Chapter 1

b

T. Poggio[†], S. Mukherjee[†‡], R. Rifkin[†], A. Rakhlin[†], and A. Verri[†]
Center for Biological and Computational Learning[†]
Center for Genome Research, Whitehed Institute[‡]
Massachusetts Institute of Technology
Cambridge, MA 02139
tp@ai.mit.edu, {sayan, rif, rakhlin}@mit.edu, verri@ai.mit.edu

Abstract In this chapter we characterize the role of b, which is the constant in the standard form of the solution provided by the Support Vector Machine technique $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$, which is a special case of Regularization Machines. In the process, we describe properties of Reproducing Kernel Hilbert Spaces induced by different classes of kernels.

Keywords: Regularization, Support Vector Machines, a, b, Reproducing Kernel Hilbert Space

Introduction

Support Vector Machines (SVMs), originally introduced by Vapnik [Vapnik, 1995] in the context of Statistical Learning Theory, can be shown to be a special case of a regularization approach [Tikhonov and Arsenin, 1977] to the ill-posed problem of regression or classification from sparse and finite data [Evgeniou et al, 2000, Wahba, 1990]. The derivation of [Evgeniou et al, 2000, Girosi et al, 1990] makes it clear that SVMs and a large body of different learning and approximation techniques, known as Regularization Networks (RNs) [Girosi et al, 1995], can be obtained from the same general principles. Note that in the past [Evgeniou et al, 2000, Girosi et al, 1995] we have used the term RN to indicate the networks arising from regularization techniques, mainly involving the (classical) quadratic loss function. In this paper we use the term RN exclusively for the subset of regularization machines (RM) techniques that involve a quadratic loss functions and call Regularization Machines the broader set of techniques (see equation 1.1 later) that includes SVMs and RNs as special cases.

The aim of this chapter[1] is to clarify the role of b, the constant term in the solution to the learning problem obtained in the standard derivation of SVMs due to Vapnik [Vapnik, 1995] and found also in some RNs. These issues might have some impact for a) exploring the theoretical connection between SVMs, RNs, and other techniques (see [Girosi, 1998] for some of the difficulties arising in the connection between SVMs for regression and a special version of Basis Pursuit Denoising) and b) for developing efficient algorithms for SVMs. In the process, we will characterize properties of a Reproducing Kernel Hilbert Space (RKHS) induced by positive definite and conditionally positive definite kernels.

This chapter is organized as follows: we motivate this study in section 1, present our analysis in section 2, and then list remarks and conclusions. Throughout the chapter we assume some familiarity with both SVMs and RNs as presented in [Evgeniou et al, 2000].

1. Motivation

Given ℓ training pairs $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_\ell, y_\ell)\}$ a regularized solution – that we call a *Regularization Machine* – to the learning problem is found by minimizing the following functional for fixed λ

$$I[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2 \tag{1.1}$$

corresponding to the minimization of the empirical loss – the first term – under capacity control – the second term. The choice of the loss function V determines the learning scheme. In classical (quadratic) Regularization Networks: $V(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$, in SVM Classification: $V(y_i, f(\mathbf{x}_i)) = |1 - y_i f(\mathbf{x}_i)|_+$ where $|x|_+ = \max\{x, 0\}$, and in SVM Regression: $V(y_i, f(\mathbf{x}_i)) = |y_i - f(\mathbf{x}_i)|_\epsilon$ where $|x|_\epsilon \equiv \max\{0, |x| - \epsilon\}$ is called ϵ-insensitive loss.

The term $\|f\|_K$ is the norm of the function f in the RKHS induced by a positive definite kernel K [Aronszajn, 1950, Wahba, 1990]. It has been known for some time (see [Girosi and Poggio, 1990] and for an unusually elegant proof [Schölkopf et al, 2001]) that the minimizer of $I[f]$ for rather general $V(\cdot)$ and in particular for RNs [Wahba, 1990, Girosi et al, 1995] and SVMs [Vapnik, 1998, Girosi, 1998, Evgeniou et al, 2000, Lin et al, 2000] belongs to the RKHS induced by K and can be written as

[1] Besides entering the Guinness Book of World Records for the shortest title.

