A GENERAL DIRECT WEIGHT OPTIMIZATION FRAMEWORK FOR NONLINEAR SYSTEM IDENTIFICATION

Jacob Roll * **Alexander Nazin** ** **Lennart Ljung** *

* *Division of Automatic Control, Linköping University, SE-581 83 Linköping, Sweden, e-mail:* roll, ljung@isy.liu.se
** *Institute of Control Sciences, Profsoyuznaya str., 65, 117997 Moscow, Russia, e-mail:* nazine@ipu.rssi.ru

Abstract: The direct weight optimization (DWO) approach is a method for finding optimal function estimates via convex optimization, applicable to nonlinear system identification. In this paper, an extended version of the DWO approach is introduced. A general function class description — which includes several important special cases — is presented, and different examples are given. A general theorem about the principal shape of the weights is also proven.

Keywords: Non-parametric identification, Function approximation, Minimax techniques, Quadratic programming, Nonlinear systems, Mean-square error

1. INTRODUCTION

A wide-spread technique to model non-linear mappings is to use basis function expansions:

$$f(\varphi(t),\theta)=\sum_{k=1}^{d}\alpha_k f_k(\varphi(t),\beta),\qquad \theta=\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \quad (1)$$

Here, $\varphi(t)$ is the regression vector, $\alpha=(\alpha_1\dots\alpha_d)^T$, $\beta=(\beta_1\dots\beta_l)^T$, and θ is the parameter vector.

A common case is that the basis functions $f_k(\varphi)$ are a priori fixed, and do not depend on any parameter β, i.e., (with $\theta_k=\alpha_k$)

$$f(\varphi(t),\theta)=\sum_{k=1}^{d}\theta_k f_k(\varphi(t))=\theta^T F(\varphi(t)) \quad (2)$$

where we use the notation

$$F(\varphi)=\begin{pmatrix}f_1(\varphi) & \dots & f_d(\varphi)\end{pmatrix}^T \quad (3)$$

That makes the fitting of the model (1) to observed data a linear regression problem, which has many advantages from an estimation point of view. The drawback is that the basis functions are not adapted to the data, which in general means that more basis functions are required (larger d). Still, this special case is very common (see, e.g., Harris *et al.* (2002), Suykens *et al.* (2002)).

Now, assume that the observed data, $\{\varphi(t),y(t)\}_{t=1}^{N}$, are generated from a system described by

$$y(t)=f_0(\varphi(t))+e(t) \quad (4)$$

where f_0 is an unknown function, $f_0:\mathcal{D}\to\mathbb{R}$, and $e(t)$ are zero-mean, i.i.d. random variables with known variance σ^2, independent of $\varphi(\tau)$ for all τ. Furthermore, suppose that we have reasons to believe that the "true" function f_0 can locally be *approximately* described by a given basis function expansion, and that we know a given bound on the approximation error. How then would we go about estimating f_0? This is the problem considered in the following. We will take a pointwise estimation approach, where we estimate f_0 for a given point φ^*. This gives rise to a Model on Demand methodology (Stenman, 1999). Similar problems have also been studied within local polynomial modelling

⋆ Corresponding author J. Roll. Tel. +46-13-281338. Fax +46-13-282622.

(Fan and Gijbels, 1996), although mostly based on asymptotic arguments.

The direct weight optimization (DWO) approach was first proposed in (Roll *et al.*, 2002) and presented in detail in (Roll, 2003; Roll *et al.*, 2005). Those presentations mainly consider differentiable functions f_0, for which a Lipschitz bound on the derivatives is given (see Examples 1 and 2 below). This paper suggests an extension to a much more general framework, which contains several interesting special cases, including the ones mentioned above. Another special case is given in Example 3 below. In Section 5, a general theorem about the structure of the optimal solutions is also given.

