Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Direct Weight Optimization Applied to Discontinuous Functions

Henrik Ohlsson*, Jacob Roll*, Anders Brun†‡, Hans Knutsson†‡, Mats Andersson†‡, Lennart Ljung*
* Div. of Automatic Control, Dept. of Electrical Eng., Linköping University, Sweden
† Div. of Medical Informatics, Dept. of Biomedical Eng., Linköping University, Sweden
‡ Center for Medical Image Science and Visualization, Linköping University, Sweden
{ohlsson, roll, ljung}@isy.liu.se {andbr, knutte, matsa}@imt.liu.se

Abstract— **The Direct Weight Optimization (DWO) approach is a nonparametric estimation approach that has appeared in recent years within the field of nonlinear system identification. In previous work, all function classes for which DWO has been studied have included only continuous functions. However, in many applications it would be desirable also to be able to handle discontinuous functions. Inspired by the bilateral filter method from image processing, such an extension of the DWO framework is proposed for the smoothing problem. Examples show that the properties of the new approach regarding the handling of discontinuities are similar to the bilateral filter, while at the same time DWO offers a greater flexibility with respect to different function classes handled.**

I. INTRODUCTION

The Direct Weight Optimization (DWO) approach [9], [11], [1] is a nonparametric estimation approach that has appeared in recent years within the field of nonlinear system identification. In its original formulation, the DWO method finds pointwise function estimates that minimize an upper bound of the mean square error (MSE), by solving a convex optimization problem. In [1], a variant was proposed, where instead the probability was minimized that an upper bound on the estimation error exceeded a given threshold.

In previous work, all function classes for which DWO has been studied have included only continuous functions. However, in many applications, e.g., when modelling hybrid systems or in image and signal processing applications, one often encounters functions containing discontinuities. Hence, it would be interesting also to be able to handle this case. If not properly adjusted for this, the DWO method (similarly to, e.g., any standard kernel smoothing) smooths out the discontinuous level changes. To avoid this, a modification inspired by the bilateral filter [12], [5] is introduced. The bilateral filter is a kernel smoothing method stemming from image processing, in which the kernel function depends both on the regression vector and the measured outputs. By introducing a similar dependence on the output in the DWO method, discontinuities can be detected and accounted for. In this paper, we limit ourselves to a smoothing problem, where we are given a measurement of the output also for the point where the function is to be estimated (ideas on how to extend this are discussed in Section VI). The results from the modified DWO and the bilateral filter turn out to have similar properties. An advantage with DWO is that it can offer a great flexibility in handling different kinds of function classes, and thus can be easily adapted to different conditions.

The paper is organized as follows: The basic problem is formulated in Section II. DWO in its original form is presented in Section III. Section IV introduces the bilateral filter and proposes how to extend DWO to handle discontinuities. The modified DWO algorithm is exemplified in Section V, and its properties are compared to the bilateral filter.

II. PROBLEM FORMULATION

Let us assume that we are given a set of data $(y(t),\varphi(t))_{t=1}^{N}$, generated from

$$y(t) = f_0(\varphi(t)) + e(t) \tag{1}$$

where f_0 is an unknown function, $f_0 : \mathbb{R}^n \to \mathbb{R}$, and $e(t)$ is independent, identically distributed (i.i.d.) white noise with variance σ^2. Consider the problem of estimating $f_0(\varphi^*)$ for a certain given regression vector φ^*. In this paper, we will assume that φ^* equals one of the values $\varphi(t)$ from the given data set, and hence that we are given a corresponding output value, which will be denoted y^*. In this case, we can regard the problem as a smoothing problem, i.e., with the purpose of eliminating the noise as well as possible. A simple alternative would be to use kernel smoothing (the Nadaraya-Watson estimator [7]) to smooth out the noise. However, with some information of the function sampled, there are more sophisticated methods giving better estimates.

The goal of this paper is to be able to handle the case when $f_0(\varphi)$ is discontinuous. This is considered in Section IV. First, however, the original version of DWO will be described.

