Preprints of the 18th IFAC World Congress
Milano (Italy) August 28 - September 2, 2011

Learning using state space kernel machines ⋆

Aleksandr Y. Aravkin * **Bradley M. Bell** ** **James V. Burke** ***
Gianluigi Pillonetto ****

* *Department of Earth and Ocean Sciences, University of British Columbia, Vancouver BC, CA (e-mail: saravkin@eos.ubc.ca).*
** *Applied Physics Laboratory, University of Washington, Seattle WA, USA(e-mail: bradbell@apl.washington.edu)*
*** *Department of Mathematics, University of Washington, Seattle, WA, USA, (e-mail: burke@math.washington.edu)*
**** *Department of Information Engineering, University of Padova, Padova, Italy, (e-mail: giapi@dei.unipd.it)*

Abstract: Reconstruction of a function from noisy data is often formulated as a regularized optimization problem whose solution closely matches an observed data set and also has a small reproducing kernel Hilbert space norm. The loss functions that measure agreement between the data and the function are often smooth (e.g. the least squares penalty), but non-smooth loss functions are of interest in many applications. Using the least squares penalty, large machine learning problems with kernels amenable to a stochastic state space representation (which we call *state space kernel machines*) have been solved using a Kalman smoothing approach. In this paper we extend this approach for state space kernel machines to the Vapnik penalty, a particularly important non-smooth penalty that is robust to outlier noise in the measurements and induces a sparse representation of the reconstructed function. We exploit the structure of such models using interior point methods that efficiently solve the functional recovery problem with the Vapnik penalty, and demonstrate the effectiveness of the method on a numerical experiment. The computational effort of our approach scales linearly with the number of data points.

Keywords: kernel based regularization; Kalman smoothing; Gaussian processes; statistical learning; estimation theory; regularization networks; support vector regression; robust estimators

1. INTRODUCTION

Minimizing a regularization functional in a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ associated with a symmetric and positive-definite kernel K has become a popular approach to reconstruct a scalar function from noisy data, e.g. see (Aronszajn, 1950; Schölkopf and Smola, 2001; Wahba, 1990; Girosi et al., 1995). In particular, regularization in $\mathscr{H}$ estimates the unknown function as

$$\tilde{f}(\cdot) = \arg\min_{f(\cdot)} \sum_{i=1}^{N} \rho\left[z_i - f(t_i)\right] + \gamma \|f(\cdot)\|^2_{\mathscr{H}}, \qquad (1)$$

where $\gamma \in \mathbb{R}^+$ is the regularization parameter, t_i is the location where the measurement z_i is collected, $\rho(\cdot)$ is the so called loss function and $\|\cdot\|_{\mathscr{H}}$ is the RKHS norm induced by the reproducing kernel $K(\cdot,\cdot)$, see (Aronszajn, 1950). The solution of (1) can also be interpreted in terms of Bayesian estimation. Under suitable assumptions on the noise model, it represents the maximum a posteriori estimate of a zero-mean Gaussian random field of covariance proportional to $K(\cdot,\cdot)$ (Hengland, 2007). This connection is described in more detail in Section 2. One of the important features of the above approach is that, even if the dimension of $\mathscr{H}$ is infinite, the solution is a linear combination of a finite number of kernel functions centered at the points $\{t_i\}$. In fact, under mild assumptions on the loss, according to the representer theorem (Wahba, 1998; Schölkopf et al., 2001) solutions to (1) can be expressed as kernel sections centered on the sampling locations, i.e.

$$\tilde{f}(\cdot) = \sum_{i=1}^{N} \tilde{c}_i K(t_i, \cdot), \qquad (2)$$

where $\tilde{c}_i$ are suitable scalars. For instance, (2) holds when $\rho(\cdot)$ is quadratic, defining a regularization network (Poggio and Girosi, 1990), or is the Vapnik's ε-insensitive loss, leading to support vector regression (Vapnik, 1998; Evgeniou et al., 2000; Gunter and Zhu, 2006). In particular, the Vapnik penalty has recently been object of considerable attention, due to its sparsity inducing properties (many $\tilde{c}_i$ in (2) turn out to be zero) and its robustness with respect to outliers relative to the quadratic loss. In many scientific fields such as telecommunications, biology and biomedicine, one is faced with the problem of reconstructing the (possibly multi-dimensional) input of a dynamic linear system from noisy output data (Bertero, 1989; De Vito et al., 2005). In addition, the system model is often formulated in state space (Pillonetto and Saccomani, 2006). Without loss of generality, in this paper we focus on single input single output systems, restricting our attention to monodimensional regression problems. We assume that the kernel $K(\cdot,\cdot)$ is given in state space form, see e.g. Section 4 in (De Nicolao and Ferrari Trecate, 2003), and for such problems we use kernel-based regularization to reconstruct $f(\cdot)$ using (1). More precisely, we model $f(\cdot)$ as a realization of a nonstationary scalar Gaussian process (sampled at times $\{t_i\}$) defined as follows:

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

$$\begin{cases} \dot{x}(t) &= F(t)x(t)+G(t)dw(t) \\ f(t) &= H(t)x(t), \end{cases} \tag{3}$$

where t is the time, $x(t) \in \mathbb{R}^n$ is the state, $w(t) \in \mathbb{R}$ is a Wiener process, and for any t the matrices $F(t) \in \mathbb{R}^{n\times n}, G(t) \in \mathbb{R}^{n\times 1}$ and $H(t) \in \mathbb{R}^{1\times n}$ are chosen so that the covariance function of $f(\cdot)$ is $K(\cdot,\cdot)$. In particular, this means that for any discrete sampling the random vector $f = [f(t_1),\ldots,f(t_N)]^{\mathrm{T}}$ is Gaussian with a positive definite autocovariance matrix $K \in \mathbb{R}^{N\times N}$, where the (i,j) entry of K is given by $K(t_i,t_j)$. Note that we use f for the vector corresponding to a discrete sampling of the function $f(\cdot)$.

We also introduce the covariance of the state $x(\cdot)$ in the state space model (3). In particular, for the discrete sampling of states

$$x = [x(t_1)^{\mathrm{T}},\ldots,x(t_N)^{\mathrm{T}}]^{\mathrm{T}}$$

we use Σ to denote the autocovariance matrix of the random vector x. Then Σ^{-1} is block tridiagonal, and we exploit this structure in Section 4.