2. MODEL AND FUNCTION CLASSES

We assume that we are given data $\{\varphi(t), y(t)\}_{t=1}^{N}$ from a system described by (4). Also assume that f_0 belongs to a function class $\mathcal{F}$ which can be "approximated" by a fixed basis function expansion (2). More precisely, let $\mathcal{F}$ be defined as follows:

Definition 1. Let $\mathcal{F} = \mathcal{F}(\mathcal{D}, \mathcal{D}_\theta, F, M)$ be the set of all functions f, for which there, for each $\varphi_0 \in \mathcal{D}$, exists a $\theta^0(\varphi_0) \in \mathcal{D}_\theta$, such that

$$\left| f(\varphi) - {\theta^0}^T(\varphi_0) F(\varphi) \right| \leq M(\varphi, \varphi_0) \quad \forall \varphi \in \mathcal{D} \quad (5)$$

We assume here that the domain $\mathcal{D}$, the parameter domain $\mathcal{D}_\theta$, the basis functions F and the non-negative upper bound M are given a priori. We can show the following lemma:

Lemma 1. Assume that $M(\varphi, \varphi_0)$ in (5) does not depend on φ_0, i.e., $M(\varphi, \varphi_0) \equiv M(\varphi)$. Then there is a $\theta^0(\varphi_0) \equiv \theta^0$ that does not depend on φ_0 either. Conversely, if $\theta^0(\varphi_0)$ does not depend on φ_0, there is an $\bar{M}(\varphi)$ that does not depend on φ_0, and that satisfies (5).

Proof: Given a function $f \in \mathcal{F}$, and for a given φ_0, there is a θ^0 satisfying (5) for all $\varphi \in \mathcal{D}$. But since M does not depend on φ_0, we can choose the same θ^0 given any φ_0, and it will still satisfy (5). Hence, θ^0 does not depend on φ_0.

Conversely, if θ^0 does not depend on φ_0, we can just let

$$\bar{M}(\varphi) = \inf_{\varphi_0} M(\varphi, \varphi_0)$$

□

In (Sacks and Ylvisaker, 1978), a function class given by Lemma 1 is called a class of *approximately linear models*. For a function f_0 of this kind, there is a vector $\theta^0 \in \mathcal{D}_\theta$, such that

$$\left| f_0(\varphi) - {\theta^0}^T F(\varphi) \right| \leq M(\varphi) \quad \forall \varphi \in \mathcal{D} \quad (6)$$

Note that Definition 1 is an extension of this function class, allowing for more natural function classes such as in Example 1 below.

Example 1. Suppose that $f_0 : \mathbb{R} \to \mathbb{R}$ is a once differentiable function with Lipschitz continuous derivative, with a Lipschitz constant L. In other words, the derivative should satisfy

$$|f_0'(\varphi + h) - f_0'(\varphi)| \leq L|h| \quad \forall\, \varphi, h \in \mathbb{R} \quad (7)$$

This could be treated by choosing the fixed basis functions as

$$f_1(\varphi) \equiv 1, \quad f_2(\varphi) \equiv \varphi \quad (8)$$

For each φ_0, f_0 satisfies (Dennis and Schnabel, 1983, Chapter 4)

$$|f_0(\varphi) - f_0(\varphi_0) - f_0'(\varphi_0)(\varphi - \varphi_0)| \leq \frac{L}{2}(\varphi - \varphi_0)^2$$

for all $\varphi \in \mathbb{R}$. In other words, (5) is satisfied with

$$\theta_1^0(\varphi_0) = f_0(\varphi_0) - f_0'(\varphi_0)\varphi_0, \quad \theta_2^0(\varphi_0) = f_0'(\varphi_0)$$
$$M(\varphi, \varphi_0) = \frac{L}{2}(\varphi - \varphi_0)^2 \quad (9)$$

◇

Example 2. A multivariate extension of Example 1 (with $f_0 : \mathbb{R}^n \to \mathbb{R}$) can be obtained by assuming that

$$\|\nabla f_0(\varphi + h) - \nabla f_0(\varphi)\|_2 \leq L\|h\|_2 \quad \forall\, \varphi, h \in \mathbb{R}^n$$

where ∇f_0 is the gradient of f_0 and $\|\cdot\|_2$ is the Euclidean norm. We get

$$\left| f_0(\varphi) - f_0(\varphi_0) - \nabla^T f_0(\varphi_0)(\varphi - \varphi_0) \right| \leq \frac{L}{2}\|\varphi - \varphi_0\|_2^2$$

for all $\varphi \in \mathbb{R}^n$, and can choose the basis functions as

$$f_1(\varphi) \equiv 1, \quad f_{1+k}(\varphi) \equiv \varphi_k \quad \forall\, k = 1, \ldots, n \quad (10)$$

