

# Unlinked Monotone Regression

**Fadoua Balabdaoui**

*Seminar für Statistik  
ETH, Zurich  
Rämistrasse 101  
8092, Zurich, Switzerland*

FADOUA.BALABDAOUI@STAT.MATH.ETHZ.CH

**Charles R. Doss**

*School of Statistics  
University of Minnesota  
Minneapolis, MN 55455, USA*

CDOSS@UMN.EDU

**Cécile Durot**

*Modal'x  
Université Paris Nanterre  
F92000, Nanterre, France*

CECILE.DUROT@GMAIL.COM

**Editor:** Gabor Lugosi

## Abstract

We consider so-called univariate unlinked (sometimes “decoupled,” or “shuffled”) regression when the unknown regression curve is monotone. In standard monotone regression, one observes a pair  $(X, Y)$  where a response  $Y$  is linked to a covariate  $X$  through the model  $Y = m_0(X) + \epsilon$ , with  $m_0$  the (unknown) monotone regression function and  $\epsilon$  the unobserved error (assumed to be independent of  $X$ ). In the unlinked regression setting one gets only to observe a vector of realizations from both the response  $Y$  and from the covariate  $X$  where now  $Y \stackrel{d}{=} m_0(X) + \epsilon$ . There is no (observed) pairing of  $X$  and  $Y$ . Despite this, it is actually still possible to derive a consistent non-parametric estimator of  $m_0$  under the assumption of monotonicity of  $m_0$  and knowledge of the distribution of the noise  $\epsilon$ . In this paper, we establish an upper bound on the rate of convergence of such an estimator under minimal assumption on the distribution of the covariate  $X$ . We discuss extensions to the case in which the distribution of the noise is unknown. We develop a second order algorithm for its computation, and we demonstrate its use on synthetic data. Finally, we apply our method (in a fully data driven way, without knowledge of the error distribution) on longitudinal data from the US Consumer Expenditure Survey.

**Keywords:** deconvolution, quantile, monotone regression, rates, shuffled, uncoupled, unlinked

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>The Minimum Contrast Estimator: Existence</b>	<b>6</b>
2.1	Setup, Terminology, and Notation . . . . .	6
2.2	Existence . . . . .	8
<b>3</b>	<b>Convergence Rates of the Minimum Contrast Estimator</b>	<b>9</b>
3.1	Convergence Rate Under Ordinary Smooth Noise . . . . .	9

3.2	Convergence Rate Under Supersmooth Noise . . . . .	14
3.3	Convergence Rate Under a Discrete Noise Distribution . . . . .	14
3.4	Convergence Rate in the Case of Different Sample Sizes . . . . .	15
3.5	Uniform Consistency . . . . .	15
<b>4</b>	<b>Estimation of Moments of <math>m_0(X)</math></b>	<b>16</b>
<b>5</b>	<b>Fenchel Optimality Conditions</b>	<b>17</b>
<b>6</b>	<b>Computation</b>	<b>18</b>
<b>7</b>	<b>Extension to the Case of Unknown Noise Distribution</b>	<b>20</b>
7.1	Estimation of the Noise Distribution in the Semi-supervised Setting . .	20
7.2	Estimation of the Noise Distribution with Longitudinal Responses . . .	20
<b>8</b>	<b>Demonstrations on Synthetic and Real Data</b>	<b>21</b>
8.1	Computations on Synthetic Data . . . . .	21
8.2	Computations on CEX Data . . . . .	22
<b>9</b>	<b>Conclusions and Directions for Future Research</b>	<b>25</b>
<b>A</b>	<b>Bounding the Integrals in (56) and (57).</b>	<b>26</b>
<b>B</b>	<b>Basic Empirical Process Theory Definitions and a Fundamental Preservation Result</b>	<b>27</b>
<b>C</b>	<b>Wasserstein Distance Lemmas</b>	<b>28</b>
<b>D</b>	<b>Gradient, Curvature, and Other Algorithmic Computations</b>	<b>28</b>
D.1	Proof of (24) . . . . .	29
D.2	Curvature Derivation . . . . .	29
D.3	Computations for Laplace Distribution . . . . .	30
D.4	Computations for Gaussian Errors . . . . .	30
D.5	Mixtures of Gaussian . . . . .	31
D.6	On Implementation of Algorithm 1 . . . . .	31
<b>E</b>	<b>Proofs</b>	<b>31</b>
E.1	A Preparatory Lemma . . . . .	31
E.2	Proofs for Section 2 . . . . .	33
E.3	Proofs for Section 3 . . . . .	38
E.4	Proofs for Section 4 . . . . .	49
<b>F</b>	<b>Proof of Proposition 7</b>	<b>53</b>

## 1. Introduction

An important part of data science is the construction of a data set; nowadays, because there are so many different entities collecting increasing amounts of data, data sets are often constructed by combining separate sub-data sets or datastreams. Also, data sets sometimes undergo some form of anonymization: this can be due to the increasing prevalence of privacy concerns, or in some cases due to concerns about having limited

data-transmission bandwidth when many separate sensors are streaming data to a central server (Pananjady et al., 2016). Thus, it is increasingly common for data scientists/analysts to want to relate variables in one data set to variables in another data set when the two data sets are *unlinked*. In this paper, we consider the problem of unlinked regression, specifically when the regression function is assumed to satisfy a monotonicity constraint.

In the standard regression setting, we have

$$Y_i = m_0(X_i) + \epsilon_i, \quad E(\epsilon_i) = 0, \quad i = 1, \dots, n, \quad (1)$$

for a random noise variable  $\epsilon_i$  that is independent of  $X_i$ . The most basic assumption of this model is that for each index  $i = 1, \dots, n$ , the *pair*  $(X_i, Y_i)$  is observed. For now, we assume that the covariates  $X_i$ ,  $i = 1, \dots, n$  are univariate random variables. A more general model than the above standard regression model is the *shuffled* regression, in which we do not get to see the pairs  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ ; rather, we only observe  $(X_1, \dots, X_n)$  and  $(Y_1, \dots, Y_n)$ , without knowing which  $X_i$  is paired with which  $Y_i$ . Thus, we have the same model as (38) except that the first equality is now only an equality in distribution and there is an unknown permutation  $\pi$  on  $\{1, \dots, n\}$  such that  $Y_i = m_0(X_{\pi(i)}) + \epsilon_i$  for all  $i$ . This happens for instance in case of anonymized data. An even more general model is the *unlinked* regression model that we consider in this paper, where again, the first equality only holds in distribution but where in addition, the  $X_i$ 's could be observed on different individuals from the  $Y_i$ 's so that the two samples are not necessarily connected through a permutation  $\pi$ . This happens for instance if the two samples have been collected independently (by separate entities). The number of observed  $X_i$ 's may even differ from the number of observed  $Y_i$ 's so we observe independent and identically distributed (i.i.d.) variables  $Y_1, \dots, Y_{n_y}$  and i.i.d. variables  $X_1, \dots, X_{n_x}$  such that

$$Y \stackrel{d}{=} m_0(X) + \epsilon \quad (2)$$

where  $Y \sim Y_i$ ,  $X \sim X_i$ , and  $\epsilon$  is independent of  $X$  with distribution function  $\Phi_\epsilon$ . In shuffled or unlinked regression models, it may seem hopeless to even try to learn the regression function  $m_0$  since  $m_0$  is in general not even identifiable (see Section 2.1 below); however, it turns out that if  $m_0$  is assumed to be monotonically increasing, as we will do in this paper, and if the error distribution is known a priori, then in fact one can construct consistent estimators of  $m_0$ . (We will return in Section 7 to discuss the case where the error distribution is unknown.)

Unlinked estimation can be considered in a variety of settings (e.g., DeGroot and Goel (1980)). It appears that unlinked monotone regression was recently introduced by Carpentier and Schlüter (2016). One motivating example discussed in Carpentier and Schlüter (2016) is about expenditure on goods or services, such as housing: the price an individual is willing to pay for housing is expected to be monotonically increasing (at least on average, if not individually) as a function of the individual's salary. However, estimating the monotonic relationship is hindered by the simple fact that the data on wages and housing transactions are often gathered by different agents. There are many other motivating examples for unlinked or shuffled regression, besides the ones already discussed. (In some examples, there may be information allowing partial matching; see our discussion point 4 in Section 9.) In flow cytometry, cells suspended in a fluid flow past a laser, and the response (scattering of light) reveals information

about the cell, which may be explained by its features (e.g., gene expression). However, the order of the cells as they pass the laser is unknown, so we are in a shuffled regression setting (Abid et al., 2017). In “image stitching” (related to the so-called pose-correspondence problem) one wants to find the unknown correspondence between point clouds constructed from multiple camera angles of the same image (Pananjady et al., 2018). Several other motivating examples are discussed by Pananjady et al. (2018) (in the context of permuted/shuffled linear regression).

The method of Carpentier and Schlüter (2016) is based on the fact that monotone unlinked regression can be rephrased as the so-called deconvolution problem, as is apparent from (2). Below, we will discuss links to this problem in more detail. Note that, as is also true in the deconvolution setting,  $m_0$  is not identifiable if we do not know the distribution of the noise  $\epsilon$  (or have at least an estimate thereof). The following simple example explains why. Suppose that  $m_0(x) = 2x$  and  $X \sim \mathcal{N}(0, 1)$  is independent of  $\epsilon \sim \mathcal{N}(0, 1)$ . Then, this model is the same as  $m_0(x) = x$  and  $X \sim \mathcal{N}(0, 1)$  independent of  $\epsilon \sim \mathcal{N}(0, 4)$ . Let  $F_0$  be the cumulative distribution function (CDF) of  $X$  and  $L_0$  the CDF of  $m_0(X)$ . If both  $m_0$  and  $F_0$  are assumed to be one-to-one, then  $L_0$  satisfies

$$L_0(w) = P(m_0(X) \leq w) = F_0 \circ m_0^{-1}(w) \quad (3)$$

for all  $w$  in the range of  $m_0$ , and we have

$$m_0 = L_0^{-1} \circ F_0. \quad (4)$$

Thus, the estimator constructed by Carpentier and Schlüter (2016) takes the form

$$\tilde{m}(x) = \tilde{L}_{n_y}^{-1} \circ \mathbb{F}_{n_x}(x) \quad (5)$$

where  $\tilde{L}_{n_y}$  is an estimator of  $L_0$  obtained by deconvolution methods and is based on the sample  $(Y_1, \dots, Y_{n_y})$  and knowledge of the distribution of  $\epsilon$ , and where  $\mathbb{F}_{n_x}$  can be taken for example to be the empirical distribution function of  $F_0$  based on the sample  $X_1, \dots, X_{n_x}$ . Thus, the estimator in (5) at a point  $x$  is equal to the deconvolution estimator of the quantile of  $m_0(X)$  corresponding to the random level  $\mathbb{F}_{n_x}(x)$ . In the presence of “contextual variables” (i.e., covariates that are paired with both  $X$  and  $Y$ ), Carpentier and Schlüter (2016) gave some consistency and rate of convergence results in their Theorem 3.2 under the assumption that  $m_0$  and  $F_0^{-1}$  belong to Hölder classes. Carpentier and Schlüter (2016) do not discuss how to choose the optimal bandwidth, as the main focus in that paper is to show how their estimation approach can be easily implemented for real data sets.

Beside Carpentier and Schlüter (2016), the only other work of which we are aware on a similar problem is the very recent article of Rigollet and Weed (2019). In their setting, the authors consider a shuffled monotone regression model in a fixed design setting but use the term “uncoupled” to describe the problem. In fact, the authors do not make any attempt to recover the unknown permutation, and focus entirely, as we do in this paper, on estimating the unknown regression function. Rigollet and Weed (2019) assume that the known distribution of the noise is sub-exponential and the true monotone function is bounded by some known constant, but they do not make any smoothness assumption on that function. Using the Wasserstein’s distance and arguments from optimal transport, they showed that their estimator converges to the truth at a rate no slower than  $\log \log n / \log n$  and that this rate is minimax when the distribution of the noise is Gaussian. Although Rigollet and Weed (2019) describe in

their Section 2.2 an algorithm for computing their monotone estimator, the authors do not present simulation results.

It is worth mentioning that more research seems to have been done in the shuffled (permuted) linear regression model than in the monotone regression model. We refer here to the work of Abid et al. (2017), Pananjady et al. (2017), Pananjady et al. (2018), and Unnikrishnan et al. (2018). The main focus in the former three papers is to find conditions on the signal-to-noise ratio that guarantee recovery of the unknown permutation. Abid et al. (2017) show that the least-squares estimator in this model is inconsistent in general, but they construct a method-of-moments type estimator and prove that this estimator is consistent assuming that  $E(X_i) \neq 0$ . A common feature of these works is to restrict attention to the case of Gaussian noise. As we show in the present paper, the noise distribution may be of fundamental importance in these problems.

As announced above, we give now some more detail to the existing link between unlinked monotone regression and estimation in deconvolution problems. As noted earlier, (4) shows clearly that there is a tight connection to quantile estimation in a deconvolution setting. In fact, consider the deconvolution setting in which one observes  $n$  i.i.d. copies of  $Y$  where  $Y = X + \epsilon$  for independent random variables  $X$  and  $\epsilon$ . In this problem, the goal is learn the distribution of  $X$  under the assumption that the distribution of  $\epsilon$  is known (such an assumption can be relaxed if this distribution can be estimated). Estimation of the distribution and quantile functions of  $X$  has been considered in Hall and Lahiri (2008), and revisited in Dattner et al. (2011, 2016) under slightly different assumptions. In particular, contrary to Hall and Lahiri (2008), no moment assumptions are made about the covariate  $X$  or  $\epsilon$  in Dattner et al. (2011, 2016). There, the smoothness of the density of  $X$  is measured in terms of belonging to Sobolev or Hölder balls. In the case where the error is ordinary smooth of order larger than  $1/2$ , Dattner et al. (2011) recover the rates of convergence that Hall and Lahiri (2008) obtained for the integrated risk when estimating the distribution function, and moreover provide new rates of convergence for the case of smoother error distributions; the square-root rate is shown to be achieved for smooth enough distribution of  $X$ . The convergence results obtained in these previous papers do not apply directly in this present paper, as we do not assume that the covariate  $X$  (from (2)) admits even a Lebesgue density, which also means that  $m_0(X)$  is not assumed to have a density.

While Carpentier and Schlüter (2016) and Rigollet and Weed (2019) are the only other articles of which we are aware on unlinked monotone regression besides the present paper, the classical isotonic regression model given in (1) is a very well-known estimation problem with a vast literature. The most known estimator in this problem is certainly the Grenander-type estimator, obtained by taking the right derivative of the greatest convex minorant of the cumulative sum diagram associated with the data  $(X_i, Y_i), i = 1, \dots, n$  (Barlow et al., 1972; Robertson et al., 1988; Groeneboom and Jongbloed, 2014). The pointwise non-standard rate of convergence of the Grenander estimator is  $n^{-1/3}$  if  $m_0$  is continuously differentiable with a non-vanishing derivative and  $\sqrt{n}$  in case  $m_0$  is locally flat (Groeneboom, 1983; Carolan and Dykstra, 1999; Zhang, 2002; Cator, 2011; Chatterjee et al., 2015). Asymptotic properties, including the pointwise limit distribution and convergence in the  $L_p$ -norms for  $p \in [1, 5/2) \cup \{\infty\}$  have been fully described in Brunk (1970), Durot (2002), Durot (2007) and Durot et al. (2012); see also Groeneboom (1985, 1989). One can also combine kernel estimation

with the monotonicity constraint to improve rates of convergence if  $m_0$  has higher orders of smoothness (Mammen, 1991; Durot and Lopuhaä, 2014).