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

for some coefficients α_i which depend on the ℓ examples and on λ. In the original formulation of SVMs due to Vapnik, the minimizer is actually written as

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{1.2}$$

with $-\frac{1}{2\lambda\ell} \leq \alpha_i \leq \frac{1}{2\lambda\ell}$ and

$$\sum_{i=1}^{\ell} \alpha_i = 0. \tag{1.3}$$

The offset parameter b is estimated (like the coefficients α_i) from the ℓ examples. From the dual formulation of the optimization problem of SVMs (see for example [Vapnik, 1995]), it can be seen that the equality constraint (1.3) is induced[2] by the form of the solution assumed in equation (1.2).

The case of standard RNs (quadratic V) is well known (see [Wahba, 1990] or [Girosi and Poggio, 1990]). If the kernel is a positive definite function (i.e., in the case of Gaussian Radial Basis Functions) one chooses the solution in the linear span of the RKHS written as $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ without any constraint on the α_i. If the kernel is a conditionally positive definite function of order 1 (i.e., in the case of piecewise linear splines) a constant term, b, is added to the solution which becomes $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$, subject to the same constraint (1.3). As shown in [Evgeniou et al, 2000] the solution $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$ with the constant b (and the equality constraint in (1.3)) can be used also in the case of a positive definite kernel K[3]

This chapter is devoted to answering the following questions: When should b be used? Is there a choice of using or not using b? What does the choice mean? Are the answers different for RNs and SVMs? How does this relate to properties of the kernel K?

[2] This property of b was not discussed in [Vapnik, 1995].

[3] This choice – rather than the usual one without b – effectively corresponds to a different (semi)norm and to a different RKHS. Equation (4.10) in [Evgeniou et al, 2000] is incorrect and should be replaced with $(K' + l\lambda I)\alpha + \mathbf{1}b = (K + l\lambda I)\alpha + \mathbf{1}b = \mathbf{y}$. The key equation (4.12) and the conclusions are however correct. This error (noticed by Steve Smale) propogates throught several other equations in [Evgeniou et al, 2000] rendering them incorrect and confusing: λ should be replaced by λl in Equations (4.3, 4.6, 4.19).

2. Analysis

2.1 Definitions

First we define (conditional) positive definiteness of kernels. More properties of shift invariant kernels, $K(x,y) = K(x-y)$, can be found in [Schoenberg, 1998, Berg et al, 1984, Micchelli, 1986].

Let X be some set, for example a subset of $\mathbb{R}^d$ or $\mathbb{R}^d$ itself. A *kernel* is a symmetric function $K : X \times X \to \mathbb{R}$.

Definition 2.1

A kernel $K(\mathbf{t}, \mathbf{s})$ *is* positive definite (pd) *if* $\sum_{i,j=1}^{n} c_i c_j K(\mathbf{t}_i, \mathbf{t}_j) \geq 0$ *for any* $n \in \mathbb{N}$ *and choice of* $\mathbf{t}_1, ..., \mathbf{t}_n \in X$ *and* $c_1, ..., c_n \in \mathbb{R}$.

An equivalent definition could be given in terms of positive semidefiniteness of the matrix $K_{ij} = K(\mathbf{t}_i, \mathbf{t}_j)$. A *pd* kernel is *strictly positive definite* if for any distinct vectors $\mathbf{t}_1, ..., \mathbf{t}_n \in X$ the above inequality holds strictly when the c_i are not all zero (in that case the matrix K_{ij} is positive definite and not just positive semidefinite).