In accordance with (9), we now get

$$\theta^0(\varphi_0) = \begin{pmatrix} f_0(\varphi_0) - \nabla^T f_0(\varphi_0)\varphi_0 \\ \nabla f_0(\varphi_0) \end{pmatrix}$$
$$M(\varphi, \varphi_0) = \frac{L}{2}\|\varphi - \varphi_0\|_2^2$$

◇

Example 3. As in (6), $M(\varphi, \varphi_0)$ and $\theta^0(\varphi_0)$ do not necessarily need to depend on φ_0. For example, we could assume that f_0 is well described by a certain basis function expansion, with a constant upper bound on the approximation error, i.e.,

$$\left| f_0(\varphi) - {\theta^0}^T F(\varphi) \right| \leq M(\varphi) \quad \forall\, \varphi \in \mathcal{D}$$

where θ^0 and $M(\varphi)$ are both constant. If the approximation error is known to vary with φ in a certain way, this can be reflected by choosing an appropriate function $M(\varphi)$.

A specific example of this kind is given by a model (linear in the parameters) with both unknown-but-bounded and Gaussian noise. Suppose that

$$y(t) = {\theta^0}^T F(\varphi(t)) + r(t) + e(t) \quad (11)$$

where $|r(t)| \leq M$ is a bounded noise term. We can then treat this as if (slightly informally)

$$f_0(\varphi(t)) = {\theta^0}^T F(\varphi(t)) + r(t) \quad (12)$$

i.e., f_0 satisfies

$$|f_0(\varphi(t)) - {\theta^0}^T F(\varphi(t))| \leq M \tag{13}$$

This case is studied in (Nazin *et al.*, 2003). Some other examples are given in (Sacks and Ylvisaker, 1978). ◇

3. CRITERION AND ESTIMATOR

Now, the problem to solve is to find an estimator $\hat{f}_N$ to estimate $f_0(\varphi^*)$ in a certain point φ^*, under the assumption $f_0 \in \mathcal{F}$ from Definition 1. A common criterion for evaluating the quality of the estimate is the *mean squared error (MSE)* given by

$$\begin{aligned} &MSE(f_0, \hat{f}_N, \varphi^*) \\ &\quad = E\left[\left(f_0(\varphi^*) - \hat{f}_N(\varphi^*)\right)^2 \,\middle|\, \{\varphi(t)\}_{t=1}^N\right] \end{aligned} \tag{14}$$

However, since the true function value $f_0(\varphi^*)$ is unknown, we cannot compute the MSE. Instead we will use a minimax approach, in which we aim at minimizing the *maximum MSE*

$$\max_{f_0 \in \mathcal{F}} MSE(f_0, \hat{f}_N, \varphi^*) \tag{15}$$

It is common to use a linear estimator in the form

$$\hat{f}_N(\varphi^*) = \sum_{t=1}^N w_t y(t) \tag{16}$$

Not surprisingly, it can be shown that when $M(\varphi, \varphi^*) \equiv 0$, the estimator obtained by minimizing the maximum MSE equals what one gets from the corresponding linear least-squares regression (see Roll *et al.* (2005)).

As we will see, sometimes when having some more prior knowledge about the function around φ^*, it will also be natural to consider an affine estimator

$$\hat{f}_N(\varphi^*) = w_0 + \sum_{t=1}^N w_t y(t) \tag{17}$$

instead of (16). This is the estimator that will be considered in the sequel. We will use the notation $w = (w_1 \dots w_N)^T$.

