qquad \forall i \in [1..\ell] \end{array} \right. \tag{10}$$

with $\tilde{G} = YV\tilde{\Lambda}V^\top Y$. To obtain this system, note that $\gamma = [V_+^\top Y\alpha, -V_-^\top Y\alpha]$. This is a classic quadratic program, providing a sparse solution α. To be able to classify unknown examples through the original kernel, we need to produce $\beta = V\gamma = V[V_+^\top Y\alpha, -V_-^\top Y\alpha]$ which can be arranged as $\beta = V\tilde{V}^\top Y\alpha$ with $\tilde{V} = [V_+, -V_-]$. One can remark that if the kernel is definite positive, V_- is empty, hence $\beta_i = y_i\alpha_i$ and $f(.) = \sum_i^\ell y_i\alpha_i k(x_i, .)$. Having a non positive kernel, the sparsity of the solution is lost. Moreover, the sign of β_i, which can be either positive or negative, is not only linked to the class of the corresponding example : examples can contribute *negatively* to the solution.

3 Decomposition of the dual system according to eigenvalues

We now show that we obtain the same resolution system when taking the dual point of view. The stabilization setting induces that the solution minimizes the cost function in some directions and maximizes it in others.

Proposition 3.1 *Let D_{min} represent the directions that should be minimized and D_{max} represent the directions that should be maximized. The optimality conditions at the saddle point can be written as follows, considering that we can decompose the problem into a min-max one (*ie. *minimize in D_{min} : lagrange multiplier are positive for a superiority constraint, maximize in D_{max} : lagrange multipliers are negative for a superiority constraint).*

$$eq.(1) \quad and \begin{cases} f(.) = \sum_{i=1}^{\ell} \alpha_i y_i k_{\mathcal{K}}(x_i, .) \\ \sum_{i=1}^{\ell} \alpha_i y_i = 0 \\ C - \alpha_i - \beta_i = 0, \quad \alpha_i \geq 0, \forall i | x_i \in D_{min} \\ C + \alpha_i - \beta_i = 0, \quad \alpha_i \leq 0, \forall i | x_i \in D_{max} \\ \beta_i \geq 0, \quad \{\forall i \in [1..\ell]\} \end{cases} \Rightarrow \begin{cases} stab_\alpha & \frac{1}{2}\alpha^\top G\alpha - e^\top\alpha & \\ st. & y^\top\alpha = 0 & \\ & 0 \leq \alpha_i \leq C & \forall i | x_i \in D_{min} \\ & -C \leq \alpha_i \leq 0 & \forall i | x_i \in D_{max} \end{cases} \tag{11}$$

where $G(i,j) = y_i y_j k_{\mathcal{K}}(x_i, x_j)$. However, the kernel matrix G is not definite positive and the dual system cannot be solved directly with classical technics. Moreover, we need to define the partition of the training points in D_{min} and D_{max}.

Eigen decomposition Let U be the column matrix of eigenvectors and Λ be the diagonal matrix of corresponding eigenvalues. We have $G = U\Lambda U^\top$, $U = [U_+, U_-]$ and $\Lambda = [\Lambda_+, 0; 0, \Lambda_-]$. Let us define $a = U^\top\alpha$. The multipliers' vector α can be written as $\alpha = Ua = U_+a_+ + U_-a_-$ with a the coordinates in the eigenvector space, a_+ (resp. a_-) the coordinates corresponding to positive (resp. negative) eigenvalues. We can rewrite the dual system:

$$\begin{cases} stab_a & \frac{1}{2}a^\top\Lambda a - e^\top Ua \\ \text{st.} & y^\top(Ua) = 0 \\ & 0 \leq U_+a_+ \leq Ce \\ & -Ce \leq U_-a_- \leq 0 \end{cases} \tag{12}$$

Let's note $\tilde{a} = [a_+; -a_-]$. We observe that $a_-^\top\Lambda_- a_- = (-a_-)^\top\Lambda_-(-a_-)$. Let $\tilde{\Lambda} = [\Lambda_+, 0; 0, -\Lambda]$.

$$\begin{cases} \min_{a_+}\max_{a_-} & \frac{1}{2}a_+^\top\Lambda_+a_+ + \frac{1}{2}a_-^\top\Lambda_-a_- - e^\top U_+a_+ - e^\top U_-a_- \\ \text{st.} & y^\top(U_+a_+ + U_-a_-) = 0 \\ & 0 \leq U_+a_+ \leq Ce \\ & -Ce \leq U_-a_- \leq 0 \end{cases} \Leftrightarrow \begin{cases} \min_a & \frac{1}{2}\tilde{a}^\top\tilde{\Lambda}\tilde{a} - e^\top U\tilde{a} \\ \text{st.} & y^\top U\tilde{a} = 0 \\ & 0 \leq U\tilde{a} \leq Ce \end{cases} \tag{13}$$