Definition 2.2

A kernel $K(\mathbf{t}, \mathbf{s})$ *is* conditionally positive definite (cpd) *of order 1 if* $\sum_{i,j=1}^{n} c_i c_j K(\mathbf{t}_i, \mathbf{t}_j) \geq 0$ *for any* $n \in \mathbb{N}$ *and choice of* $\mathbf{t}_1, ..., \mathbf{t}_n \in X$ *and* $c_1, ..., c_n \in \mathbb{R}$ *subject to the constraint* $\sum_{i=1}^{n} c_i = 0$. *It is* strictly conditionally positive definite *if* $\sum_{i,j=1}^{n} c_i c_j K(\mathbf{t}_i, \mathbf{t}_j) > 0$.

2.2 Integral operators

We consider the integral operator L_K on $L_2(X, \nu)$ defined by

$$\int_X K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\nu(\mathbf{x}') = g(\mathbf{x}) \tag{1.4}$$

where X is a compact subset of $\mathbb{R}^n$ and ν a Borel measure. We assume K to be continuous. Thus the integral operator is compact [Cucker and Smale, 2001]. Note that K pd (definition 2.1) is equivalent [Mercer, 1909] to L_K positive that is

$$\int_X K(\mathbf{t}, \mathbf{s}) f(\mathbf{t}) f(\mathbf{s}) d\nu(\mathbf{t}) d\nu(\mathbf{s}) \geq 0 \tag{1.5}$$

for all $f \in L_2(X, \nu)$.

2.3 Mercer's theorem

The key tool in our analysis is the result published by Mercer in 1909 [Mercer, 1909, Courant and Hilbert, 1962].

Theorem 2.1 *A symmetric,* pd *kernel* $K : X \times X \to \mathbb{R}$*, with* X *a compact subset of* $\mathbb{R}^n$ *has the expansion*

$$K(\mathbf{x},\mathbf{x}') = \sum_{q=1}^{\infty} \mu_q \phi_q(\mathbf{x})\phi_q(\mathbf{x}') \qquad (1.6)$$

where the convergence is in $L_2(X,\nu)$. If the measure ν on X is non-degenerate in the sense that open sets have positive measure everywhere, then the convergence is absolute and uniform and the $\phi(\mathbf{x})$ are continuous on X (see [Cucker and Smale, 2001] and Smale, pers. comm.). The ϕ_q are the orthonormal eigenfunctions of the integral equation

$$\int_X K(\mathbf{x},\mathbf{x}')\phi(\mathbf{x})d\nu(\mathbf{x}) = \mu\,\phi(\mathbf{x}'). \qquad (1.7)$$

Using Mercer's theorem we distinguish three cases, depending on the properties of the kernel.

1 *Strictly positive case:* the kernel is *strictly pd* and all eigenvalues (note that there are an infinite number of eigenvalues) of the integral operator L_K are strictly positive.

2 *Degenerate case:* the kernel $K(\mathbf{t},\mathbf{s})$ positive definite but only a finite number of eigenvalues of the integral operator L_K are strictly positive, the rest being zero (see [Courant and Hilbert, 1962]).

3 *Conditionally strictly positive case:* the kernel $K(\mathbf{t},\mathbf{s})$ is *conditionally positive definite* and all the eigenvalues of the integral operator L_K are positive with only a finite number being non-positive. Notice that for cpd kernels, the kernel K can be made into a positive definite kernel K' by subtracting the terms $\mu_q\phi_q(\mathbf{x})\phi_q(\mathbf{x}')$ belonging to negative eigenvalues [Courant and Hilbert, 1962].

2.4 Reproducing Kernel Hilbert Spaces

The RKHS induced by K is equivalent (see [Cucker and Smale, 2001]) to the Hilbert space of the functions spanned by $\Phi = \{\phi_1(\mathbf{x}), ...\}$,

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} c_q \phi_q(\mathbf{x}),$$

equipped with the scalar product $< f, g >= \sum_{i=1}^{\infty} \frac{c_q d_q}{\mu_q}$ where $f(\mathbf{x}) = \sum_{i=1}^{\infty} c_q\phi_q(\mathbf{x})$ and $g(\mathbf{x}) = \sum_{i=1}^{\infty} d_q\phi_q(\mathbf{x})$, and finite norm in the RKHS $\|f\|_K^2 = \sum_{q=1}^{\infty} c_q^2/\mu_q$, where the sums for the norm and scalar products are over terms with nonzero μ_q. Note that it is possible to prove directly that the RKHS is independent of the measure ν (assumed positive everywhere), as observed by Smale and Cucker (though the $\phi(\mathbf{x})$ and the μ_q in equation (1.7) are not).