In this paper, we investigate the unlinked monotone regression following the method introduced by Edelman (1988) for estimating the mixing distribution in a mixture problem with Normal noise. Let  $(X_1, \theta_1), \dots, (X_n, \theta_n)$  be independent random vectors such that  $X_i | \theta_i \sim \mathcal{N}(\theta_i, 1)$ , that is, conditionally on  $\theta_1, \dots, \theta_n$  the random variables  $X_1, X_2, \dots, X_n$  are generated from the unknown distributions  $\Phi(\cdot - \theta_1), \Phi(\cdot - \theta_2), \dots, \Phi(\cdot - \theta_n)$ , where  $\Phi$  denotes the cumulative distribution function of  $\mathcal{N}(0, 1)$ . The approach of Edelman (1988) consists of finding the vector  $(\tilde{\theta}_1, \dots, \tilde{\theta}_n)$  which minimizes the integrated difference between  $n^{-1} \sum_{i=1}^n \Phi(\cdot - \theta_i)$  and the empirical distribution  $\mathbb{F}_n$  based on the observations  $X_i, i = 1, \dots, n$  among all vectors  $(\theta_1, \dots, \theta_n)$ . As Edelman (1988) already noted, the normal distribution can be replaced by other distributions, which is exactly what we do here. The merits of this approach are the facts that it does not depend on some bandwidth, and that it is easily implementable. The link between our problem and the work Edelman (1988) is made clear in Section 2. There, we introduce our estimator, which we call the minimum contrast estimator, and we establish its existence and some of its important features in the case of equal sample sizes for the responses and covariates. In Section 3, we establish rates of convergence under some fairly general conditions on the distribution function of the covariates. In that section, we assume that the noise distribution is known and distinguish between three cases for its smoothness: (1) ordinary smooth, (2) supersmooth and (3) discrete with finite number of support points. The convergence rates in cases (1) and (2) are derived using some classical Fourier transform techniques, while in case (3) a very different approach is employed, which makes use of a recent result in Meis and Mammen (2020). In the proofs, we use a conversion device which allows us to link the convergence rate of the estimator to that of its generalized inverse, an interesting result in its own right. Since our method allows for different sample sizes for the responses and covariates, we extend the construction of the estimator to this case and derive the corresponding convergence rate in the three aforementioned smoothness cases. In Section 4, we show that the estimator achieves the parametric rate in estimating the moments of  $m_0(X)$  in case (3) and also when the noise is known to be uniformly distributed on a compact. Although our estimator cannot be shown to be unique, we prove in Section 5 that any solution has to satisfy a necessary optimizing condition. This condition is used to develop a gradient-descent algorithm to compute the estimator, see Section 6. In Section 7, we discuss how our method can be extended to the more realistic situation where the noise distribution is unknown. In Section 8, we illustrate our approach through synthetic and real data. We finish this manuscript with some concluding remarks and future research directions; see Section 9. Technical proofs are deferred to an Appendix.

## 2. The Minimum Contrast Estimator: Existence

### 2.1 Setup, Terminology, and Notation

Let  $m_0$  be the monotone function appearing in the model in (2), and in which we are interested. In the model that we consider, the response  $Y$  has the same distribution as the convolution of  $m_0(X)$  and the noise  $\epsilon$  with  $m_0$  monotone non-decreasing. Note for the case where  $m_0$  is non-increasing it is enough to consider  $-Y$  instead of  $Y$  and all our results will still apply. Denote by  $\mathcal{M}$  the set of all bounded non-decreasing and

right continuous functions defined on  $[0, 1]$ . This class accomodates for the assumption made in the sequel that the covariate  $X \in [0, 1]$  almost surely.

It is worth mentioning that without the monotonicity assumption, the function  $m_0$  is not identifiable in general, even in the simple case where  $m_0(X)$  is observed without noise, that is in the case where  $Y = m_0(X)$  is observed. Precisely, the distribution of  $X$  together with the distribution of  $m_0(X)$  does not pin down the function  $m_0$ . A simple counterexample (that was suggested by a referee) is where  $X$  has the uniform distribution on  $[0, 1]$ :

$$m_1(x) = x \text{ and } m_2(x) = 1 - 2|x - \frac{1}{2}|.$$

Then, it is easy to see that both  $m_1(X)$  and  $m_2(X)$  have the uniform distribution on  $[0, 1]$ . On the other hand, if  $m_0$  is monotone, then it can be determined on the support of  $X$  using knowledge about the distribution of  $X$  and  $m_0(X)$ . This identifiability property is proved in the following proposition.

**Proposition 1.** *Let  $X$  be a real valued random variable with a continuous distribution on  $[0, 1]$ , and let  $m_1$  and  $m_2$  be non-decreasing and right-continuous functions on the support of  $X$ . If  $m_1(X)$  has the same distribution as  $m_2(X)$ , then  $m_1 = m_2$  on the support of  $X$ .*

Here, we describe the working framework of our estimation approach. We observe two independent samples  $(X_1, \dots, X_n)$  and  $(Y_1, \dots, Y_n)$  such that the following holds.

**Assumption A0.**  $X_1, \dots, X_n \stackrel{iid}{\sim} F_0$ ,  $0 \leq X_i \leq 1$  almost surely, and  $Y_1, \dots, Y_n \stackrel{iid}{\sim} H_0$ . The unobserved error  $\epsilon$  is independent of  $X$ , satisfies  $E(\epsilon) = 0$ , and has CDF  $\Phi_\epsilon$ . Moreover,  $m_0 \in \mathcal{M}$ .

Note that there are no restrictions on the relationships between the  $X_i$ 's and the  $Y_j$ 's, other than the equality in distribution in (2). Then,  $F_0$ ,  $H_0$ , and  $\Phi_\epsilon$  are the true CDF's of  $X_i$ ,  $Y_i$ , and  $\epsilon_i$ , respectively. Let  $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{[X_i, \infty)}(x)$  and  $\mathbb{H}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{[Y_i, \infty)}(x)$  where  $x \mapsto \mathbb{1}_A(x)$  is the indicator function for the set  $A$ . Also, let  $\|\cdot\|_\infty$  be the supremum norm, i.e.  $\|m\|_\infty = \sup_{t \in [0, 1]} |m(t)|$ . For  $K > 0$ , let  $\mathcal{M}_K$  be the set of functions  $m \in \mathcal{M}$  such that  $\|m\|_\infty \leq K$ . For  $m \in \mathcal{M}$ ,  $m^{-1}$  denotes the *generalized inverse* of  $m$ , i.e.

$$m^{-1}(y) := \inf \{x \in [0, 1] : m(x) \geq y\} \quad (6)$$

where the infimum of an empty set is defined to be 1. Hence, we have  $m^{-1}(y) = 1$  for all  $y > m(1)$ .

In this section, we assume that  $\Phi_\epsilon$  is known. This assumption will be relaxed in Section 7. Also, although we take the respective sizes of the samples of covariates and responses to be equal, our method can be easily adapted to the case where these sizes are different. In that case, the convergence rate of our estimator is driven by the minimum of the sample sizes; see Subsection 3.4.

Let  $\mathcal{F}$  be the set of all distribution functions on  $\mathbb{R}$ . A contrast function  $\mathcal{C}$  defined on the Cartesian product  $\mathcal{F} \times \mathcal{F}$  is any non-negative function such that  $\mathcal{C}(F_1, F_2) = 0$  if and only if  $F_1 = F_2$ . Consider the estimator

$$\hat{m}_n = \operatorname{argmin}_{m \in \mathcal{M}} \mathcal{C} \left( \mathbb{H}_n, n^{-1} \sum_{i=1}^n \Phi_\epsilon(\cdot - m(X_i)) \right) \quad (7)$$

provided that a minimizer exists (in which case it is not necessarily unique). Since the criterion depends on  $m$  only through its values at the observations  $X_i, i = 1, \dots, n$ , it follows that any candidate for the minimization problem in (7),  $m$  in  $\mathcal{M}$ , can be well replaced by the non-decreasing function  $\tilde{m}$  such that

$$\tilde{m}(t) = \begin{cases} m(X_{(1)}), & \text{for } t \in [0, X_{(1)}] \\ m(X_{(n)}), & \text{for } t \in [X_{(n)}, 1] \end{cases}$$

and  $\tilde{m}$  is constant between the remaining order statistics such that  $\tilde{m}$  is right continuous (by definition of  $\mathcal{M}$ ) and coincides with  $m$  at every  $X_{(i)}, i = 1, \dots, n$ . Here as is customary,  $X_{(1)} \leq \dots \leq X_{(n)}$  denote the order statistics corresponding to  $X_1, \dots, X_n$ .

In addition, note that  $\hat{m}_n$  does not have to be unique at the data points and thus  $\hat{m}_n$  denotes any solution of the minimization problem.

## 2.2 Existence

In the sequel, we consider the following contrast function

$$\mathcal{C}(F_1, F_2) = \int_{\mathbb{R}} \left\{ F_1(y) - F_2(y) \right\}^2 dy$$

whenever this integral is finite, which is the case when  $F_1$  and  $F_2$  are distribution functions of random variables admitting finite expectations. The choice of such contrast function is mainly motivated by application of the Parseval-Plancherel's Theorem. The estimator we consider here is reminiscent of the one studied in Edelman (1988) for deconvoluting a distribution function from a Gaussian noise. However, our goal here is different since the main target in our problem is the monotone transformation  $m_0$ .

Before starting the analysis of the estimator, we establish first its existence. Denote

$$\mathbb{M}_n(m) = \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - m(X_i)) \right\}^2 dy$$

and let

$$\mathbb{M}(m) = \int_{\mathbb{R}} \left\{ H_0(y) - \int_{\mathbb{R}} \Phi_{\epsilon}(y - m(x)) dF_0(x) \right\}^2 dy$$

be its deterministic counterpart. Then the minimizer  $\hat{m}_n$  in (7) (if it exists) can also be written as

$$\hat{m}_n \in \underset{m \in \mathcal{M}}{\operatorname{argmin}} \mathbb{M}_n(m). \quad (8)$$

The following assumptions will be needed.

**Assumption A1.** For some  $K_0 \in [0, \infty)$ , we have  $\|m_0\|_{\infty} = K_0$ .

**Assumption A2.** The distribution function  $\Phi_{\epsilon}$  is continuous on  $\mathbb{R}$ .

**Proposition 2.** Let Assumption A0 hold. Then,

1.  $\mathbb{M}_n(m)$  is finite for any  $m \in \mathcal{M}$  for all  $n \geq 1$ ;
2. If Assumption A1 also holds then  $\mathbb{M}(m)$  is finite for any  $m \in \mathcal{M}$ ;

3. If Assumptions A1 and A2 also hold, then there exists at least a solution to (8) that is piecewise constant and right-continuous, for all  $n \geq 1$ .
4. If Assumptions A1 and A2 also hold, and  $\epsilon$  has a bounded support, then with probability 1, there exists at least a solution to (8) that is piecewise constant, right-continuous and bounded in the sup-norm by a deterministic constant for all  $n \geq 1$ . This deterministic constant can be taken to be equal to  $K_0 + 2$  in case  $\|\epsilon\|_\infty \leq 1$  with probability 1.

Note that parts 3 and 4 of the proposition give existence but not necessarily uniqueness of  $\widehat{m}_n$ . In fact, it can be seen from the proof of Proposition 2 that if  $\widehat{m}_n$  is a solution to (8), then any monotone function that coincides with  $\widehat{m}_n$  at the observed covariates  $X_1, \dots, X_n$  gives another solution to (8). In what follows, we will consider a solution  $\widehat{m}_n$  that takes constant values between successive covariates and that is right continuous. This choice is consistent with the way the Grenander-type estimator is defined, that is, the estimator in the classical monotone regression estimation problem.

### 3. Convergence Rates of the Minimum Contrast Estimator

In this section, we will give upper bounds on the convergence rate of the minimum contrast estimator defined above. Not surprisingly, this rate depends on the smoothness of the noise distribution. More specifically, we will consider the following cases for the smoothness: (1) ordinary smooth, (2) supersmooth and (3) discrete with a finite support. In the whole section, we assume that the Assumptions A0, A1 and A2 hold. By Proposition 2, this guarantees that the minimization problem in (8) admits a piecewise constant and right-continuous solution. We will denote  $\widehat{m}_n$  any such solution. Also, we will require the following assumption about the distribution of the design points.

**Assumption A3.** *The common distribution function  $F_0$  of the covariates  $X_1, \dots, X_n$  is continuous.*

For the smoothness cases (1) and (2), we will need the following notation:

$$\psi_F(x) = \int_{\mathbb{R}} e^{itx} dF(t) \quad (9)$$

for any distribution function  $F$  on  $\mathbb{R}$ , and

$$\phi_g(x) = \int_{\mathbb{R}} e^{itx} g(t) dt \quad (10)$$

denotes the Fourier transform of  $g$ , whenever  $g$  is integrable. Note that when  $F$  is absolutely continuous with density  $f$ , then it follows immediately that  $\psi_F = \phi_f$ .

#### 3.1 Convergence Rate Under Ordinary Smooth Noise

We start with the ordinary smooth case for the noise distribution described in the next assumption.

**Assumption A4.** *The distribution function  $\Phi_\epsilon$  is absolutely continuous with a 0-mean ordinary smooth density  $f_\epsilon$  in the sense that*

$$\frac{d_0}{|t|^\beta} \leq |\phi_{f_\epsilon}(t)| \leq \frac{d_1}{|t|^\beta} \quad (11)$$

as  $|t| \rightarrow \infty$ , for some  $\beta > 0$  and constants  $d_0 > 0, d_1 > 0$ .

Some comments are in order. Assumption A4 is common in deconvolution problems, see for instance Dattner et al. (2011). The positive real number  $\beta$  is usually referred to as the order of smoothness. Known examples include the Exponential distribution with any scale parameter for which  $\beta = 1$ , Gamma distribution with shape parameter  $\alpha > 0$  and scale  $\gamma > 0$  for which  $\beta = \alpha$ , the Laplace distribution with  $\beta = 2$ , and more generally symmetric Gamma distributions (the distribution of  $X - X'$  where  $X$  and  $X'$  are i.i.d.  $\sim \text{Gamma}(\alpha, \gamma)$ ) in which case we have  $\beta = 2\alpha$ . See for example the examples given in Fan (1991) after (1.4). We plot in Figure 3 in the appendix several gamma densities with a variety of shape parameters, to show the “decreasing smoothness:” as the shape parameter goes to 0, the density has an increasing spike at 0.

Our main theorem below provides the rate of convergence of the  $L_1$ -error with respect to the distribution of the design points,  $F_0$ , over a given interval  $[a, b] \subset [0, 1]$  provided that the estimator is stochastically bounded on that interval. Sufficient conditions for this boundedness are given below.

The following assumption is crucial for deriving the convergence rate of  $\widehat{m}_n$ . It will be also needed below in the supersmooth case.

**Assumption A5.** *Assume that there exists  $T^*$  such that  $|\phi_{f_\epsilon}(t)| \geq |\phi_{f_\epsilon}(T)| > 0$  for all  $T > T^*$  and  $|t| \leq T$ .*

Now, we we are able to state the main result of this subsection.

**Theorem 1.** *1. Suppose that Assumptions A0 to A5 hold. Let  $[a, b] \subset [0, 1]$  be a fixed interval such that*

$$\widehat{m}_n(a) = O_P(1) \text{ and } \widehat{m}_n(b) = O_P(1). \quad (12)$$

*Then, it holds that*

$$\int_a^b |\widehat{m}_n(x) - m_0(x)| dF_0(x) = O_P(n^{-1/(2(2\beta+1))}). \quad (13)$$

*2. If the Assumptions A0 to A5 hold, then the claims in (12) hold for all  $a$  and  $b$  such that  $0 < F_0(a) \leq F_0(b) < 1$ .*

We give below the main steps of the proof and conclude this subsection with some comments about the rates in Theorem 1. Details of the proof are postponed to Section E.3. Let us define

$$\widehat{\mathbb{L}}_n(w) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) \leq w\}} \text{ and } \mathbb{L}_n(w) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{m_0(X_i) \leq w\}} \quad (14)$$

for all  $w \in \mathbb{R}$ . Using deconvolution arguments we show closeness of the two processes in the following proposition.

**Proposition 3.** *Assume that Assumptions A4 and A5 hold. Then,*

$$E \left[ \int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \right] = O(n^{-1/(2\beta+1)}).$$

Next, using entropy arguments from empirical process theory, we show in the following proposition that on intervals  $[A, B]$  that are possibly random, the empirical processes  $\widehat{\mathbb{L}}_n$  and  $\mathbb{L}_n$  are close to their population counterparts  $\widehat{L}_n^0$  and  $L_0$  respectively, where for all  $w \in \mathbb{R}$ ,

$$\widehat{L}_n^0(w) = \int \mathbb{1}_{\{\widehat{m}_n(x) \leq w\}} dF_0(x) \text{ and } L_0(w) = \int \mathbb{1}_{\{m_0(x) \leq w\}} dF_0(x). \quad (15)$$

More precisely, we derive the convergence rate of the associated  $L_2$ -error integrated on such interval  $[A, B]$ . Note that the first two claims in the proposition hold under the only assumptions that  $X_1, \dots, X_n$  are i.i.d. and  $\widehat{m}_n$  is taken from (8). In fact, it can be seen from the proof that these two claims continue to hold with  $\widehat{m}_n$  replaced by any monotone estimator.

