Technical report from Automatic Control at Linköpings universitet

Direct Weight Optimization in Statistical Estimation and System Identification

Alexander V. Nazin, Jacob Roll, Lennart Ljung, Ion Grama

Division of Automatic Control

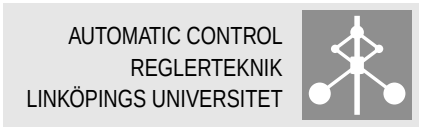
E-mail: nazine@ipu.rssi.ru, roll@isy.liu.se, ljung@isy.liu.se, ion.grama@univ-ubs.fr

14th November 2007

Report no.: LiTH-ISY-R-2831
Accepted for publication in SICPRO'08

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se



Abstract

The Direct Weight Optimization (DWO) approach to statistical estimation and the application to nonlinear system identification has been proposed and developed during the last few years. Computationally, the approach is typically reduced to a convex (e.g., quadratic or conic) program, which can be solved efficiently. The optimality or sub-optimality of the obtained estimates, in a minimax sense w.r.t. the estimation error criterion, can be analyzed under weak a priori conditions. The main ideas of the approach are discussed here and an overview of the obtained results is presented.

Keywords: Statistical estimation, Nonparametric identification, Minimax techniques, Convex programming, Nonlinear systems, Estimation error

Direct Weight Optimization in Statistical Estimation and System Identification

A. V. Nazin*, J. Roll†, L. Ljung†, I. Grama‡

Abstract

The Direct Weight Optimization (DWO) approach to statistical estimation and the application to nonlinear system identification has been proposed and developed during the last few years. Computationally, the approach is typically reduced to a convex (e.g., quadratic or conic) program, which can be solved efficiently. The optimality or sub-optimality of the obtained estimates, in a minimax sense w.r.t. the estimation error criterion, can be analyzed under weak a priori conditions. The main ideas of the approach are discussed here and an overview of the obtained results is presented.

1 Introduction

Identification of nonlinear systems is a very broad and diverse field. Very many approaches have been suggested, attempted and tested. See among many references, e.g., [20, 6, 22, 17, 23, 3]. In this paper we represent a new perspective on nonlinear system identification, which we call *Direct Weight Optimization, DWO*. It is based on postulating an estimator that is linear in the observed outputs and then determining the weights in this estimator by direct optimization of a suitably chosen (min-max) criterion. The presented results on regression function estimation and on system identification are published at greater length in [16]; see also [11, 18, 12]. A recent paper [1] should be noted where a recursive DWO method for nonlinear system identification based on minimal probability criterion is proposed. Moreover, we also extend the DWO approach here to a classic statistical problem of probability density function (pdf) estimation from an observed i.i.d. sample. The extension is based on reducing the problem to a regression function estimation and on further application of the developed DWO ideas.

*Institute of Control Sciences, RAS, 65 Profsoyuznaya, Moscow 117997, Russia, e-mail: nazine@ipu.rssi.ru. The work of the first author has been partly supported by Russian Foundation for Basic Research via grant RFBR 06-08-01474. The first author also gratefully acknowledges the Division of Automatic Control, Linköping University, and the Laboratoire de Mathématiques et Application des Mathématiques, Université de Bretagne Sud, for their invitations.

†Div. of Automatic Control, Linköping University, SE-58183 Linköping, Sweden, e-mail: roll, ljung@isy.liu.se

‡LMAM, Université de Bretagne Sud, CERYC – Campus Tohannic, BP 573, F-56017 Vannes CEDEX, France, e-mail: ion.grama@univ-ubs.fr

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

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

Here, $\varphi(t)$ is the regression vector, $\alpha=(\alpha_1,\ldots,\alpha_d)^T$, $\beta=(\beta_1,\ldots,\beta_l)^T$, 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 $\theta_k=\alpha_k$)

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

where we use the notation

$$F(\varphi)=\big(f_1(\varphi),\quad\ldots,\quad f_d(\varphi)\big)^T \tag{3}$$

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., [6], [22]).

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

$$y(t)=f_0(\varphi(t))+e(t) \tag{4}$$

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 [21]. Similar problems have also been studied within local polynomial modelling [5], although mostly based on asymptotic arguments.

The DWO approach was first proposed in [17] and presented in detail in [14, 18]. Those presentations mainly consider differentiable functions f_0, for which a Lipschitz bound on the derivatives is given (see Examples 1 and 2 below). In Sections 2–5 we suggest an extension to a much more general framework, which contains several interesting special cases, including the ones mentioned above. In Section 5, a general theorem about the structure of the optimal solutions is also given. Sections 6–8[1]are devoted to the DWO approach application for estimating approximately linear functions (see [10] for extensions and further details). Their objective is twofold. We first find the MSE minimax lower bound among arbitrary estimators (Subsection 7.1). Then we study both the DWO-optimal weights and the DWO-optimal MSE upper bound; the latter is further

[1]The results of Sections 2–8 have been jointly obtained by L. Ljung, J. Roll, and A. Nazin during the visit of the latter to Linköping University (Sweden) in 2002–2005.

compared with the MSE minimax lower bound (Subsection 7.2). Experiment design issues are also studied (Section 8). As we will see, some of the results obtained here hold for an arbitrary fixed design $\{\varphi(t)\}$ and a fixed number of observations N while others are of asymptotic consideration, as $N \to \infty$, and of equidistant (or uniform random) design. Particularly, under equidistant design the upper and lower bounds coincide when $|\varphi^*| < 1/6$ which implies the DWO-optimal weights are positive. An extension of DWO approach to pdf estimation is represented in Section 9. It may be treated as an optimal method of smoothing the initially undersmoothed kernel estimates of an unknown pdf from a Lipschitz a priori given class, for a finite sample size n. Asymptotic properties are also studied in order to compare with classic results. In particular, it is demonstrated that the resulting DWO pdf estimator possesses asymptotically optimal rate of convergence when $nh^3 \to 0$, where h stands for a window size (bandwidth). Thus, the DWO pdf estimator can be treated as an approximation for its optimal linear counterpart and, in this sense, represents its easier countable version. Some particular studies and examples are moved into Appendices. Finally, conclusions are given in Section 12.

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

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

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 should also remark that $\theta^0(\varphi_0)$ in (5) depends on f. 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 [19], 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

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

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

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

This could be treated by choosing the fixed basis functions as

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

For each φ_0, f_0 satisfies [4, 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

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

$\lozenge$

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

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

In accordance with (9), we now get

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

$\lozenge$

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

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

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

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

i.e., f_0 satisfies

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

This case is studied in Sections 6–8. Some other examples are given in [19]. $\Diamond$

3 Criterion and estimator

Now, the problem to solve is to find an estimator $\widehat{f}_N$ 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

$$MSE(f_0, \widehat{f}_N, \varphi^*) = E\left[\left(f_0(\varphi^*) - \widehat{f}_N(\varphi^*)\right)^2 \;\middle|\; \{\varphi(t)\}_{t=1}^N\right]$$

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*

$$\max_{f_0 \in \mathcal{F}} MSE(f_0, \widehat{f}_N, \varphi^*) \tag{14}$$

It is common to use a linear estimator in the form

$$\widehat{f}_N(\varphi^*) = \sum_{t=1}^N w_t y(t) \tag{15}$$

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 [18]).

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

$$\widehat{f}_N(\varphi^*) = w_0 + \sum_{t=1}^N w_t y(t) \tag{16}$$

instead of (15). This is the estimator that will be considered in the sequel. We will use the notation $w = (w_1, \ldots, w_N)^T$ for the vector of weights. Note that (16) represents a nonparametric estimator, since the parameter number N is in fact the number of samples (see, e.g., [7]). Such a problem was studied in [19], where a DWO-related method was also proposed.

Under assumption (4), the MSE can be written

$$
\begin{aligned}
MSE(f_0, \widehat{f}_N, \varphi^*) &= E\left[\left(w_0 + \sum_{t=1}^{N} w_t(f_0(\varphi(t)) + e(t)) - f_0(\varphi^*)\right)^2\right] \\
&= \left(w_0 + \sum_{t=1}^{N} w_t\left(f_0(\varphi(t)) - {\theta^0}^T(\varphi^*)F(\varphi(t))\right)\right. \\
&\quad + {\theta^0}^T(\varphi^*)\left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*)\right) \\
&\quad \left. + {\theta^0}^T(\varphi^*)F(\varphi^*) - f_0(\varphi^*)\right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2
\end{aligned} \tag{17}
$$

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 (17). 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

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

3.1 A general computable upper bound on the maximum MSE

In general, the upper bound (18) 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[2] 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:

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

[2]In fact, we do not really need $G(w)$ to be convex; what we need is that the upper bound in (19) is convex on W.

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

$$MSE(f_0, \widehat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) \right.$$
$$\left. + \left| w_0 + \bar{\theta}^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \right| + G(w) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{19}$$

Note that this upper bound just contains known quantities, and thus is computable for any given w_0 and w. Note also that it is easily minimized with respect to w_0, giving

$$w_0 = -\bar{\theta}^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \tag{20}$$

and yielding the estimator

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

The upper bound on the maximum MSE thus reduces to

$$MSE(f_0, \widehat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + G(w) + M(\varphi^*, \varphi^*) \right)^2 \tag{21}$$
$$+ \sigma^2 \sum_{t=1}^{N} w_t^2, \quad w \in W$$

In the following, we will assume that w_0 is chosen according to (20).

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 (17) could be arbitrarily large, unless the middle sum is eliminated. This is done by requiring that

$$\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \tag{22}$$

We then get the following upper bound:

$$MSE(f_0, \widehat{f}_N, \varphi^*) \leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \tag{23}$$

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

The upper bound (23) can now be minimized with respect to w under the constraints (22). By introducing slack variables we can formulate the optimization problem as a convex quadratic program (QP) [2]:

$$\begin{aligned} \min_{w,s} \quad & \left(\sum_{t=1}^{N} s_t M(\varphi(t), \varphi^*) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \end{aligned} \tag{24}$$

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:

$$\begin{aligned} \min_{w,s} \quad & \frac{L^2}{4} \left(\sum_{t=1}^{N} s_t \|\widetilde{\varphi}(t)\|_2^2 \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t = 1, \quad \sum_{t=1}^{N} w_t \widetilde{\varphi}(t) = 0 \end{aligned} \tag{25}$$

Note that, in this case, when the weights w are all non-negative, the upper bound (23) 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

$$\begin{aligned} \min_{w,s} \quad & M^2 \left(\sum_{t=1}^{N} s_t \right)^2 + \sigma^2 \sum_{t=1}^{N} s_t^2 \\ \text{subj. to} \quad & s_t \geq \pm w_t \\ & \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) = 0 \end{aligned} \tag{26}$$

♢

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

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

$$\|\theta^0(\varphi^*) - \bar{\theta}\|_p \leq R \tag{27}$$

where $1 \leq p \leq \infty$. Using the Hölder inequality, we can see from (18) and (20) that the MSE is bounded by

$$\begin{aligned} MSE(f_0, \widehat{f}_N, \varphi^*) &\leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) \right. \\ &\quad \left. + \left| (\theta^0(\varphi^*) - \bar{\theta})^T \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) \right| + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \\ &\leq \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) \right. \\ &\quad \left. + R \left\| \sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right\|_q + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \end{aligned} \tag{28}$$

where

$$q = \begin{cases} \infty & p = 1 \\ 1 & p = \infty \\ 1 + \frac{1}{p-1} & \text{otherwise} \end{cases} \tag{29}$$

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) [2]. Comparing to the general case of Section 3.1, 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.,

$$-\theta^b \preccurlyeq \theta^0(\varphi^*) - \bar{\theta} \preccurlyeq \theta^b \tag{30}$$

– where $\preccurlyeq$ denotes componentwise inequality – which after a simple normalization can be written in the form (27) 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

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

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 [14, Chapter 5] for details. $\lozenge$

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 (17) 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 [14, Section 6.2].