III. DIRECT WEIGHT OPTIMIZATION

Consider the smoothing problem described in Section II, and let us assume that we know that f_0 belongs to some function class $\mathscr{F}$. In previous work on DWO (see, e.g., [10]), a standard assumption on $\mathscr{F}$ has been that the member functions of $\mathscr{F}$ locally can be approximately described by a given basis function expansion, and that we can give an upper bound on the approximation error. More precisely, $\mathscr{F}$ is defined as follows (see [10]):

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

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

Here, $F(\cdot)$ is a vector of given basis functions, while ${\theta^0}^T(\varphi_0)F(\varphi)$ is a local (unknown) approximation of $f(\varphi)$ around φ_0, and $M(\varphi,\varphi_0)$ is a given upper bound on the approximation error. Figure 1 illustrates the definition for a case when ${\theta^0}^T(\varphi_0)F(\varphi)$ is linear and the bound $M(\varphi,\varphi_0)$ is quadratic. Note that the local approximation does not need to be explicitly computed; it is enough to know the set of basis functions, F, and how well those can locally approximate f.

Examples of function classes that can be formulated in this way include the class of functions with Lipschitz continuous gradients with a given Lipschitz constant L (in fact, it can be shown that this function class is obtained when selecting F and M as in Figure 1). Also systems with both stochastic and unknown-but-bounded noise terms can be handled within the DWO framework. For more details and examples of function classes covered by Definition 1, see e.g., [10].

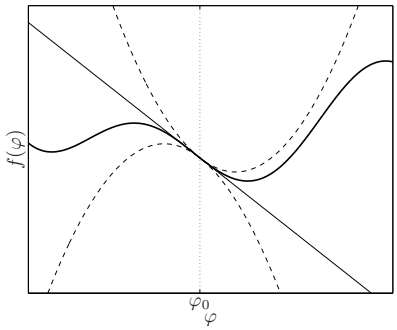

Fig. 1. Illustration of Definition 1: The true function $f(\varphi)$ (thick line), the local approximation ${\theta^0}^T(\varphi_0)F(\varphi)$ (thin line), and the bounds ${\theta^0}^T(\varphi_0)F(\varphi)\pm M(\varphi,\varphi_0)$ (dashed). Here, $F(\varphi)$ and $M(\varphi,\varphi_0)$ are chosen as $F(\varphi)=[1\ \varphi]^T$ and $M(\varphi,\varphi_0)=(\varphi-\varphi_0)^2$.

Now, given this assumption that $f_0\in\mathscr{F}$, with $\mathscr{F}$ defined as in Definition 1, how would we estimate $f_0(\varphi^*)$ for a given point φ^*? The idea behind DWO is to estimate $f_0(\varphi^*)$ by postulating that the estimate should be linear in $y(t)$, i.e.[1],

$$\hat{f}_0(\varphi^*)=\sum_{t=1}^{N} w_t y(t) \tag{2}$$

and determine the weights $w=(w_1,\ldots,w_N)$ by minimizing an upper bound on the maximum mean-squared error (MSE), i.e.,

$$\begin{aligned} &mMSE(\varphi^*,w) \qquad (3)\\ &=\sup_{f_0\in\mathscr{F}} E\left[\left(f_0(\varphi^*)-\hat{f}_0(\varphi^*)\right)^2\,\middle|\,(\varphi(t))_{t=1}^N\right]\\ &=\sup_{f_0\in\mathscr{F}} E\left[\left(f_0(\varphi^*)-\sum_{t=1}^{N} w_t\left(f_0(\varphi(t))+e(t)\right)\right)^2\,\middle|\,(\varphi(t))_{t=1}^N\right] \end{aligned}$$

where in the last expression we have replaced $\hat{f}_0(\varphi^*)$ by using (1) and (2).