Consider the following notable example: when $f(\cdot)$ in (3) is the integral of $w(\cdot)$, the kernel related to the cubic smoothing splines is obtained, see Remark 2 in (De Nicolao and Ferrari Trecate, 2003). For a more involved example, when the reproducing kernel $K(\cdot,\cdot)$ is radial (in our context, depends only on $(|t_1-t_2|)$), the state-space representation model (3) is derived in (De Nicolao and Ferrari Trecate, 2001). However, the state space model above is much more general, and allows non-stationary kernel functions. Notice also that, in view of Runge's theorem (Rudin, 1987), the state space representation allows us to approximate a very wide class of kernels with arbitrary precision. For instance, the popular Gaussian kernel can be approximated by the so-called modified Bessel kernels (Williams and Vivarelli, 1998), which are representable by model (3).

When the loss function $\rho(\cdot)$ in the model (1) is quadratic, and the model can be represented using (3), then the solution (2) can be obtained using Kalman smoothing with a number of operations that scales linearly with the size of the training set, see (De Nicolao and Ferrari Trecate, 2001). Computational efficiency is crucial in these problems since the regularization parameter γ is iteratively adjusted to achieve an estimate for f that satisfies the chosen fitness criterion (e.g. cross validation or maximum likelihood). We extend these results to a wider class of loss functions and focus in particular on the Vapnik penalty. A new class of learning machines is defined using the state space models (3) which we call *state space kernel machines*. The paper proceeds as follows. In Section 2 we extend the connection between kernel-based regularization and continuous-time Kalman smoothing to general losses $\rho(\cdot)$, whereas currently the connection is known in the literature only for quadratic loss functions, see e.g. (Kohn et al., 1993). In Section 3 we take a detailed look at the Vapnik penalty, reformulating it in a way that enables the application of interior point methods. In Section 4 we formulate the optimization problem for state space kernel machines and demonstrate how to solve it with interior point methods. We also show that the work required scales linearly with the number of sampled points. In Section 5 we present a simulated example of functional estimation where the model (3) is a spline model. Results are obtained and compared using both a classical Kalman smoother relying upon the L_2 loss function and the proposed smoother which incorporates the Vapnik penalty. The paper ends with a few concluding remarks.

2. CONNECTION WITH BAYESIAN REGULARIZATION AND KALMAN SMOOTHING

In this section, we describe how to use Kalman smoothing for general loss functions to solve the functional recovery problem (1). In our notation, vectors are column vectors. If a and b are random vectors, $\mathbf{p}(a)$ is the probability density for a, $\mathbf{E}[a]$ is the expected value of a,

$$\mathbf{V}(a,b) = \mathbf{E}\left[(a-\mathbf{E}[a])(b-\mathbf{E}[b])^{\mathrm{T}}\right]$$

is the covariance of a with b, $\mathbf{V}(a) = \mathbf{V}(a,a)$ is the variance of a, and $\max_a \mathbf{p}(a|b)$ is the maximum (with respect to a) of the conditional probability density for a given b. We begin with two lemmas which are instrumental in proving Proposition 2.4, the main result of this section. The proof of the first lemma relies upon well known properties of joint Gaussian vectors, see e.g. (Anderson and Moore, 1979), and is therefore omitted.

Lemma 2.1. Suppose that u and y are jointly Gaussian random vectors. It follows that the maximum of $\mathbf{p}(y|u)$ with respect to y does not depend on the value of u and is given by

$$\max_y \mathbf{p}(y|u) = \exp\left(-\frac{1}{2}\det\left\{2\pi\left[\mathbf{V}(y)-\mathbf{V}(y,u)\mathbf{V}(u)^{-1}\mathbf{V}(u,y)\right]\right\}\right).$$

Lemma 2.2. Suppose that u and y are jointly Gaussian random vectors, and that z is a random vector for which $\mathbf{p}(z|u,y) = \mathbf{p}(z|u)$. Given a realization of z, define corresponding estimates for u and y by

$$(\hat{u},\hat{y}) = \arg\max_{u,y} \mathbf{p}(z,u,y),$$

where we assume that the solution $(\hat{u},\hat{y})$ is unique. It follows that

$$\hat{u} = \arg\max_u \mathbf{p}(z|u)\mathbf{p}(u) \quad \text{and}$$
$$\hat{y} = \arg\max_y \mathbf{p}(y|\hat{u}) = \mathbf{E}(y|\hat{u}) .$$

By $\mathbf{p}(z,u,y)$ we mean the joint density for all three random vectors, and by $\mathbf{p}(y|\hat{u})$ we mean the density for y given that $u = \hat{u}$.

Proof: We have

$$\begin{aligned} \mathbf{p}(z,u,y) &= \mathbf{p}(z|u,y)\ \mathbf{p}(u,y) \\ &= \mathbf{p}(z|u,y)\ \mathbf{p}(y|u)\ \mathbf{p}(u) \\ &= \mathbf{p}(z|u)\ \mathbf{p}(u)\ \mathbf{p}(y|u) . \end{aligned}$$

It follows from the definition of $(\hat{u},\hat{y})$, and the last equation above, that

$$\hat{y} = \arg\max_y \mathbf{p}(y|\hat{u}) = \arg\max_y \mathbf{p}(z,\hat{u},y) , \tag{4}$$

which proves the second assertion of this lemma.