Algorithm EigNPSVM Solving the non positive SVM is quite simple in this setting : find $\tilde{\alpha}$ according to system (14), and then come back to α as follows: $\tilde{\alpha} = U\tilde{a}$, $U^\top\tilde{\alpha} = \tilde{a}$. Deduce a with $a = \tilde{a}$ and $a(m) = -\tilde{a}(m)$ where m indicate the position of negative eigenvalues, then $\alpha = Ua$. Let $\tilde{G} = U\tilde{\Lambda}U^\top$.

$$\begin{cases} \min_{\tilde{\alpha}} & \frac{1}{2}\tilde{\alpha}^\top\tilde{G}\tilde{\alpha} - e^\top\tilde{\alpha} \\ \text{st.} & y^\top\tilde{\alpha} = 0 \\ & 0 \leq \tilde{\alpha} \leq C \end{cases} \tag{14}$$

This algorithm is limited due to its need for eigenvalues/eigenvectors, which requires to compute the complete kernel and decompose it. We can reduce the complexity by computing an approximate version of the decomposition, taking only the largest absolute eigenvalues. The other limitation concerns the evaluation time in the original space, since the final solution α is not sparse (the solution is sparse only in $\tilde{\alpha}$).

4 Illustration, discussion and conclusion

With the *tanh* kernel. The hyperbolic tangent kernel $k(x_i, x_j) = tanh(a * x_i' x_j + b)$ is a popular kernel for SVM, even though it is not definite positive. It is well known [9] that for some range of parameters, one can find a solution which is correct in the sense of usual optimization, *ie. minimization*. For some very simple experiments on a checker dataset, we illustrate in figure 1 the ability of the proposed method to solve the stabilization problem for any range of parameter values of the *tanh* kernel. The resulting kernel matrix is highly non positive, the largest eigenvalues is negative ($\min_i(\lambda_i) = -124.80$, $\max_i(\lambda_i) = 19.31$). In this experiments, we also show that using only a partial eigen-decomposition of the kernel matrix leads to correct results.

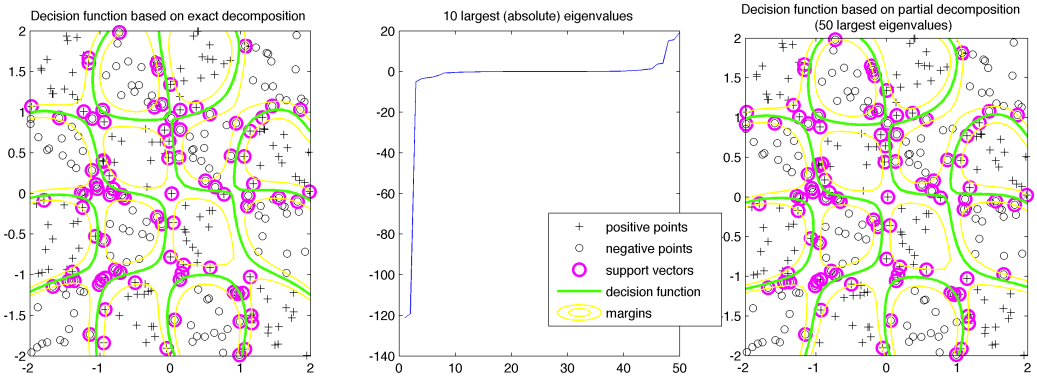

Figure 1: Results of EigNPSVM on a simple checker problem. The left figure shows the solution with a complete eigen decomposition of the kernel. The middle figure represents the 50 highest (in absolute values) eigenvalues from the kernel eigen decomposition. The right figure shows the resulting classifier using EigNPSVM with only the first 50 eigenvalues/eigenvectors. Those figures illustrate the ability of EigNPSVM to solve a highly non positive SVM. On both left and right figure, the support vectors are represented by large pink circles : they correspond to $\tilde{\alpha}$ in system (14).

Observations on the final solution The proposed algorithm has a non sparse final solution, for which coefficients α_i can be negative. We interpret this as an effect of the stabilization setting, in which some components contribute to the minimization and some others contribute to the maximization. If the min-max problem is quite clear once decomposed according to the sign of the eigenvalues, it is not easy to see in the original space. On figure 2, we show an interesting output of EigNPSVM. On this figure, we want to point out the fact that training points associated to positive α_i (resp. negative) are grouped together in the original space. This observation was a motivation to the definition of D_{min} and D_{max} in definition (**??**).