2.5 Density of a RKHS

We characterize density properties of a RKHS. In particular, we ask under which condition is a RKHS dense in $L_2(X,\nu)$ or $C(X)$[4]. The answer below was developed starting from separate observations by Zhou [personal communication], Girosi [personal communication] and Smale [personal communication]. (In the following we assume ν to be the Lebesgue measure.) The following statements follow for the three cases above:

1 in the *strictly positive case* the RKHS is infinite dimensional and dense in $L_2(X,\nu)$. It is also dense in $C(X)$ with the topology of uniform convergence (Zhou, in preparation).

2 in the *degenerate case* the RKHS is finite dimensional and not dense in $L_2(X,\nu)$; the null space of the operator L_K is infinite dimensional.

3 in the *conditionally strictly positive case* the RKHS associated with K' is infinite dimensional and the null space of the operator L_K is finite dimensional. The RKHS is not dense in $L_2(X,\nu)$ but when completed with a finite number of polynomials of appropriate degree can be made to be dense in $L_2(X,\nu)$ and in $C(X)$.

2.6 Regularization Networks (including SVMs) for regression:

In regression, given sparse data it is natural and desirable to be able to approximate the unknown function under the most general conditions, such as all functions in $L_2(X,\nu)$. From this perspective we look at possible solutions of (1.1) for the three cases above.

1 *Strictly positive case*: in this case the solution

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x},\mathbf{x}_i) \tag{1.8}$$

is dense in $L_2(X,\nu)$ and in $C(X)$: b is not needed. Note that this is a different kind of SVM from the one originally proposed by Vapnik, even if Vapnik's loss functions are used. However, the solution with b that Vapnik originally proposed, is also valid since a positive definite kernel K is also *cpd*. It is easy to check (see the following cpd case) that using the solution with b is equivalent to using the cpd kernel $K'(\mathbf{x},\mathbf{y}) = K - \mu_1\phi_1(\mathbf{x})\phi_1(\mathbf{y})$ in the stabilizer

[4] $C(X)$ is dense in $L_2(X,\nu)$

term of equation (1.1) with a solution $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha'_i K'(\mathbf{x}, \mathbf{x}_i) + b$. Somewhat surprisingly, it follows that

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha'_i K'(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^{\ell} \alpha'_i K(\mathbf{x}, \mathbf{x}_i) + b. \tag{1.9}$$

Thus, in this case, both solutions (1.8) and (1.9) are dense in $L_2(X, \nu)$. Notice that they correspond, respectively, to the minimizers of following functionals, each one using a different prior on the function space

$$\begin{aligned} I[f] &= \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_K^2 \\ I[f] &= \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{K'}^2. \end{aligned}$$

In the RN case with quadratic $V(\cdot)$ the minimization of the two different $I[f]$ yields the linear equations (see [Evgeniou et al, 2000]): $(K + \ell\lambda I)\alpha = \mathbf{y}$ and $(K + \ell\lambda I)\alpha' + \mathbf{1}b = \mathbf{y}$ subject to $\mathbf{1}\alpha = 0$. Thus in the standard RN case it is possible to compute one solution from the other since $\alpha - \alpha' = (K + \ell\lambda I)^{-1}\mathbf{1}b$. In the SVM case the two different solutions correspond to the minima of two different QP problems and the relation between α and (α', b) cannot be given in closed form.

2 *Degenerate case*: in this case the regularization solution is not dense in $L_2(X)$ with or without the addition of a polynomial of finite degree. In other words, with a finite dimensional kernel it is in general impossible to approximate arbitrarily well a continuous function on a bounded interval. This is the case for polynomial kernels of the form $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^n$ often used in SVMs. The use of b here is therefore even more arbitrary (or more dependent on prior knowledge about the specific problem), since it does not restore density.