Define

$$\bar{u} = \arg\max_u \mathbf{p}(z|u)\mathbf{p}(u) , \qquad \bar{y} = \arg\max_y \mathbf{p}(y|\bar{u}) . \tag{5}$$

Using (4), Lemma 2.1, and (5) we have

$$\begin{aligned} \mathbf{p}(z,\hat{u},\hat{y}) &= \mathbf{p}(z|\hat{u})\ \mathbf{p}(\hat{u})\ \mathbf{p}(\hat{y}|\hat{u}) \\ &= \mathbf{p}(z|\hat{u})\ \mathbf{p}(\hat{u})\ \mathbf{p}(\bar{y}|\bar{u}) \\ \mathbf{p}(z,\hat{u},\hat{y}) &\le \mathbf{p}(z|\bar{u})\ \mathbf{p}(\bar{u})\ \mathbf{p}(\bar{y}|\bar{u}) \\ &\le \mathbf{p}(z,\bar{u},\bar{y}) \end{aligned}$$

The reverse inequality follows from the definition of $(\hat{u},\hat{y})$. Thus equality holds and it follows from the uniqueness assump-

tion that $\bar{u} = \hat{u}$. This proves the first assertion of the lemma and thereby completes the proof. □

Remark 2.3. For some applications of Lemma 2.2, there is a deterministic matrix Y such that $y = Yu$. In this case $\mathbf{p}(y|u)$ is a delta function concentrated at u and $\hat{y} = Y\hat{u}$. For other applications, there is a non-zero deterministic matrix U such that $u = Uy$. In this case, the matrix $\mathbf{V}(y) - \mathbf{V}(y,u)\mathbf{V}(u)^{-1}\mathbf{V}(u,y)$ in Lemma 2.1 is singular and

$$\hat{y} = \underset{y}{\operatorname{argmax}} \; p(y) \text{ subject to } \hat{u} = Uy \quad .$$

Fix a sequence of time points $\{t_1, \ldots, t_N\}$ and define

$$x_i = x(t_i), \quad H_i = H(t_i), \quad f_i = f(t_i) = H_i x_i, \quad x_i^{i+1} = \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}$$

To simplify the notation, we set $\gamma = 1$ in (1). For any function $v : \mathbb{R} \to \mathbb{R}^n$, we use the notation $v(\cdot)$ for the function and

$$v = [v(t_1)^{\mathrm{T}}, v(t_2)^{\mathrm{T}}, \cdots, v(t_N)^{\mathrm{T}}]^{\mathrm{T}}$$

for the sample vector $v \in \mathbb{R}^{nN}$. In addition, we use the notation $H(\cdot)$ for the measurement function and $H \in \mathbb{R}^{N \times Nn}$ for the matrix

$$H = \begin{pmatrix} H_1 & 0 & \cdots & 0 \\ 0 & H_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & H_N \end{pmatrix}$$

The following proposition states the relationship between Kalman smoothing and regularization in RKHS.

Proposition 2.4. Let $f(\cdot)$ be a zero-mean Gaussian process with prior distribution given by (3) [1] . Let $\rho : \mathbb{R} \to \mathbb{R}$ be a loss function such that

$$\mathbf{p}(z|f) \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\rho(z_i - f_i)\right] . \tag{6}$$

Let $\hat{x}$ be the maximum a posteriori (MAP) estimate of x conditional on z, i.e.

$$\hat{x} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(z_i - H_i x_i) + \frac{1}{2} x^{\mathrm{T}} \Sigma^{-1} x, \tag{7}$$

where Σ is the covariance of the state vector x and Σ^{-1} is block tridiagonal. Then, the solution of (1) is given for any $\tau \in \mathbb{R}$ by

$$\tilde{f}(\tau) = \begin{cases} H_i \, \hat{x}_i & , \; \tau = t_i \\ H(\tau) \, \mathbf{V}\left[x(\tau), x_i^{i+1}\right] \, \mathbf{V}\left(x_i^{i+1}\right)^{-1} \, \hat{x}_i^{i+1} & , \; t_i < \tau < t_{i+1} \\ H(\tau) \, \mathbf{V}[x(\tau), x_1] \, \mathbf{V}(x_1)^{-1} \, \hat{x}_1 & , \; \tau < t_1 \\ H(\tau) \, \mathbf{V}[x(\tau), x_N] \, \mathbf{V}(x_N)^{-1} \, \hat{x}_N & , \; t_N > \tau \, , \end{cases} \tag{8}$$

where i is understood to be in $\{1, \ldots, N\}$ where it appears in the expression above.

Proof: The random vectors $f \in \mathbb{R}^N$ and $x \in \mathbb{R}^{nN}$ are jointly Gaussian, because $f = Hx$. In addition $\mathbf{p}(z|f,x) = \mathbf{p}(z|f)$. Hence we can apply Lemma 2.2 with $u = f$ and $y = x$ (see Remark 2.3). This lemma defines $\hat{f}$ and $\hat{x}$ by

$$[\hat{f}, \hat{x}] = \underset{f,x}{\operatorname{argmax}} \, \mathbf{p}(z, f, x)$$

Using $f = Hx$, we can can remove f from the optimization problem above and get

$$\hat{x} = \underset{x}{\operatorname{argmax}} \, \mathbf{p}(z, x) = \underset{x}{\operatorname{argmax}} \, \mathbf{p}(z|x)\mathbf{p}(x)$$
$$\hat{f} = H\hat{x}$$

This shows that $\hat{x}$ is the MAP estimator for x. Hence we can complete the proof by showing that (8) is true.

We begin by showing the $\hat{f} = \tilde{f}$. According to Lemma 2.2 we also have

$$\hat{f} = \underset{f}{\operatorname{argmax}} \, \mathbf{p}(z|f)\mathbf{p}(f).$$

The function $K(\cdot,\cdot)$ is the covariance for $f(\cdot)$ and $K \in \mathbb{R}^{N \times N}$ the autocovariance of the sample vector $f \in \mathbb{R}^N$. Taking twice the negative log of the right hand side in the equation above gives

$$\hat{f} = \underset{f}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(z_i - f_i) + f^{\mathrm{T}} K^{-1} f.$$

Define $c = K^{-1} f$. It follows that $\hat{f} = K\hat{c}$ where $\hat{c}$ solves the problem

$$\text{minimize} \sum_{i=1}^{N} \rho\left[z_i - \sum_{j=1}^{N} K(t_i, t_j) c_j\right] + c^{\mathrm{T}} K c.$$

Note that $c^{\mathrm{T}} K c = \|f(\cdot)\|^2_{\mathscr{H}}$ where

$$f(\cdot) = \sum_{j=1}^{N} K(\cdot, t_j) c_j.$$

It follows from the representer theorem (2), and $\gamma = 1$, that $\tilde{c} = \hat{c}$ and hence $\tilde{f} = \hat{f}$. Thus we have proved the following assertion:

Assertion: Equation (8) is true for $\tau \in \{t_1, \ldots, t_N\}$.