[1]If some prior knowledge about the value of θ^0 is given, the estimate should instead be affine in $y(t)$ [10]. Note also that this assumption is not very restrictive. For instance, any least-squares estimation with fixed basis functions gives an estimate that is linear in $y(t)$.

Minimizing the maximum MSE in (3) with respect to w is a convex problem. However, depending on the function class $\mathscr{F}$, the supremum in (3) may be very difficult to compute. Instead, we can give an upper bound to minimize, which leads to the following optimization problem [10]:

$$\begin{aligned} &\min_{w,s}\ \left(\sum_{t=1}^{N}|w_t|M(\varphi(t),\varphi^*)+M(\varphi^*,\varphi^*)\right)^2+\sigma^2\sum_{t=1}^{N}w_t^2 \qquad (4)\\ &\text{subj. to}\ \sum_{t=1}^{N} w_t F(\varphi(t))-F(\varphi^*)=0 \end{aligned}$$

This problem can easily be rewritten as a QP [3], which can be solved efficiently.

Interpreting the role of $M(\varphi(t),\varphi^*)$ intuitively, we can say that a large value of $M(\varphi(t),\varphi^*)$ means that there is a large uncertainty in the relation between the values of $f(\varphi^*)$ and $f(\varphi(t))$. Hence, the information contained in $y(t)$ is not of great value when estimating $f(\varphi^*)$, and the corresponding weight w_t should be small.

In practice, $M(\varphi(t),\varphi^*)$ is often unknown and has to be estimated or selected as a design choice. How this can be done is discussed in [8].

IV. HANDLING DISCONTINUOUS FUNCTIONS

The above algorithm has been shown to handle a number of different function classes containing continuous functions. However, the question on how to describe a class of piecewise continuous functions in a way that suits DWO is still open. Here, we will describe one possible extension, inspired by bilateral filters.

A. The Bilateral Filter

The bilateral filter [12], [6] is commonly seen as a quite ad-hoc method to filter noisy piecewise constant signals. However, connections to weighted least squares have been shown (see [6]). The bilateral filter removes noise and forms an estimate of the noisy signal by a weighted sum of the neighboring points, just like DWO. The bilateral filter is a simple extension of the classical shift invariant convolution filter:

$$\hat{y}(x)=\sum_{x'\in\Omega_x} w_x(x')y(x-x')$$

where w_x is a weighting function with support corresponding to the set Ω_x, and $\sum_{\Omega_x} w_x(x')=1$. To take into account jumps in the signal, the weights are allowed to depend also on y or more precisely, the difference between $y=y(x)$ and $y'=y(x')$. This is done by introducing a weighting function w_y. The bilateral filter output is given by:

$$\hat{y}(x)=\left(\sum_{\Omega_x} w_x(x')\,w_y(y-y')\right)^{-1}\sum_{\Omega_x} w_x(x')\,w_y(y-y')\,y(x-x')$$

The weighting functions are commonly chosen to be Gaussian, i.e.

$$w_x(x)=e^{-\frac{\|x\|^2}{2\sigma_x^2}}, \qquad w_y(y)=e^{-\frac{\|y\|^2}{2\sigma_y^2}}.$$

The result is that only points that are close in both the x- and the y-space will be accumulated in the sum, i.e., the filter is no longer shift invariant but adapts to the local situation.

B. Extending DWO to Handle Discontinuities

Inspired by the bilateral filter, let us now return to DWO. As we have seen, the key to handling abrupt changes in the outputs for the bilateral filter was to include a dependence of $y(t)-y^*$ for the weights. For DWO, since the weights are computed through optimization of a criterion that depends on the assumed function class, the closest parallel to the bilateral filter would be to include knowledge of the measured outputs $y(t)$ and y^* in the description of the function class. This could be done in several ways. Here, we choose to let the approximation error bound M depend on the observed outputs. The optimization problem (4), from which we compute w, is therefore modified as follows:

$$\begin{aligned}\min_{w,s}\ &\left(\sum_{t=1}^{N}|w_t|M(y(t),y^*,\varphi(t),\varphi^*)+M(y^*,y^*,\varphi^*,\varphi^*)\right)^2\\ &+\sigma^2\sum_{t=1}^{N}w_t^2\\ \text{subj. to}\ &\sum_{t=1}^{N}w_tF(\varphi(t))-F(\varphi^*)=0\end{aligned}\tag{5}$$