Under assumptions (4), the MSE can be written

$$\begin{aligned} &MSE(f_0, \hat{f}_N, \varphi^*) \\ &= E\left[\left(w_0 + \sum_{t=1}^N w_t(f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2\right] \\ &= \left(w_0 + \sum_{t=1}^N w_t\left(f_0(\varphi(t)) - {\theta^0}^T(\varphi^*)F(\varphi(t))\right)\right. \\ &\quad + {\theta^0}^T(\varphi^*)\left(\sum_{t=1}^N w_t F(\varphi(t)) - F(\varphi^*)\right) \\ &\quad \left. + {\theta^0}^T(\varphi^*)F(\varphi^*) - f_0(\varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^N w_t^2 \end{aligned} \tag{18}$$

Instead of estimating $f_0(\varphi^*)$, one could also estimate a (any) linear combination $B^T\theta^0(\varphi^*)$ of $\theta^0(\varphi^*)$, e.g., ${\theta^0}^T(\varphi^*)F(\varphi^*)$ (cf. Definition 1).

Example 4. Consider the function class of Example 1, and suppose that we would like to estimate $f_0'(\varphi^*)$. From (9) we know that $f_0'(\varphi^*) = \theta_2^0(\varphi^*)$, and so we can use $B = \begin{pmatrix}0 & 1\end{pmatrix}^T$. ◇

In the sequel, we will mostly assume that $f_0(\varphi^*)$ is to be estimated, and hence that the MSE is written according to (18). However, with minor adjustments, all of the following computations and results hold also for estimation of $B^T\theta^0(\varphi^*)$.

By using Definition 1, we get

$$\begin{aligned} &MSE(f_0, \hat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^N |w_t| M(\varphi(t), \varphi^*)\right. \\ &\quad + \left|w_0 + {\theta^0}^T(\varphi^*)\left(\sum_{t=1}^N w_t F(\varphi(t)) - F(\varphi^*)\right)\right| \\ &\quad \left. + M(\varphi^*, \varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^N w_t^2 \end{aligned} \tag{19}$$

3.1 A general computable upper bound on the maximum MSE

In general, the upper bound (19) is not computable, since ${\theta^0}^T(\varphi^*)$ is unknown. However, assume that we know a matrix A, a vector $\bar{\theta} \in \mathcal{D}_\theta$ and a non-negative, convex [1] function $G(w)$, such that for

$$w \in W \triangleq \left\{w \,\middle|\, A\left(\sum_{t=1}^N w_t F(\varphi(t)) - F(\varphi^*)\right) = 0\right\}$$

the following inequality holds:

$$\left|(\theta^0(\varphi^*) - \bar{\theta})^T\left(\sum_{t=1}^N w_t F(\varphi(t)) - F(\varphi^*)\right)\right| \leq G(w)$$

Then we can get an upper bound on the maximum MSE (for $w \in W$)

$$\begin{aligned} &MSE(f_0, \hat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^N |w_t| M(\varphi(t), \varphi^*)\right. \\ &\quad + \left|w_0 + \bar{\theta}^T\left(\sum_{t=1}^N w_t F(\varphi(t)) - F(\varphi^*)\right)\right| \\ &\quad \left. + G(w) + M(\varphi^*, \varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^N w_t^2 \end{aligned} \tag{20}$$

Note that this upper bound just contains known quantities, and thus is computable for any given

[1] In fact, we do not really need $G(w)$ to be convex; what we need is that the upper bound in (20) is convex on W.

w_0 and w. Note also that it is easily minimized with respect to w_0, giving

$$w_0 = -\bar{\theta}^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \tag{21}$$

and yielding the estimator

$$\hat{f}_N(\varphi^*) = \bar{\theta}^T F(\varphi^*) + \sum_{t=1}^{N} w_t \left(y(t) - \bar{\theta}^T F(\varphi(t)) \right)$$

The upper bound on the maximum MSE thus reduces to

$$MSE(f_0, \hat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + G(w) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2, \ w \in W \tag{22}$$

In the following, we will assume that w_0 is chosen according to (21).

Depending on the nature of $\mathcal{D}_\theta$, the upper bound on the maximum MSE may take different forms. Some examples are given in the following subsections.