Consider the case where $\tau \notin \{t_1, \ldots, t_N\}$. The prior for $f(\tau) \in \mathbb{R}$ and $x \in \mathbb{R}^{nN}$ is jointly Gaussian. In addition we have

$$\mathbf{p}[z|x, f(\tau)] = \mathbf{p}(z|x) \quad \text{and} \quad [\hat{x}, \hat{f}(\tau)] = \underset{x, f(\tau)}{\operatorname{argmax}} \, \mathbf{p}[z, x, f(\tau)].$$

Thus we can apply Lemma 2.2 with $u = x$ and $y = f(\tau)$ to conclude that

$$\hat{f}(\tau) = \underset{f(\tau)}{\operatorname{argmax}} \, \mathbf{p}[f(\tau)|\hat{x}] = \mathbf{E}[f(\tau)|\hat{x}] = H(\tau) \, \mathbf{E}[x(\tau)|\hat{x}].$$

In addition, we can apply Lemma 2.2 with $u = f$ and $y = f(\tau)$ to conclude that

$$\hat{f}(\tau) = \underset{f(\tau)}{\operatorname{argmax}} \, \mathbf{p}[f(\tau)|\hat{f}] = \underset{f(\tau)}{\operatorname{argmax}} \, \mathbf{p}[f(\tau)|\tilde{f}] = \tilde{f}(\tau).$$

Note that the definition for $\hat{f}(\tau)$ in this application of Lemma 2.2 has the same value as when $u = x$ and $y = f(\tau)$. Also note that $\hat{f} = \tilde{f}$ was shown by the assertion above. Finally note that this sequence of equalities shows the following assertion:

Assertion: $\tilde{f}(\tau) = \hat{f}(\tau)$ for all τ.

We now restrict our attention to the case where there is an $i \in \{1, \ldots, N-1\}$ such that $t_i < \tau < t_{i+1}$. Using the Markov property, we have

$$\mathbf{E}[x(\tau)|\hat{x}] = \mathbf{E}[x(\tau)|\hat{x}_i^{i+1}] = \mathbf{V}[x(\tau), x_i^{i+1}] \mathbf{V}\left(x_i^{i+1}\right)^{-1} \hat{x}_i^{i+1} \, ,$$
$$\tilde{f}(\tau) = \hat{f}(\tau) = H(\tau) \, \mathbf{V}[x(\tau), x_i^{i+1}] \mathbf{V}\left(x_i^{i+1}\right)^{-1} \hat{x}_i^{i+1} \, .$$

This proves the assertion in equation (8) for the case where $t_1 \leq \tau \leq t_N$. The other two cases $\tau < t_1$ and $t_N < \tau$ can be obtained following the same reasoning as in the paragraph above. This completes the proof. □

[1] This corresponds to assuming, just to simplify the exposition, that $H(0)x(0)$ is a zero-mean normal random variable.

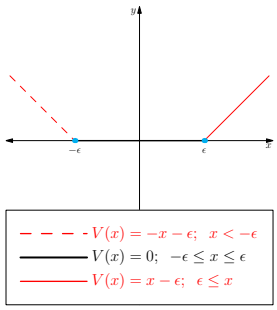

Fig. 1. Vapnik Penalty

The above result shows that given $\hat{x}$, the solution to (1) can be computed using equation (8). In other words, the estimates of the state at sampling locations are sufficient to reconstruct the function $\hat{f}(\cdot)$ over its entire domain. Moreover, the optimization problem (7) exploits the block tridiagonal structure of Σ^{-1}

In the case where $\rho(\cdot)$ is quadratic, (De Nicolao and Ferrari Trecate, 2001, Theorem 1) provides a procedure for obtaining the weight vector $\hat{c}$ at the same time as obtaining the Kalman smoother estimate $\hat{f}$. That procedure has $O(Nn^3)$ computational complexity. In the sections below we present an interior point algorithm that obtains the same complexity when $\rho(\cdot)$ is the Vapnik loss.

3. KALMAN SMOOTHING WITH THE VAPNIK PENALTY

Here and below, we assume $-\log \mathbf{p}(z|x)$ in (6) is the Vapnik loss function. The Vapnik loss function with parameter ε, also known as the ε-insensitive loss function, is displayed in Figure 1 and given by

$$\begin{aligned}\rho(y) &= \begin{cases} y-\varepsilon & \text{if } y \ge \varepsilon \\ 0 & \text{if } y \in [-\varepsilon,+\varepsilon] \\ y+\varepsilon & \text{if } y \le -\varepsilon \end{cases} \\ &= \sup_{a\in[0,1]} a(y-\varepsilon) \;+\; \sup_{b\in[0,1]} b(-y-\varepsilon).\end{aligned}$$

Applying Proposition 2.4 to the Vapnik case, the MAP objective as a function of $x \in \mathbb{R}^{nN}$ is

$$\begin{aligned}\frac{1}{2}x^{\mathrm{T}}\Sigma^{-1}x + \sum_{i=1}^{N} \Big[&\sup_{u^+(i)\in[0,1]} u_i^+(z_i - H_i x_i - \varepsilon) \\ &+ \sup_{u^-(i)\in[0,1]} u_i^-(-z_i + H_i x_i - \varepsilon)\Big],\end{aligned} \tag{9}$$

where $u^+ \in \mathbb{R}^N$, $u^- \in \mathbb{R}^N$, and $u^+(i) = u_i^+$, $u^-(i) = u_i^-$ are used to avoid two levels of subscripting. Note that this objective is strictly convex, and so has a unique global minimum.

Lemma 3.1. Suppose that $x \in \mathbb{R}^{nN}$ and there are vectors u^+, u^- s^+, s^- p^+, p^- q^+, q^- are all in $\mathbb{R}^N_+$, such that the following conditions hold for $i = 1,\ldots,N$:

$$\Sigma^{-1}x - H^{\mathrm{T}}(u^+ - u^-) = 0 \quad , \tag{10}$$

$$s_i^+ p_i^+ = 0,\; u_i^+ q_i^+ = 0,\; s_i^+ + u_i^+ = 1,\; p_i^+ - q_i^+ = z_i - H_i x_i - \varepsilon, \tag{11}$$

$$s_i^- p_i^- = 0,\; u_i^- q_i^- = 0,\; s_i^- + u_i^- = 1,\; p_i^- - q_i^- = -z_i + H_i x_i - \varepsilon. \tag{12}$$

It follows that this choice of x above minimizes (9) over $\mathbb{R}^{nN}$.