Another case is given by the following theorem. The function classes in, e.g., [9] and [19] 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 (18) is tight and attained by any function in $\mathcal{F}$ satisfying*

$$f_0(\varphi(t)) = {\theta^0}^T F(\varphi(t)) + \gamma \operatorname{sgn}(w_t) M(\varphi(t)) \tag{31}$$

and

$$f_0(\varphi^*) = {\theta^0}^T F(\varphi^*) - \gamma M(\varphi^*) \tag{32}$$

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 (31) and (32). But this follows, since plugging in (31) into (5) gives

$$M(\varphi(t)) \leq M(\varphi(t))$$

and similarly for (32), so (5) is satisfied for all these points.

Replacing $f_0(\varphi(t))$ and $f_0(\varphi^*)$ in (17) by the expressions in (31) and (32), respectively, now shows that the bound is tight. $\square$

In general, however, the bound (18) 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*

$$\begin{aligned} \min_{w} \quad & \left(\sum_{t=1}^{N} |w_t| M(\varphi(t), \varphi^*) + G(w) + M(\varphi^*, \varphi^*) \right)^2 + \sigma^2 \sum_{t=1}^{N} w_t^2 \qquad (33) \\ \text{subj. to} \quad & A \left(\sum_{t=1}^{N} w_t F(\varphi(t)) - F(\varphi^*) \right) = 0 \end{aligned}$$

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

$$\begin{aligned} w_k^* = & \left(\mu^T A F(\varphi(k)) - g \left(M(\varphi(k), \varphi^*) + \nu_k \right) \right)_+ \\ & - \left(-\mu^T A F(\varphi(k)) + g \left(-M(\varphi(k), \varphi^*) + \nu_k \right) \right)_+ \qquad (34) \end{aligned}$$

where $(a)_+ = \max\{a, 0\}$ and $\nu = (\nu_1 \dots \nu_N)^T$ is a subgradient of $G(w)$ at the point $w = w^*$ [13],

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

Proof. The proof is based on a special version of the Karush-Kuhn-Tucker (KKT) conditions [13, Cor. 28.3.1] and can be found in [15]. □

6 DWO for approximately linear functions

We now study the DWO approach to estimating a regression function for the class of approximately linear functions, i.e., functions whose deviation from an affine function is bounded by a known constant. Upper and lower bounds for the asymptotic maximum MSE are given below, some of which also hold in the non-asymptotic case and for an arbitrary fixed design. Their coincidence is then studied. Particularly, under mild conditions, it can be shown that there is always an interval in which the DWO-optimal estimator is optimal among all estimators. Experiment design issues are also studied.

Let us study the particular problem of estimating an unknown univariate function $f_0 : [-0.5, 0.5] \to \mathbb{R}$ at a fixed point $\varphi^* \in [-0.5, 0.5]$ from the given dataset $\{\varphi(t), y(t)\}_{t=1}^N$ with equation (4), i.e.,

$$y(t) = f_0(\varphi(t)) + e(t), \quad t = 1, \dots, N \qquad (35)$$

where $\{e(t)\}_{t=1}^N$ is a random sequence of uncorrelated, zero-mean Gaussian variables with a known constant variance $Ee^2(t) = \sigma^2 > 0$.

Here, DWO for the class of approximately linear functions is studied. This class $\mathcal{F}_1(M)$ consists of functions whose deviation from an affine function is bounded by a known constant $M > 0$ (cf Example 3):

$$\mathcal{F}_1(M) = \left\{ f : [-0.5, 0.5] \to \mathbb{R} \,\middle|\, f(\varphi) = \theta_1 + \theta_2 \varphi + r(\varphi), \theta \in \mathbb{R}^2, \, |r(\varphi)| \leq M \right\} \qquad (36)$$

The DWO-estimator $\widehat{f}_N(\varphi^*)$ is defined as in (15), i.e.,

$$\widehat{f}_N(\varphi^*) = \sum_{t=1}^{N} w_t y(t) \tag{37}$$

where the weights $w = (w_1, \ldots, w_N)^T$ are chosen to minimize an upper bound on $U_N(w)$ on the worst-case MSE:

$$U_N(w) \geq \sup_{f_0 \in \mathcal{F}_1(M)} E_{f_0} \left(\widehat{f}_N(\varphi^*) - f_0(\varphi^*) \right)^2 \tag{38}$$

It can be shown [16] that the RHS of (38) is infinite unless the following constraints are satisfied:

$$\sum_{t=1}^{N} w_t = 1, \quad \sum_{t=1}^{N} w_t \varphi(t) = \varphi^* \tag{39}$$

Under these constraints, on the other hand, we can choose the following upper bound to minimize:

$$U_N(w) = \sigma^2 \sum_{t=1}^{N} w_t^2 + M^2 \left(1 + \sum_{t=1}^{N} |w_t| \right)^2 \to \min_w \tag{40}$$

See [16] for further details.

A solution to the convex optimization problem (40), (39) is denoted by w^*, and its components w_t^* are called the DWO-optimal weights. The corresponding estimate is also called DWO-optimal.

The main study below is devoted to an arbitrary *fixed design* $\{\varphi(t)\}_{t=1}^N$ having at least two different regressors $\varphi(t)$. We also assume that $\varphi(t) \neq \varphi^*$, $t = 1, \ldots, N$, for the sake of simplicity. Further details are then given for equidistant design, i.e.,

$$\varphi(t) = -0.5 + t/N, \quad t = 1, \ldots, N \tag{41}$$

We also discuss the extension to uniform *random design* when regressors $\varphi(t)$ are uniformly distributed on $[-0.5, 0.5]$, i.i.d. random variables, and $\{e(t)\}_{t=1}^N$ being independent of $\{\varphi(t)\}_{t=1}^N$.

7 DWO-estimator: Upper and Lower Bounds

The results in this section may be immediately extended also to multivariate functions $f : D \subset \mathbb{R}^d \to \mathbb{R}$. However, for the sake of simplicity, we consider below the case of $d = 1$.

7.1 Minimax Lower Bound

Consider an arbitrary estimator $\widetilde{f}_N = \widetilde{f}_N(y_1^N, \varphi_1^N)$ for $f_0(\varphi^*)$, i.e., an arbitrary measurable function of the observation vectors $y_1^N = (y(1), \ldots, y(N))^T$ and $\varphi_1^N = (\varphi(1), \ldots, \varphi(N))^T$. Introduce

$$e_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$$

and the shifted regressors

$$\widetilde{\varphi}(t) = \varphi(t) - \varphi^*.$$

Assertion 1. *For any $N > 1$, any estimator $\widetilde{f}_N$, and an arbitrary fixed design the following lower bound holds true:*

$$\sup_{f_0 \in \mathcal{F}_1(M)} E_{f_0}(\widetilde{f}_N - f_0(\varphi^*))^2 \geq 4M^2 + e_1^T J_N^{-1} e_1. \tag{42}$$

Here the information matrix

$$J_N = \frac{1}{\sigma^2} \sum_{t=1}^{N} \begin{pmatrix} 1 & \widetilde{\varphi}(t) \\ \widetilde{\varphi}(t) & \widetilde{\varphi}^2(t) \end{pmatrix} \tag{43}$$

is supposed to be invertible (i.e., there are at least two different $\varphi(t)$ in the data set). Particularly, under equidistant design (41)*, as $N \to \infty$,*

$$\sup_{f_0 \in \mathcal{F}_1(M)} E_{f_0}(\widetilde{f}_N - f_0(\varphi^*))^2 \geq 4M^2 + \frac{\sigma^2}{N}\left(1 + 12\varphi^{*2}\right) + O\left(N^{-2}\right) \tag{44}$$

Proof. See [12] and/or [10]. □

Remark 1. The result of (44) is presented in asymptotical form. However, the term $O\left(N^{-2}\right)$ in (44) can be given explicitly as a function of N.

Remark 2. The same MSE minimax lower bound (44) can be obtained for the uniform random design (and $f_0 \in \mathcal{F}_1(M)$), even non-asymptotically, for any $N > 1$ with the term $O\left(N^{-2}\right) \equiv 0$ in (44); see [12] for details.

Remark 3. Assertion 1 may be extended to non-Gaussian i.i.d. noise sequences $\{e(t)\}$ having a regular probability density function $q(\cdot)$ for $e(t)$. Then, as is seen from the proof, the noise variance σ^2 in (43) and (44) should be changed for the inverse Fisher information $I^{-1}(q)$ where

$$I(q) = \int \frac{q'^2(u)}{q(u)}\, du \tag{45}$$

7.2 DWO-Optimal Estimator

Following the DWO approach we are to minimize the MSE upper bound (40) subject to the constraints (39). The solution to this optimization problem as well as its properties appear to be dependent of φ^*. It turns out that there arise two different cases which are studied below separately.

7.2.1 Positive Weights

When all the DWO-optimal weights are positive, the following assertion shows that the lower bound is then reached.

Assertion 2. *Let $N > 1$, and $\{\varphi(t)\}_{t=1}^N$ be a fixed design where J_N given by* (43) *is invertible, i.e., there are at least two different $\varphi(t)$. Assume that all the DWO-optimal weights w_t^* are positive. Then the DWO-optimal upper bound for the function class* (36) *equals*

$$U_N(w^*) = 4M^2 + e_1^T J_N^{-1} e_1 \tag{46}$$

Particularly, when

$$|\varphi^*| < 1/6 \tag{47}$$

the equidistant design (41) *reduces* (46) *to*

$$U_N(w^*) = 4M^2 + \left(1 + 12{\varphi^*}^2\right)\left(\sigma^2 N^{-1} + O\left(N^{-2}\right)\right) \tag{48}$$

as $N \to \infty$*, with the DWO-optimal weights*

$$w_t^* = \frac{1 + 12\varphi^*\varphi(t)}{N}\left(1 + O\left(N^{-1}\right)\right),\ t = 1, \ldots, N \tag{49}$$

being positive for sufficiently large N*.*

Proof. When the DWO-optimal solution w^* only contains positive components, it is easy to see from (40), (39) that the following optimization problem will have the same optimal solution:

$$\sum_{t=1}^{N} w_t^2 \to \min_{w} \tag{50}$$

subject to the constraints (39). Moreover, the inverse statement holds: If the solution w^{opt} to the optimization problem (50), (39) has only positive components, then $w^* = w^{opt}$.

Now, to prove (46), one needs to minimize $\|w\|_2^2$ subject to the constraints (39). Applying the Lagrange function technique, we arrive at

$$w_t^* = \lambda + \mu\,\widetilde{\varphi}(t), \quad t = 1, \ldots, N \tag{51}$$

with

$$\begin{pmatrix}\lambda\\ \mu\end{pmatrix} = \left(\sum_{t=1}^{N}\begin{pmatrix}1 & \widetilde{\varphi}(t)\\ \widetilde{\varphi}(t) & \widetilde{\varphi}^2(t)\end{pmatrix}\right)^{-1}\begin{pmatrix}1\\ 0\end{pmatrix} = \frac{1}{D_N}\sum_{t=1}^{N}\begin{pmatrix}\widetilde{\varphi}^2(t)\\ -\widetilde{\varphi}(t)\end{pmatrix}, \tag{52}$$

$$D_N = N\sum_{t=1}^{N}\widetilde{\varphi}^2(t) - \left(\sum_{t=1}^{N}\widetilde{\varphi}(t)\right)^2 \tag{53}$$

Thus, from (43) and (52) follows

$$\sum_{t=1}^{N} {w_t^*}^2 = \lambda = \frac{1}{D_N}\sum_{t=1}^{N}\widetilde{\varphi}^2(t) = \frac{1}{\sigma^2}\,e_1^T J_N^{-1} e_1 \tag{54}$$

and we arrive at (46) assuming all the DWO-optimal weights w_t^* are positive. For the equidistant design (41), the results (48)–(49) now follow from straightforward calculations. □

Notice that for Gaussian $e(t)$ the DWO-optimal upper bound (46) coincides with the minimax lower bound (42) which means *minimax optimality* of the DWO-estimator among all estimators, not only among linear ones. For non-Gaussian $e(t)$, similar optimality may be proved in a minimax sense over the class $\mathcal{Q}(\sigma^2)$ of all the densities $q(\cdot)$ of $e(t)$ with bounded variances

$$Ee^2(t) \le \sigma^2 \tag{55}$$

As is well known, condition (55) implies

$$I(q) \geq \sigma^{-2} \tag{56}$$

Hence, see Remark 3, the lower bound

$$\begin{aligned} \sup_{q \in \mathcal{Q}(\sigma^2)} \sup_{f_0 \in \mathcal{F}_1(M)} & E_{f_0}(\widetilde{f}_N - f_0(\varphi^*))^2 \\ & \geq 4M^2 + e_1^T J_N^{-1} e_1 \end{aligned} \tag{57}$$

follows directly from that of (42) with the same matrix J_N as in (43).