3.2 The case $\mathcal{D}_\theta = \mathbb{R}^d$

If nothing is known about $\theta^0(\varphi^*)$, the MSE (18) could be arbitrarily large, unless the middle sum is eliminated. This is done by requiring that

$$\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \tag{23}$$

We then get the following upper bound:

$$MSE(f_0, \hat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{24}$$

Comparing to the general case in Section 3.1, this corresponds to $A = I$ and $G(w) = 0$.

The upper bound (24) can now be minimized with respect to w under the constraints (23). By introducing slack variables we can formulate the optimization problem as a convex quadratic program (QP) (Boyd and Vandenberghe, 2004):

$$\begin{aligned} \min_{w,s} \quad & \left(\sum_{t=1}^{N} s_t M(\varphi(t), \varphi^*) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \end{aligned} \tag{25}$$

Example 5. Let us continue with the function class in Example 2. For this class, with $\mathcal{D}_\theta = \mathbb{R}^{n+1}$ and with the notation $\widetilde{\varphi} = \varphi - \varphi^*$, we get the following QP to minimize:

$$\begin{aligned} \min_{w,s} \quad & \frac{L^2}{4} \left(\sum_{t=1}^{N} s_t \|\widetilde{\varphi}(t)\|_2^2 \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t = 1 \\ & \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{26}$$

Note that, in this case, when the weights w are all non-negative, the upper bound (24) is tight and attained by a paraboloid. ◇

Example 6. For the type of systems defined by (11), with $\mathcal{D}_\theta = \mathbb{R}^d$, we would probably like to estimate ${\theta^0}^T F(\varphi^*)$ rather than the artificial $f_0(\varphi^*)$. In this case, the QP becomes

$$\begin{aligned} \min_{w,s} \quad & M^2 \left(\sum_{t=1}^{N} s_t \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \end{aligned} \tag{27}$$

◇

3.3 $\mathcal{D}_\theta$ with p-norm bound

Now suppose we know that $\theta^0(\varphi^*)$ is bounded by

$$\|\theta^0(\varphi^*) - \bar{\theta}\|_p \leq R \tag{28}$$

where $1 \leq p \leq \infty$. Using the Hölder inequality, we can see that the MSE is bounded by

$$\begin{aligned} MSE(f_0, \hat{f}_N, \varphi^*) &\leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + \left| (\theta^0(\varphi^*) - \bar{\theta})^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \right| + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ &\leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + R \left\| \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right\|_q + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \end{aligned} \tag{29}$$

where

$$q = \begin{cases} \infty & p = 1 \\ 1 & p = \infty \\ 1 + \dfrac{1}{p-1} & \text{otherwise} \end{cases} \tag{30}$$

The upper bound is convex in w and can efficiently be minimized. In particular, we can note that if $p = 1$ or $p = \infty$, the optimization problem can be written as a QP. If $p = 2$, we can instead transform the optimization problem into a second-order cone program (SOCP) (Boyd and Vandenberghe, 2004). Comparing to the general case, we get $A = 0$ and

$$G(w) = R \left\| \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right\|_q$$

A special case of interest is if we know some bounds on $\theta^0(\varphi^*)$, i.e.,

$$-\theta^b \preccurlyeq \theta^0(\varphi^*) - \bar{\theta} \preccurlyeq \theta^b \tag{31}$$

– where $\preccurlyeq$ denotes componentwise inequality – which after a simple normalization can be written in the form (28) with $p = \infty$.

3.4 Polyhedral $\mathcal{D}_\theta$

In case $\mathcal{D}_\theta$ can be described by a polyhedron, we can make a relaxation to get a semidefinite program (SDP). This can be done using the S-procedure, but will not be considered further here.

3.5 Combinations of the above

The different shapes of $\mathcal{D}_\theta$ can easily be combined. For instance, a subset of the parameters $\theta_k^0(\varphi^*)$ may be unbounded, while a few may be bounded componentwise, and yet another subset would be bounded in 2-norm. This case would give an SOCP to minimize.