**Proposition 4.** *If  $X_1, \dots, X_n$  are i.i.d., then for all random variables  $A < B$  (that may depend on  $n$ ) it holds that*

$$\int_A^B \left( \widehat{\mathbb{L}}_n(w) - \widehat{L}_n^0(w) \right)^2 dw \leq (B - A) O_P(1/n),$$

$$\int_A^B \left( \mathbb{L}_n(w) - L_0(w) \right)^2 dw \leq (B - A) O_P(1/n),$$

where  $O_P(1/n)$  is uniform in  $A$  and  $B$ . Moreover, if  $B - A = O_P(n^{2\beta/(2\beta+1)})$  and Assumptions A4 and A5 holds, then

$$\int_A^B \left( \widehat{L}_n^0(w) - L_0(w) \right)^2 dw = O_P(n^{-1/(2\beta+1)}). \quad (16)$$

The following proposition makes the connection between the above error and a squared distance between the inverse functions of  $\widehat{m}_n$  and  $m_0$  composed with the distribution function  $F_0$  of the  $X_i$ 's; a rate of convergence of that squared distance is derived. We recall that for  $m \in \mathcal{M}$ , the inverse of  $m$  is defined by (6), where the infimum of an empty set is defined to be 1.

**Proposition 5.** *Under Assumptions A3, A4, and A5 for all random variables  $A < B$  such that  $B - A = O_P(1)$  it holds that*

$$\begin{aligned} \int_A^B (F_0 \circ \widehat{m}_n^{-1}(w) - F_0 \circ m_0^{-1}(w))^2 dw &= \int_A^B \left( \widehat{L}_n^0(w) - L_0(w) \right)^2 dw \\ &= O_P(n^{-1/(2\beta+1)}). \end{aligned}$$

The last step in the proof makes the connection between the above squared distance and the  $L_1$ -error of  $\widehat{m}_n$ .

**Proposition 6.** *Suppose that Assumption A3 holds, and let  $[a, b] \subset [0, 1]$  be a fixed interval. Then it holds that*

$$\int_a^b |\widehat{m}_n(x) - m_0(x)| dF_0(x) \leq \left( (B_n - A_n) \int_{A_n}^{B_n} (F_0 \circ \widehat{m}_n^{-1}(x) - F_0 \circ m_0^{-1}(x))^2 dx \right)^{1/2}$$

where  $A_n = m_0(a) \wedge \widehat{m}_n(a)$  and  $B_n = m_0(b) \vee \widehat{m}_n(b)$ .

We conclude the subsection with some comments about the convergence rate obtained in Theorem 1.

- When  $\beta = 1/4$ , then the  $L_1$ -rate of convergence obtained for our estimator matches the well-known  $n^{1/3}$ -rate of the Grenander estimator in the classical isotonic regression (where the link between the responses and covariates is known). If  $\beta < 1/4$ , then the rate is strictly better than the cubic rate, which may appear as a contradiction. However, there is one major difference between the regular isotonic regression model and the one we consider here: we assume the error distribution to be known, which is not the case in regular isotonic regression. Thus, it seems that a transitional regime occurs in the rate of convergence in case the noise distribution is known. In the unlinked regression setting which we study in this paper, if the error distribution is known to be a centered (or symmetric) Gamma with shape and scale parameters  $\alpha \in (0, 1/4)$  (or  $\alpha \in (0, 1/8)$ ) and  $\lambda > 0$  then the smoothness parameter is  $\beta = \alpha$  (or  $\beta = 2\alpha$ ) belongs to  $(0, 1/4)$  and the rate of convergence of our estimator will be faster than  $n^{1/3}$ .
- Note that if we let  $\beta \rightarrow 0$ , then the centered or symmetric Gamma will converge to a Dirac at 0 and the rate of convergence will approach the parametric rate  $\sqrt{n}$ . In fact, it can be shown that the rate of convergence is precisely the parametric rate if the error distribution is a Dirac at 0, i.e. if  $\epsilon = 0$  with probability one. Indeed, when  $\epsilon = 0$  with probability 1, we have  $\Phi_\epsilon(y - \hat{m}_n(X_i)) = \mathbb{1}_{\hat{m}_n(X_i) \leq y}$ . If  $\hat{m}_n$  is taken to be a minimum contrast estimator that is bounded in the sup-norm by  $K_0 + 2$  (such a solution is known to exist with probability 1 by Proposition 2), then we have by definition of  $\hat{m}_n$

$$\int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \mathbb{1}_{\hat{m}_n(X_i) \leq y} \right\}^2 dy \leq \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \mathbb{1}_{m_0(X_i) \leq y} \right\}^2 dy$$

implying that

$$\begin{aligned} & \int_{\mathbb{R}} \left\{ n^{-1} \sum_{i=1}^n \mathbb{1}_{\hat{m}_n(X_i) \leq y} - n^{-1} \sum_{i=1}^n \mathbb{1}_{m_0(X_i) \leq y} \right\}^2 dy \\ & \leq 4 \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \mathbb{1}_{m_0(X_i) \leq y} \right\}^2 dy \\ & \leq 8 \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - H_0(y) \right\}^2 dy + 8 \int_{\mathbb{R}} \left\{ H_0(y) - n^{-1} \sum_{i=1}^n \mathbb{1}_{m_0(X_i) \leq y} \right\}^2 dy. \end{aligned}$$

The first term was already shown to be  $O_P(n^{-1})$ ; see the proof of Proposition 3. Moreover, the true distribution  $H_0$  of the  $Y_i$ 's is that of  $m_0(X)$ , which implies that the second term is also  $O_P(n^{-1})$  since its expectation is equal to  $n^{-1} \int \text{Var}(\mathbb{1}_{m_0(X) \leq y}) dy = n^{-1} \int H_0(y)(1 - H_0(y)) dy$  where the integral  $\int H_0(y)(1 - H_0(y)) dy$  is known to be finite; see Appendix A. Hence,

$$\int_{\mathbb{R}} \left\{ n^{-1} \sum_{i=1}^n \mathbb{1}_{\hat{m}_n(X_i) \leq y} - n^{-1} \sum_{i=1}^n \mathbb{1}_{m_0(X_i) \leq y} \right\}^2 dy = O_P(n^{-1}). \quad (17)$$

It follows from the rate obtained in (17) and the empirical process arguments as in the proof of Proposition 4 that

$$\begin{aligned} \int_{-K_0-2}^{K_0+2} \left( \widehat{L}_n^0(w) - L_0(w) \right)^2 dw &= \int_{-K_0-2}^{K_0+2} \left( F_0 \circ \widehat{m}_n^{-1}(w) - F_0 \circ m_0^{-1}(w) \right)^2 dw \\ &= O_P(n^{-1}). \end{aligned}$$

This in turn implies by Proposition 6 (with  $a = 0$ ,  $b = 1$ , and  $[A_n, B_n] \subseteq [-K_0 - 2, K_0 + 2]$ ) that

$$\int_0^1 |\widehat{m}_n(x) - m_0(x)| dF_0(x) = O_P(n^{-1/2}).$$

- The parametric rate obtained for a noise that is Dirac at 0 can be generalized to the case where  $\epsilon$  is supported on a finite set of points. This is proved in Theorem 3.
- It may be considered as unsatisfactory that the conclusions were stated with  $O_P$  notation, so we provide now more precise bounds. The bounds can be obtained by closely inspecting the proof of Theorem 1; details are omitted. It can be seen from the proofs that, with  $A_n$  and  $B_n$  the stochastically bounded random variables taken from Proposition 6, we have

$$\begin{aligned} &\int_a^b |\widehat{m}_n(x) - m_0(x)| dF_0(x) \\ &\leq \left( (B_n - A_n) \left[ 3 \int_{A_n}^{B_n} \left( \widehat{L}_n(w) - L_n(w) \right)^2 dw + \frac{6(B_n - A_n)}{n} \|\mathbb{G}_n\|_{\mathcal{I}}^2 \right] \right)^{1/2} \\ &\leq \left( 3(B_n - A_n) \int_{\mathbb{R}} \left( \widehat{L}_n(w) - L_n(w) \right)^2 dw \right)^{1/2} \\ &\quad + (B_n - A_n) \left( \frac{6\|\mathbb{G}_n\|_{\mathcal{I}}^2}{n} \right)^{1/2} \end{aligned} \tag{18}$$

where  $\|\mathbb{G}_n\|_{\mathcal{I}}^2$  is a random variable with finite expectation. The expectation is bounded above by an unknown absolute constant that is connected to the entropy measure of the set of monotone functions on  $\mathbb{R} \rightarrow [0, 1]$ , as well as to the absolute constants that emerge from the empirical process theory. Hence, the second summand in the right-hand term is of the parametric order  $n^{-1/2}$ , and can be seen as a systematic error that is typically negligible as compared to the first summand. This means that the rate of convergence of the estimator is driven by the integral in the first summand. For this integral, we have

$$E \left[ \int_{\mathbb{R}} \left( \widehat{L}_n(w) - L_n(w) \right)^2 dw \right] \leq K n^{-1/(2\beta+1)},$$

where  $K$  depends on the parameters of the model. One can choose for instance

$$K := (2\beta + 1) \left( \frac{24}{d_0} (2/(\beta\pi))^{2\beta} (2K_0 + E|\epsilon|) \right)^{1/(2\beta+1)}.$$

It is worth mentioning that the decomposition in (18) still holds with  $\widehat{m}_n$  replaced (in the definitions of  $\widehat{L}_n$ ,  $A_n$  and  $B_n$ ) by any estimator in  $\mathcal{M}$ .

### 3.2 Convergence Rate Under Supersmooth Noise

The main arguments used above for an ordinary smooth noise continue to apply to the supersmooth case. In this setting, Assumption A4 should be replaced by the following one.

**Assumption A4'.** *The distribution function  $\Phi_\epsilon$  is absolutely continuous with a 0-mean supersmooth density  $f_\epsilon$  in the sense that*

$$d_0|t|^\alpha \exp(-|t|^\beta/\gamma) \leq |\phi_{f_\epsilon}(t)| \leq d_1|t|^\alpha \exp(-|t|^\beta/\gamma)$$

as  $|t| \rightarrow \infty$ , for some  $\alpha > 0, \beta > 0$  and constants  $d_0 > 0, d_1 > 0$ .

In the above definition we provide for supersmoothness we deviate from the one given by Fan (1991) in (1.3) by taking the same exponent  $\alpha$  in the lower and upper bound, for simplicity.

**Theorem 2.** *Under the Assumptions A0–A3, A4', and A5, we have for any  $[a, b] \subset (0, 1)$  such that  $0 < F_0(a) \leq F_0(b) < 1$  that*

$$\int_a^b |\widehat{m}_n(x) - m_0(x)| dF_0(x) = O_P((\log n)^{-1/(2\beta)}).$$

For instance, in the case of a Gaussian noise, in which case  $\beta = 2$ , the rate is  $1/(\log n)^{1/4}$ , which matches the conclusion of Edelman (1988) right after the proof of his Theorem 1. On the other hand, this rate is slower than the minimax rate,  $\log \log n / \log n$ , obtained by Rigollet and Weed (2019) in the shuffled regression problem. Note however that the setting studied by Rigollet and Weed (2019) is the shuffled regression problem, whereas we consider the unlinked regression problem (recall we explain the difference in our introduction). It is not known whether the two problems share the same minimax rates; or, rather, it is not known if the minimax lower bound derived in Rigollet and Weed (2019) applies to the unlinked problem, since the unlinked problem is statistically more difficult than the shuffled problem. Therefore, it is currently unknown if the rate we derived is in fact minimax suboptimal or not.

### 3.3 Convergence Rate Under a Discrete Noise Distribution

In this section, we consider the case where  $\epsilon$  is supported on a finite set of points. We prove that in that case, the minimum contrast estimator achieves the parametric rate in the  $L_1$ -loss.

**Theorem 3.** *Suppose that Assumptions A0 to A2 hold, and that  $\epsilon$  is supported on a finite set of points. If  $\widehat{m}_n$  is a solution to (8) that is bounded in the sup-norm by a deterministic constant (which exists with probability 1 in view of Proposition 2), then*

$$\int_0^1 |\widehat{m}_n(x) - m_0(x)| dF_0(x) = O_P(n^{-1/2}).$$

### 3.4 Convergence Rate in the Case of Different Sample Sizes

In this section we briefly consider the case where one observes  $Y_1, \dots, Y_{n_y}$  and  $X_1, \dots, X_{n_x}$  with possibly different sample sizes  $n_x \neq n_y$ . In that case, the estimator is defined as

$$\hat{m}_{n_x, n_y} = \operatorname{argmin}_{m \in \mathcal{M}} \int_{\mathbb{R}} \left\{ \mathbb{H}_{n_y}(y) - n_x^{-1} \sum_{i=1}^{n_x} \Phi_{\epsilon}(y - m(X_i)) \right\}^2 dy$$

where  $\mathbb{H}_{n_y}$  denotes the empirical distribution function corresponding to the sample  $Y_1, \dots, Y_{n_y}$ . The asymptotic here has to be understood in the sense that both sample sizes  $n_x$  and  $n_y$  go to infinity. This means that  $n_x \wedge n_y \rightarrow \infty$ , where  $n_x \wedge n_y$  denotes the infimum between  $n_x$  and  $n_y$ . In the following theorem, we give the upper bound on the convergence rate of our minimum contrast estimator in the three regimes (1) ordinary smooth, (2) supersmooth and (3) discrete with finite number of support points.

**Theorem 4.** *1. Suppose that the Assumptions A0, A1, A2, A3, and A5 hold true. Let  $[a, b] \subset (0, 1)$  be any fixed interval such that  $0 < F_0(a) \leq F_0(b) < 1$ .*

- *If Assumption A4 holds true, then as  $n_x, n_y \rightarrow \infty$ , we have*

$$\hat{m}_{n_x, n_y}(a) = O_P(1) \text{ and } \hat{m}_{n_x, n_y}(b) = O_P(1) \quad (19)$$

and

$$\int_a^b |\hat{m}_{n_x, n_y}(x) - m_0(x)| dF_0(x) = O_P((n_x \wedge n_y)^{-1/(2(2\beta+1))}).$$

- *If Assumption A4' holds true, then as  $n_x, n_y \rightarrow \infty$ , (19) holds, and*

$$\int_a^b |\hat{m}_{n_x, n_y}(x) - m_0(x)| dF_0(x) = O_P(\log(n_x \wedge n_y)^{-1/(2\beta)}).$$

2. *Suppose that the Assumptions A0, A1, A2 hold true. If  $\epsilon$  is supported on a finite number of points, then with probability 1 there exists a solution  $\hat{m}_{n_x, n_y}$  which is bounded by a deterministic constant. Furthermore,*

$$\int_0^1 |\hat{m}_{n_x, n_y}(x) - m_0(x)| dF_0(x) = O_P((n_x \wedge n_y)^{-1/2})$$

as  $n_x, n_y \rightarrow \infty$ .

### 3.5 Uniform Consistency

The convergence rates for the minimum contrast estimator we derived above are obtained for the  $L_1$ -norm. Thus, a natural question is whether these global rates also hold pointwise. Although the exact answer to this question is still unknown we can provide at least an intermediate result which shows that the estimator is pointwise consistent, even uniformly provided that the true monotone regression function,  $m_0$ , is continuous. For simplicity, we assume that  $n_x = n_y = n$ .