Typically, $M(y(t),y^*,\varphi(t),\varphi^*)$ will be chosen as a sum of two terms, one depending on $|\varphi(t)-\varphi^*|$ and one depending on $|y(t)-y^*|$.

Intuitively, when $|y(t)-y^*|$ is large, we would suspect that there might be a discontinuity between them, and by increasing M we reduce the influence of $y(t)$ on the estimate $\hat{f}_0(\varphi^*)$.

We can also interpret the modification in terms of function classes. From this perspective, let us use the following type of function class instead of Definition 1:

***Definition** 2:* Let $\mathscr{F}=\mathscr{F}(\mathscr{D},\mathscr{D}_\theta,F,M)$ be the set of all functions $f:\mathscr{D}\to\mathbb{R}$ such that for each $\varphi_0\in\mathscr{D}$, there exists a $\theta^0(\varphi_0)\in\mathscr{D}_\theta$, such that

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

Now, to be able to formulate an optimization problem that is easy to handle, we approximate $f(\varphi(t))$ and $f(\varphi^*)$ in $M(f(\varphi(t)),f(\varphi^*),\varphi(t),\varphi^*)$ with the observed $y(t)$ and y^*, respectively, and thus we obtain (5).

C. A Comparison Between DWO and the Bilateral Filter

As will be seen in the examples, the effect of applying the extended version of DWO is very similar to the results given by an application of the bilateral filter. Given that the bilateral filter is created to filter piecewise constant functions and the similarities between the M of the extended DWO and the filter kernel of the bilateral filter, this is not very surprising. In fact, it can be shown that there is a choice of M giving exactly the same weights as used by the bilateral filter. However, since DWO offers a flexibility in the choice of assumed function classes, it is more general than the bilateral filter in its basic form, and is therefore able to give a better performance for a wider variety of function classes, which can be seen in the example section next.

V. EXAMPLES

Example 1: To study the properties of the extended DWO approach, let us first consider a simple one-dimensional example. Here, 120 measurements were collected from the following function:

$$f(\varphi(t))=\begin{cases}1 & \text{if } -1\le\varphi(t)<-0.7,\\ 0 & \text{if } -0.7\le\varphi(t)<1,\\ 1 & \text{if } 1\le\varphi(t)<2,\\ 1-(\varphi(t)-2) & \text{if } 2\le\varphi(t)<3,\\ 0.1(\varphi(t)-3)^2 & \text{if } 3\le\varphi(t)<5,\end{cases}$$
$$\varphi\in U(-1,5).$$

The measurements $y(t)$ were made noisy by adding normally distributed white noise with a standard deviation of 0.05. An outlier, $y(45)=0$, was added as well.

Figure 2 shows the measurements along with filtered measurements. As can be seen, adding just the knowledge of $y-y_0$ to M improves the result considerably. Notice also the differences between the results using different basis functions. The differences are quite intuitive. Constant basis function handles the piecewise constant parts best while using linear basis functions makes the best job on the linear parts.

Figure 3 shows the corresponding weights for one of the points close to one of the discontinuities ($\varphi^*=1.02$). Nicely, the weights corresponding to points belonging to the other side of the discontinuity are smaller than the weights belonging to the same side as the considered point.