Example 7. Consider Example 2, and suppose that $\varphi^* = 0$. If we, e.g., would know that

$$|f_0(0) - a| \le \delta, \qquad \|\nabla f_0(0) - b\|_2 \le \Delta$$

this would mean that θ_1^0 is bounded within an interval, and that $\begin{pmatrix} \theta_2^0 & \dots & \theta_{n+1}^0 \end{pmatrix}$ is bounded in 2-norm. We could then find appropriate weights w by solving an SOCP. See (Roll, 2003, Chapter 5) for details. ◇

4. MINIMIZING THE EXACT MAXIMUM MSE

In the previous section, we have derived upper bounds on the maximum MSE, which can be efficiently computed and minimized. It would also be interesting to investigate under what conditions the exact maximum MSE can be minimized. In these cases we get the exact, nonasymptotic minimax estimator.

First, note that the MSE (18) for a fixed function f_0 is actually convex in w_0 and w (namely, a quadratic positive semidefinite function; positive definite if $\sigma > 0$). Furthermore, since the maximum MSE is the supremum (over $\mathcal{F}$) of such convex functions, *the maximum MSE is also convex in w_0 and w!*

However, the problem is to compute the supremum over $\mathcal{F}$ for fixed w_0 and w. This is often a nontrivial problem, and we might have to resort to the upper bounds given in the previous section.

In some cases, though, the maximum MSE is actually computable. One case is when considering the function class in Example 1. It can be shown that for each given weight vector w, there is a function attaining the maximum MSE. This function can be constructed explicitly, and hence, we can calculate the maximum MSE. For more details and simulation results, see (Roll, 2003, Section 6.2).

Another case is given by the following theorem. The function classes in, e.g., (Legostaeva and Shiryaev, 1971) and (Sacks and Ylvisaker, 1978) fall into this category.

Theorem 1. Assume that M and θ^0 in (5) do not depend on φ_0. Then, if $\varphi^* \neq \varphi(t)$, $t = 1, \dots, N$, and w is chosen such that $\varphi(t) = \varphi(\tau) \Rightarrow \operatorname{sgn}(w_t) = \operatorname{sgn}(w_\tau)$ for all $t, \tau = 1, \dots, N$, the inequality (19) is tight and attained by any function in $\mathcal{F}$ satisfying

$$f_0(\varphi(t)) = {\theta^0}^T F(\varphi(t)) + \gamma \operatorname{sgn}(w_t) M(\varphi(t)) \tag{32}$$

and

$$f_0(\varphi^*) = {\theta^0}^T F(\varphi^*) - \gamma M(\varphi^*) \tag{33}$$

where

$$\gamma = \operatorname{sgn}\left(w_0 + {\theta^0}^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \right)$$

Here we define $\operatorname{sgn}(0)$ to be 1.

Proof: We first need to observe that there exist functions in $\mathcal{F}$ satisfying (32) and (33). But this follows, since plugging in (32) into (5) gives

$$M(\varphi(t)) \le M(\varphi(t))$$

and similarly for (33), so (5) is satisfied for all these points.

Replacing $f_0(\varphi(t))$ and $f_0(\varphi^*)$ in (18) by the expressions in (32) and (33), respectively, now shows that the bound is tight. □

In general, however, the bound (19) might not be tight.

5. AN EXPRESSION FOR THE WEIGHTS

An interesting property of the solutions to the DWO problems given in Section 3 is that where the bound $M(\varphi, \varphi_0)$ on the approximation error is large enough, the weights will become exactly equal to zero. In fact, we can prove the following theorem:

Theorem 2. Suppose that $\sigma^2 > 0$. If the optimization problem

$$\begin{aligned} \min_w \quad & \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + G(w) \right. \\ & \left. + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ \text{subj. to} \quad & A \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) = 0 \end{aligned} \tag{34}$$

is feasible, there is a μ and a $g \geq 0$ such that the optimal solution w^* is given by

$$\begin{aligned} w_k^* = & \left(\mu^T A F(\varphi(k)) - g \left(M(\varphi(k), \varphi^*) + \nu_k \right) \right)_+ \\ & - \left(-\mu^T A F(\varphi(k)) + g \left(-M(\varphi(k), \varphi^*) + \nu_k \right) \right)_+ \end{aligned} \tag{35}$$

where $(a)_+ = \max\{a, 0\}$ and ν is a subgradient of $G(w)$ (Rockafellar, 1970),

$$\begin{aligned} \nu \in \partial G(w^*) \triangleq \{ & v \in \mathbb{R}^N | \\ & v^T (w' - w^*) + G(w^*) \leq G(w') \quad \forall w' \in \mathbb{R}^N \} \end{aligned}$$