**Theorem 5.** *Suppose that the assumptions of Theorem 1, or 2 or 3 hold, and let  $\mathcal{S}_0$  denote the support of  $F_0$ . Let  $a, b \in [0, 1]$  be any points such that  $\hat{m}_n(a)$  and  $\hat{m}_n(b)$  are each  $O_P(1)$ . Let  $C$  be any compact set in the interior of  $[a, b] \cap \mathcal{S}_0$ . Assume  $m_0$  is continuous on  $(0, 1)$ . Then it holds that*

$$\sup_{x \in C} |\hat{m}_n(x) - m_0(x)| = o_P(1).$$

Note that in the case where  $\epsilon$  is compactly supported, the result above implies that if  $m_0$  is continuous on  $[0, 1]$

$$\sup_{x \in [0, 1] \cap \mathcal{S}_0} |\widehat{m}_n(x) - m_0(x)| = o_P(1)$$

for any estimator  $\widehat{m}_n$  that is bounded in the sup-norm by a deterministic constant. Recall such an estimator exists with probability 1 (see Claim 4 of Proposition 2). Finally, note that we cannot hope to extend the convergence outside  $\mathcal{S}_0$  since  $m_0$  is only identifiable on this set; see Proposition 1.

#### 4. Estimation of Moments of $m_0(X)$

In this section, we showcase that our minimum contrast estimator can achieve the  $\sqrt{n}$ -rate for estimating certain smooth functionals of  $m_0$ . In doing so, we restrict attention to estimating moments of  $m_0(X)$  and the cases where either  $\epsilon$  is discrete with a finite number of points in the support or when it is uniformly distributed over a compact. Modulo some scaling, we can assume without loss of generality that the support is a subset of  $[-1, 1]$  in both cases. For simplicity, we assume that the sample sizes of the covariates and responses are equal.

Note that boundedness of  $m_0$  implies that  $m_0(X)$  admits finite moments of any order. Furthermore, in case the noise distribution is compactly supported, as assumed here in this section, all moments of the response  $Y$  are finite and we have

$$E(Y^k) = \sum_{j=0}^k \binom{k}{j} E[m_0^j(X)] E(\epsilon^{k-j}).$$

Since the distribution of  $\epsilon$  is known, the moments of the error can be exactly computed. Then, replacing the moments of  $Y$  by the corresponding empirical estimators in the above formula yields a  $\sqrt{n}$ -consistent estimator of  $E[m_0^k(X)]$ . The minimum contrast estimator offers an alternative (and also a direct) way for estimating the moments since one can simply take the natural choice  $\int \widehat{m}_n^k d\mathbb{F}_n$ , that is

$$\frac{1}{n} \sum_{i=1}^n \widehat{m}_n^k(X_i).$$

Our result below shows that the latter estimator is converging at the parametric rate.

**Theorem 6.** *Assume that Assumptions A0 to A2 hold, and that either  $\epsilon$  is supported on a finite set of points or uniformly distributed on some compact. Let  $\widehat{m}_n$  denote a solution to (8) which is piecewise constant, right-continuous and bounded in the sup-norm by a deterministic constant, for all  $n \geq 1$  (which exists with probability 1), and  $\mathbb{F}_n$  the empirical CDF based on  $X_1, \dots, X_n$ . The following holds true.*

- *If  $\epsilon$  is supported on a finite set of points, then for all integers  $k \geq 1$*

$$\left| \int_0^1 \widehat{m}_n^k(x) d\mathbb{F}_n(x) - \int_0^1 m_0^k(x) dF_0(x) \right| = O_P(n^{-1/2}), \quad (20)$$

- *If  $\epsilon$  is uniformly distributed on a compact, then*

$$\left| \int_0^1 \widehat{m}_n(x) d\mathbb{F}_n(x) - \int_0^1 m_0(x) dF_0(x) \right| = O_P(n^{-1/2}).$$

Note that when  $\epsilon$  is a uniformly distributed noise (the second case in the above theorem), the convergence rate of  $\widehat{m}_n$  cannot be obtained from the results obtained in Section 3. Indeed, considering without loss of generality the case where the support is  $[-1, 1]$ , a uniform distribution does not belong to the ordinary smooth nor to supersmooth categories since in this case

$$|\phi_\epsilon(t)| = \left| \frac{\sin(t)}{t} \right|, \quad \text{for } t \neq 0$$

and hence  $|\phi_\epsilon(t)|$  cannot be bounded below by  $d_0/|t|^\beta$  for some  $d_0, \beta > 0$  nor by  $d_0|t|^\alpha \exp(-|t|^\beta/\gamma)$  for some  $d_0, \alpha, \beta, \gamma > 0$ . Thus, the convergence rate of  $\widehat{m}_n$  for such a noise distribution is still an open problem. Nevertheless, Theorem 6 shows clearly that our estimator behaves reasonably well when the goal is estimation of the first moment of  $m_0(X)$ .

## 5. Fenchel Optimality Conditions

In view of the computational section below, we derive in this section the optimality conditions related to the optimization problem defining the estimator  $\widehat{m}_n$ . Recall that by Assumption A0, the covariates  $X_1, \dots, X_n$  are assumed to belong to  $[0, 1]$ . In the following, we denote by  $\widehat{m}_n$  a piecewise constant and right-continuous solution to (8), see Proposition 2. For some  $1 \leq p \leq n$ , we write  $\widehat{m}_{n,(1)} < \dots < \widehat{m}_{n,(p)}$  for the distinct values taken by  $\widehat{m}_n$  on  $[0, 1]$ . We will use the following assumption on the density  $f_\epsilon$ .

**Assumption A6.** *The density  $f_\epsilon$  is continuously differentiable such that*

$$\sup_{t \in \mathbb{R}} |f'_\epsilon(t)| \leq D,$$

for some constant  $D > 0$ , and  $\int_{\mathbb{R}} |f'_\epsilon(t)| dt < \infty$ .

**Proposition 7.** *Let  $\widehat{m}_{n,(1)} < \dots < \widehat{m}_{n,(p)}$  be the distinct values of the estimator  $\widehat{m}_n$ . Let Assumption A6 hold. Then, for any  $k \in \{1, \dots, p\}$*

$$\int_{\mathbb{R}} \left( \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) f_\epsilon(y - \widehat{m}_{n,(k)}) dy = 0. \quad (21)$$

Furthermore, this condition can be equivalently re-written as

$$1 - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(Y_i - \widehat{m}_{n,(k)}) = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \Phi_\epsilon(y - \widehat{m}_n(X_i)) f_\epsilon(y - \widehat{m}_k) dy \quad (22)$$

for  $k = 1, \dots, p$ .

**Remark 1.** *The alternative form in (22) gives a useful way of verifying numerically the second equality condition via numerical integration. Indeed, with  $m = \widehat{m}_n(X_i)$  and  $m' = \widehat{m}_k$ , the integral on the right side of (22) takes the form*

$$\int_{\mathbb{R}} \Phi_\epsilon(y - m) f_\epsilon(y - m') dy = B(m' - m),$$

where for all  $m \in \mathbb{R}$ ,  $B(m) = E\Phi_\epsilon(\epsilon + m) = \int \Phi_\epsilon(y) f_\epsilon(y - m) dy$ . Explicit formulas of  $B(m)$ , for  $m \in \mathbb{R}$ , can be even found for some distributions such as Laplace or Gaussian; see Subsections D.3 and D.4.

**Remark 2.** Consider the case where  $\hat{m}_n$  takes one unique value, denoted  $\hat{m}_{1,(1)}$ . Then the right side of (22) equals  $\int_{\mathbb{R}} \Phi_{\epsilon}(y - \hat{m}_{1,(1)}) f_{\epsilon}(y - \hat{m}_{1,(1)}) dy = \int_{\mathbb{R}} \Phi_{\epsilon}(y) f_{\epsilon}(y) dy$  which equals  $E(\Phi_{\epsilon}(\epsilon)) = E(U) = 1/2$  where  $U$  is Uniform(0, 1).

## 6. Computation

Recall that our goal is to minimize  $\mathbb{M}_n$ . Write  $\mathbf{m} := (m_1, \dots, m_n)$ , so that the objective function can be written as

$$\mathbb{M}_n(\mathbf{m}) := \int_{\mathbb{R}} \left\{ \mathbb{H}_n - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(\cdot - m_i) \right\}^2 \quad (23)$$

(by a slight abuse of notation). To minimize  $\mathbb{M}_n$ , we can compute an unconstrained minimizer  $\tilde{\mathbf{m}}$  of  $\mathbb{M}_n$  (i.e., we do not force  $\tilde{m}_i \leq \tilde{m}_{i+1}$  for  $i = 1, \dots, n-1$ ), and then the overall solution would be given by reordering the entries of  $\tilde{\mathbf{m}}$  so that it is nondecreasing; i.e.,  $\hat{m}_n(X_{(i)}) := \tilde{m}_{(i)}$  (where  $\tilde{m}_{(1)} \leq \dots \leq \tilde{m}_{(n)}$ ). The gradients  $\frac{\partial}{\partial m_i} \mathbb{M}_n(\mathbf{m})$  can be computed using that

$$\frac{\partial}{\partial m_i} \mathbb{M}_n(\mathbf{m}) = 2n^{-1} - 2n^{-2} \sum_{\alpha=1}^n \{ \Phi_{\epsilon}(Y_{\alpha} - m_i) + B(m_i - m_{\alpha}) \} \quad (24)$$

where for all  $m \in \mathbb{R}$ ,  $B(m) = E\Phi_{\epsilon}(\epsilon + m)$ , see Appendix D.1. In Appendix D we show how to derive the gradient of  $\mathbb{M}_n$  when  $\Phi_{\epsilon}$  is either a Gaussian or a Laplace distribution. Thus, we can consider using a (first order) gradient descent algorithm for computation.

However, because the objective function is symmetric in all its  $m_i$  components, at any stationary point at which some of the  $m_i$ 's are equal, no gradient-based method can tell if we could improve the objective function by allowing the equal  $m_i$ 's to take separate values. Thus, we will consider a second order method to solve this problem: we will compute a second derivative in a direction related to separating  $m_i$  into two distinct values.

Let  $\mathbb{M}_{n,p}$  denote the objective function parameterized such that it takes  $2p$  arguments and the regression function  $m$  is represented by  $p$  values  $\mathbf{m} := (m_1, \dots, m_p)$  and  $p$  weights  $\boldsymbol{\pi} := (\pi_1, \dots, \pi_p)$  ( $\sum_{\alpha=1}^p \pi_{\alpha} = 1$ ), so that  $\mathbb{M}_{n,p}(\boldsymbol{\pi}, \mathbf{m}) = \int_{\mathbb{R}} (\mathbb{H}_n(y) - \sum_{\alpha=1}^p \pi_{\alpha} \Phi_{\epsilon}(y - m_{\alpha}))^2 dy$ .

The following algorithm is an active set type of algorithm. At a point  $(\tilde{\boldsymbol{\pi}}_p, \tilde{\mathbf{m}}_p) \in \mathbb{R}^{2p}$ , define

$$\tilde{\boldsymbol{\pi}}_{p+1} := (\tilde{\pi}_1, \dots, \lambda \tilde{\pi}_i, (1 - \lambda) \tilde{\pi}_i, \tilde{\pi}_{i+1}, \dots, \tilde{\pi}_p), \text{ and } \tilde{\mathbf{m}}_{p+1} := (\tilde{m}_1, \dots, \tilde{m}_i, \tilde{m}_i, \dots, \tilde{m}_p),$$

and consider  $\mathbb{M}_{n,p+1}(\tilde{\boldsymbol{\pi}}_{p+1}, \tilde{\mathbf{m}}_{p+1})$ , where  $\lambda \in [0, 1]$ . Let  $\theta := (m_i - m_{i+1})/2$  and, for  $1 \leq i \leq p$ , let

$$\mathfrak{C}_{i,p} := \partial^2 / \partial \theta^2 \mathbb{M}_{n,p+1}(\tilde{\boldsymbol{\pi}}_{p+1}, \tilde{\mathbf{m}}_{p+1}). \quad (25)$$

This is the curvature in the direction in which we separate out the  $i$ th component into two separate components. Note that  $\mathfrak{C}_{i,p} \equiv \mathfrak{C}_{i,p}(\tilde{\boldsymbol{\pi}}_p, \tilde{\mathbf{m}}_p)$  depends on  $(\tilde{\boldsymbol{\pi}}_p, \tilde{\mathbf{m}}_p)$  but we will suppress that dependence below for succinctness. Also, a priori,  $\mathfrak{C}_{i,p}$  depends on  $\lambda$ . However we can see from (29) that in fact  $\mathfrak{C}_{i,p}$  is minimized by taking  $\lambda = 1/2$  always, so we do this from now on. We will use  $\mathfrak{C}_{i,p}$  when  $(\tilde{\boldsymbol{\pi}}_p, \tilde{\mathbf{m}}_p)$  is a stationary point.

With the above setup, we describe our active set type of algorithm in Algorithm 1 below. Algorithm 1 includes a call to a generic subroutine to find the optimum value of  $\mathbf{m}$  given a fixed length  $p$  and a fixed value of  $\boldsymbol{\pi}$  above (corresponding to “counts” in the algorithm). This is referred to as the “fixed- $p$ -subroutine” in the algorithm. Our suggested implementation is to use a “trust-region” second-order method for this generic subroutine, see (Fletcher, 1987, Section 5.1) or (Nocedal and Wright, 1999, Section 4.2) for details of a trust region method. Algorithm 1 also includes a call to a subroutine, for collapsing (approximately) unique entries, described in Algorithm 2. To optimize a function, the method requires to be able to compute the function, its gradient, and its Hessian. The latter two we have derived closed forms for, for error distributions that are either Gaussian, Laplace, or mixtures of Gaussians. We use numerical integration to compute the objective function itself at this point. In the algorithms, we use the notation  $\mathbf{m}_{i:j}$  to denote the subvector of  $\mathbf{m}$  given by the indices  $\{i, \dots, j\}$ .

---

**Algorithm 1:** Active set algorithm
 

---

**input** :  $p^{(0)} \in \mathbb{R}$ ,  $\mathbf{m}^{(0)} := (m_1^{(0)}, \dots, m_p^{(0)})$ ,  
 counts $^{(0)} := (\text{counts}_1^{(0)}, \dots, \text{counts}_p^{(0)})$  where  $p = p^{(0)}$ , and Tolerance  
 parameter  $\text{eps}$ , Stepsize  $\eta$

**output:**  $\hat{m}_n(X_{(1)}), \dots, \hat{m}_n(X_{(n)})$

**while** *end criterion not met* **do**

- Do fixed- $p$ -subroutine( $\mathbf{m}^{(i-1)}$ , counts $^{(i-1)}$ ): Find fixed- $p^{(i)}$  and fixed-counts optimal  $\mathbf{m}$ , and assign to  $\mathbf{m}^{(i)}$ ;
- Do activate-constraints-subroutine( $\mathbf{m}^{(i)}$ , counts $^{(i-1)}$ ,  $\text{eps}$ ): run Algorithm 2 which collapses the non-unique entries in  $\mathbf{m}^{(i)}$ , and store the output in  $\mathbf{m}^{(i)}$  and counts $^{(i)}$ , and let  $p^{(i)}$  be the new (smaller) number of unique entries;
- Compute  $\mathfrak{C}_{j,p}$  (see (25)) for each  $j = 1, \dots, p$ ;
- if**  $\min_j \mathfrak{C}_{j,p} \geq 0$  **then**
  - End algorithm;
- else**
  - $k \leftarrow \text{argmin}_i \mathfrak{C}_{i,p}$ ;
  - $p^{(i)} \leftarrow p^{(i)} + 1$ ;
  - counts $^{(i)} \leftarrow (\text{counts}_1^{(i)}, \dots, \text{counts}_k^{(i)}/2, \text{counts}_k^{(i)}/2, \dots, \text{counts}_p^{(i)})$ ;
  - $\mathbf{m}^{(i)} \leftarrow (m_1^{(i)}, \dots, m_k^{(i)} - \eta, m_k^{(i)} + \eta, \dots, m_p^{(i)})$ ;
- $i \leftarrow i + 1$ ;

**end**

**/\* Reconstruct full length solution \*/**

The solution vector is given by the (unique, sorted) elements  $m_i^{(K)}$ ,  
 $i = 1, \dots, p_K$ , each repeated counts $_i^{(K)}$  times, respectively, where  $K$  is the number of iterations run.

---

In Appendix D.6 we provide a few comments about practical implementation of the algorithm.