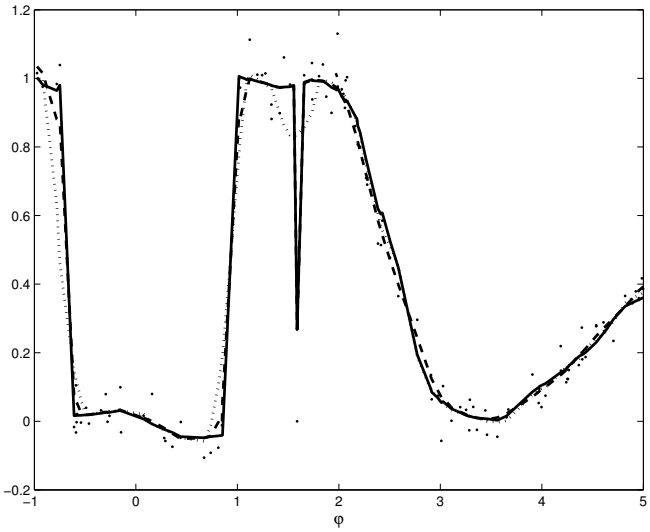

Fig. 2. The result after having applied DWO to a one dimensional example. Dots: measured y values; solid line: constant basis function, $M(y,y_0,\varphi,\varphi_0)=2(\varphi-\varphi_0)^2+\frac{1}{2}(y-y_0)^2$; dashed line: linear basis functions, $M(y,y_0,\varphi,\varphi_0)=2(\varphi-\varphi_0)^2+\frac{1}{2}(y-y_0)^2$; dotted line: ordinary DWO for continuous function classes, with constant basis function, $M(\varphi,\varphi_0)=5(\varphi-\varphi_0)^2$.

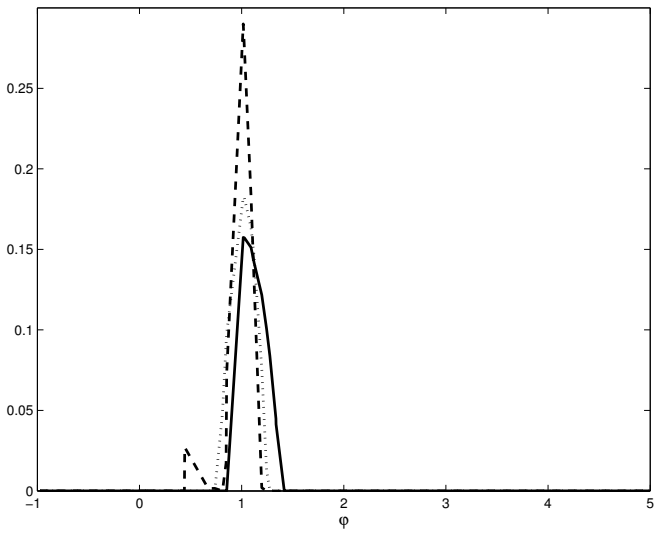

Fig. 3. Weights for one of the points filtered in Figure 2. Solid line: constant basis function, $M(y,y_0,\varphi,\varphi_0) = 2(\varphi-\varphi_0)^2 + \frac{1}{2}(y-y_0)^2$; dashed line: linear basis functions, $M(y,y_0,\varphi,\varphi_0) = 2(\varphi-\varphi_0)^2 + \frac{1}{2}(y-y_0)^2$; dotted line: ordinary DWO for continuous function classes, with constant basis function, $M(\varphi,\varphi_0) = 5(\varphi-\varphi_0)^2$.