Proof: The proof is based on a special version of the Karush-Kuhn-Tucker (KKT) conditions (Rockafellar, 1970, Cor. 28.3.1) and can be found in (Roll and Ljung, 2004). □

6. CONCLUSIONS

In this paper, we have given a rather general framework, in which the DWO approach can be used for function estimation at a given point. As we have seen from Theorem 2, if the true function can only locally be approximated well by the basis F (i.e., if M is (enough) large far away from φ^*), we get a finite bandwidth property, i.e., the weights corresponding to data samples far away will be zero.

The field is far from being completed. The following list gives some suggestions for further research:

- Different special cases of the general function class given here should be studied further.
- It would also be interesting to study the asymptotic behavior of the estimators, as $N \to \infty$. This has been done for special cases in (Roll *et al.*, 2002; Nazin *et al.*, 2003).
- Another question is what properties $\hat{f}_N(\varphi^*)$ has as a function of φ^*. It is easy to see that $\hat{f}_N$ might not belong to $\mathcal{F}$, due to the noise. From this, two questions arise: What happens on average, and is there a simple (nonlinear) method to improve the estimate in cases where $\hat{f}_N(\varphi^*) \notin \mathcal{F}$?
- In practice, we might not know the function class or the noise variance, and estimation of σ and some function class parameters (such as the Lipschitz constant L in Example 1) may become necessary. One idea on how to do this is presented in (Juditsky *et al.*, 2004). Note that for a function class like in Example 1, we only need to know (or estimate) the ratio L/σ, not the parameters themselves.
- In some cases, explicit expressions for the weights could be given, as was done for the function class in Example 1 in (Roll, 2003, Section 3.2.2).

REFERENCES

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Dennis, Jr., J. E. and R. B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Harris, C., X. Hong and Q. Gan (2002). *Adaptive Modelling, Estimation and Fusion from Data: A Neurofuzzy Approach*. Springer-Verlag.

Juditsky, A., A. Nazin, J. Roll and L. Ljung (2004). Adaptive DWO estimator of a regression function. In: *NOLCOS´04*. Stuttgart.

Legostaeva, I. L. and A. N. Shiryaev (1971). Minimax weights in a trend detection problem of a random process. *Theory of Probability and its Applications* **16**(2), 344–349.

Nazin, A., J. Roll and L. Ljung (2003). A study of the DWO approach to function estimation at a given point: Approximately constant and approximately linear function classes. Technical Report LiTH-ISY-R-2578. Dept. of EE, Linköping Univ., Sweden.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press. Princeton, NJ.

Roll, J. (2003). Local and Piecewise Affine Approaches to System Identification. PhD thesis. Dept. of EE, Linköping Univ., Sweden.

Roll, J., A. Nazin and L. Ljung (2002). A non-asymptotic approach to local modelling. In: *The 41st IEEE Conference on Decision and Control*. pp. 638–643.

Roll, J. and L. Ljung (2004). Extending the direct weight optimization approach. Technical Report LiTH-ISY-R-2601. Dept. of EE, Linköping Univ., Sweden.

Roll, Jacob, Alexander Nazin and Lennart Ljung (2005). Nonlinear system identification via direct weight optimization. *Automatica* **41**(3), 475–490.

Sacks, J. and D. Ylvisaker (1978). Linear estimation for approximately linear models. *The Annals of Statistics* **6**(5), 1122–1137.

Stenman, A. (1999). Model on Demand: Algorithms, Analysis and Applications. PhD thesis. Dept. of EE, Linköping Univ., Sweden.

Suykens, J. A. K., T. van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). *Least Squares Support Vector Machines*. World Scientific. Singapore.