---

**Algorithm 2:** Activate constraints: group (approximately) non-unique entries in  $\mathbf{m}$

---

**input** :  $\mathbf{m} := (m_1, \dots, m_p)$ , counts := (counts<sub>1</sub>, ..., counts<sub>p</sub>),  $eps$  (tolerance parameter)  
**output**:  $\mathbf{m}^{\text{new}} \in \mathbb{R}^{\tilde{p}}$ , counts<sup>new</sup>  $\in \mathbb{R}^{\tilde{p}}$ , where  $1 \leq \tilde{p} \leq p$   
newidx  $\leftarrow 1$ , begidx  $\leftarrow 1$ ;  
**for**  $j \leftarrow 2$  **to**  $p + 1$  **do**  
    **if**  $(j == p + 1)$  *OR*  $(m_j - m_{\text{begidx}} > eps)$  **then**  
         $m_{\text{newidx}}^{\text{new}} \leftarrow \text{mean}(m_{\text{begidx}:(j-1)})$ ;  
        counts<sub>newidx</sub><sup>new</sup>  $\leftarrow$  sum of counts<sub>begidx:(j-1)</sub>;  
        begidx  $\leftarrow j$ ;  
        newidx  $\leftarrow$  newidx + 1;  
    **end**  
     $\tilde{p} \leftarrow$  length of  $\mathbf{m}^{\text{new}}$ ;  
**end**

---

## 7. Extension to the Case of Unknown Noise Distribution

### 7.1 Estimation of the Noise Distribution in the Semi-supervised Setting

In general, full knowledge of the distribution of  $\epsilon$  might not be available which means that one needs to estimate it. In this case, it may be possible to collect a sample of  $\epsilon$ 's,  $\epsilon_1^*, \dots, \epsilon_N^*$ , from a separate data source. These can be used to construct an estimate of  $\Phi_\epsilon$  which can be then plugged into the objective function. The sample of  $\epsilon$ 's does not necessarily need to be independent of the  $Y$  or  $X$  samples (note, for instance, Dattner et al. 2016). There are a variety of ways one may arrive at the sample of  $\epsilon$ 's. In some cases, the main data set may consist of unlinked covariates and responses, but there may be a smaller (or sub) data set of linked/paired covariates and responses,  $(X_1^*, Y_1^*), \dots, (X_N^*, Y_N^*)$ . In this case, one may run the traditional monotone regression on this subset to obtain a monotone estimator  $\hat{m}_N^*$  from which one can compute the estimated residuals by  $\hat{\epsilon}_i^* := Y_i^* - \hat{m}_N^*(X_i^*)$ ,  $i = 1, \dots, N$ . In general, the previously-described setting might be considered to be one of *semi-supervised learning*, where only a part of the data is unlinked. It would be useful with such data to learn from all of it simultaneously. This may be possible using the M-estimation framework we have proposed in this paper, but we leave an investigation of that question for future research.

### 7.2 Estimation of the Noise Distribution with Longitudinal Responses

Another common framework in which we may want to estimate  $\Phi_\epsilon$  from data is the one where we have repeated (or longitudinal) observations on the response  $Y$ . Assume we observe  $X_1, \dots, X_{n_x}$  as before and also  $Y_{1,j}, \dots, Y_{n_y,j}$ , where we take  $j \in \{1, 2\}$  (for simplicity). We will impose the assumption that the distribution of  $\epsilon$  is symmetric around 0. We also assume that  $Y_{i,j} = m_0(\tilde{X}_i) + \epsilon_{i,j}$  (for some  $\tilde{X}_i$  which does not belong to our data set and need not be observed), where  $\epsilon_{i,1}$  is independent of  $\epsilon_{i,2}$ , and both are independent of all other error terms and all  $X$  variables. Then, as in Carroll et al. (2006) and Dattner et al. (2016), we can let  $Y_i^* := (Y_{i,1} + Y_{i,2})/2 = m_0(\tilde{X}_i) + \epsilon'_i$  where  $\epsilon'_i := (\epsilon_{i,1} + \epsilon_{i,2})/2$ . If we let  $\epsilon_i^* := (Y_{i,1} - Y_{i,2})/2 = (\epsilon_{i,1} - \epsilon_{i,2})/2$ , then it follows

from the assumption of 0-symmetry that  $\epsilon_i^* \sim \epsilon'_i$ . Then, we can use  $X_1, \dots, X_{n_x}$  and  $Y_1^*, \dots, Y_{n_y}^*$  as our unlinked data and  $\epsilon_1^*, \dots, \epsilon_{n_y}^*$  to estimate  $\Phi_{\epsilon'}$ .

Note that computing the estimator of  $m$  in practice as described in Algorithm 1 generally requires computation of derivatives of  $\mathbb{M}_n$ , e.g., the gradients  $\frac{\partial}{\partial m_i} \mathbb{M}_n(\mathbf{m})$ , where we use the same slight abuse of notation as in (23). The gradient depends on  $\Phi_\epsilon$  and is given by (24). Hence, in the case where  $\Phi_\epsilon$  is unknown, the gradient cannot be computed directly and has to be replaced by an appropriate estimator. In the setting of longitudinal responses, we have

$$\frac{\partial}{\partial m_i} \mathbb{M}_n(\mathbf{m}) = 2n^{-1} - 2n^{-2} \sum_{\alpha=1}^n \{\Phi_{\epsilon'}(Y_\alpha^* - m_i) + B(m_i - m_\alpha)\}$$

where  $\Phi_{\epsilon'}$  denotes the common distribution function of  $\epsilon'_1, \dots, \epsilon'_n$  and where for all  $m \in \mathbb{R}$ ,  $B(m) = E(\Phi_{\epsilon'}(\epsilon' + m))$ . With  $\hat{\Phi}$  the empirical distribution function based on the sample  $\epsilon_1^*, \dots, \epsilon_{n_y}^*$ , the gradient can be estimated by

$$2n^{-1} - 2n^{-2} \sum_{\alpha=1}^n \left\{ \hat{\Phi}(Y_\alpha^* - m_i) + \hat{B}(m_i - m_\alpha) \right\}$$

where  $\hat{B}(m) = n_y^{-1} \sum_{i=1}^{n_y} \hat{\Phi}(\epsilon_i^* + m)$ . The advantage of this estimator is that it does not require any choice of tuning parameters.

## 8. Demonstrations on Synthetic and Real Data

### 8.1 Computations on Synthetic Data

In this subsection we present simulation studies for our method and compare our minimum contrast estimator to the deconvolution method of Carpentier and Schlüter (2016). We also compare to classical/linked (oracle) isotonic regression (which uses matching information that the other estimators do not use). We will use mean-squared errors (MSE's) for comparison: for an estimator  $\hat{m}$ , we report  $n^{-1} \sum_{i=1}^n (\hat{m}(X_{(i)}) - m_0(X_{(i)}))^2$ . There are 2 output tables, Tables 2 and 3, containing Monte Carlo estimates of MSE's. The sample sizes are taken to be  $n = 100$  and  $n = 1000$  for the first and second tables respectively. In both tables, we used 1000 Monte Carlo replications. We used 5 different true mean functions  $m_0$ . They are (up to translation and additive constants), together with the abbreviations that denote them, gathered in Table 1. In the tables, our estimator is “UL BDD” (where “UL” stands for “unlinked”), Car-

$m_0$	Abbreviation
$x$	“lin”
0	“const”
$2\mathbb{1}_{[0,5]}(x) + 8\mathbb{1}_{[5,10]}(x)$	“step2”
$5\mathbb{1}_{[10/3,20/3]}(x) + 10\mathbb{1}_{[20/3,10]}(x)$	“step3”
$(x^4\mathbb{1}_{(0,5]}(x) - x^4\mathbb{1}_{[-5,0)}(x))/120$	“power”

Table 1: The true monotone regression function used in the simulations and their abbreviations.

	UL BDD	UL CS	L mono
lin, Laplace	0.31	0.27	0.16
const, Laplace	0.18	0.25	0.05
step2, Laplace	0.33	2.84	0.09
step3, Laplace	0.43	2.74	0.12
power, Laplace	0.29	0.47	0.13
lin, Gauss	0.48	0.78	0.16
const, Gauss	0.10	1.23	0.05
step2, Gauss	0.19	2.48	0.09
step3, Gauss	0.32	2.59	0.12
power, Gauss	0.43	0.59	0.14

Table 2: Monte Carlo'd MSE's,  $n = 100$ .

pentier and Schlüter (2016)'s is “UL CS”, and isotonic regression is “L mono” (for linked regression, that is, based on the classical case where all covariates and responses are perfectly linked). Our simulations were performed taking both Laplace and Gaussian errors with standard deviation 1 (and both unlinked methods are well-specified). In Figure 1 we present the output from a single Monte Carlo run, so that the true functions along with sample data and estimates can be visualized. The 10 plots in the figure are all based on  $n = 100$  samples for each of 10 settings: the 5 monotone regression functions  $m_0$  of Table 1 with Laplace distributed errors (left column) and Gaussian errors (right column).

To implement the deconvolution method of Carpentier and Schlüter (2016), we needed to monotone the estimated CDF so that we could compute its generalized inverse. Carpentier and Schlüter (2016) do not mention a specific method for doing this; we chose to replace  $\hat{F}(x)$  with  $\max(\hat{F}(y) : y \leq x)$ . The deconvolution estimator at  $X_{(1)}$  or  $X_{(n)}$  was occasionally unstable because of the steepness of the deconvolution-estimated quantile function. In that case, we have dropped those values out without a noticeable effect. Furthermore, the bandwidth for the deconvolution estimator of the CDF was chosen using the bootstrap method of Delaigle and Gijbels (2004).

From the output one can see that it is not always the case that the Gaussian noise is harder for our method than Laplace is. This is an interesting finding because it is known that deconvolution is harder with Gaussian noise than it is with Laplace; see e.g. Fan (1991). This suggests that, although unlinked regression is tightly connected to deconvolution, considering the problem from this point of view may not be the most efficient approach. The output also shows that the CS estimator performs poorly especially when  $m_0$  has discontinuities. In general our estimator is competitive with the CS deconvolution estimator and in some cases is significantly better.

## 8.2 Computations on CEX Data

Figure 2 shows plots based on the United States' Consumer Expenditure Survey (CEX). The CEX survey has detailed data on both income and the expenditure patterns of so-called “U.S. consumer units” (roughly, households), see Ruggles et al. (2020). The CEX consists of two surveys, the “Interview” and the “Diary”; the data we use here come from the former. Since the CEX survey has data on both income and expenditures, we can use regular (matched / standard) regression techniques. We

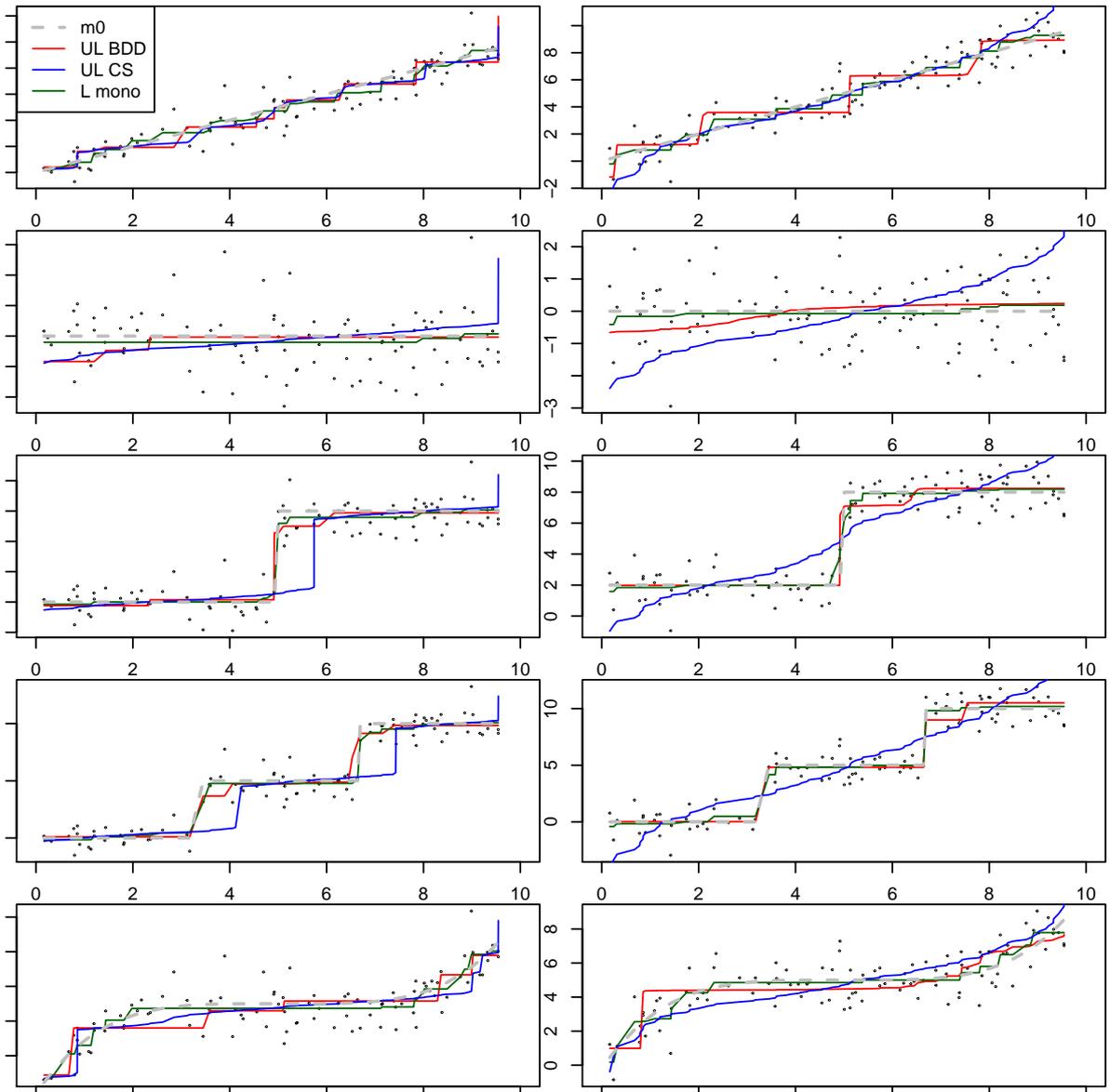


Figure 1: Output from a single Monte Carlo simulation, with  $n = 100$ , and  $X_i$  i.i.d. uniform on  $[0, 10]$ . The left column has Laplace errors and the right has Gaussian errors, both with standard deviation 1. The dotted gray line is the true  $m_0$ , the red line is our minimum contrast estimator, the blue line is the deconvolution estimator of Carpentier and Schlüter (2016), and the green line is a classical/linked isotonic regression.

MSE's	UL BDD	UL CS	L mono
lin, Laplace	0.10	0.10	0.03
const, Laplace	0.06	0.13	0.01
step2, Laplace	0.12	3.58	0.01
step3, Laplace	0.15	3.34	0.02
power, Laplace	0.09	0.36	0.03
lin, Gauss	0.29	0.18	0.03
const, Gauss	0.03	0.70	0.01
step2, Gauss	0.07	1.04	0.01
step3, Gauss	0.13	1.40	0.02
power, Gauss	0.25	0.25	0.03

Table 3: Monte Carlo'd MSE's,  $n = 1000$ .

compare a U.S. consumer unit's food expenditure to its income by regressing the former on the latter. Moreover, by simply ignoring the  $X$ - $Y$  pairing information we can also use our approach detailed above for the unlinked setting, and then compare the results obtained with both approaches. Note that we prefer here to use data that are matched, for the sake of being able to “validate” our results, but there are many settings where matching naturally lacks; for instance a firm may be able to gather information on an individual consumer's expenditures on the firm's products, but the firm would not be able to know the individual's income information. They would be able to access that expenditure information (at least nationally) through the CEX, which creates a data set with unlinked covariates and responses.