Proof: For a convex set $A \subset \mathbb{R}$, and $a \in A$, we use $\mathcal{N}(A,a)$ to denote the normal cone to A at the point a; e.g.,

$$\mathcal{N}([0,1],a) = \begin{cases} (-\infty,0] & \text{if } a = 0 \\ 0 & \text{if } a \in (0,1) \\ [0,+\infty) & \text{if } a = 1 \end{cases} .$$

See (Rockafellar, 1970) for more details on normal cones in general. The optimality conditions for minimizing (9) are condition (10) and

$$z_i - H_i x_i - \varepsilon \in \mathcal{N}\left([0,1], u_i^+\right) \qquad (i = 1,\cdots N) \tag{13}$$

$$-z_i + H_i x_i - \varepsilon \in \mathcal{N}\left([0,1], u_i^-\right) \qquad (i = 1,\cdots N). \tag{14}$$

See (Rockafellar and Wets, 1998) for a discussion of these optimality conditions. We will show that condition (11) implies condition (13). The fact that condition (12) implies condition (14) can be demonstrated in a similar manner and hence this will complete the proof.

The optimality condition in (13) is equivalent to the following three conditions for $i = 1,\ldots,N$:

$$z_i - H_i x_i - \varepsilon < 0 \Rightarrow u_i^+ = 0 \quad , \tag{15}$$

$$z_i - H_i x_i - \varepsilon = 0 \Rightarrow u_i^+ \in [0,1] \quad , \tag{16}$$

$$z_i - H_i x_i - \varepsilon > 0 \Rightarrow u_i^+ = 1 \quad . \tag{17}$$

Hence it will suffice to show that the conditions (11) imply the three conditions above. We divide this demonstration into the three cases corresponding to the three conditions above:

(1) Suppose that $z_i - H_i x_i - \varepsilon < 0$: It follows from (11) that $q_i^+ > 0$ and $u_i^+ = 0$. Hence (15) holds.
(2) Suppose that $z_i - H_i x_i - \varepsilon = 0$: It follows from (11) that $u_i^+ \in [0,1]$. Hence (16) holds.
(3) Suppose that $z_i - H_i x_i - \varepsilon > 0$: It follows from (11) that $p_i^+ > 0$, $s_i^+ = 0$, and $u_i^+ = 0$. Hence (17) holds.

This completes the proof of this lemma. □

We minimize (9) using an interior point method (see Kojima et al. (1991); Nemirovskii and Nesterov (1994)). The main idea is to relax the complementarity conditions in (11) and (12) to obtain (18) below, and solve this system using a damped Newton's method. The relaxation parameter μ is driven to 0, yielding the unique solution to (9). All of the implementation details are given below.

Define $1_N \in \mathbb{R}^N$ to be the vector with all of its components equal to one. Given a $v^+ \in \mathbb{R}^N_+$ ($v^- \in \mathbb{R}^N_+$) define $V_+ \in \mathbb{R}^{N\times N}_+$ ($V_- \in \mathbb{R}^{N\times N}_+$) to be the diagonal matrix with with diagonal $v^+ \in \mathbb{R}^N_+$ ($v^- \in \mathbb{R}^N_+$). This is correspondence between lower case and upper case letters applies only to vectors with the superscript plus or minus. Fix a relaxation parameter μ and define $F_\mu : \mathbb{R}^{8N+nN} \to \mathbb{R}^{8N+nN}$ by

$$F_\mu \begin{pmatrix} p^+ \\ q^+ \\ u^+ \\ s^+ \\ p^- \\ q^- \\ u^- \\ s^- \\ x \end{pmatrix} = \begin{bmatrix} p^+ - q^+ - z + Hx + \varepsilon 1_N \\ Q_+U_+1_N - \mu 1_N \\ u^+ + s^+ - 1_N \\ P_+S_+1_N - \mu 1_N \\ p^- - q^- + z - Hx + \varepsilon 1_N \\ Q_-U_-1_N - \mu 1_N \\ u^- + s^- - 1_N \\ P_-S_-1_N - \mu 1_N \\ \Sigma^{-1}x - H^{\mathrm{T}}(u^+ - u^-) \end{bmatrix} := \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \end{bmatrix} . \tag{18}$$

The last line in the definition of F_μ is the residual equation (10). Expressions r_2, r_4, r_6 and r_8 in the definition of F_μ are μ-relaxed versions of the residuals in the first two equations of (11) and (12), while r_1, r_3, r_5, and r_7 are the residuals in the last two

equations of (11) and (12). Thus if u^+, u^- s^+, s^- p^+, p^- q^+, q^- are all in $\mathbb{R}^N_+$, and

$$F_0(p^+,q^+,u^+,s^+,p^-,q^-,u^-,s^-,x) = 0,$$

then x is the unique minimizer of (9). Solutions to the equation $F_\mu = 0$ in (18) can be obtained via a damped Newton's method by solving

$$\begin{bmatrix} I & -I & 0 & 0 & 0 & 0 & 0 & 0 & H \\ 0 & U_+ & Q_+ & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & I & 0 & 0 & 0 & 0 & 0 \\ S_+ & 0 & 0 & P_+ & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I & -I & 0 & 0 & -H \\ 0 & 0 & 0 & 0 & 0 & U_- & Q_- & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & I & 0 \\ 0 & 0 & 0 & 0 & S_- & 0 & 0 & P_- & 0 \\ 0 & 0 & -H^{\mathrm{T}} & 0 & 0 & 0 & H^{\mathrm{T}} & 0 & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \Delta p^+ \\ \Delta q^+ \\ \Delta u^+ \\ \Delta s^+ \\ \Delta p^- \\ \Delta q^- \\ \Delta u^- \\ \Delta s^- \\ \Delta x \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \end{bmatrix}, \tag{19}$$

where the matrix on the left hand side of (19) is $F_\mu^{(1)}$.