3 *Conditionally strictly positive case*: in this case [5] the solution

$$f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^{\ell} \alpha_i K_{pd}(\mathbf{x}, \mathbf{x}_i) + b$$

is dense in $L_2(X, \nu)$ and also on $C(X)$ when the b term is included. We define the *pd* kernel $K_{pd} = K - \mu_0 \phi_1 \phi_1$ where $-\mu_0 \phi_1 \phi_1$ is a positive constant (we assume here, since we are dealing with cpd kernels of degree 1, that $\phi_1(\mathbf{x})\phi_1(\mathbf{x}')$ corresponds to a constant). The

[5] For simplicity we consider in the following cpd kernels of order 1 only.

stabilizer term in $I[f]$ is then formally interpreted (compare [Aronszajn, 1950]) as $\|f\|_K^2 = \sum_{q=2}^{\infty} \frac{c_q^2}{\mu_q}$, that is as $\|f\|_K^2 = \|f\|_{K_{pd}}^2 = \|P_{K_{pd}} f\|_{K_{pd}}^2$, where $P_{K_{pd}}$ projects f into the RKHS induced by K_{pd} (see [Wahba, 1990]). To obtain solutions that are dense in $L_2(X, \nu)$ we consider solutions of the form $f(\mathbf{x}) = \sum_{q=2}^{\infty} c_q \phi_q(\mathbf{x}) + b$. These are functions in the RKHS induced by the *pd* kernel K_{pd} completed with the constants (which are not in the RKHS). Taking derivatives of $I[f]$ with respect to the coefficients c_q and b and setting them equal to zero (following [Girosi, 1998]) we get $c_q = \mu_q \sum_{i=1}^{\ell} \alpha_i \phi_q(\mathbf{x}_i)$ where $\alpha_i = \frac{1}{\lambda} V'(y_i - f(\mathbf{x}_i))$, subject to the constraint

$$\sum_{i=1}^{\ell} \alpha_i = 0. \tag{1.10}$$

Therefore the minimizer of $I[f]$ is

$$\begin{aligned} f(\mathbf{x}) &= \sum_{q=2}^{\infty} c_q \phi_q(\mathbf{x}) + b = \sum_{i=1}^{\ell} \alpha_i \sum_{q=2}^{\infty} \mu_q \phi_q(\mathbf{x}_i) \phi_q(\mathbf{x}) + b \\ &= \sum_{i=1}^{\ell} \alpha_i K_{pd}(\mathbf{x}, \mathbf{x}_i) + b = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \end{aligned}$$

where we have used expansion 1.6 and, in the last step, constraint 1.10. By interpreting $V'(\cdot)$ in a generalized sense, the proof is valid for a broad class of $V(\cdot)$ (see [Girosi et al, 1990]) and for the non-differentiable $V(\cdot)$ used by SVM regression and classification see appendix B.2 of [Girosi, 1998]. Thus cpd kernels can be used not only for standard RNs but also for SVMs[6]. In both cases the term b is needed in the solution in order to approximate functions in $L_2(X, \nu)$ or $C(X)$.

3. Regression and Classification: b or not b?

Let us consider here only the *strictly positive case* for K. In regression the general solution does not have a b term. However, it is possible to use a positive definite K *and* the constant b. The latter choice is effectively the choice of a different kernel and a different feature space relative to the initial K used in the standard solution without b: the constant feature ”disappears” from the RKHS norm and therefore is not “penalized”.

[6] This extension – well known for regularization since at least a decade – was correctly suggested in [Smola et al, 1998] in an otherwise confusing paper.

This choice may be reasonable but only when specific prior information is available about the problem. For instance, there may be regression problems in which shifts of f by a constant should not be penalized. This is especially true for the binary classification framework originally considered by Vapnik. Only the sign of the function f found by the SVM is used for classification. A constant b plays therefore the role of a threshold; using a solution of the form $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$, without penalizing b in the stabilizer, corresponds to the reasonable assumption that there is no privileged value – such as 0 – for the classification threshold.