From (51)–(54) we can derive a necessary and sufficient condition for the DWO-optimal weights to be positive, which can be explicitly written as

$$\sum_{t=1}^{N} \varphi^2(t) - \varphi^* \sum_{t=1}^{N} \varphi(t) > \frac{1}{2} \left| \sum_{t=1}^{N} \varphi(t) - N\varphi^* \right| \tag{58}$$

At least one point always satisfies (58), namely

$$\varphi^* = \frac{1}{N} \sum_{t=1}^{N} \varphi(t), \tag{59}$$

assuming that J_N is non-degenerate. Thus, inequality (58) defines an interval of all those points φ^* for which the DWO-optimal estimator is minimax optimal among all the estimators.

The exact (non-asymptotic) DWO-optimal weights w_t^* will depend linearly on $\varphi(t)$, as directly seen from (51). Note also, that the analytic study of this subsection was possible to carry out since for the considered case the DWO-optimal weights are all positive, which led to a simpler, equivalent optimization problem (50), (39), having also a positive solution w^*. When there are also non-positive components in the solution of the problem (40), (39), an explicit analytic treatment is more difficult; it is considered below via approximating sums by integrals, for the equidistant design. In general, it follows as a special case of Theorem 2 that the weights satisfy

$$w_t^* = \max\{\lambda_1 + \mu\widetilde{\varphi}(t), 0\} + \min\{\lambda_2 + \mu\widetilde{\varphi}(t), 0\} \tag{60}$$

for some constants $\lambda_1 < \lambda_2$ and μ.

7.2.2 Both positive and non-positive weights

In order to understand (at least on a qualitative level) what may happen when w^{opt} contains both positive and negative components, let us assume equidistant design (41) and introduce the piecewise constant kernel functions $K_w : [-0.5, 0.5] \to \mathbb{R}$ which correspond to an admissible vector w:

$$K_w(\varphi) = \sum_{t=1}^{N} \mathbf{1}\{\varphi(t-1) < \varphi \leq \varphi(t)\}\, N w_t, \quad t = 1, \ldots, N$$

where $\varphi_0 = -0.5$ and $\mathbf{1}\{\cdot\}$ stands for indicator. Now one may apply the following representations for the sums from (40), (39):

$$\sum_{t=1}^{N} |w_t| = \int_{-0.5}^{0.5} |K_w(u)|\, du \tag{61}$$

$$\sum_{t=1}^{N} w_t^2 = \frac{1}{N}\int_{-0.5}^{0.5} K_w^2(u)\,du \tag{62}$$

$$\sum_{t=1}^{N} w_t = \int_{-0.5}^{0.5} K_w(u)\,du \tag{63}$$

$$\sum_{t=1}^{N} w_t\,\varphi(t) = \int_{-0.5}^{0.5} uK_w(u)\,du + O\left(N^{-1}\right) \tag{64}$$

Thus, the initial optimization problem (40), (39) may asymptotically, as $N \to \infty$, be rewritten in the form of the following variational problem:

$$U_N(K) = \frac{\sigma^2}{N}\int_{-0.5}^{0.5} K^2(u)\,du + M^2\left(1 + \int_{-0.5}^{0.5} |K(u)|\,du\right)^2 \to \min_K \tag{65}$$

subject to constraints

$$\int_{-0.5}^{0.5} K(u)\,du = 1, \quad \int_{-0.5}^{0.5} u\,K(u)\,du = \varphi^*. \tag{66}$$

Minimization in (65) is now meant to be over the admissible set D_0 that is the set of all piecewise continuous functions $K : [-0.5, 0.5] \to \mathbb{R}$ meeting constraints (66). The solution to this problem is represented in the following assertion.

Assertion 3. *Let $1/6 < \varphi^* < 1/2$. Then the asymptotically DWO-optimal kernel*

$$K^*(u) = \frac{1}{h}\left(1 + \frac{2}{h}(u - \Delta)\right)\mathbf{1}\{a \le u \le 0.5\} \tag{67}$$

with

$$h = \frac{3}{2}(1 - 2\varphi^*), \quad \Delta = \frac{6\varphi^* - 1}{4}, \quad a = 3\varphi^* - 1 \tag{68}$$

The DWO-optimal MSE upper bound

$$U_N(K^*) = 4\,M^2 + \frac{\sigma^2}{N}\frac{8}{9(1 - 2\varphi^*)}, \tag{69}$$

and the approximation to w^* is given by

$$w_t^* \approx \frac{1}{N}K^*(\varphi_t) \tag{70}$$

Proof. See [10]. □

It is easily seen from (65) that asymptotically, as $N \to \infty$, the influence of the first summand in the RHS (65) becomes negligible, compared to the second one. Hence, we first need to minimize

$$U_N^{(2)}(K) = \int_{-0.5}^{0.5} |K(u)|\,du \to \min_{K \in D_0} \tag{71}$$

However, the solution to (71) is not unique, and it is attained on any non-negative kernel $K \in D_0$. A useful example of such a kernel is the uniform kernel function

$$K_{uni}^*(u) = \frac{1}{1 - 2\varphi^*}\mathbf{1}\{|u - \varphi^*| \le 1 - \varphi^*\}\,. \tag{72}$$

Here and below in the current subsection we assume that $0 \le \varphi^* < 1/2$, for the concreteness. It is straightforward to verify that $K^*_{uni} \in D_0$, and

$$U_N^{(1)}(K^*_{uni}) = \int_{-0.5}^{0.5} K^2(u)\,du = \frac{1}{1-2\varphi^*}. \tag{73}$$

Let us compare this value $U_N^{(1)}(K^*_{uni})$ with that of $U_N^{(1)}(K^*)$ where the DWO-optimal kernel is known for $|\varphi^*| \le 1/6$ to be

$$K^*(u) = (1 + 12\varphi^* u)\,\mathbf{1}\{|u| \le 1/2\} \tag{74}$$

The latter equation corresponds to (49) and may be obtained directly from (65)–(66) in a similar manner. Thus,

$$U_N^{(1)}(K^*) = 1 + 12{\varphi^*}^2. \tag{75}$$

Figure 1 shows $U_N^{(1)}$ for the different kernels, as functions of φ^*.

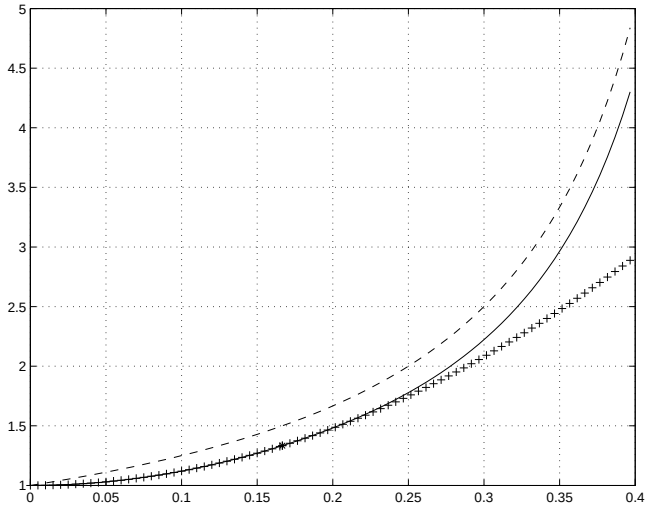


Figure 1: $U_N^{(1)}$ for DWO-optimal (solid) and uniform DWO-suboptimal (dashed) kernels; their minimax lower bound $1 + 12{\varphi^*}^2$ is represented by plus signs; the point $\varphi^* = 1/6$ is marked by a star.

Eq. (60) indicates that an optimal kernel K^* might also contain a negative part. However, asymptotically (as $N \to \infty$), that may not occur since otherwise the main term of the MSE upper bound (65) — the second summand of the RHS (65) — is not minimized.

8 Experiment Design

Let us now briefly consider some experiment design issues. We first find and study the optimal design for a given estimation point $\varphi^* \in (-0.5, 0.5)$ which minimizes the lower bound (42). Then a similar minimax solution is given for $|\varphi^*| \le \delta$ with a given $\delta \in (0, 0.5)$.

8.1 Fixed $\varphi^* \in (-0.5, 0.5)$

Let us fix $\varphi^* \in (-0.5, 0.5)$ and minimize the lower bound (42) with respect to $\{\varphi(t)\}_{t=1}^N$ From (43), (52)–(54) follows that we are to minimize

$$\lambda = \left(N - \frac{\left(\sum_{t=1}^N \widetilde{\varphi}(t)\right)^2}{\sum_{t=1}^N \widetilde{\varphi}^2(t)} \right)^{-1} \tag{76}$$

which is equivalent to

$$\begin{aligned} &\frac{(S_N - N\varphi^*)^2}{V_N - 2\varphi^* S_N + N\varphi^{*2}} \to \min_{|\varphi(t)|\le 1/2}, \\ &S_N = \sum_{t=1}^N \varphi(t), \quad V_N = \sum_{t=1}^N \varphi^2(t) \end{aligned} \tag{77}$$

Thus, the minimum in (77) equals zero and is attained on any design which meets the condition

$$\frac{1}{N} S_N = \varphi^*. \tag{78}$$

One might find a design which maximizes V_N subject to (78), arriving at the one of the form, for instance, $\varphi(t) = \pm 0.5$ with

$$\#\{\varphi(t) = 0.5\} = \frac{N}{2}(1 + 2\varphi^*) \tag{79}$$

and corresponding for $\#\{\varphi(t) = -0.5\}$, assuming the value in RHS (79) is an integer. Since $\lambda = 1/N$ and $\mu = 0$ in (51), and the DWO-optimal weights are uniform, $w_t^* = 1/N$. Hence, the upper and lower bounds coincide and equal

$$U_N(w^*) = 4M^2 + \frac{\sigma^2}{N} \tag{80}$$

In general, however, the RHS of (79) is a non-integer. Then, one might take an integer part in (79), that is put $\#\{\varphi(t) = 0.5\} = \lfloor 0.5N(1 + 2\varphi^*) \rfloor$ and $\#\{\varphi(t) = -0.5\} = N - \#\{\varphi(t) = 0.5\}$, correcting also the value $\varphi(t) = 0.5$ by a term $O(1/N)$. Hence, we will have an additional term $O(N^{-2})$ in the RHS (80).