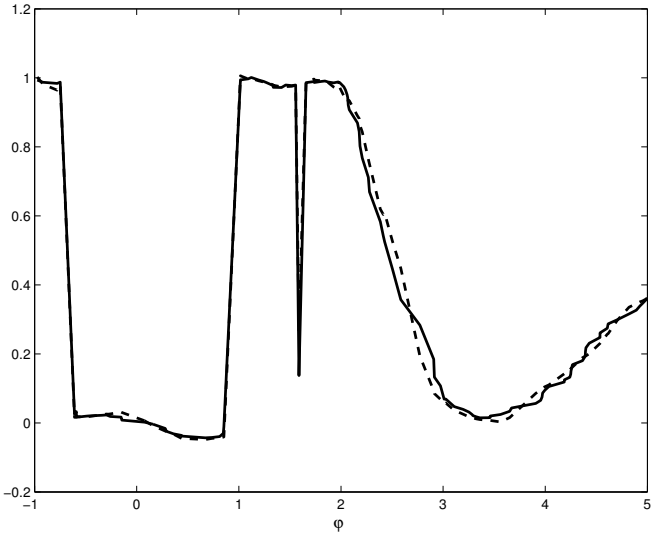

Fig. 4. Bilateral filter and DWO applied to a one dimensional example. A constant basis function together with the M inspired by bilateral filters were used. Dashed line: DWO; solid line: the bilateral filter.

Figure 4 shows both the results from applying the DWO, using a constant basis function and the extended M, and the bilateral filter. There is not much difference for $\varphi^* < 2$ between the two. Figure 5 shows the corresponding weights for one of the filtered points ($\varphi^* = 1.02$), close to one of the discontinuities. For $\varphi^* > 2$ though, the assumption that the sampled function is piecewise constant is not very good, and the performance of the bilateral filter is hence degraded. In contrast, the DWO method still performs well even though using a constant basis function, thanks to the M function which allows a certain deviation from the basis function expansion. Pay attention to the stair-like estimate of the bilateral filter (also called the staircasing effect, see [4]) and the considerably more smooth by the DWO for $\varphi^* > 3$.

The behavior around the outlier deserves attention. When using DWO for a continuous function class, the algorithm tries to smooth the function around the outlier, at the cost of having the neighboring estimates being affected. The discontinuous DWO, on the other hand, treats the outlier as if it belongs to a piece of its own, thus only affecting the neighboring points to a very small extent. A similar behavior can be seen for the bilateral filter (see Figure 4).

Example 2: Another interesting application area is images. In many applications, images of interest contain edges and can be modelled by piecewise continuous functions. In Figure 6, different filtering methods are tested to remove noise from a star-shaped image, containing both sharp and faint edges separating areas of gradually changing gray-levels. The two DWO filters perform better than the lowpass filter and comparable to the bilateral filter. The performance

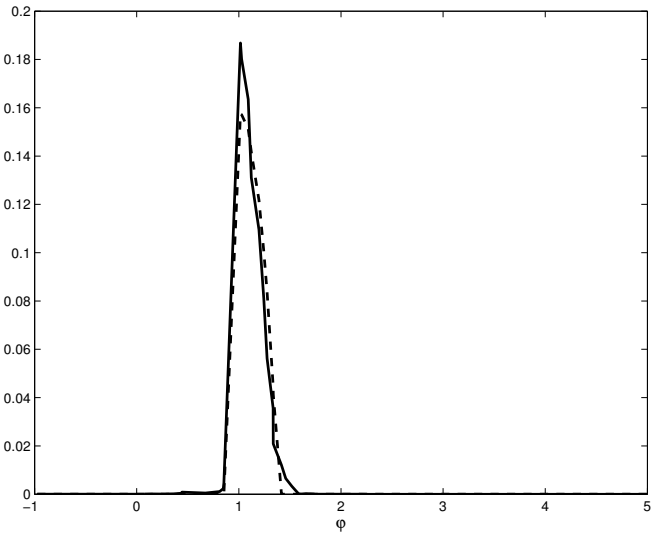

Fig. 5. Weights for one of the filtered points, close to one of the discontinuities, from the bilateral filter and extended DWO (constant basis function). Dashed line: DWO; solid line: the bilateral filter.

of the methods was evaluated via MSE gain[2]. Slightly better MSE gain can be obtained using DWO compared to the bilateral filter (6.86 for DWO with constant basis functions, 7.45 for DWO with linear basis functions, 6.85 for the bilateral filter and 2.82 using a lowpass filter). These results are encouraging but less pronounced than for the previous one-dimensional experiment. It is however clear that the preservation of edges in DWO is similar to the bilateral filter, in particular when comparing the different filters to each other in Figure 7.