We consider the interview data only from the second quarter of 2018, for which there are approximately 6000 respondents. We narrow this down to 2164 respondents who provided the relevant information, had income no larger than \$250,000, and reported a non-negative response for both income and food expenditure. The survey actually follows each individual for four consecutive quarters, but we only included those who were surveyed in both quarter 2 and quarter 3. The “residuals” were computed as described in the previous section: for each individual  $i$ , we computed  $\tilde{\epsilon}_i := (Y_{i,1} - Y_{i,2})/2$  where  $Y_{i,1}$  is the quarter 2 response and  $Y_{i,2}$  is the quarter 3 response. The error distribution is assumed to be Laplace distributed with  $\lambda = \sqrt{\hat{\sigma}^2}/2$  where  $\hat{\sigma}^2$  is the empirical variance of  $\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_{2164}$ . We chose to assume that the errors follow the Laplace distribution rather than to use the full method detailed in Subsection 7.2 because the former is much faster to run. Choosing the Laplace distribution was based on observing that it fits much better the distribution of the residuals than the Gaussian one. In Figure 2, the “UL BDD” line is the unlinked monotone minimum contrast estimator proposed in this paper. This estimate is *fully* data driven, using no oracle (matching) information. The “L mono” line corresponds to monotone regression estimator based on the matched data; i.e., the Grenander-type estimator. Similarly the “L linear” line is a linear regression estimator based on the matched data. The “UL CS” line corresponds to the deconvolution estimator of Carpentier and Schlüter (2016) (using the same choice of  $\lambda$  we used in our method). We also implemented a type of “unlinked oracle” estimator in which we used a Gaussian-mixture as the error distribution, labeled “UL-oracle BDD”: For this estimator we used the residuals from the matched monotone regression to estimate  $\Phi_\epsilon$  (so this is oracle information which would not be available in a true unmatched problem). We used a four component

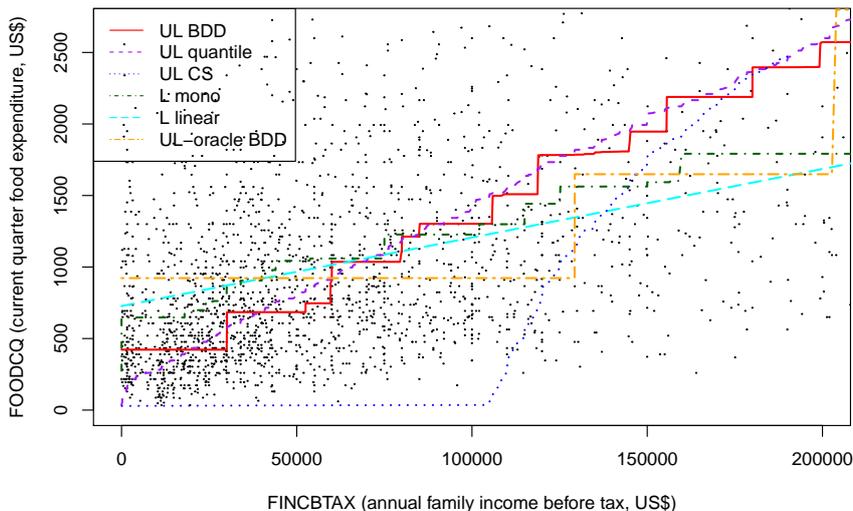


Figure 2: CEX Interview Survey: Family income vs. Food expenditure

Gaussian mixture (with no variance constraint) to approximate the error distribution, which was fit using an EM algorithm to converge to a local maximum (it is well known that the global maximum is infinite). The estimate has weights  $(.56, .05, .33, .06)$ , means/locations  $(-345, -416, 326, 1710)$  and standard deviations  $(322, 75, 562, 1286)$ . The mixture distribution is a much better fit to the residual distribution than either a single Gaussian or Laplace distribution, since the residual distribution is multimodal and heavy tailed. This estimator of  $\Phi_\epsilon$  is much more dispersed than the fully data driven one; for instance, the former has standard deviation 743 instead of 266 for the latter one. This leads to differences in these two estimators. Finally, the “UL quantile” line is based on matching the empirical quantiles of the  $Y$  and  $X$  samples: it is simply given by (connecting linearly) the points  $(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)})$ .

Our “UL BDD” estimator is somewhat accurate although it does differ noticeably from the oracle isotonic regression as well as the “UL-oracle BDD” estimator. The estimator of Carpentier and Schlüter (2016) does very poorly on this data set. We suspect that inaccuracy in choice of the error distribution causes difficulty for both of the unlinked estimators, which is of course expected.

## 9. Conclusions and Directions for Future Research

In this paper, we have presented a general method for unlinked regression with a monotonic regression function, and developed basic theory for the resulting estimator. We believe our approach will generalize to other (identifiable) unlinked regression settings. We have introduced a variant of an active set algorithm for computing the estimator, and demonstrated its use on a real data example in a fully data driven way in which we estimated the unknown error distribution. There are many remaining questions about this problem and about our method that future work could answer.

1. Our current study is restricted to the case of a univariate predictor. Studying both theory and practice when dimension is larger than 1 will be an important avenue for future work. In the case of linear regression, several works (Abid et al. (2017);

Pananjady et al. (2017, 2018); Unnikrishnan et al. (2018)) have already begun this study, although those works focus mostly on the case of Gaussian noise.

2. Finding (minimax) lower bounds for the rate of convergence seems to be hard to obtain in our setting. Such bounds are needed for a more complete theoretical understanding of the problem setting. In earlier literature, the closest result we are aware of is the one obtained in Rigollet and Weed (2019), who provide a minimax lower bound in the case of a Gaussian noise in the shuffled monotone regression. However, it is not known if the minimax lower bound derived by these authors applies to our problem, since the unlinked problem is statistically more difficult than the shuffled problem.
3. One of the major differences in unlinked regression from linked regression is that in the former the specification of the error distribution is crucially important. As shown in the introduction, if the error distribution is unknown then the model is not even identifiable. It would be helpful to understand the general properties of unlinked regression models and of our method in particular when one has partial but incomplete knowledge of the error distribution (e.g., some moment parameters can be estimated well but the full distribution is not known precisely).
4. Carpentier and Schlüter (2016) allowed for so-called “contextual variables”; for instance, if the unit  $i$  of observation is an individual, both  $Y_i$  and  $X_i$  may be paired with a contextual variable  $Z_{i,Y}$  and  $Z_{i,X}$  such as the individual’s age. One may “match”  $Y$ ’s and  $X$ ’s with equal (or similar) ages, and then one may consider unmatched regression on these partially matched data sets. This is what Carpentier and Schlüter (2016) proposed in the case of discrete, perfectly (noiselessly) observed contextual variables. More broadly, one may use so-called linkage methods (Herzog et al., 2007) to partially link  $Y$  and  $X$  (effectively reducing the noise level) when the contextual variables are not as idealized, and then perform linked regression. This methodology could be broadly useful in the linkage literature and warrants further study.

## Acknowledgements

The second author is supported in part by NSF Grant DMS-1712664. The third author is supported in part by MME-DII (ANR11-LBX-0023-01) and by the FP2M federation (CNRS FR 2036).

## Appendix A. Bounding the Integrals in (56) and (57).

Using Assumptions A0 and A1 we can write that

$$\begin{aligned} \int_{-\infty}^0 H_0(y)dy &= \int_{-\infty}^0 \int_{\mathbb{R}} \Phi_{\epsilon}(y - m_0(x))dF_0(x)dy \\ &\leq \int_{-\infty}^0 \Phi_{\epsilon}(y + K_0)dy < \infty, \end{aligned}$$

(note that this also follows from  $E(|Y|) < \infty$  and integration by parts.) Similarly it can be shown that  $\int_0^{\infty} (1 - H_0(y))dy < \infty$ . Hence,  $I_1 \leq \int_0^{\infty} (1 - H_0(y))dy + \int_{-\infty}^0 H_0(y)dy <$

$\infty$ . Also,

$$\begin{aligned}
 \int_{\mathbb{R}} (\Phi_{\epsilon}(y - m_0(x)) - H_0(y))^2 dy &\leq 2 \int_{-\infty}^0 (\Phi_{\epsilon}(y - m_0(x))^2 + H_0(y)^2) dy \\
 &\quad + \int_0^{\infty} (1 - H_0(y))^2 dy \\
 &\leq 2 \int_{-\infty}^0 \Phi_{\epsilon}(y + K_0) dy + 2 \int_{-\infty}^0 H_0(y) dy \\
 &\quad + \int_0^{\infty} (1 - H_0(y)) dy \\
 &< \infty,
 \end{aligned}$$

as shown above; this implies, by Fubini's Theorem, that  $I_2 < \infty$ .  $\square$

## Appendix B. Basic Empirical Process Theory Definitions and a Fundamental Preservation Result

For a (possibly random) signed measure  $Q$  on a (measurable) space  $\mathcal{X}$  and a measurable function  $f$  on  $\mathcal{X}$ , we denote  $Qf := \int_{\mathcal{X}} f dQ$ . For some class of functions  $\mathcal{G}$ , we can define

- its  $\epsilon$ -covering number  $N(\epsilon, \mathcal{G}, \|\cdot\|)$  with respect to some norm  $\|\cdot\|$  is defined as the smallest integer  $N > 0$  such that there exists  $g_1, \dots, g_N$  satisfying that for any  $g \in \mathcal{G}$ , there exists  $i \in \{1, \dots, N\}$  such that

$$\|g - g_i\| < \epsilon,$$

- its  $\epsilon$ -bracketing number  $N_B(\epsilon, \mathcal{G}, \|\cdot\|)$  with respect to some norm  $\|\cdot\|$  is defined as the smallest integer  $N > 0$  such that there exist pairs  $(h_1, k_1), \dots, (h_N, k_N)$  satisfying that for any  $g \in \mathcal{G}$ , there exist  $i \in \{1, \dots, N\}$  such that  $h_i \leq g \leq k_i$  and

$$\|k_i - h_i\| < \epsilon.$$

From the definition of the covering and bracketing numbers it can be easily shown (van der Vaart and Wellner, 1996, pp. 83–84) that if  $\|\cdot\|$  is an  $L_p$  norm, for some  $1 \leq p \leq \infty$ , then for any  $\delta > 0$ ,

$$N(\delta, \mathcal{G}, \|\cdot\|) \leq N_B(2\delta, \mathcal{G}, \|\cdot\|). \quad (26)$$

Also, if the class  $\mathcal{G}$  admits an envelope  $F$ , then define for  $\eta > 0$  the number

$$J(\eta, \mathcal{G}) = \sup_Q \int_0^{\eta} \sqrt{1 + \log N(\delta \|F\|_{Q,2}, \mathcal{G}, L_2(Q))} d\delta \quad (27)$$

where the supremum is taken over all discrete probability measures  $Q$  such that  $\|F\|_{Q,2} := \left( \int |F(x)|^2 dQ(x) \right)^{1/2} < \infty$ .

We finish this section by the following preservation result.

**Proposition 8.** *Let  $\|\cdot\|$  be some norm, and  $\mathcal{G}_1$  and  $\mathcal{G}_2$  two classes of functions. For fixed  $\lambda_1, \lambda_2$  such that  $(\lambda_1, \lambda_2) \neq (0, 0)$ , define the class*

$$\lambda_1 \mathcal{G}_1 + \lambda_2 \mathcal{G}_2 = \{h = \lambda_1 g_1 + \lambda_2 g_2 : (g_1, g_2) \in \mathcal{G}_1 \times \mathcal{G}_2\}.$$

Then, for any  $\epsilon > 0$

$$N(\epsilon, \lambda_1 \mathcal{G}_1 + \lambda_2 \mathcal{G}_2, \|\cdot\|) \leq N(\epsilon(|\lambda_1| + |\lambda_2|)^{-1}, \mathcal{G}_1, \|\cdot\|) \times N(\epsilon(|\lambda_1| + |\lambda_2|)^{-1}, \mathcal{G}_2, \|\cdot\|).$$

**Proof** [Proof of Proposition 8.] Fix  $\epsilon > 0$ . Let  $h = \lambda_1 g_1 + \lambda_2 g_2 \in \lambda_1 \mathcal{G}_1 + \lambda_2 \mathcal{G}_2$ ,  $N_1 = N(\epsilon(|\lambda_1| + |\lambda_2|)^{-1}, \mathcal{G}_1, \|\cdot\|)$  and  $N_2 = N(\epsilon(|\lambda_1| + |\lambda_2|)^{-1}, \mathcal{G}_2, \|\cdot\|)$ . We assume in the sequel that both  $N_1$  and  $N_2$  are finite since otherwise, the inequality in Proposition 8 is trivial. Then, there exists a pair  $(i, j) \in \{1, \dots, N_1\} \times \{1, \dots, N_2\}$  and  $(g_{1,i}, g_{2,j})$  such that  $\|g_1 - g_{1,i}\| < \epsilon(|\lambda_1| + |\lambda_2|)^{-1}$  and  $\|g_2 - g_{2,j}\| < \epsilon(|\lambda_1| + |\lambda_2|)^{-1}$ . Then, by the triangle inequality we have that

$$\begin{aligned} \|h - \lambda_1 g_{1,i} - \lambda_2 g_{2,j}\| &= \|\lambda_1(g_1 - g_{1,i}) + \lambda_2(g_2 - g_{2,j})\| \\ &\leq |\lambda_1| \|g_1 - g_{1,i}\| + |\lambda_2| \|g_2 - g_{2,j}\| < \epsilon \end{aligned}$$

which completes the proof. ■

## Appendix C. Wasserstein Distance Lemmas

Recall that  $W_1(F, G)$  denotes the first Wasserstein distance between two probability distributions with distribution functions  $F$  and  $G$ . The following is a well-known representation of the Wasserstein-1 distance in one dimension (see, e.g., Bobkov and Ledoux (2019), or e.g., Proposition 2 of Meis and Mammen (2020)).

**Proposition 9.** *Let  $F, G$  be distribution functions on  $\mathbb{R}$ , each having finite first moment. Then*

$$W_1(F, G) = \int_{\mathbb{R}} |F(x) - G(x)| dx.$$

The following is Proposition 3 of Meis and Mammen (2020).

**Proposition 10.** *Let  $F, G$  be two distribution functions supported on  $[0, V]$  and suppose  $H$  is a distribution function supported on a finite set of points. Then*

$$W_1(F, G) \leq C(V, H) W_1(F \star H, G \star H)$$

where  $C(V, H) > 0$  depends only on  $V, H$ .

## Appendix D. Gradient, Curvature, and Other Algorithmic Computations

In this section, we prove (24) and give an explicit formula of  $B(m)$  in the case of Laplace, Gaussian, and Gaussian-mixture distributions.

### D.1 Proof of (24)

We have

$$\begin{aligned}
 \frac{\partial}{\partial m_i} \mathbb{M}_n(\mathbf{m}) &= \int_{\mathbb{R}} 2 \left( \mathbb{H}_n(y) - n^{-1} \sum_{\alpha=1}^n \Phi_\epsilon(y - m_\alpha) \right) n^{-1} f_\epsilon(y - m_i) dy \\
 &= 2n^{-2} \sum_{\alpha=1}^n \int_{\mathbb{R}} \left\{ \mathbb{1}_{[Y_\alpha, \infty)}(y) f_\epsilon(y - m_i) - \Phi_\epsilon(y - m_\alpha) f_\epsilon(y - m_i) \right\} dy \quad (28) \\
 &= 2n^{-2} \sum_{\alpha=1}^n \left\{ \int_{Y_\alpha - m_i}^{\infty} f_\epsilon(y) dy - \int_{\mathbb{R}} \Phi_\epsilon(y - m_\alpha + m_i) f_\epsilon(y) dy \right\},
 \end{aligned}$$

and (24) follows.