8.2 Minimax DWO-optimal Design

Assume now $|\varphi^*| \le \delta$ with $0 < \delta \le 0.5$, and, instead of (77), let us find a design solving

$$\max_{|\varphi^*|\le\delta} \frac{(S_N - N\varphi^*)^2}{V_N - 2\varphi^* S_N + N\varphi^{*2}} \to \min_{|\varphi(t)|\le 1/2} \tag{81}$$

The maximum in (64) can be explicitly calculated which reduces (64) to

$$\frac{(|S_N| + N\delta)^2}{V_N + 2\delta|S_N| + N\delta^2} \to \min_{|\varphi(t)|\le 1/2} \tag{82}$$

Evidently, the RHS function in (82) is monotone decreasing w.r.t. V_N and monotone increasing w.r.t. $|S_N|$. Hence, the minimum in (81) would be attained

if $V_N = N/4$ (that is its upper bound) and if $S_N = 0$. Assuming that N is even, these extremal values for V_N and $|S_N|$ are attained under the symmetric design $\varphi(t) = \pm 0.5$ with

$$\#\{\varphi(t) = 0.5\} = \#\{\varphi(t) = -0.5\} = \frac{N}{2} \tag{83}$$

This design ensures the minimax of the DWO-optimal MSE

$$\min_{|\varphi(t)|\le 1/2} \max_{|\varphi^*|\le\delta} U_N(w^*) = 4M^2 + \frac{\sigma^2}{N}\,(1 + 4\delta^2) \tag{84}$$

Particularly, for $\delta = 1/2$,

$$\min_{|\varphi(t)|\le 1/2} \max_{|\varphi^*|\le 1/2} U_N(w^*) = 4M^2 + \frac{2\sigma^2}{N} \tag{85}$$

Putting $\delta = 0$ in (84) yields (80) with $\varphi^* = 0$.

Now, if we apply this design for an arbitrary $\varphi^* \in (-0.5, 0.5)$, we arrive at the DWO-optimal MSE

$$U_N(w^*) = 4M^2 + \frac{\sigma^2}{N}\,\left(1 + 4{\varphi^*}^2\right) \tag{86}$$

with the DWO-optimal weights

$$w_t^* = \frac{1}{N}\,(1 + 4\varphi^*\varphi(t)) \tag{87}$$

which are all positive. Hence, the upper bound (86) coincides with the lower bound (42), and the DWO estimator with weights (87) is minimax optimal for any $\varphi^* \in (-0.5, 0.5)$. For the odd sample size N, one may slightly correct the design, arriving at an additional term $O(N^{-2})$ in the RHS (86), similarly to the previous subsection.

9 DWO-estimator for pdf

Below in Sections 9–11[3], we apply the DWO approach to smooth the initially undersmoothed kernel estimates of an unknown probability density function (pdf) from a Lipschitz a priori given class, for a finite sample size n. Asymptotic properties are also studied in order to compare with classic results. In particular, it is demonstrated that the resulting DWO pdf estimator possesses asymptotically optimal rate of convergence when $nh^3 \to 0$, where h stands for a window size (bandwidth). Thus, the DWO pdf estimator can be treated as an approximation for its optimal linear counterpart and, in this sense, represents its easier countable version.

[3]The results of those Sections as well of the Appendices A–B have been jointly obtained by I. Grama and A. Nazin during the visit of the latter to LMAM/UBS (Vannes, France) in May–June 2007.

9.1 Problem Statement via DWO

9.1.1 Notations and assumptions

Let $\{X_1, \dots, X_n\}$ be a sample of n i.i.d. random variables having Lipschitz pdf $p : [0,1] \to \mathbb{R}_+$, i.e.,

$$|p(t) - p(s)| \le L|t - s| \,.$$

Introduce a partition of the pdf support $[0,1]$ on m intervals (bins) of the same size

$$h = 1/(2m) \,,$$

the points $a_k = (1 + 2k)h$ being the intervals' centers, $k = 1, \dots, m$. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function with a support $\operatorname{supp} K = [-1, +1]$ and

$$\int K(t)\, dt = 1 \,. \tag{88}$$

Assume in what follows that we are to estimate pdf p at a fixed point $x \in [0,1]$,

$$p(x) > 0 \,. \tag{89}$$

Remark 1. Non-equally sized partition can also be treated, as well as extensions to another smoothness classes, different auxiliary estimates $\widehat{p}_k$, etc.

9.1.2 Kernel estimates and their aggregate

Introduce kernel (auxiliary) pdf estimates at points a_k, i.e.,

$$\widehat{p}_k = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - a_k}{h}\right) , \quad k = 1, \dots, m \,. \tag{90}$$

Consequently, their aggregate at a point $x \in [0,1]$ is defined as follows:

$$\widehat{p}(x) = \sum_{k=1}^{m} w_k(x)\, \widehat{p}_k \tag{91}$$

with weights $w_k = w_k(x)$ summing to 1, that is

$$\sum_{k=1}^{m} w_k(x) = 1 \,. \tag{92}$$

9.1.3 Associated regression model

One may treat estimates $\widehat{p}_k$ as the observations in the related regression model with a biased noise [19], that is

$$\widehat{p}_k = p(x) + b_k(x) + \xi_k \tag{93}$$

with the bias term

$$b_k(x) = E\{\widehat{p}_k\} - p(x) \tag{94}$$

and stochastic error

$$\xi_k = \widehat{p}_k - E\{\widehat{p}_k\} \,. \tag{95}$$

The bias term is bounded over the Lipschitz class of pdf's as follows:

$$\begin{aligned} |b_k(x)| &= \left| \frac{1}{h} E\{K\left(\frac{X_1 - a_k}{h}\right)\} - p(x) \right| && (96) \\ &\leq \left| \frac{1}{h} \int p(u)\, K\left(\frac{u - a_k}{h}\right) du - p(a_k) \right| + |p(a_k) - p(x)| \\ &\leq \int_{-1}^{1} |p(a_k + ht) - p(a_k)|\, |K(t)|\, dt + L|a_k - x| \\ &\leq L\rho_k(x) && (97) \end{aligned}$$

where

$$\begin{aligned} \rho_k(x) &\triangleq |a_k - x| + hC_1\,, && (98) \\ C_1 &\triangleq \int_{-1}^{1} |tK(t)|\, dt\,. && (99) \end{aligned}$$

Notice, that the stochastic errors ξ_k are correlated, see below. Their variances are evaluated as follows:

$$\begin{aligned} \sigma_k^2 &= E\{\widehat{p}_k^2\} - (E\{\widehat{p}_k\})^2 && (100) \\ &= \frac{1}{(nh)^2} \left(nE\{K^2\left(\frac{X_1 - a_k}{h}\right)\} + n(n-1) \left(E\{K\left(\frac{X_1 - a_k}{h}\right)\} \right)^2 \right) \\ &\quad - \left(\frac{1}{h} E\{K\left(\frac{X_1 - a_k}{h}\right)\} \right)^2 \\ &= \frac{1}{nh} \left(\int_{-1}^{1} K^2(t)\, p(a_k + ht)\, dt - h \left(\int_{-1}^{1} K(t)\, p(a_k + ht)\, dt \right)^2 \right). && (101) \end{aligned}$$

Particularly, under $h \to 0$ one holds

$$\sigma_k^2 = \frac{p(a_k)}{nh}(1 + O(h)) \int_{-1}^{1} K^2(t)\, dt \qquad (102)$$

where the term $O(h)$ does not depend explicitly on n.

9.1.4 Bias of the estimation error

The estimation error for aggregate $\widehat{p}(x)$ follows from (91)–(95) to be

$$\widehat{p}(x) - p(x) = \sum_{k=1}^{m} w_k(x)\, (b_k(x) + \xi_k)\,. \qquad (103)$$

Thus, its bias

$$b(x) \triangleq \sum_{k=1}^{m} w_k(x)\, b_k(x) \qquad (104)$$

is bounded, due to (96)–(97), as follows

$$|b(x)| \leq L \sum_{k=1}^{m} |w_k(x)|\, \rho_k(x)\,. \qquad (105)$$

Now evaluate the stochastic term

$$\xi(x) \triangleq \sum_{k=1}^{m} w_k(x)\,\xi_k\,. \tag{106}$$

9.1.5 Variance of the estimation error

The variance of the stochastic error (106) may be written as follows:

$$\sigma^2(x) \triangleq E\{\xi^2(x)\} = w^T(x)Bw(x)\,. \tag{107}$$

We denoted here the vector of weights $w(x) \triangleq (w_1(x),\ldots,w_m(x))^T$ and covariance matrix B for the random vector $\xi \triangleq (\xi_1,\ldots,\xi_m)^T$, that is $B = \|\beta_{kl}\|_{m\times m}$ with the diagonal entries $\beta_{kk} = \sigma_k^2$ evaluated in (100)–(102). Evaluate now the non-diagonal entries β_{kl}, $k \neq l$. Notice, that

$$K\left(\frac{X_i - a_k}{h}\right)\cdot K\left(\frac{X_i - a_l}{h}\right) = 0$$

with probability 1. Hence, similarly to (100)–(101) one may write

$$\begin{aligned}\beta_{kl} &= \frac{1}{(nh)^2}\,n(n-1)\,E\{K\left(\frac{X_1 - a_k}{h}\right)\}\,E\{K\left(\frac{X_1 - a_l}{h}\right)\} \qquad (108)\\ &\quad -\frac{1}{h^2}\,E\{K\left(\frac{X_1 - a_k}{h}\right)\}\,E\{K\left(\frac{X_1 - a_l}{h}\right)\}\\ &= -\frac{1}{n}\int_{-1}^{1} K(t)\,p(a_k + ht)\,dt\int_{-1}^{1} K(t)\,p(a_l + ht)\,dt\,. \qquad (109)\end{aligned}$$

Particularly, under $h \to 0$ one holds

$$\beta_{kl} = -\frac{1}{n}p(a_k)p(a_l)\,(1 + O(h)) \tag{110}$$

where $O(h)$ does not depend explicitly on n.

9.1.6 MSE Upper Bound and Quadratic Program

The Mean-Square Error is now written

$$MSE(x) = b^2(x) + \sigma^2(x)\,. \tag{111}$$

Substituting (105) and (107) one obtains an MSE Upper Bound as follows:

$$MSE(x) \le L^2\left(\sum_{k=1}^{m}|w_k(x)|\,\rho_k(x)\right)^2 + w^T(x)Bw(x)\,. \tag{112}$$

Thus, the DWO leads to the following Optimization Problem (OP):

$$\min_{w\in\mathbb{R}^m} L^2\left(\sum_{k=1}^{m}|w_k|\,\rho_k(x)\right)^2 + w^T Bw \tag{113}$$

subject to the constraint

$$\sum_{k=1}^{m} w_k = 1\,. \tag{114}$$

Since matrix B depends on the unknown pdf, we call the OP (113)–(114) the Oracle Optimization Problem (OOP).

The OOP may equivalently be reduced to a Quadratic Program (QP) in a standard manner by introducing auxiliary variables s_k, $k = 1, \ldots, m$, as well as $2m$ additional inequality constraints as follows:

$$s_k \geq w_k\,, \quad s_k \geq -w_k\,, \qquad k = 1, \ldots, m\,. \tag{115}$$

In other words, $s_k \geq |w_k|$. Introducing the auxiliary variable vector

$$s = (s_1, \ldots, s_m)^T$$

one may write the related QP:

$$\min_{(s,w)\in\mathbb{R}^m\times\mathbb{R}^m} L^2 \left(\sum_{k=1}^{m} s_k\, \rho_k(x) \right)^2 + w^T B w \tag{116}$$

subject to constraints

$$\sum_{k=1}^{m} w_k \;=\; 1\,, \tag{117}$$

$$s_k - w_k \;\geq\; 0\,, \tag{118}$$

$$s_k + w_k \;\geq\; 0\,. \tag{119}$$

Thus, OOP of the type (113)–(114) may be effectively solved numerically on a computer with modern software, subject to a given matrix B.

Below we assume that matrix B is positive definite, $B > 0$. This implies that the OOP (113)–(114) is to minimize a strongly convex function over the hyperplane (114). Thus, the problem (113)–(114) has a unique solution. Since B depends on unknown pdf, we give the following definitions.

Definition 2. Let vector $w^* = (w_1^*, \ldots, w_m^*)^T$ be the solution of OOP (113)–(114) for a given point $x \in [0, 1]$. Then the weights w_k^*, $k = 1, \ldots, m$, are called oracle DWO-weights (for the point x).

Definition 3. Let the estimate $\widehat{p}(x)$ be defined by (91) under the oracle DWO-weights $w_k = w_k^*$, $k = 1, \ldots, m$ for a given point $x \in [0, 1]$. Then $\widehat{p}(x)$ is called the oracle DWO-estimate for pdf at the point x.