Acknowledgments

We wish to thank Federico Girosi for many discussions and a legacy of insights, Steve Smale for very useful suggestions and Ding-Xie Zhou for quickly providing the answer to our question on conditions for density of RKHS and for much help afterwards. We acknowledge the comments of some members of the NIPS community, showing unsurprising ignorance of relevant, older classical mathematical literature.

References

[Aronszajn, 1950] N. Aronszajn "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 686, pp. 337-404, 1950.

[Berg et al, 1984] C. Berg, J.P.R. Christensen, and P. Ressel *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions* Springer, 1984.

[Courant and Hilbert, 1962] R. Courant and D. Hilbert *Methods of Mathematical Physics, Vol 2* Interscience, 1962.

[Cucker and Smale, 2001] F. Cucker and S. Smale On the Mathematical Foundations of Learning, *Bull. Amer. Math. Soc.*, vol. 39, pp. 1–49, 2002.

[Evgeniou et al, 2000] T. Evgeniou, M. Pontil, and T. Poggio "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, vol. 13, pp. 1–50, 2000.

[Freiss et al, 1998] T. Friess, N. Cristianini, and C. Campbell "The Kernel-Adatron: A fast and simple learning procedure for Support Vector Machines," *Proceedings of the 15th International Conference in Machine Learning*, pages 188-196, 1998.

[Girosi, 1998] F. Girosi "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, pp. 1455–1480, 1998.

[Girosi et al, 1990] F. Girosi and T. Poggio and B. Caprile Extensions of a Theory of Networks for Approximation and Learning: outliers and Negative Examples, Advances in Neural information processings systems 3, R. Lippmann and J. Moody and D. Touretzky, Morgan Kaufmann, 1991, San Mateo, CA.

[Girosi and Poggio, 1990] F. Girosi and T. Poggio "Networks and the Best Approximation Property," *Biological Cybernetics*, vol. 63, pp. 169–176, 1990.

[Girosi et al, 1995] F. Girosi, M. Jones, and T. Poggio "Regularization theory and neural network architectures," *Neural Computation,* vol. 7, pp. 219–269, 1995.

[Lin et al, 2000] Y. Lin, Y. Lee, and G. Wahba *Support Vector Machines for Classification in Nonstandard Situations* TR 1016, March 2000. To appear, Machine Learning

[Mercer, 1909] J. Mercer "Functions of positive and negative type and their connection with the theory of integral equations," *Phil.los. Trans. Roy. Soc. London Ser. A*, vol. 209, pp. 415–446, 1909.

[Micchelli, 1986] C.A. Micchelli "Interpolation of scattered data: distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, pp. 11–22, 1986.

[Poggio et al., 2001] T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin and A. Verri " b," CBCL Paper 198/AI Memo 2001-011, Massachusetts Institute of Technology, Cambridge, MA, July 2001

[Schoenberg, 1998] I.J. Schoenberg "Metric spaces and completely monotone functions" *Ann. of Math.*,vol. 39, pp.811–841, 1938.

[Schölkopf et al, 2001] B. Schölkopf and A. Smola and R. Herbrich "A Generalized Representer Theorem" *Computational Learning Theory*, vol.14, pp. 416–426, 2001.

[Smola et al, 1998] A. Smola and B. Schölkopf and K.R. Müller "The connection between Regularization Operators and Support Vector Kernels" *Neural Networks*, vol.11, num. 4, pp. 637–649, 1998.

[Tikhonov and Arsenin, 1977] A. N. Tikhonov and V. Y. Arsenin Solutions of Ill-posed Problems W. H. Winston, 1977

[Vapnik, 1995] V.N. Vapnik The Nature of Statistical Learning Theory Springer, 1995.

[Vapnik, 1998] V.N. Vapnik Statistical Learning Theory Wiley, 1998.

[Wahba, 1990] G. Wahba *Spline models for observational data* SIAM, 1990.