Lemma 3.2. We are given values for the variables p^+, q^+, u^+, s^+, p^-, q^-, u^-, s^-, all in $\mathbb{R}^N_+$, $x \in \mathbb{R}^{nN}$, and the parameters $\mu \in \mathbb{R}_+$, $\varepsilon \in \mathbb{R}_+$, $z \in \mathbb{R}^N$, $H \in \mathbb{R}^{N\times nN}$, and $\Sigma^{-1} \in \mathbb{R}^{nN\times nN}$.

The algorithm (20) given below computes the values Δp^+, Δq^+, Δu^+, Δs^+, Δp^-, Δq^-, Δu^-, Δs^-, all in $\mathbb{R}^N$ and $\Delta x \in \mathbb{R}^{nN}$ that solve equation (19). Note $\{r_i\}$, $i = 1,\ldots,9$, are defined in (18).

$$\begin{aligned} T_+ &= S_+^{-1}P_+ + U_+^{-1}Q_+ \\ T_- &= S_-^{-1}P_- + U_-^{-1}Q_- \\ \hat{r}_4 &= S_+^{-1}r_4 - r_1 - U_+^{-1}r_2 + U_+^{-1}Q_+r_3 \\ \hat{r}_8 &= S_-^{-1}r_8 - r_5 - U_-^{-1}(r_6 - Q_-r_7) \\ \hat{r}_9 &= r_9 + H^{\mathrm{T}}(r_3 - T_+^{-1}\hat{r}_4 - r_7 + T_-^{-1}\hat{r}_8) \\ \Delta x &= [\Sigma^{-1} + H^{\mathrm{T}}(T_+^{-1} + T_-^{-1})H]^{-1}\hat{r}_9 \\ \Delta s^- &= T_-^{-1}(\hat{r}_8 - H\Delta x) \\ \Delta u^- &= r_7 - \Delta s^- \\ \Delta q^- &= U_-^{-1}(r_6 - Q_-\Delta u^-) \\ \Delta p^- &= r_5 + \Delta q^- + H\Delta x \\ \Delta s^+ &= T_+^{-1}(\hat{r}_4 + H\Delta x) \\ \Delta u^+ &= r_3 - \Delta s^+ \\ \Delta q^+ &= U_+^{-1}(r_2 - Q_+\Delta u^+) \\ \Delta p^+ &= r_1 + \Delta q^+ + H\Delta x\,. \end{aligned} \tag{20}$$

Proof: We now provide row operations necessary to reduce the matrix $F_\mu^{(1)}$ to block upper triangular form. In addition, we track the changes to the right hand side in (19). Note that only rows 4, 8, and 9 of $F_\mu^{(1)}$ need to be modified to implement the reduction. To save space, after a series of row operations, we present the final row and corresponding right hand side. We use the symbol $\leftarrow$ to indicate assignment:

$$\begin{aligned} \mathrm{row}_4 &\leftarrow \mathrm{row}_4 - S_+\mathrm{row}_1 \\ \mathrm{row}_4 &\leftarrow \mathrm{row}_4 - S_+U_+^{-1}\mathrm{row}_2 \\ \mathrm{row}_4 &\leftarrow \mathrm{row}_4 + S_+U_+^{-1}Q_+\mathrm{row}_3 \\ \mathrm{row}_4 &\leftarrow S_+^{-1}\mathrm{row}_4. \end{aligned}$$

The final result for row_4 and the corresponding right hand side, which we label $\hat{r}_4$, are given by

$$\begin{aligned} \mathrm{row}_4 &= [0\ 0\ 0\ T_+\ 0\ 0\ 0\ 0\ -H] \\ \hat{r}_4 &= S_+^{-1}r_4 - r_1 - U_+^{-1}r_2 + U_+^{-1}Q_+r_3, \end{aligned}$$

where $T_+ \in \mathbb{R}^{N\times N}$ is defined by $T_+ = S_+^{-1}P_+ + U_+^{-1}Q_+$.

The reduction of row 8 is analogous:

$$\begin{aligned} \mathrm{row}_8 &\leftarrow \mathrm{row}_8 - S_-\mathrm{row}_5 \\ \mathrm{row}_8 &\leftarrow \mathrm{row}_8 - S_-U_-^{-1}\mathrm{row}_6 \\ \mathrm{row}_8 &\leftarrow \mathrm{row}_8 + S_-U_-^{-1}Q_-\mathrm{row}_7 \\ \mathrm{row}_8 &\leftarrow S_-^{-1}\mathrm{row}_8. \end{aligned}$$

The final form for row_8 and the corresponding right hand side, which we label $\hat{r}_8$, are given by

$$\begin{aligned} \mathrm{row}_8 &= [0\ 0\ 0\ 0\ 0\ 0\ 0\ T_-\ H] \\ \hat{r}_8 &= S_-^{-1}r_8 - r_5 - U_-^{-1}r_6 + U_-^{-1}Q_-r_7, \end{aligned}$$

where $T_- \in \mathbb{R}^{N\times N}$ is defined by $T_- = S_-^{-1}P_- + U_-^{-1}Q_-$.

The final modifications are to row 9:

$$\begin{aligned} \mathrm{row}_9 &\leftarrow \mathrm{row}_9 + H^{\mathrm{T}}\mathrm{row}_3 \\ \mathrm{row}_9 &\leftarrow \mathrm{row}_9 - H^{\mathrm{T}}T_+^{-1}\mathrm{row}_4 \\ \mathrm{row}_9 &\leftarrow \mathrm{row}_9 - H^{\mathrm{T}}\mathrm{row}_7 \\ \mathrm{row}_9 &\leftarrow \mathrm{row}_9 + H^{\mathrm{T}}T_-^{-1}\mathrm{row}_8. \end{aligned}$$

The final forms of row_9 and the corresponding right hand side, which we label $\hat{r}_9$, are given by

$$\begin{aligned} \mathrm{row}_9 &= \left[0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ [\Sigma^{-1} + H^{\mathrm{T}}(T_-^{-1} + T_+^{-1})H]\right] \\ \hat{r}_9 &= r_9 + H^{\mathrm{T}}r_3 - H^{\mathrm{T}}T_+^{-1}\hat{r}_4 - H^{\mathrm{T}}r_7 + H^{\mathrm{T}}T_-^{-1}\hat{r}_8. \end{aligned}$$

We now have a block upper triangular system which is equivalent to (19). In the remaining steps, we solve this system for $\Delta x, \Delta s^-, \Delta u^-, \Delta q^-, \Delta p^-, \Delta s^+, \Delta u^+, \Delta q^+, \Delta p^+$. These calculations, together with definitions of $r_1 : r_9$ in (19) and the definitions of $T^+, T^-, \hat{r}_4, \hat{r}_8, \hat{r}_9$ provided above comprise the algorithm (20).