Lemma 2. *Let $\rho_k(x) > 0$ for all $k = 1, \ldots, m$. A vector $w^* \in \mathbb{R}^m$ is a solution to OOP (113)–(114) iff there exists vector $s^* \in \mathbb{R}^m$ such that the pair (s^*, w^*) is a solution to QP (116)–(119) with $s_k^* = |w_k^*|$, $k = 1, \ldots, m$. Particularly, if $B > 0$ then both the problems have unique solutions.*

Proof is a direct consequence of the inequality

$$L^2 \left(\sum_{k=1}^{m} |w_k|\, \rho_k(x) \right)^2 + w^T B w \leq L^2 \left(\sum_{k=1}^{m} s_k\, \rho_k(x) \right)^2 + w^T B w \tag{120}$$

holding true for all pairs $(s, w) \in \mathbb{R}^m \times \mathbb{R}^m$ subject to constraints (117)–(119) and turning into the exact equality iff $s_k = |w_k|$ for all $k = 1, \ldots, m$. ■

10 Approximate analysis of the OOP (113)–(114)

As it is demonstrated in (100)–(102) and (108)–(110), matrix B is approximately diagonal for a small h, namely

$$B = \frac{\|K\|_2^2}{nh}(D + \mathcal{O}(h)) \tag{121}$$

with a diagonal matrix $D \triangleq \mathrm{diag}\{p(a_1), \dots, p(a_m)\}$ and a symmetric matrix $\mathcal{O}(h)$ (i.e., its norm is of the order $O(h)$).

Remark 2. One finally could change D for its approximate $\mathrm{diag}\{\widetilde{p}_1, \dots, \widetilde{p}_m\}$ with sufficiently good estimates $\widetilde{p}_1, \dots, \widetilde{p}_m$. Another option is to use upper bound for D. We are studying both options further on.

Let us neglect the term $\mathcal{O}(h)$ in (121). This may be explained from a continuous dependence of the minimum value in (113)–(114) on the matrix B. Then the OOP (113)–(114) becomes

$$\min_{w \in \mathbb{R}^m} L^2 \left(\sum_{k=1}^{m} |w_k| \, \rho_k(x) \right)^2 + \kappa^2 \sum_{k=1}^{m} p(a_k) \, w_k^2 \tag{122}$$

subject to constraint

$$\sum_{k=1}^{m} w_k = 1 \tag{123}$$

where

$$\kappa \triangleq \frac{1}{\sqrt{nh}} \|K\|_2 \, . \tag{124}$$

Assertion 4. *Due to positiveness of $\rho_k(x)$ and $p(a_k)$, $k = 1, \dots, m$, the solution to OP (122)–(123) may not contain negative entries, i.e., related DWO-weights are all non-negative.*

Proof Introduce

$$U(w) \triangleq L^2 \left(\sum_{k=1}^{m} |w_k| \, \rho_k(x) \right)^2 + \kappa^2 \sum_{k=1}^{m} p(a_k) \, w_k^2 \, . \tag{125}$$

Let $w \in \mathbb{R}^m$ be such a point that meets constraint (123) and has some negative entries. Evidently, it has also positive entries. The latter may be assumed to be first, say ℓ, entries $w_1, \dots, w_\ell$, without loss of generality. Thus, $w_k \geq 0$ for all $k = 1, \dots, \ell$, and $w_k < 0$ otherwise, and constraint (123) implies

$$S_\ell \triangleq \sum_{k=1}^{\ell} w_k = 1 - \sum_{k=\ell+1}^{m} w_k > 1 \, . \tag{126}$$

Therefore, the weight vector $\widetilde{w} \triangleq \frac{1}{S_\ell}(w_1, \dots, w_\ell, 0, \dots, 0)^T \in \mathbb{R}^m$ meets the constraint (123), and

$$U(\widetilde{w}) = \frac{1}{S_\ell} \left(L^2 \left(\sum_{k=1}^{\ell} w_k \, \rho_k(x) \right)^2 + \kappa^2 \sum_{k=1}^{\ell} p(a_k) \, w_k^2 \right) < U(w) \, , \tag{127}$$

the admissible point $\widetilde{w}$ is "better" then w. The contradiction ends the proof. ■

Analytic solution for the OP (122)–(123)

Now use the Assertion 4 and assume, without loss of generality, that the first ℓ^* DWO-weights are positive and the rest are all zeros. Let us introduce integer variable $\ell \in [2, \ell^*]$. In other words, consider minimization of $U(w)$ (122) subject to both constraint (123) and $w_k = 0$ for all $k > \ell$. Therefore, one may write $|w_k| = w_k$ in (122) and arrive at Lagrange function

$$\mathcal{L}(w, \lambda) = L^2 \left(\sum_{k=1}^{\ell} w_k \, \rho_k(x) \right)^2 + \kappa^2 \sum_{k=1}^{\ell} p(a_k) \, w_k^2 - \lambda \left(\sum_{k=1}^{\ell} w_k - 1 \right) \tag{128}$$

with a Lagrange multiplier λ. The partial derivative

$$\frac{\partial \mathcal{L}}{\partial w_k} = 2L^2 \left(\sum_{j=1}^{\ell} w_j \, \rho_j(x) \right) \rho_k(x) + 2\kappa^2 p(a_k) \, w_k - \lambda \, . \tag{129}$$

Due to optimizing a quadratic function over a hyperplane, this leads to the necessary and sufficient conditions for the OP solution:

$$2L^2 \left(\sum_{j=1}^{\ell} w_j \, \rho_j(x) \right) \rho_k(x) + 2\kappa^2 p(a_k) \, w_k - \lambda \;\; = \;\; 0 \, , \quad k = 1, \ldots, \ell, \tag{130}$$

$$\sum_{j=1}^{\ell} w_j \;\; = \;\; 1 \, . \tag{131}$$

In order to find the solution to this system of linear equations, we first sum in (130). This gives

$$\lambda = 2 \left(L^2 r \ell \overline{\rho} + \kappa^2 \overline{p} \right) \tag{132}$$

with

$$r \;\; \triangleq \;\; \frac{1}{\ell} \sum_{k=1}^{\ell} w_k \, \rho_k(x) \, , \tag{133}$$

$$\overline{\rho} \;\; \triangleq \;\; \frac{1}{\ell} \sum_{k=1}^{\ell} \rho_k(x) \, , \tag{134}$$

$$\overline{p} \;\; \triangleq \;\; \frac{1}{\ell} \sum_{k=1}^{\ell} w_k \, p(a_k) \, . \tag{135}$$

Furthermore, from (130)–(135) one obtains

$$w_k = \frac{1}{p(a_k)} \left(\overline{p} + r\ell \frac{L^2}{\kappa^2} (\overline{\rho} - \rho_k(x)) \right) , \quad k = 1, \ldots, \ell \tag{136}$$

with

$$r \;\; = \;\; \frac{1}{\ell} \, \frac{\overline{\rho/p}}{\overline{1/p} + \dfrac{\ell L^2}{\kappa^2} \left(\overline{1/p} \cdot \overline{\rho^2/p} - \overline{\rho/p}^{\,2} \right)} \, , \tag{137}$$

$$\overline{p} \;\; = \;\; \frac{1}{\ell} \, \frac{1 + \dfrac{\ell L^2}{\kappa^2} \left(\overline{\rho^2/p} - \overline{\rho} \cdot \overline{\rho/p} \right)}{\overline{1/p} + \dfrac{\ell L^2}{\kappa^2} \left(\overline{1/p} \cdot \overline{\rho^2/p} - \overline{\rho/p}^{\,2} \right)} \, , \tag{138}$$

where

$$\overline{1/p} \triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{1}{p(a_k)}\,, \tag{139}$$

$$\overline{\rho/p} \triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{\rho_k(x)}{p(a_k)}\,, \tag{140}$$

$$\overline{\rho^2/p} \triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{\rho_k^2(x)}{p(a_k)}\,. \tag{141}$$

Now, we check that the weights (136) are all positive, with r being defined by (137). Thus, the positiveness of the DWO-weights $w_1,\ldots,w_\ell$ given by (136) is equivalent to the inequality

$$\max_{k=1,\ldots,\ell}\rho_k < \frac{\overline{\rho^2/p}+\dfrac{\kappa^2}{\ell L^2}}{\overline{\rho/p}}\,. \tag{142}$$

This inequality may only hold true for a sufficiently large $\kappa^2/(\ell L^2)$ since one always has

$$\overline{\rho/p}\max_{k=1,\ldots,\ell}\rho_k > \overline{\rho^2/p}\,. \tag{143}$$

Finally note, that the Lagrange multiplier (132) with r, ρ, and $\overline{p}$ being defined by (133)–(135) equals the double minimum value in the OP (113), (114). Indeed, multiplying the kth equation in (130) by w_k and summing over $k = 1,\ldots,\ell$ we arrive at the desired

Assertion 5. *Let the positive DWO-weights be $w_1,\ldots,w_\ell$ and λ be the Lagrange multiplier related to a saddle point of function (128). Then the double minimum value in the OP (113)–(114) equals λ, i.e.,*

$$2L^2\left(\sum_{k=1}^{\ell}w_k\,\rho_k(x)\right)^2 + 2\kappa^2\sum_{k=1}^{\ell}p(a_k)\,w_k^2 = \lambda\,. \tag{144}$$

Remark 3. Eq.(136) shows that DWO-weights depend explicitly only on $\rho_k(x)$ and $p(a_k)$. However, they do depend on all the values of $\rho_i(x)$ and $p(a_i)$, $i = 1,\ldots,m$ via parameters (134), (137)–(141). These formulas can be useful for theoretical studies of DWO-weights and the estimate oracle risk, see the Appendix A. As for the estimate calculation, we recommend to apply the related OP numerical solution.

An example of oracle DWO-weights is presented below in the Appendix A.

11 Links To Optimal Linear pdf Estimation

Rewrite the considered above two-step-estimator (90)–(91) as a linear combination of auxiliary kernel estimators, i.e.,

$$\widehat{p}(x) = \sum_{k=1}^{m}w_k(x)\,\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-a_k}{h}\right) \tag{145}$$

$$= \frac{1}{n}\sum_{i=1}^{n}W(X_i) \tag{146}$$

with the following weighting function

$$W(u) \triangleq \sum_{k=1}^{m} w_k(x)\,\frac{1}{h}K\left(\frac{u-a_k}{h}\right). \tag{147}$$

These equations show that $\widehat{p}(x)$ is a linear pdf estimators. Below we are to demonstrate that the considered above oracle DWO-estimators generating $\widehat{p}(x)$ via optimization of a related DWO-risk can be treated as an approximate to the related optimal linear pdf estimator.

General consideration

Let us consider a class of linear pdf estimators of the following type

$$\widetilde{p}(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} W(X_i)\,. \tag{148}$$

Here kernel function $W : [0,1] \to \mathbb{R}$ may also depend on x and n. Moreover, assume

$$\int_0^1 W(u)\,du = 1\,. \tag{149}$$

Therefore, the estimate bias is

$$\begin{aligned} b_W(x) &\triangleq \mathbb{E}(\widetilde{p}(x) - p(x)) && (150)\\ &= \int_0^1 W(u)(p(u)-p(x))\,du && (151)\end{aligned}$$

with the upper bounds

$$\begin{aligned} |b_W(x)| &\le \int_0^1 |W(u)|\cdot|p(u)-p(x)|\,du && (152)\\ &\le L\int_0^1 |W(u)|\cdot|u-x|\,du\,, && (153)\end{aligned}$$

and the estimate variance

$$\begin{aligned} \sigma^2_W(x) &\triangleq \mathbb{E}(\widetilde{p}(x) - \mathbb{E}\widetilde{p}(x))^2 && (154)\\ &= \frac{1}{n}\left(\int_0^1 W^2(u)\,p(u)\,du - \left(\int_0^1 W^2(u)\,p(u)\,du\right)^2\right) && (155)\\ &\le \frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du\,. && (156)\end{aligned}$$

The estimation Mean Square Error may be bounded from (150)–(156) as follows, for instance,

$$\begin{aligned} MSE(W,x) &= b^2_W(x) + \sigma^2_W(x) && (157)\\ &\le L^2\left(\int_0^1 |W(u)|\cdot|u-x|\,du\right)^2 + \frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du && (158)\end{aligned}$$

Remark 4. A tighter upper bounds follow from (150)–(156), i.e.,

$$\begin{aligned} MSE(W,x) &\le \left(\int_0^1 |W(u)|\cdot|p(u)-p(x)|\,du\right)^2 && (159)\\ &\quad + \frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du - \left(\int_0^1 W^2(u)\,p(u)\,du\right)^2 && (160)\\ &\le \left(\int_0^1 |W(u)|\cdot|p(u)-p(x)|\,du\right)^2 && (161)\\ &\quad + \frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du\,. && (162) \end{aligned}$$

They can lead to the oracles with different properties.