[2]The MSE gain is defined as the ratio between the MSE for the non-filtered image and the MSE for the filtered image. A high MSE gain is therefore desirable. See e.g. [5].

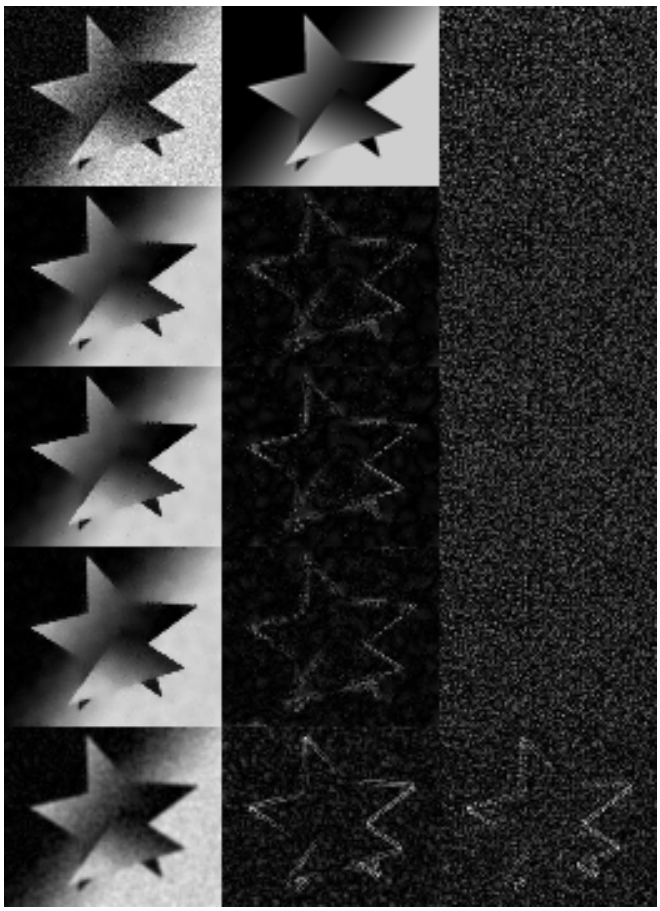

Fig. 6. First row: (l) True image + noise, (m) true image, (r) noise. Second row: DWO with linear basis functions, (l) Filtered image, (m) abs(true image - filtered image), (r) abs(true image + noise - filtered image). Third row: DWO with constant basis functions, (l) Filtered image, (m) abs(true image - filtered image), (r) abs(true image + noise - filtered image). Fourth row: bilateral filtering, (l) Filtered image, (m) abs(true image - filtered image), (r) abs(true image + noise - filtered image). Fifth row: Gaussian filtering, (l) Filtered image, (m) abs(true image - filtered image), (r) abs(true image + noise - filtered image). The same color mapping has been used for all subimages, with the exception that all images showing differences between two images, and the noise image, have been scaled by a factor 3.

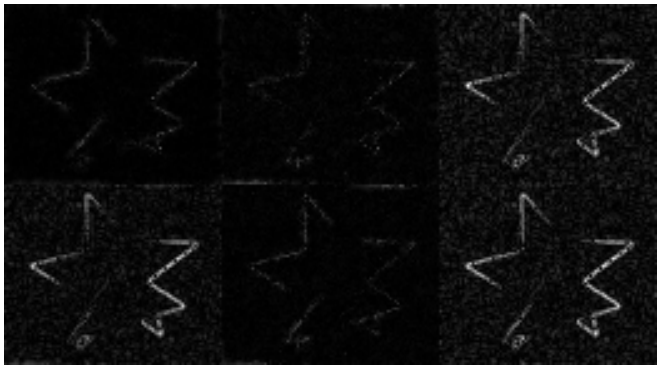

Fig. 7. A comparison between filters, abs(filter1 - filter2): Top row: DWO linear vs DWO constant, DWO linear vs bilateral, DWO linear vs Gaussian. Bottom row: Bilateral vs Gaussian, DWO constant vs bilateral, DWO constant vs Gaussian. Clearly the extensions of DWO are very similar to the bilateral filter. In all comparisons, the largest differences are found close to the edges. The same color mapping and scaling have been used for all subimages.