### D.2 Curvature Derivation

It is convenient to re-parameterize the objective function as was done in the algorithm development in the main text. Let  $\widehat{\mathbb{H}}_{\mathbf{m}}(y) := \sum_{j=1}^p \pi_j \Phi_\epsilon(y - m_j)$ , and then recall

$$\mathbb{M}_{n,p}(\mathbf{m}) := \int (\mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y))^2 dy$$

where  $\pi_j \geq 0$  and  $\sum_{j=1}^p \pi_j = 1$ . (Here  $n = n_y$ , and  $n_x$  is defined implicitly in terms of the  $\pi_j$  and  $p$ .) Then, as derived above but in the new notation, we have

$$\begin{aligned}
 \frac{\partial}{\partial m_i} \mathbb{M}_{n,p}(\mathbf{m}) &= 2\pi_i \int (\mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y)) f_\epsilon(y - m_i) dy \\
 &= 2\pi_i \left( n^{-1} \sum_{\alpha=1}^n 1 - \Phi(Y_\alpha - m_i) - \sum_{j=1}^p \pi_j E\Phi_\epsilon(\epsilon + m_i - m_j) \right) \\
 &= 2\pi_i \left( 1 - n^{-1} \sum_{\alpha=1}^n \Phi_\epsilon(Y_\alpha - m_i) - \sum_{j=1}^p \pi_j B(m_i - m_j) \right)
 \end{aligned}$$

where again  $B(m) := E\Phi_\epsilon(\epsilon + m)$ . Then (assuming  $i \neq j$ )

$$\begin{aligned}
 \frac{\partial^2}{\partial m_i^2} \mathbb{M}_{n,p}(\mathbf{m}) &= 2\pi_i \left( \sum_{\alpha=1}^n n^{-1} f_\epsilon(Y_\alpha - m_i) - \sum_{j \neq i}^p \pi_j B'(m_i - m_j) \right) \\
 \frac{\partial^2}{\partial m_i \partial m_j} \mathbb{M}_{n,p}(\mathbf{m}) &= 2\pi_i \pi_j B'(m_i - m_j).
 \end{aligned}$$

Thus, in order to compute the curvature (i.e.,  $\partial^2/\partial\theta^2$ ), it suffices to compute  $B'$  (which we do below in the three cases we consider).

One more set of calculations that are useful are the following; we have

$$\begin{aligned}
 \frac{\partial}{\partial \theta} \mathbb{M}_n(\mathbf{m}) &= \frac{2}{\sqrt{2}} \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \sum_{\alpha=1}^p \pi_\alpha \Phi_\epsilon(y - m_\alpha) \right) (\pi_i f_\epsilon(y - m_i) - \pi_j f_\epsilon(y - m_j)) dy, \\
 \frac{\partial^2}{\partial \theta^2} \mathbb{M}_n(\mathbf{m}) &= \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_{\mathbf{m}}(y) - \mathbb{H}_n(y) \right) \times \\
 &\quad (\pi_i f'_\epsilon(y - m_i) + \pi_j f'_\epsilon(y - m_j)) + (\pi_i f_\epsilon(y - m_i) - \pi_j f_\epsilon(y - m_j))^2 dy. \quad (29)
 \end{aligned}$$

### D.3 Computations for Laplace Distribution

Let  $\lambda > 0$  and assume that

$$\Phi_\epsilon(z) := \begin{cases} 2^{-1}e^{-|z|/\lambda} & \text{if } z \leq 0 \\ 1 - 2^{-1}e^{-z/\lambda} & \text{if } z > 0 \end{cases}$$

and let  $f_\epsilon(y) := \Phi'_\epsilon(y) = e^{-|z|/\lambda}/2\lambda$  for  $z \in \mathbb{R}$ . We compute  $B(m)$  for  $m \in \mathbb{R}$ . We have

$$B(m) = \int_{-\infty}^0 (4\lambda)^{-1} e^{y/\lambda} e^{-|y-m|/\lambda} dy + \int_0^{\infty} (1 - 2^{-1}e^{-y/\lambda}) e^{-|y-m|/\lambda} (2\lambda)^{-1} dy$$

which equals

$$\begin{aligned} & \int_{-\infty}^{m \wedge 0} (4\lambda)^{-1} e^{(y-m/2)2/\lambda} dy + \int_{m \wedge 0}^0 (4\lambda)^{-1} e^{m/\lambda} dy \\ & + \int_0^{m \vee 0} (2\lambda)^{-1} (e^{(y-m)/\lambda} - 2^{-1}e^{-m/\lambda}) dy \\ & + \int_{m \vee 0}^{\infty} \frac{1}{2\lambda} e^{-(y-m)/\lambda} - \frac{1}{4\lambda} e^{-(y-m/2)2/\lambda} dy. \end{aligned}$$

If  $m \leq 0$ , then

$$B(m) = \frac{1}{8}e^{m/\lambda} + \frac{|m|}{4\lambda}e^{m/\lambda} + 0 + \frac{3}{8}e^{m/\lambda}. \quad (30)$$

If  $m \geq 0$ , then

$$B(m) = \frac{1}{8}e^{-m/\lambda} + 0 + \left( \frac{1}{2} - \frac{1}{4\lambda}e^{-m/\lambda}(2\lambda + m) \right) + \left( \frac{1}{2} - \frac{1}{8}e^{-m/\lambda} \right). \quad (31)$$

This gives an explicit formula for  $B(m)$ .

We can then compute that for any  $m \in \mathbb{R}$  (including  $m = 0$ ),  $B'(m) = e^{-|m|/\lambda}(\lambda + |m|)/(4\lambda^2)$ . The calculations for  $B'(m)$  are as follows. For  $m \leq 0$ , we have

$$\begin{aligned} B'(m) &= \frac{1}{2\lambda}e^{m/\lambda} = ((4\lambda)^{-1}e^{m/\lambda} + m(4\lambda^2)^{-1}e^{m/\lambda}) = e^{m/\lambda} \left( \frac{2\lambda - \lambda - m}{4\lambda^2} \right) \\ &= e^{-|m|/\lambda} \frac{\lambda + |m|}{4\lambda^2}. \end{aligned}$$

And for  $m \geq 0$  we have

$$B'(m) = - \left( \frac{1}{4\lambda}e^{-m/\lambda} - \frac{2\lambda + m}{4\lambda^2}e^{-m/\lambda} \right) = e^{-m/\lambda} \left( \frac{2\lambda + m}{4\lambda^2} - \frac{\lambda}{4\lambda^2} \right) = e^{-m/\lambda} \frac{m + \lambda}{4\lambda^2}. \quad (32)$$

### D.4 Computations for Gaussian Errors

Now we consider the case where, for some  $\sigma > 0$ ,  $\Phi_\epsilon = \Phi(\cdot/\sigma)$  is the cumulative distribution function of a  $N(0, \sigma^2)$  random variable. It turns out we can write  $B(\cdot)$  in terms of  $\Phi$ : by Corollary 1 of Ellison (1964),  $B(m) = E\Phi_\epsilon(N(m, \sigma^2)/\sigma) = \Phi(m/\sigma\sqrt{2})$ . Thus we also have

$$\begin{aligned} B'(m) &= N(0, \sigma^2) \text{ density} = \frac{1}{\sqrt{4\pi\sigma^2}} e^{-m^2/4\sigma^2}, \\ B''(m) &= -\frac{2m}{(2\sigma)^3\sqrt{\pi}} e^{-m^2/4\sigma^2}. \end{aligned}$$

### D.5 Mixtures of Gaussian

This again relies on Corollary 1 of Ellison (1964). Recall  $\Phi$  is the CDF of a  $N(0, 1)$  variable. Assume for an integer  $L \geq 1$ , locations  $\mu_i \in \mathbb{R}$ , and weights  $\lambda_i$  summing to 1 that

$$\Phi_\epsilon = \sum_{i=1}^L \lambda_i \Phi \left( \frac{\cdot - \mu_i}{\sigma_i} \right).$$

Then

$$B(m) = E\Phi_\epsilon(\epsilon + m) = \sum_{i,j} \lambda_i \lambda_j E\Phi \left( \frac{Y - \mu_i + m}{\sigma_i} \right) \quad (33)$$

where  $Y \sim N(\mu_j, \sigma_j^2)$ , so  $\frac{Y - \mu_i + m}{\sigma_i} \sim N((\mu_j - \mu_i + m)/\sigma_i, \sigma_j^2/\sigma_i^2)$  and thus (33) equals (by Corollary 1 of Ellison (1964))

$$\sum_{i,j} \lambda_i \lambda_j \Phi \left( \frac{(m - \mu_i + \mu_j)/\sigma_i}{\sqrt{1 + \sigma_j^2/\sigma_i^2}} \right) = \sum_{i,j} \lambda_i \lambda_j \Phi \left( \frac{\mu_j - \mu_i + m}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right).$$

We then have that

$$B'(m) = \sum_{i,j} \lambda_i \lambda_j \phi \left( \frac{\mu_j - \mu_i + m}{\sqrt{\sigma_i^2 + \sigma_j^2}} \right) \frac{1}{\sqrt{\sigma_i^2 + \sigma_j^2}}.$$

### D.6 On Implementation of Algorithm 1

A few remarks about the implementation of Algorithm 1 are as follows. The entries of the initial counts vector should sum to  $n_x$ ; this relationship is then preserved throughout the algorithm. Generally the initializer for  $\mathbf{m}$  is taken to be a constant (e.g., the median of the responses with  $p^{(0)} = 1$ ). In the algorithm we did not take care to force the counts to be integers when we divided by 2, but this can easily be done and should be done for easy interpretability. As end criterion, one can iterate for a fixed number  $K$  of steps, or one can iterate until a stopping rule (e.g., the objective function decrease is smaller than a fixed tolerance level) is satisfied. A heuristic choice for the parameter  $\text{eps}$  is  $\text{eps} = (Y_{(n)} - Y_{(1)})/(n^{1/3}\sigma)$  where  $\sigma^2$  is the variance of  $\epsilon$ , and  $n^{1/3}$  is motivated by properties of classical isotonic regression.

## Appendix E. Proofs

In this section we provide our proofs.

### E.1 A Preparatory Lemma

We begin with a lemma that will be used several times in the proofs of the main results.

**Lemma 3.** *Let  $m \in \mathcal{M}$ . If  $F_0$  is continuous, then*

$$\int_{\mathbb{R}} \left( \int \mathbb{1}_{\{m(x) \leq w\}} dF_0(x) - F_0 \circ m^{-1}(w) \right)^2 dw = 0. \quad (34)$$

**Proof** [Proof of Lemma 3] Recall that  $m^{-1}$  is defined by (6) where the infimum of an empty set is defined to be 1. If the set in (6) is non-empty, then the infimum is achieved by right-continuity of  $m$ . Hence, we have  $m \circ m^{-1}(y) \geq y$  for all  $y \leq m(1)$ . Now, consider  $x \in [0, 1]$  and  $y \leq m(1)$  such that  $m(x) \geq y$ . Since the infimum in (6) is achieved this implies that  $x \geq m^{-1}(y)$ . Conversely, if we have  $x \geq m^{-1}(y)$  then monotonicity of  $m$  implies that  $m(x) \geq m \circ m^{-1}(y)$  where as mentioned above,  $m \circ m^{-1}(y) \geq y$ . It follows that for all  $x \in [0, 1]$  and  $y \leq m(1)$  we have the equivalence

$$m(x) \geq y \Leftrightarrow x \geq m^{-1}(y). \quad (35)$$

Now, consider  $y > m(1)$ . The set in (6) is empty and therefore,  $m^{-1}(y) = 1$  by definition. The left-hand inequality in (35) does not hold if  $x \in [0, 1]$ , and the right-hand inequality does not hold neither if  $x < 1$  since  $m^{-1}(y) = 1$ . This mean that the above equivalence holds for all  $x \in [0, 1]$  and  $y \in \mathbb{R}$ . Let  $X$  be a random variable with distribution function  $F_0$ . Since  $P(X = 1) = 0$  by assumption, it follows that for all  $w \in \mathbb{R}$ , we have

$$P(m(X) < w) = P(X < m^{-1}(w)) \quad (36)$$

and therefore,

$$P(m(X) \leq w) - P(X \leq m^{-1}(w)) = P(m(X) = w) - P(X = m^{-1}(w))$$

where the second probability on the right hand side equals zero since  $X$  has a continuous distribution function. It follows that

$$\begin{aligned} \int_{\mathbb{R}} (P(m(X) \leq w) - F_0 \circ m^{-1}(w))^2 dw &= \int_{\mathbb{R}} P(m(X) = w)^2 dw \\ &= 0. \end{aligned} \quad (37)$$

To see why the preceding equality holds true, note that since the distribution function of  $X$ ,  $F_0$ , is assumed to be continuous, then it follows that

$$\int_{\mathbb{R}} P(m(X) = w)^2 dw = \int_{\mathcal{W}} P(X \in [a(w), b(w)])^2 dw$$

where  $\mathcal{W}$  is the set of point  $w \in \mathbb{R}$  such that there exist  $x \neq x'$  that satisfy  $m(x) = m(x') = w$ , and for  $w \in \mathcal{W}$ ,  $a(w) < b(w)$  are such that  $m$  takes the constant value  $w$  on  $[a(w), b(w))$ , and  $a(w) = m^{-1}(w)$ . Using the well-known fact that a monotone function admits at most countably many constant parts, the set  $\mathcal{W}$  is at most countable and therefore,

$$\int_{\mathbb{R}} P(m(X) = w)^2 dw \leq \int_{\mathcal{W}} dw = \lambda(\mathcal{W}) = 0$$

where  $\lambda$  denotes the Lebesgue measure on  $\mathbb{R}$ . Lemma 3 follows from (37) since

$$P(m(X) \leq w) = \int \mathbb{1}_{\{m(x) \leq w\}} dF_0(x).$$

■

## E.2 Proofs for Section 2

**Proof of Proposition 1.** Let  $\bar{F}(u) = P(X \geq u)$  for all  $u \in \mathbb{R}$ . It follows from (36), that holds for all  $w \in \mathbb{R}$  and  $m \in \mathcal{M}$ , that

$$P(m_1(X) \geq t) = P(X \geq m_1^{-1}(t)) \quad (38)$$

$$= \bar{F} \circ m_1^{-1}(t) \quad (39)$$

for all  $t \in \mathbb{R}$ . Since  $m_1(X)$  has the same distribution as  $m_2(X)$ , this implies that

$$\bar{F} \circ m_1^{-1} = \bar{F} \circ m_2^{-1}. \quad (40)$$

It follows from the definition (6) of the inverse of a function  $m \in \mathcal{M}$ , where we recall that the infimum of an empty set is defined to be one, that  $m_j^{-1}(y) = 0$  for all  $y \leq m_j(0)$  and  $m_j^{-1}(y) = 1$  for all  $y > m_j(1)$ . Hence, we define the inverse of the non-increasing left-continuous function  $\bar{F} \circ m_j^{-1}$  as

$$(\bar{F} \circ m_j^{-1})^{-1}(t) = \sup\{y \in [m_j(0); m_j(1)], \bar{F} \circ m_j^{-1}(y) \geq t\}$$

for all  $t \in \mathbb{R}$ , with the convention that the supremum of an empty set is equal to  $m_j(0)$ . Our aim is to derive from (40) that the inverses of  $\bar{F} \circ m_1^{-1}$  and  $\bar{F} \circ m_2^{-1}$  are equal. This is not an immediate consequence of the equality in (40) since the definition of the inverse function of  $\bar{F} \circ m_j^{-1}$  involves the function  $m_j$  in addition to the function  $\bar{F} \circ m_j^{-1}$ . However, we show below that the dependence on  $m_j$  can be removed by restricting attention to a restricted support.