Let us study the oracle defined by MSE upper bound (157)–(158).

Definition 4. $W1$-oracle is a minimizer to MSE upper bound (157)–(158)

$$\begin{aligned} U_{W1}(p,W) &\triangleq L^2\left(\int_0^1 |W(u)|\cdot|u-x|\,du\right)^2 + \frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du && (163)\\ &\to \min_{W(\cdot)} \end{aligned}$$

subject to constraint (149). In other words, it returns the $W1$-oracle weighting function $W_1^*:[0,1]\to\mathbb{R}_+$.

The $W1$-oracle is correctly defined since functional U_{W1} (163) is strongly convex (w.r.t. L_2-norm). Remind that we assume that p is Lipschitz continuous, and $p(x)>0$.

Assertion 6. *Let function $\rho:[0,1]\to\mathbb{R}$ be a.s. positive, constant $\kappa>0$, and W^* minimizes the functional*

$$U_\rho(p,W) \triangleq \left(L^2\int_0^1 |W(u)|\,\rho(u)\,du\right)^2 + \kappa^2\int_0^1 W^2(u)\,p(u)\,du \qquad (164)$$

subject to constraint (149). Then $W^*\ge 0$ a.s. (w.r.t. Lebesgue measure).

Proof is similar to that of Assertion 4. Let $W:[0,1]\to\mathbb{R}$ be such a function that meets constraint (149) and has negative values over a subset $S\in[0,1]$ of a non-zero Lebesgue measure. Evidently, it has also positive values at some points of $\overline{S}\triangleq[0,1]\setminus S$, i.e., $W(p,u)\ge 0$ for all $u\in\overline{S}$, and $W(p,u)<0$ otherwise; constraint (149) consequently implies

$$I_+ \triangleq \int_{\overline{S}} W(u)\,du = 1 - \int_S W(u)\,du > 1\,. \qquad (165)$$

Therefore, the "positive-part" weighting function

$$\widetilde{W}(u) \triangleq \frac{1}{I_+}[W(u)]_+ \qquad (166)$$

meets the constraint (149), and

$$U_\rho(p,\widetilde{W}) = \frac{1}{I_+}\left(L^2\left(\int_{\overline{S}}|W(u)|\,\rho(u)\,du\right)^2+\kappa^2\int_{\overline{S}}W^2(u)\,p(u)\,du\right) \quad (167)$$

$$< U_\rho(p,W)\,. \quad (168)$$

Admissible weight function $\widetilde{W}$ is "better" then W. The Assertion is proved. ■

Corollary 1. $W1$-oracle weighting function is a.s. non-negative.

Proof follows directly for $\kappa^2 = 1/n$ and

$$\rho(u) \triangleq |u-x|\,. \blacksquare \quad (169)$$

Corollary 2. $W1$-oracle is equivalent to the minimizer of a quadratic functional, w.r.t. $W(\cdot)\geq 0$, that is

$$U_{W+}(p,W) = L^2\left(\int_0^1 W(u)\,|u-x|\,du\right)^2+\frac{1}{n}\int_0^1 W^2(u)\,p(u)\,du \quad (170)$$

$$\to \min_{W(\cdot)\geq 0} \quad (171)$$

subject to constraint (149) as well.

Proof is evident since $|W(u)| = W(u)$ iff $W(u)\geq 0$. ■

Introduce a "reduced" functional U_{WS} by taking the integrals in (170) over a subset $S\subseteq[0,1]$, that is

$$U_{WS}(p,W) = L^2\left(\int_S W(u)\,|u-x|\,du\right)^2+\frac{1}{n}\int_S W^2(u)\,p(u)\,du\,. \quad (172)$$

Consider auxiliary problem of minimizing $U_{WS}(p,W)$ w.r.t. $W: S\to\mathbb{R}$ subject to constraint

$$\int_S W(u)\,du = 1\,. \quad (173)$$

Corollary 2 leads to the following property of the $W1$-oracle weighting function, say W^*. Introduce the support of $W^*:[0,1]\to\mathbb{R}_+$, that is

$$S^* \triangleq \operatorname{supp} W^*\,. \quad (174)$$

Then $W^*: S^*\to\mathbb{R}_+$ remains to be the minimizer of a reduced functional $U_{WS^*}(p,W)$ w.r.t. W subject to the unique constraint (173) with $S = S^*$.

Corollary 3. Let S be such a subset of $[0,1]$ that functional (172) attains its minimum w.r.t. function $W: S\to\mathbb{R}$ subject to the unique constraint (173) on a non-negative function $W^0: S\to\mathbb{R}_+$. Determine $W^0\equiv 0$ over subset $\overline{S}\triangleq[0,1]\setminus S$. Then $W^0:[0,1]\to\mathbb{R}_+$ is the $W1$-oracle weighting function.

Proof is straightforward. ■

Given subset S of $[0,1]$, the minimizer of $U_{WS}(p,W)$ w.r.t. $W: S\to\mathbb{R}$ subject to constraint (173) may be easily found by Lagrange multipliers technique. Hence, we are looking for a saddle point (W,λ) of a Lagrange functional

$$\mathcal{L}(W,\lambda) \triangleq U_{WS}(p,W)-\lambda\left(\int_S W(u)\,du-1\right), \quad (175)$$

and arrive at

$$W(u) = \frac{1}{p(u)} \left(\mu - \frac{L^2}{\kappa^2} \, r \, \rho(u) \right) \tag{176}$$

where $\rho(u) = |u - x|$,

$$\mu \triangleq \frac{\lambda}{2\kappa^2} = \frac{\left(1 + \frac{L^2}{\kappa^2} \int_S \frac{\rho^2(u)}{p(u)} \, du\right) \left(\int_S \frac{du}{p(u)}\right)^{-1}}{1 + \frac{L^2}{\kappa^2} \left[\int_S \frac{\rho^2(u)}{p(u)} \, du - \left(\int_S \frac{\rho(u)}{p(u)} \, du\right)^2 \left(\int_S \frac{du}{p(u)}\right)^{-1}\right]}, \tag{177}$$

$$r = \frac{\left(\int_S \frac{\rho(u)}{p(u)} \, du\right) \left(\int_S \frac{du}{p(u)}\right)^{-1}}{1 + \frac{L^2}{\kappa^2} \left[\int_S \frac{\rho^2(u)}{p(u)} \, du - \left(\int_S \frac{\rho(u)}{p(u)} \, du\right)^2 \left(\int_S \frac{du}{p(u)}\right)^{-1}\right]}, \tag{178}$$

and the Lagrange multiplier $\lambda = 2\kappa^2 \mu$. Moreover, the value of $\lambda/2$ gives minimum for $U_{WS}(p, W)$ in the considered variation problem.

In some particular cases, the formulas (176)–(178) may lead to the explicit analytic representation for the $W1$-oracle. Below in the Appendix B, we illustrate it for the uniform pdf as well as for a hat pdf.

12 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 regression function can only locally be approximated well by the basis F (i.e., if M is (enough) large far away from φ^* and $g > 0$), we get a finite bandwidth property, i.e., the weights corresponding to data samples far away will be zero.

Furthermore, the DWO approach has been studied for the class of approximately linear functions, as defined by (36). A lower bound on the maximum MSE for any estimator was given, and it was shown that this bound is attained by the DWO estimator if the DWO-optimal weights are all positive. This means that the DWO estimator is optimal among all estimators for these cases. As we can see from (58)–(59), there is always at least one φ^* (and hence an interval) for which this is the case, as long as the information matrix is non-degenerate. For the optimal experiment designs considered in Section 8, the corresponding DWO estimators are always minimax optimal.

The field of DWO regression function estimation 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. This has been done for special cases in [17, 10].
- Another question is what properties $\widehat{f}_N(\varphi^*)$ has as a function of φ^*. It is easy to see that $\widehat{f}_N$ 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 $\widehat{f}_N(\varphi^*) \notin \mathcal{F}$?

- 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 [8]. 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 [14, Section 3.2.2].

Similar items remain for the area of DWO-estimation of pdf. However, there is another open problem in the latter area that is MSE upper bound dependence on the unknown pdf; see (112), for instance, where matrix $B = \|\beta_{kl}\|_{m\times m}$ has its entries as defined in (108)–(109). This is a well-known difficulty in linear pdf estimation, and one may overcome it by plugging an auxiliary pdf estimate in, e.g., minimax pdf estimate. A detailed study of the properties of the resulting estimator represents an open problem of the authors' further interests.

Appendix A

Example A.1: Oracle DWO-weights for the uniform pdf and the central estimation point

Let us study the DWO-weights (136) for the case of the uniform pdf, $p(t) = \mathbf{1}\{t \in [0,1]\}$, for the sake of simplicity. Therefore, (134)–(135) imply $p(a_k) = 1$ and $\overline{p} = 1/\ell$. By Assertion 4, one ought now to minimize

$$U(w) \triangleq L^2 \left(\sum_{k=1}^{m} w_k \, \rho_k(x) \right)^2 + \kappa^2 \sum_{k=1}^{m} w_k^2 \tag{179}$$

over the simplex

$$\Theta_m \triangleq \left\{ w \in \mathbb{R}^m : \; \sum_{k=1}^{m} w_k = 1 \,, \; \forall \, w_j \geq 0 \right\} . \tag{180}$$

This implies that all the non-zero DWO-weights w_k relate to the smallest coefficients $\rho_k(x)$. As earlier, denote the number of positive DWO-weights by ℓ. In order to further simplify our consideration, we put the estimation point $x = 1/2$. Hence, by (98)

$$\overline{\rho} = \frac{1}{\ell} \sum_{k=1}^{\ell} |a_k - 1/2| + C_1 h \,. \tag{181}$$

In order to have the smallest coefficients $\rho_k(x)$, we take the ℓ points a_k symmetrically w.r.t. $1/2$. Moreover, we further assume for the sake of concreteness that m is even (a similar analysis may be given for odd m); then ℓ is even as well, by symmetric arrangement. Therefore,

$$\overline{\rho} = \frac{2}{\ell} \sum_{i=1}^{\ell/2} h(2i-1) + C_1 h = \frac{h\ell}{2} + C_1 h \,. \tag{182}$$

Furthermore, equations (139)–(141) become

$$\overline{1/p} \triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{1}{p(a_k)} = 1\,, \tag{183}$$

$$\overline{\rho/p} \triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{\rho_k(x)}{p(a_k)} = \overline{\rho} = \frac{h\ell}{2} + C_1 h\,, \tag{184}$$

and

$$\begin{aligned}\overline{\rho^2/p} &\triangleq \frac{1}{\ell}\sum_{k=1}^{\ell}\frac{\rho_k^2(x)}{p(a_k)} = \frac{2}{\ell}\sum_{i=1}^{\ell/2} h^2(2i-1+C_1)^2 \\ &= h^2\left(\frac{1}{3}(\ell+1)(\ell+2) + (C_1-1)(\ell+2) + (C_1-1)^2\right). \end{aligned}\tag{185}$$