This completes the proof of the lemma. □

The corresponding system of equations is block upper triangular, and its solution depends on inversion of the matrices U_+, U_-, S_+ , S_-, T_+, T_-, and $\Sigma^{-1} + H^{\mathrm{T}}(T_+^{-1} + T_-^{-1})H$.

The relaxed complementarity conditions in (18) are

$$\begin{aligned} Q_+U^+1_N = \mu 1_N \quad , \quad P_+S^+1_N = \mu 1_N \quad , \\ Q_-U^-1_N = \mu 1_N \quad , \quad P_-S^-1_N = \mu 1_N \quad . \end{aligned}$$

For $\mu > 0$, solutions of the equation above are strictly positive; i.e., the solutions satisfy

$$(p^+,q^+,u^+,s^+,p^-,q^-,u^-,s^-) > 0.$$

Primal-dual interior point methods use predictor and corrector steps to follow the solution of $F_\mu = 0$ as $\mu > 0$ descends to zero. These methods keep the vectors u^+, u^-, s^+, and s^- strictly positive. Hence the diagonal matrices U_+, U_-, S^+ and S^- can be easily inverted. The vectors p^+, and q^+ are also strictly positive, so the diagonal matrix

$$T_+ = S_+^{-1}P_+ + U_+^{-1}Q_+$$

can be easily inverted. In addition, the vectors p^- and q^- are strictly positive, so the diagonal matrix

$$T_- = S_-^{-1}P_- + U_-^{-1}Q_-$$

can be easily inverted. Finally, consider inverting the matrix

$$T = \Sigma^{-1} + H^{\mathrm{T}}(T_+^{-1} + T_-^{-1})H.$$

The $nN \times nN$ matrix Σ^{-1} is block tridiagonal with blocks of size $n \times n$. It can be inverted in $O(n^3N)$ operations using (Bell, 2000, Lemma 6). Furthermore, the $nN \times nN$ matrix $H^{\mathrm{T}}(T_+^{-1} + T_-^{-1})H$ is block diagonal with positive semi-definite blocks of size $n \times n$. It follows that T can also be inverted in $O(n^3N)$ operations using (Bell, 2000, Lemma 6). Thus, notably, the computational complexity of the Vapnik based smoother is linear in the number of time steps, as that of the classical Kalman smoother that relies upon quadratic losses.

4. NUMERICAL EXAMPLE

In this section we test the new Kalman smoother that incorporates the Vapnik's loss via a simulated example. The unknown function, taken from (Dinuzzo et al., 2007), is given by

$$f(t) = e^{\sin(8t)}$$

and has to be reconstructed from 2000 noisy samples collected uniformly over the unit interval. The measurement noise v_k was generated using a mixture of two normals with $p = 0.1$ denoting the fraction from each normal; i.e.,

$$v_k \sim (1-p)\mathbf{N}(0, 0.25) + p\mathbf{N}(0, 25).$$

Data are displayed as dots in Fig. 2. Note that the purpose of the second component of the normal mixture is to simulate outliers in output data. Note also that any points exceeding vertical axis limits are plotted at Fig. 2 to improve readability.
The initial condition $f(0) = 1$ is assumed to be known, while the difference of the unknown function from the initial condition (i.e. $f(\cdot) - 1$) is the second component of the two-dimensional state vector $x(t)$ and corresponds to the integral of the first state component, which is modeled as Brownian motion. To be more specific, letting $\Delta t = 1/2000$, the process model for the mean of x_k given x_{k-1} is

$$F_k(x_{k-1}) = \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} x_{k-1} \,,$$

while the autocovariance of x_k given x_{k-1} is

$$Q_k = \lambda^2 \begin{bmatrix} \Delta t & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t^3/3 \end{bmatrix} \,,$$

(Jazwinski, 1970; Oksendal, 2005), where λ^2 is an unknown scale factor to be estimated from the data. It corresponds to the inverse of the regularization parameter γ that appears in (1). The measurement model for the mean of z_k given x_k is

$$h_k(x_k) = (0,1)x_k = x_{2,k} \,,$$

where $x_{2,k}$ denotes the second component of x_k.
In order to estimate the two unknown parameters λ and ε characterizing the Vapnik loss, the 2000 measurements were randomly split into training and validation sets of 1300 and 700 data points, respectively. For each value of λ^2 and ε contained in a 10×20 grid on $[0.01, 10000] \times [0, 1]$, with λ^2 logarithmically spaced, the function estimate was rapidly obtained by the new smoother applied to the training set. Then, the relative average prediction error on the validation set was computed, see Fig. 3. The parameters leading to the best prediction were $\lambda^2 = 2.15e3$ and $\varepsilon = 0.45$, which give a sparse solution defined by less of 400 support vectors. For the sake of comparison, we also derived the solution of (1) using a quadratic loss function (by implementing a classical Kalman smoother) with γ estimated following the same validation strategy described above. Differently from the Vapnik penalty, the quadratic loss does not induce any sparsity, so that, in this case, the number of support vectors equals the size of the training set.

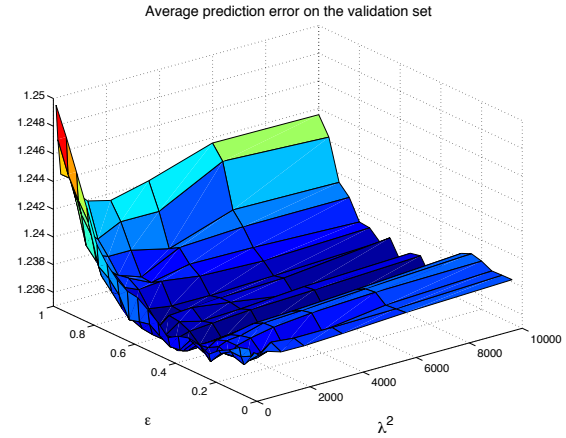

Fig. 3. Estimation of the smoothing filter parameters using the Vapnik loss. Average prediction error on the validation data set as a function of the variance process λ^2 and ε

The left and right panels of Fig. 2 report the function estimate adopting the quadratic and the Vapnik loss, respectively. It is clear that the Gaussian estimate is heavily affected by the outliers. In contrast, the estimate coming from the Vapnik based smoother performs well over the entire time period, being virtually unaffected by the presence of large outliers.