In this example, M was chosen as

$$M(y,y_0,\varphi,\varphi_0) = \frac{L_\varphi}{2}(\varphi-\varphi_0)^2 + \frac{L_y}{2}m(|y-y_0|/d),$$

where $m(x)$ is a Huber-like norm;

$$m(x) = \begin{cases} x^\alpha & \text{if } x \le 1, \\ \beta x - (\beta - 1) & \text{if } x > 1. \end{cases}$$

The scale parameter d was chosen so that two measurements from the same level would be given a small contribution to the uncertainty bound, despite the differences caused by noise, while the gap between measurements from two different levels should cause a large contribution. d should therefore be chosen smaller than the smallest difference between two neighboring levels, and large enough so that additions made by the noise will be smaller than d. Too small L_φ and L_y will make the image blurry and too large will not smooth out the noise.

VI. DISCUSSION AND CONCLUSIONS

The paper has shown how DWO can be extended to handle discontinuities in a similar way as the bilateral filter known from image processing. The results for both one-dimensional signals and images show that the DWO framework can offer better noise reduction than the bilateral filter, in particular through a more rigorous modelling of the signal, while at the same time preserving edges similarly to the bilateral filter. The approach can also be immediately applied to higher-dimensional problems.

The extended DWO method, as well as the bilateral filter, is a smoothing filter, and can currently only find function estimates at points where measurements are available. In order to apply the method to prediction problems, we must find a way to handle the case also when no measured y^* is given. One way would be to use DWO in its original form (or some other local modelling method) to get a preliminary estimate $\hat{y}^*$, and then use the extended DWO to handle possible discontinuities. In image processing, this would also be a way to tackle the inpainting problem, which has previously been discussed by, e.g., [2].

VII. ACKNOWLEDGMENTS

This work was supported by the Strategic Research Center MOVIII, funded by the Swedish Foundation for Strategic Research, SSF.

REFERENCES

[1] Er-Wei Bai and Yun Liu. Recursive direct weight optimization in nonlinear system identification: A minimal probability approach. *IEEE Transactions on Automatic Control*, 52(7):1218–1231, July 2007.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[4] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. The staircasing effect in neighborhood filters and its solution. *IEEE Trans Image Process*, 15(6):1499–505, 2006.

[5] M. Elad. On the origin of the bilateral filter and ways to improve it. *ieee-ip*, 11(10):1141–1151, October 2002.
[6] M. Elad. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on Image Processing*, 11(10):1141–1151, October 2002.
[7] W. Härdle. *Applied Nonparametric Regression*. Number 19 in Econometric Society Monographs. Cambridge University Press, 1990.
[8] Jacob Roll. Piecewise linear solution paths with application to direct weight optimization. To appear in Automatica, 2008.
[9] Jacob Roll, Alexander Nazin, and Lennart Ljung. A non-asymptotic approach to local modelling. In *The 41st IEEE Conference on Decision and Control*, pages 638–643, Las Vegas, December 2002.
[10] Jacob Roll, Alexander Nazin, and Lennart Ljung. A general direct weight optimization framework for nonlinear system identification. In *16th IFAC World Congress on Automatic Control*, pages Mo–M01–TO/1, Prague, September 2005.
[11] Jacob Roll, Alexander Nazin, and Lennart Ljung. Nonlinear system identification via direct weight optimization. *Automatica*, 41(3):475–490, March 2005.
[12] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proc. 6th Int. Conf. Computer Vision*, pages 839–846, New Delhi, India, 1998.