In order to evaluate the parameter r from (137), we first write using (183)–(185)

$$\overline{1/p}\cdot\overline{\rho^2/p} - \overline{\rho/p}^{\,2} = h^2\left(\frac{\ell^2}{12} - \frac{1}{3}\right). \tag{186}$$

So, equation (137) gives

$$r = \frac{1}{\ell}\,\frac{h\left(\dfrac{\ell}{2} + C_1\right)}{1 + \dfrac{\ell L^2}{\kappa^2}h^2\left(\dfrac{\ell^2}{12} - \dfrac{1}{3}\right)}\,, \tag{187}$$

and from (132) one may write now

$$\frac{\lambda}{2} = L^2 r\ell h\left(\frac{\ell}{2} + C_1\right) + \kappa^2\frac{1}{\ell}\,. \tag{188}$$

Parameter ℓ is evaluated from the equation (136) as follows

$$\ell = \#\{w_k > 0\} = \#\left\{\frac{1}{\ell} + r\ell\frac{L^2}{\kappa^2}(\overline{\rho} - \rho_k(1/2)) > 0\right\} \tag{189}$$

Renumbering the points $a_k > 1/2$ by index $i = 1, 2, \ldots, \ell/2$, we have

$$a_{k(i)} = 1/2 + h(2i-1)\,, \tag{190}$$

$$\rho_{k(i)}(x) = h(2i-1) + C_1 h\,, \tag{191}$$

and

$$w_{k(i)} = \frac{1}{\ell} + r\ell\frac{L^2}{\kappa^2}h\left(\frac{\ell}{2} + 1 - 2i\right), \quad i = 1, 2, \ldots, \ell/2\,. \tag{192}$$

Thus, the minimal weight is the one for $i = \ell/2$, that

$$\min_i w_{k(i)} = w_{k(\ell/2)} = \frac{1}{\ell} - r\ell\frac{L^2}{\kappa^2}h\left(\frac{\ell}{2} - 1\right). \tag{193}$$

Finally, one has to minimize $\lambda/2$ (188) by even $\ell = 2, 4, \dots, m$ subject to inequality $w_{k(\ell/2)} \geq 0$ where r is defined by (187):

$$\min_{even\ \ell=2,\dots,m} \left[L^2 r \ell h \left(\frac{\ell}{2} + C_1 \right) + \kappa^2 \frac{1}{\ell} \right] \tag{194}$$

subject to

$$0 \leq \frac{1}{\ell} - r\ell \frac{L^2}{\kappa^2} h \left(\frac{\ell}{2} - 1 \right) \tag{195}$$

where by (187), (124)

$$r = \frac{1}{\ell} \frac{h\left(\frac{\ell}{2} + C_1 \right)}{1 + \frac{\ell L^2}{\kappa^2} h^2 \left(\frac{\ell^2}{12} - \frac{1}{3} \right)}, \tag{196}$$

$$h = \frac{1}{2m}, \tag{197}$$

$$\kappa^2 = \frac{1}{nh} \|K\|_2^2 . \tag{198}$$

Observe that inequality (195) with r from (196) gives

$$g(\ell) \leq 6\|K\|_2^2 L^{-2} n^{-1} h^{-3} \tag{199}$$

where the function

$$g(x) \triangleq x(x + 3C_1 - 1)(x - 2), \quad x \geq 2, \tag{200}$$

is monotone increasing; therefore, the inverse function $g^{-1} : [0, \infty) \to [2, \infty)$ exists. Hence, the admissible set for minimization (194)–(195) includes all even integer numbers ℓ meeting the inequalities

$$2 \leq \ell \leq \min\{m,\, g^{-1}(6\|K\|_2^2 L^{-2} n^{-1} h^{-3})\}. \tag{201}$$

Let us further assume that

$$6\|K\|_2^2 L^{-2} n^{-1} h^{-3} \leq g(m) \tag{202}$$

or, equivalently, by (197),

$$n \geq \frac{48\|K\|_2^2}{L^2} \frac{m^3}{g(m)}. \tag{203}$$

To ensure (203), we may restrict ourselves by $m \geq 4$ which imply

$$\frac{m^3}{g(m)} < \frac{8}{3}. \tag{204}$$

So, (203) holds for

$$n \geq \frac{2^7 \|K\|_2^2}{L^2} \quad \text{and} \quad m \geq 4 \tag{205}$$

which means that n and m are large enough. Thus, the inequality (201) may be specified as follows:

$$2 \le \ell \le \ell^* \triangleq 2\lfloor 0.5 g^{-1}(6\|K\|_2^2 L^{-2} n^{-1} h^{-3})\rfloor \,. \tag{206}$$

Finally, one may note that the function in square brackets of (194) with r from (196) decreases when even $\ell \in [2, \ell^*]$ increases, being a minimum of the strictly convex function $U(w)$ (179) over a set which widens with growing ℓ (see the beginning of subsection 10 for the details). Hence, the minimum in (194)–(195) attained at

$$\ell = \ell^* \triangleq 2\lfloor 0.5 g^{-1}(6\|K\|_2^2 L^{-2} n^{-1} h^{-3})\rfloor \,. \tag{207}$$

implies (203).

Substituting $\ell = \ell^*$ into (194), (196) gives

$$\min \lambda = L^2 \frac{h^2 \left(\ell^* + 1\right)^2}{2 + \dfrac{L^2}{3} n h^3 \ell^* ((\ell^*)^2 - 4)} + \frac{1}{n h \, \ell^*} \,. \tag{208}$$

Let us study the asymptotic in (208) assuming such a window choice that

$$n h^3 \to 0 \,. \tag{209}$$

In particular, the sample size n may be fixed while $h \to 0$; another possibility is $h = o\left(n^{-1/3}\right)$ as $n \to \infty$. Assumption (209) reduces (207) to asymptotics as follows:

$$\ell^* h \asymp \left(\frac{6\|K\|_2^2}{L^2}\right)^{1/3} n^{-1/3} \,. \tag{210}$$

Substituting (210) to (208) leads to

$$\min \frac{\lambda}{2} \asymp \left(\frac{L^2}{6\|K\|_2^2}\right)^{1/3} n^{-2/3} \,. \tag{211}$$

Remark 5. The results of the Example corroborate that the DWO oracle pdf estimate possesses the optimal rate of convergence.

Appendix B

Example B.1: $W1$-oracle weighting function for the uniform pdf

Let us illustrate the technique above for analytically finding the $W1$-oracle weighting function (176)–(178) for the case of the uniform pdf, $p(t) = \mathbf{1}\{t \in [0, 1]\}$.

Central estimation point

First, we consider estimation point $x = 1/2$, for the sake of simplicity. Consequently, $\rho(u) = |u - 0.5|$. One can easily see that it is suffice to consider subsets

S of the type $S = [0.5 - \Delta, 0.5 + \Delta],\ 0 < \Delta < 0.5$. Hence, the integrals

$$\int_S \frac{du}{p(u)} = 2\Delta\,, \tag{212}$$

$$\int_S \frac{\rho(u)}{p(u)}\,du = \Delta^2\,, \tag{213}$$

$$\int_S \frac{\rho^2(u)}{p(u)}\,du = \frac{2}{3}\,\Delta^3\,, \tag{214}$$

and the parameters μ and r from (177)–(178) become

$$\mu = \frac{1 + \dfrac{2L^2}{3\kappa^2}\,\Delta^3}{1 + \dfrac{L^2}{6\kappa^2}\,\Delta^3}\,(2\Delta)^{-1}\,, \tag{215}$$

$$r = \frac{0.5\Delta}{1 + \dfrac{L^2}{6\kappa^2}\,\Delta^3}\,, \tag{216}$$

with the minimum value for $U_{WS}(p, W)$ in the considered variation problem being equal to

$$\frac{\lambda}{2} = \frac{\dfrac{\kappa^2}{2\Delta}\left(1 + \dfrac{2L^2}{3\kappa^2}\,\Delta^3\right)}{1 + \dfrac{L^2}{6\kappa^2}\,\Delta^3}\,. \tag{217}$$

Weighting function (176) is non-negative over interval S iff

$$\Delta \le \frac{\kappa^2}{L^2}\,\frac{\mu}{r} = \frac{\kappa^2}{L^2\Delta^2}\,\left(1 + \frac{2L^2}{3\kappa^2}\,\Delta^3\right)\,. \tag{218}$$

This gives the maximal interval S related to

$$\Delta = \Delta_{\max} \triangleq \left(\frac{3\kappa^2}{L^2}\right)^{1/3} = \left(\frac{3}{L^2 n}\right)^{1/3}\,, \tag{219}$$

belonging to $(0, 0.5)$ under sufficiently large Lipschitz constant

$$L > 2\sqrt[3]{3}\,\kappa = \frac{2\sqrt[3]{3}}{\sqrt{n}}\,, \tag{220}$$

or, equivalently, under sufficiently large sample size

$$n > \left(\frac{2\sqrt[3]{3}}{L}\right)^2\,. \tag{221}$$

Assumptions (220)–(221) lead here to the triangular weighting function

$$W(u) = \begin{cases} \mu - \dfrac{L^2}{\kappa^2}\,r\,|u - 0.5|\,, & |u - 0.5| \le \Delta_{\max}\,, \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{222}$$

with $\kappa^2 = 1/n$,

$$\mu = \Delta_{\max}^{-1} = \left(\frac{L^2 n}{3}\right)^{1/3}, \tag{223}$$

$$r = \frac{1}{3}\Delta_{\max}, \tag{224}$$

and

$$\frac{\lambda}{2} = \left(\frac{L^2}{3}\right)^{1/3} n^{-2/3}. \tag{225}$$

An example of the triangular weighting function (222) for $L = 2$ and $n = 500$ is depicted on Figure 2 by a solid line; here $\Delta_{\max} \approx 0.1145$.

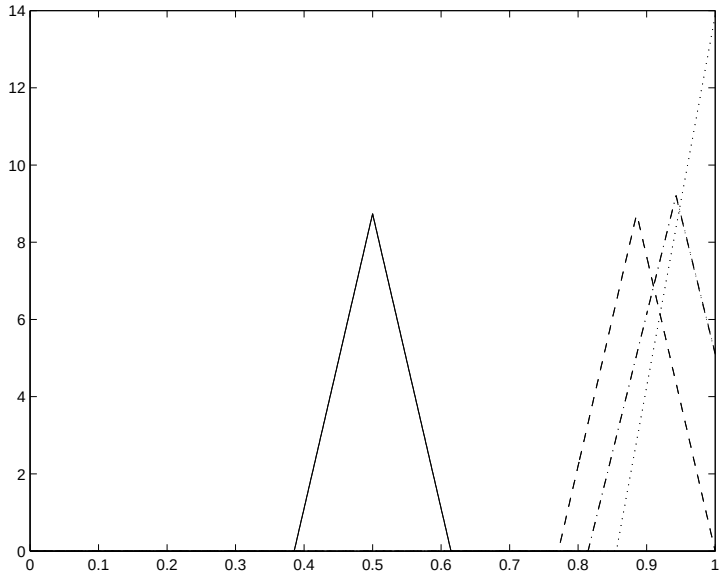


Figure 2: Weighting functions (176) for the uniform pdf $p(t) = \mathbf{1}\{t \in [0,1]\}$, $n = 500$, and for different estimation points: $x = 0.5$ (solid line), $x = 0.8855$ (dashed line), $x = 0.943$ (dash-dot line), and $x = 1$ (dotted line).

Remark 6. Notice, that the results (210) and (225), being applied to the rectangular kernel where $\|K\|_2^2 = 0.5$, asymptotically coincide with that of (211) and (219), respectively. This corroborates that the DWO-oracle approximates its continuous counterpart that is $W1$-oracle, at least asymptotically, as $nh^3 \to 0$.

Other estimation points

Evidently, the shape of triangular weighting function remains for the case of the uniform pdf when the estimation point x moves from the center up to the distance $0.5 - \Delta_{\max}$. For instance, the dashed line on the Figure 2 relates to $L = 2$, $n = 500$, and to the maximal shift of the estimation point to right when the triangular weighting function reaches the boundary of the interval $[0,1]$ without changing the shape; here $x = 1 - \Delta_{\max} \approx 0.8855$.