Discussion and conclusion In this on-going research, we want to solve the non positive SVM (and similar kernelized algorithms) using the stabilization setting, which is to our mind the actual problem to be solved. However the stabilization of a quadratic program is not a common task, even less when it is constrained. We know that a non positive kernel leads us to work in a Kreĭn space but so far it did not really help in the definition of a pertinent solver. In this study, we show that we can formulate the non positive SVM as a quadratic program in which examples can contribute either positively or negatively to the solution. The interpretation is that the cost function, to be stabilized, has to be minimized according to some components and maximized according to others. From the spectral analysis of the kernel, we deduce a simple algorithm that solves the stabilization problem which consists in modifying the kernel to make it positive definite, solving a classic quadratic program under box constraints, and using the eigenvector matrix to transpose the (sparse) solution into the original kernel space. This procedure has a cost : first, it requires to do the eigen-decomposition of

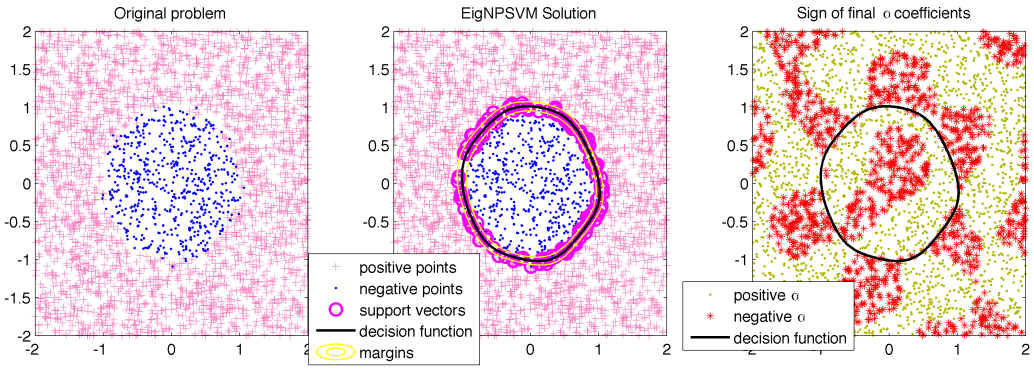

Figure 2: Observation on the positive or negative contribution of the final α_i in the solution. On the left figure, we draw the simple binary classification problem. On the middle one, the result of EigNPSVM is shown (using a $\tanh$ kernel). On the right figure, we represent the sign of the contribution α_i of each training point x_i for the learnt decision function. We observe that some areas contain only positive contribution while the rest contains only negative contribution: there is no overlap. Moreover, we can see that the label of the training points do not influence on those areas.

the kernel matrix, which implies that the kernel matrix has to be fully computed ; second, the final solution is non sparse and it slows down the test of new examples. The next step of this research is the definition of an algorithm that avoids the spectral analysis. To do so we think that we should use the observation that there are some areas of positive/negative contributing points in the example space : knowing those area, we could apply an active set algorithm. Moreover, we are convinced that there exists an equivalent sparse solution to the non sparse one we obtain.

References

[1] Sabri Boughorbel, Jean-Philippe Tarel, and Francois Fleuret. Non-mercer kernels for svm object recognition. In *In British Machine Vision Conference (BMVC*, pages 137–146, 2004.

[2] Jianhui Chen and Jieping Ye. Training svm with indefinite kernels. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 136–143. ACM, 2008.

[3] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 625–632. MIT Press, 2001.

[4] Corinna Cortes, Patrick Haffner, and Mehryar Mohri. Positive definite rational kernels. In *In Proceedings of The 16th Annual Conference on Computational Learning Theory (COLT 2003*, pages 41–56. Springer, 2003.

[5] Bernard Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:482–492, 2005.

[6] Babak Hassibi and Ali H. Sayed andThomas Kailath. *Indefinite-quadratic estimation and control: a unified approach to H2 and H [infinity] theories*, volume 16. 1999.

[7] R. Luss and A. d'Aspremont. Support Vector Machine Classification with Indefinite Kernels. *Mathematical Programming Computations*, 2009.

[8] Cheng Soon Ong, Xavier Mary, Stéphane Canu, and Alexander J. Smola. Learning with non-positive kernels. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 81, New York, NY, USA, 2004. ACM.

[9] Hsuan tien Lin and Chih-Jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical report, 2003.

[10] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y, 1995.

[11] Yiming Ying, Colin Campbell, and Mark Girolami. Analysis of svm with indefinite kernels. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2205–2213. 2009.