5. CONCLUSIONS

We have described an approach to solve a large class of machine learning, named state space kernel machines, and have focused on the non-smooth Vapnik penalty as the loss function of interest. In practise, the Vapnik loss function has two desirable features illustrated in the numerical experiment: robustness to outliers in the measurement data, and sparsity in the final representation of the reconstructed function. The interior point approach we use, together with Proposition 2.4, allows the development of learning machine algorithms for such losses with a time complexity that scales linearly with the number of samples. In the near future, we plan to extend the obtained results to more general losses, such as the soft ε-insensitive and the Huber's one.

ACKNOWLEDGEMENTS

This research has been partially supported by the PRIN Project "Sviluppo di nuovi metodi e algoritmi per l'identificazione, la stima Bayesiana e il controllo adattativo e distribuito", by the Progetto di Ateneo CPDA090135/09 funded by the University of Padova and by the European Community's Seventh Framework Programme under agreement n. FP7-ICT-223866-FeedNetBack.

REFERENCES

Anderson, B.D.O. and Moore, J.B. (1979). *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. of the American Mathematical Society*, 68, 337–404.

Bell, B.M. (2000). The marginal likelihood for parameters in a discrete Gauss-Markov process. *IEEE Transactions on Signal Processing*, 48(3), 626–636.

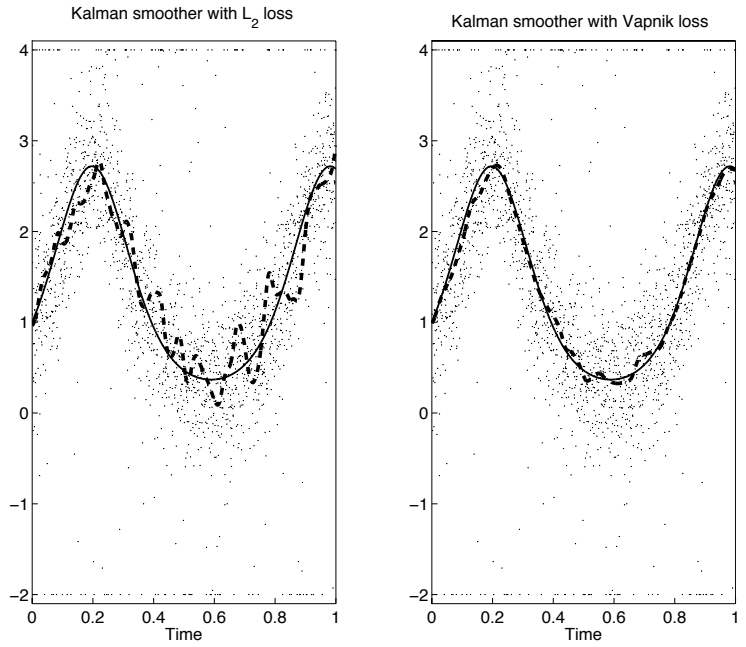

Fig. 2. Simulation: measurements $(\cdot)$ with outliers set to 4 or -2, true function (continuous line), smoothed estimate using either the quadratic loss (dashed line, left panel) or the Vapnik's ε-insensitive loss (dashed line, right panel)

Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75, 1–120.

De Nicolao, G. and Ferrari Trecate, G. (2001). Regularization networks: fast weight calculation via kalman filtering. *IEEE Transactions on Neural Networks*, 12, 228–235.

De Nicolao, G. and Ferrari Trecate, G. (2003). Regularization networks for inverse problems: a state space approach. *Automatica*, 39, 669–676.

De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., and Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6, 883–904.

Dinuzzo, F., Neve, M., and De Nicolao, G. (2007). On the Representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, 8, 2467–2495.

Evgeniou, T., Pontil, M., and Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13, 1–150.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architecture. *Neural Computation*, 7, 219–269.

Gunter, L. and Zhu, J. (2006). Computing the solution path for the regularized support vector regression. In Y. Weiss, B. Schölkopf, and J. Platt (eds.), *Advances in Neural Information Processing Systems 18*, 483–490. MIT Press, Cambridge, MA.

Hengland, M. (2007). Approximate maximum a posteriori with Gaussian process priors. *Constructive Approximation*, 26, 205–224.

Jazwinski, A. (1970). *Stochastic Processes and Filtering Theory*. Dover Publications, Inc.

Kohn, R., Ansley, C., and Wong, C. (1993). Nonparametric spline regression with prior information. *Biometrika*, 80(1), 75–88.

Kojima, M., Megiddo, N., Noma, T., and Yoshise, A. (1991). *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, volume 538 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Germany.

Nemirovskii, A. and Nesterov, Y. (1994). *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, USA.

Oksendal, B. (2005). *Stochastic Differential Equations*. Springer, sixth edition.

Pillonetto, G. and Saccomani, M. (2006). Input estimation in nonlinear dynamic systems using differential algebra concepts. *Automatica*, 42, 2117–2129.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78, 1481–1497.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press.

Rockafellar, R.T. and Wets, R.J.B. (1998). *Variational Analysis*, volume 317 of *A Series of Comprehensive Studies in Mathematics*. Springer.

Rudin, W. (1987). *Real and complex analysis.* McGraw-Hill.

Schölkopf, B., Herbrich, R., and Smola, A.J. (2001). A generalized representer theorem. *Neural Networks and Computational Learning Theory*, 81, 416–426.

Schölkopf, B. and Smola, A.J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). The MIT Press.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, New York, NY, USA.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.

Wahba, G. (1998). Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin.

Williams, C. and Vivarelli, F. (1998). Upper and lower bounds on the learning curve for Gaussian processes. Technical Report NCRG/98/015, Aston University, Birmingham, U.K.