However, what happens with $W(\cdot)$ when x becomes closer to the boundary? How does the influence of the latter change the weighting function shape? Let $x > 1/2$ being "close" to 1, for the sake of concreteness; the case $0 < x < 1/2$ can be obtained by symmetry. Applying technique of section 11 one now need to look for the "boundary" subsets $S = [x - \Delta, 1] \in [0, 1]$. Simple calculations lead to

$$W(u) = \begin{cases} \mu - \dfrac{L^2}{\kappa^2}\, r\, |u - x|\,, & u \geq x - \Delta\,, \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{226}$$

with $\kappa^2 = 1/n$, and the parameters μ and r from (177)–(178) where the integrals

$$\int_S \frac{du}{p(u)} \;=\; 1 - x + \Delta\,, \tag{227}$$

$$\int_S \frac{\rho(u)}{p(u)}\, du \;=\; \frac{1}{2}\,\left(\Delta^2 + (1-x)^2\right)\,, \tag{228}$$

$$\int_S \frac{\rho^2(u)}{p(u)}\, du \;=\; \frac{1}{3}\,\left(\Delta^3 + (1-x)^3\right)\,. \tag{229}$$

Additional condition $W(x - \Delta) = 0$ let to determine Δ, i.e.,

$$\mu = \frac{L^2}{\kappa^2}\, r\Delta\,. \tag{230}$$

Thus, we arrive at the following cubic equation w.r.t. Δ:

$$\frac{1}{6}\,\Delta^3 + \frac{1}{2}\,\Delta\,(1-x)^2 - \frac{\kappa^2}{L^2} - \frac{1}{3}\,(1-x)^3 = 0\,. \tag{231}$$

For instance, if we take $\Delta_{\max} \approx 0.1145$ from two cases above and put $x = 1 - 0.5\Delta_{\max} \approx 0.943$, we get weighting function (226) depicted on Figure 2 by a dash-dot line.

Finally, we considered $x = 1$. In this case equation (231) gives a solution $\Delta = (nL^2/6)^{-1/3}$; hence, $\Delta \approx 0.1442$. The weighting function for this case is depicted on Figure 2 by a dotted line. The minimal DWO-risk (170) is as follows:

$$\frac{\lambda}{2} = \frac{1}{3}\,\left(\frac{6L}{n}\right)^{2/3}\,. \tag{232}$$

Example B.2: $W1$-oracle weighting function for a hat pdf and the central estimation point

Let us continue to illustrate the technique above for analytically finding the $W1$-oracle weighting function (176)–(178), now for the case of hat pdf, i.e.,

$$p(t) = 2 - p_0 - 4(1 - p_0)|t - 0.5|\,, \quad t \in [0, 1]\,. \tag{233}$$

We put parameter

$$p_0 = 1 - 0.25L \tag{234}$$

to ensure L to be the true Lipschitz constant of p. Figure 3 represents an example for hat pdf with $p_0 = 0.5$, or $L = 2$.

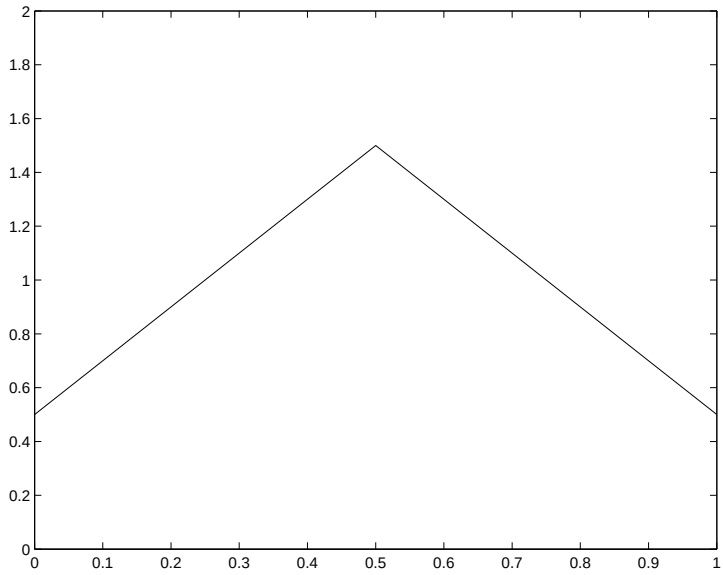


Figure 3: Hat pdf with $p_0 = 0.5$, or $L = 2$.

We consider the central estimation point, $x = 1/2$, for the sake of simplicity; $\rho(u) = |u - 0.5|$. These assumptions reduce our consideration to the subsets $S = [0.5 - \Delta, 0.5 + \Delta],\ 0 < \Delta < 0.5$. Hence, the integrals (212)–(214) become

$$\int_S \frac{du}{p(u)} = b \log \frac{a}{a-\Delta}\,, \tag{235}$$

$$\int_S \frac{\rho(u)}{p(u)}\,du = b\left(a \log \frac{a}{a-\Delta} - \Delta\right), \tag{236}$$

$$\int_S \frac{\rho^2(u)}{p(u)}\,du = b\left(a^2 \log \frac{a}{a-\Delta} - a\Delta - \frac{\Delta^2}{2}\right), \tag{237}$$

where

$$a \triangleq \frac{(2-p_0)}{4(1-p_0)}\,, \qquad b \triangleq \frac{1}{2(1-p_0)}\,. \tag{238}$$

Parameters μ and r can now be calculated from (177)–(178) subject to the additional condition (230), which is reduced to the equation

$$1 + \frac{L^2}{\kappa^2}\left(\int_S \frac{\rho^2(u)}{p(u)}\,du - \Delta \int_S \frac{\rho(u)}{p(u)}\,du\right) = 0\,. \tag{239}$$

Substituting (236)–(237) into (239) leads to the following equation for determining Δ

$$\frac{\Delta^2}{2} - a\Delta + \frac{\kappa^2}{bL^2} + a(a-\Delta)\log\frac{a}{a-\Delta} = 0\,. \tag{240}$$

It is easily verified that the LHS (240) represents a monotone decreasing function of $\Delta \in [0, a)$ which decreases from positive $\frac{\kappa^2}{bL^2}$ down to $-\infty$. Hence, equation

(240) possesses a unique solution $\Delta \in [0, a)$ defining μ, r, and

$$W(u) = \begin{cases} \dfrac{0.5\, b}{a - |u - 0.5|} \left(\mu - \dfrac{L^2}{\kappa^2}\, r\, |u - 0.5| \right), & |u - 0.5| \le \Delta\,, \\ 0\,, & \text{otherwise}\,. \end{cases} \tag{241}$$

Remind that $\kappa^2 = 1/n$. Finally, the minimum value for $U_{WS}(p, W)$ in the considered variation problem equals

$$\frac{\lambda}{2} = \kappa^2 \mu\,. \tag{242}$$

Weighting function (241) for hat pdf (233) with $p_0 = 0.5$, $L = 2$, $n = 500$ is depicted on Figure 4. One may see that it looks very similar to, but does not coincide with, a triangular weighting function.

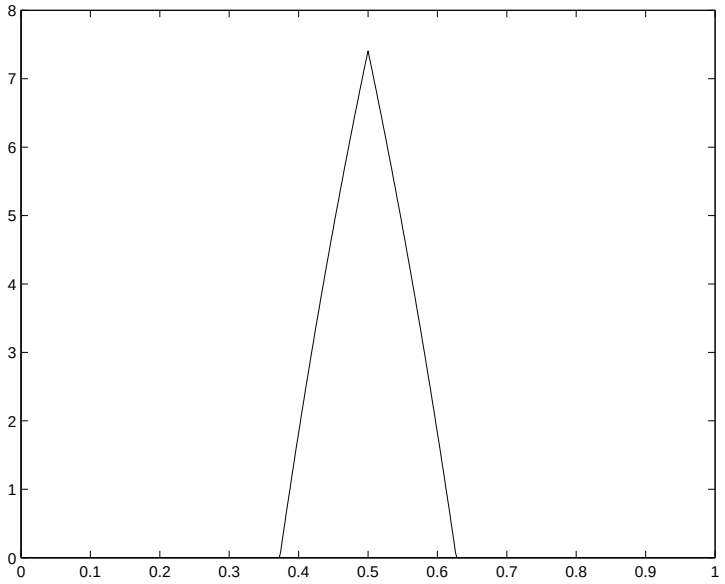


Figure 4: Weighting function (241) for the hat pdf (233) with $p_0 = 0.5$, $L = 2$, $n = 500$, and $x = 0.5$.

References

[1] E.-W. Bai and Y. Liu. Recursive Direct Weight Optimization in Nonlinear System Identification: A Minimal Probability Approach. *IEEE Trans. Automatic Control*, 52(7):1218–1231, July 2007.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] S. Chen and S. A. Billings. Neural networks for nonlinear dynamic system modeling and identification. *Int. J. Control*, 56(2):319–346, August 1992.

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

[5] J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, 1996.

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

[7] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, and Q. Zhang. Nonlinear black-box modeling in system identification: Mathematical foundations. *Automatica*, 31(12):1724–1750, 1995.

[8] A. Juditsky, A. Nazin, J. Roll, and L. Ljung. Adaptive DWO estimator of a regression function. In *NOLCOS'04*, Stuttgart, September 2004.

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

[10] A. Nazin, J. Roll, and L. Ljung. 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, December 2003.

[11] A. Nazin, J. Roll, and L. Ljung. Direct weight optimization for approximately linear functions: Optimality and design. In *14th IFAC Symposium on System Identification*, Newcastle, Australia, Mar 2006.

[12] A. Nazin, J. Roll, and L. Ljung. Direct weight optimization in nonlinear function estimation and system identification. In *Proceedings of the VI International Conference "System Identification and Control Problems" SICPRO'07, Moscow, 29 Jan – 1 Feb 2007.*

[13] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[14] J. Roll. *Local and Piecewise Affine Approaches to System Identification*. PhD thesis, Dept. of EE, Linköping Univ., Sweden, April 2003.

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

[16] J. Roll, A. Nazin, and Ljung L. A general direct weight optimization framework for nonlinear system identification. In *16th IFAC World Congress on Automatic Control*, pages Mo–M01–TO/1, Prague, Sep 2005.

[17] J. Roll, A. Nazin, and L. Ljung. A non-asymptotic approach to local modelling. In *The 41st IEEE Conference on Decision and Control*, pages 638–643, December 2002.

[18] J. Roll, A. Nazin, and L. Ljung. Nonlinear system identification via direct weight optimization. *Automatica*, 41(3):475–490, March 2005.

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

[20] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Y. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, 31(12):1691–1724, 1995.

[21] A. Stenman. *Model on Demand: Algorithms, Analysis and Applications*. PhD thesis, Dept. of EE, Linköping Univ., Sweden, 1999.

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

[23] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer-Verlag, London, 1997.

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2007-11-14

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English

☐ _____

Rapporttyp
Report category

☐ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☒ Övrig rapport
☐ _____

ISBN
—

ISRN
—

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

LiTH-ISY-R-2831

URL för elektronisk version

`http://www.control.isy.liu.se`

Titel
Title

Direct Weight Optimization in Statistical Estimation and System Identification

Författare
Author

Alexander V. Nazin, Jacob Roll, Lennart Ljung, Ion Grama

Sammanfattning
Abstract

The Direct Weight Optimization (DWO) approach to statistical estimation and the application to nonlinear system identification has been proposed and developed during the last few years. Computationally, the approach is typically reduced to a convex (e.g., quadratic or conic) program, which can be solved efficiently. The optimality or sub-optimality of the obtained estimates, in a minimax sense w.r.t. the estimation error criterion, can be analyzed under weak a priori conditions. The main ideas of the approach are discussed here and an overview of the obtained results is presented.

Nyckelord
Keywords

Statistical estimation, Nonparametric identification, Minimax techniques, Convex programming, Nonlinear systems, Estimation error