We define the generalized inverse of  $\bar{F}$  by

$$\bar{F}^{-1}(t) = \sup\{u \in [0, 1], \bar{F}(u) \geq t\}$$

for all  $t \in \mathbb{R}$ , with the convention that the supremum of an empty set is equal to zero. Similar to the proof of Lemma 3, it can be proved using that  $\bar{F}$  is non-increasing and left-continuous that the equivalence

$$\bar{F}(u) \geq t \iff \bar{F}^{-1}(t) \geq u \quad (41)$$

holds for all  $u \in (0, 1]$  and  $t \in \mathbb{R}$ . Combining this with (35) (that holds for all  $x \in [0, 1]$  and  $y \in \mathbb{R}$ ), we obtain that the equivalence

$$\bar{F} \circ m_j^{-1}(y) \geq t \iff m_j \circ \bar{F}^{-1}(t) \geq y \quad (42)$$

holds for all  $t > 0$  and  $y > m_j(0)$ . For  $y \leq m_j(0)$ , the inequalities on both sides of the equivalence in the previous display hold true for all  $t \leq 1$  since in that case,  $\bar{F} \circ m_j^{-1}(y) = \bar{F}(0) = 1$ . This means that the equivalence in (42) holds for all  $t \in (0, 1]$  and  $y \in \mathbb{R}$ . Now, consider  $t \in (0, 1]$  such that  $t < \bar{F} \circ m_j^{-1}(m_j(0))$ . Since  $m_j^{-1}(m_j(0)) = 0$ , this means that  $t < \bar{F}(0)$  where  $\bar{F}(0) = 1$ . Otherwise said, we consider  $t \in (0, 1)$ . Because  $t < \bar{F} \circ m_j^{-1}(m_j(0))$ , we have

$$(\bar{F} \circ m_j^{-1})^{-1}(t) = \sup\{y \leq m_j(1), \bar{F} \circ m_j^{-1}(y) \geq t\}.$$

Moreover, the inequality  $\bar{F} \circ m_j^{-1}(y) \geq t$  cannot hold for  $y > m_j(1)$  since  $t > 0$  and therefore,

$$(\bar{F} \circ m_j^{-1})^{-1}(t) = \sup\{y \in \mathbb{R}, \bar{F} \circ m_j^{-1}(y) \geq t\}.$$

Combining this with (40) proves that

$$(\bar{F} \circ m_1^{-1})^{-1}(t) = (\bar{F} \circ m_2^{-1})^{-1}(t)$$

for all  $t \in (0, 1)$ . Using the equivalence in (42) (that holds for all  $t \in (0, 1]$  and  $y \in \mathbb{R}$ ), we also have

$$\begin{aligned} (\bar{F} \circ m_j^{-1})^{-1}(t) &= \sup\{y \in \mathbb{R}, m_j \circ \bar{F}^{-1}(t) \geq y\} \\ &= m_j \circ \bar{F}^{-1}(t). \end{aligned}$$

Hence,

$$m_1 \circ \bar{F}^{-1}(t) = m_2 \circ \bar{F}^{-1}(t)$$

for all  $t \in (0, 1)$ . This in turn implies that  $m_1 = m_2$  on the support of  $X$  since the range of  $\bar{F}^{-1}$  is the support of  $X$ .  $\square$

**Proof of Proposition 2.** *Proof of Claim 1.* Recall that any element  $m \in \mathcal{M}$  is bounded, and hence there exists  $K > 0$  such that  $\|m\|_\infty \leq K$ . Denote by  $Y_{(1)} \leq \dots \leq Y_{(n)}$  the order statistics corresponding to  $Y_1, \dots, Y_n$ . We have for all  $y < Y_{(1)}$  that  $\mathbb{H}_n(y) = 0$ . Moreover, it follows from monotonicity of  $\Phi_\epsilon$  and  $m$  that  $0 \leq \Phi_\epsilon(y - m(X_i)) \leq \Phi_\epsilon(y - m(X_{(1)})) \leq 1$  for all  $i \in \{1, \dots, n\}$  and therefore,

$$\begin{aligned} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 &= n^{-2} \left( \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right)^2 \\ &\leq \Phi_\epsilon(y - m(X_{(1)}))^2 \\ &\leq \Phi_\epsilon(y - m(X_{(1)})). \end{aligned}$$

Now, existence of expectation of  $\epsilon$  implies that

$$\int_{-\infty}^c \Phi_\epsilon(t) dt < \infty, \quad \text{and} \quad \int_c^\infty (1 - \Phi_\epsilon(t)) dt < \infty \quad (43)$$

for arbitrary  $c \in \mathbb{R}$  and therefore,

$$\begin{aligned} \int_{-\infty}^{Y_{(1)}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 dy &\leq \int_{-\infty}^{Y_{(1)}} \Phi_\epsilon(y - m(X_{(1)})) dy \\ &= \int_{-\infty}^{Y_{(1)} - m(X_{(1)})} \Phi_\epsilon(y) dy \\ &< \infty. \end{aligned} \quad (44)$$

Similarly,  $\mathbb{H}_n(y) = 1$  for  $y > Y_{(n)}$  and hence

$$\begin{aligned} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 &= \left( 1 - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right)^2 \\ &\leq \left( 1 - \Phi_\epsilon(y - m(X_{(n)})) \right)^2 \\ &\leq 1 - \Phi_\epsilon(y - m(X_{(n)})). \end{aligned}$$

Combined with (43), this proves that

$$\int_{Y_{(n)}}^{\infty} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - m(X_i)) \right\}^2 dy < \infty. \quad (45)$$

Since the integrand is bounded, (44) and (45) yield

$$\int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - m(X_i)) \right\}^2 dy < \infty,$$

which proves that  $\mathbb{M}_n(m)$  is finite.

*Proof of Claim 2.* Assume that Assumption A1 holds and consider  $m \in \mathcal{M}$ . Then, we have that

$$\begin{aligned} \mathbb{M}(m) &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} \left( \Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - m(x)) \right) dF_0(x) \right\}^2 dy \\ &= \int_{[0, \infty)} \left\{ \int_{\mathbb{R}} \left( 1 - \Phi_{\epsilon}(y - m(x)) - (1 - \Phi_{\epsilon}(y - m_0(x))) \right) dF_0(x) \right\}^2 dy \end{aligned} \quad (46)$$

$$+ \int_{(-\infty, 0]} \left\{ \int_{\mathbb{R}} \left( \Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - m(x)) \right) dF_0(x) \right\}^2 dy. \quad (47)$$

We further bound above the integral in (46) by

$$\begin{aligned} &\leq 2 \int_{[0, \infty)} \left\{ \int_{\mathbb{R}} (1 - \Phi_{\epsilon}(y - m(x))) dF_0(x) \right\}^2 dy \\ &+ 2 \int_{[0, \infty)} \left\{ \int_{\mathbb{R}} (1 - \Phi_{\epsilon}(y - m_0(x))) dF_0(x) \right\}^2 dy \\ &\leq 4 \int_{[0, \infty)} \left( 1 - \Phi_{\epsilon}(y - \max(K_0, K)) \right)^2 dy \end{aligned} \quad (48)$$

where we recall that  $K \geq \|m\|_{\infty}$ . The latter integral is finite using again (43). Similarly, we argue that the integral in (47) can be also bounded above by

$$4 \int_{(-\infty, 0]} \left( \Phi_{\epsilon}(y + \max(K_0, K)) \right)^2 dy < \infty.$$

This completes the proof that  $\mathbb{M}(m)$  is finite.

*Proof of Claim 3.* Using again that  $\mathbb{H}_n(y) = 0$  for all  $y < Y_{(1)}$ , together with monotonicity of  $\Phi_{\epsilon}$  and  $m$ , we have that

$$\begin{aligned} \mathbb{M}_n(m) &\geq \int_{-\infty}^{Y_{(1)}} n^{-2} \left( \sum_{i=1}^n \Phi_{\epsilon}(y - m(X_i)) \right)^2 dy \\ &\geq n^{-2} \int_{-\infty}^{Y_{(1)}} \Phi_{\epsilon}(y - m(X_{(1)}))^2 dy \\ &= n^{-2} \int_{-\infty}^{Y_{(1)} - m(X_{(1)})} \Phi_{\epsilon}(t)^2 dt \\ &\rightarrow \infty, \text{ if } m(X_{(1)}) \rightarrow -\infty. \end{aligned}$$

Similarly, for  $y \geq Y_{(n)}$ , it holds that

$$\begin{aligned} \mathbb{M}_n(m) &\geq n^{-2} \int_{Y_{(n)}}^{\infty} (1 - \Phi_{\epsilon}(y - m(X_{(n)})))^2 dy \\ &\geq n^{-2} \int_{Y_{(n)} - m(X_{(n)})}^{\infty} (1 - \Phi_{\epsilon}(t))^2 dt \\ &\rightarrow \infty, \text{ if } m(X_{(n)}) \rightarrow \infty. \end{aligned}$$

Hence, there exists some  $K > 0$  (which may depend on  $n$ ) such that any candidate  $m \in \mathcal{M}$  for the minimization problem in (8) should satisfy  $-K \leq m(X_{(1)}) \leq \dots \leq m(X_{(n)}) \leq K$ . By identifying an element  $m \in \mathcal{M}_K$  by the corresponding vector  $\theta = (m(X_{(1)}), \dots, m(X_{(n)}))^T$ , it is easy to see that the original minimization problem is equivalent to minimizing

$$\tilde{\mathbb{M}}_n(\theta) =: \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - \theta_i) \right\}^2 dy$$

on the compact finite dimensional subset

$$\mathcal{S}_K =: \{(\theta_1, \dots, \theta_n)^T \in \mathbb{R}^n : -K \leq \theta_1 \leq \dots \leq \theta_n \leq K\}.$$

Now, the function  $\tilde{\mathbb{M}}_n$  is continuous on  $\mathcal{S}_K$  since for any sequence  $(\theta_p)_{p \geq 0}$  in  $\mathcal{S}_K^{\mathbb{N}}$  converging (in any distance) to  $\theta \in \mathcal{S}_K$ , the sequence of functions

$$y \mapsto \left( \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - \theta_{p,i}) \right)^2$$

converges pointwise by continuity of  $\Phi_{\epsilon}$  (see Assumption A2) to the limit

$$y \mapsto \left( \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - \theta_i) \right)^2. \quad (49)$$

Also, for  $y \in \mathbb{R}$ , we have that (49) is no larger than

$$\begin{cases} (n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y + K))^2, & \text{for } y < Y_{(1)} \\ 4, & \text{for } Y_{(1)} \leq y \leq Y_{(n)} \\ \left(1 - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - K)\right)^2, & \text{for } y > Y_{(n)} \end{cases}$$

where the function on the right side can be shown to be integrable using similar arguments as above. By the Lebesgue dominated convergence theorem, it follows that

$$\lim_{p \rightarrow \infty} \tilde{\mathbb{M}}_n(\theta_p) = \tilde{\mathbb{M}}_n(\theta).$$

Thus,  $\tilde{\mathbb{M}}_n$  admits at least a minimizer in  $\mathcal{S}_K$ ,  $\hat{\theta}_n$  say. We conclude that  $\mathbb{M}_n$  admits at least a minimizer  $\hat{m}_n$  which is bounded by  $K$ , and such that  $(\hat{m}_n(X_{(1)}), \dots, \hat{m}_n(X_{(n)}))^T = \hat{\theta}_n$ . The values of the minimizer being given by  $\hat{\theta}_n$  at the observed covariates  $X_1, \dots, X_n$ ,

any monotone interpolation of these values gives a solution to (8). In particular, there exists a solution  $\widehat{m}_n$  that takes constant values between successive covariates and that is right continuous.

*Proof of Claim 4.* Without loss of generality, and possibly changing scale, we can assume that  $\epsilon$  is supported on  $[-1, 1]$ . We show below that there exists at least a solution to (8) that is bounded in sup-norm by  $K_0 + 2$ , where  $K_0$  is taken from Assumption A1. For an arbitrary  $m \in \mathcal{M}$ , we define the truncated version  $\bar{m}$  by

$$\bar{m}(x) = \begin{cases} K_0 + 2 & \text{if } m(x) \geq K_0 + 2 \\ -K_0 - 2 & \text{if } m(x) \leq -K_0 - 2 \\ m(x) & \text{otherwise} \end{cases}.$$

In the following, we place ourselves in the event  $\|\epsilon\|_\infty \leq 1$  which occurs with probability 1. Consider  $y > K_0 + 1$ . Since  $|Y_i| \leq K_0 + 1$  for all  $i$ , we then have  $\mathbb{H}_n(y) = 1$ , and

$$\Phi_\epsilon(y - m(X_i)) = \Phi_\epsilon(y - \bar{m}(X_i)) = 1$$

for all  $X_i$ 's such that  $m(X_i) \leq -K_0 - 2$ . Also, for all  $X_i$ 's such that  $m(X_i) \geq K_0 + 2$  we have that

$$\Phi_\epsilon(y - m(X_i)) \leq \Phi_\epsilon(y - \bar{m}(X_i)) \leq 1.$$

This implies that

$$\begin{aligned} & \int_{K_0+1}^{\infty} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \bar{m}(X_i)) \right\}^2 dy \\ & \leq \int_{K_0+1}^{\infty} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 dy. \end{aligned} \quad (50)$$

Similarly, it can be shown that

$$\begin{aligned} & \int_{-\infty}^{-K_0-1} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \bar{m}(X_i)) \right\}^2 dy \\ & \leq \int_{-\infty}^{-K_0-1} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 dy. \end{aligned} \quad (51)$$

Now, consider  $y$  such that  $|y| \leq K_0 + 1$ . If for some  $i$  we have  $y > m(X_i) + 1$  (or  $y < m(X_i) - 1$ ) then we have that

$$\Phi_\epsilon(y - m(X_i)) = \Phi_\epsilon(y - \bar{m}(X_i)).$$

Indeed, if  $y > m(X_i) + 1$ , then  $m(X_i) < K_0$ . In case  $m(X_i) > -K_0 - 2$  we have  $m(X_i) = \bar{m}(X_i)$  and  $\Phi_\epsilon(y - m(X_i)) = \Phi_\epsilon(y - \bar{m}(X_i)) = 1$ . If  $m(X_i) \leq -K_0 - 2$ , then  $\bar{m}(X_i) = -K_0 - 2$  and hence  $y - \bar{m}(X_i) \geq 1$  implying again that  $\Phi_\epsilon(y - m(X_i)) = \Phi_\epsilon(y - \bar{m}(X_i)) = 1$ . Similar arguments can be used in case  $y < m(X_i) - 1$ .

Now, the equality in the above display holds also if  $|y - m(X_i)| \leq 1$  since in that case,  $|m(X_i)| \leq K_0 + 2$ , implying that  $\bar{m}(X_i) = m(X_i)$ . Combining this with (50) and (51) shows that

$$\begin{aligned} & \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \bar{m}(X_i)) \right\}^2 dy \\ & \leq \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - m(X_i)) \right\}^2 dy. \end{aligned}$$

From Claim 3 in Proposition 2, there exists at least a solution to (8) and from the arguments above, its truncated version also is a solution. Hence, there exists at least a solution that is bounded in the sup-norm by  $K_0 + 2$  with probability 1. This completes the proof of the proposition.  $\square$

### E.3 Proofs for Section 3

We first prove the propositions in Section 3 and finish with the proof of Theorems 1, 2, 3, 4 and 5.

**Proof of Proposition 3.** Let  $\widehat{\mathbb{H}}_n$  and  $\mathbb{H}_n^0$  be the distribution functions defined as

$$\widehat{\mathbb{H}}_n(y) = \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)), \text{ and } \mathbb{H}_n^0(y) = \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - m_0(X_i)), \quad (52)$$

for  $y \in \mathbb{R}$ . Recall the Plancherel's identity for Fourier transforms: for a function  $g \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ , where  $L_1(\mathbb{R})$  and  $L_2(\mathbb{R})$  denote respectively the set of integrable, and the set of square integrable functions from  $\mathbb{R}$  to  $\mathbb{R}$  with respect to the Lebesgue measure it holds that

$$\int_{\mathbb{R}} g(x)^2 dx = \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_g(x)|^2 dx$$

where  $\phi_g$  is defined in (10). If  $F_1$  and  $F_2$  are two distribution functions with finite expectations, it follows using integration by parts that

$$\psi_{F_2}(x) - \psi_{F_1}(x) = -ix \int_{\mathbb{R}} (F_2(t) - F_1(t)) e^{itx} dt$$

implying that

$$\phi_{F_2 - F_1}(x) = i \frac{\psi_{F_2}(x) - \psi_{F_1}(x)}{x} \quad (53)$$

for  $x \neq 0$ . Moreover, if  $F_1$  and  $F_2$  have finite expectations then

$$\int_{-\infty}^0 F_j(x) dx < \infty \text{ and } \int_0^{\infty} (1 - F_j(x)) dx < \infty,$$

for  $j \in \{1, 2\}$ , implying that  $F_1 - F_2 \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ . Therefore, the Plancherel identity implies that

$$\int_{\mathbb{R}} (F_2(x) - F_1(x))^2 dx = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{1}{x^2} |\psi_{F_2}(x) - \psi_{F_1}(x)|^2 dx.$$

We apply below this identity with  $F_1$  and  $F_2$  replaced respectively by  $\widehat{\mathbb{L}}_n$  and  $\mathbb{L}_n$ , defined in (14). Note that the two corresponding distributions have finite expectations since they are supported on a finite set. Hence,

$$\int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{1}{t^2} \left| \psi_{\widehat{\mathbb{L}}_n}(t) - \psi_{\mathbb{L}_n}(t) \right|^2 dt.$$

By Assumption A5, we can find  $T^* > 0$  such that  $|\phi_{f_\epsilon}(t)| \geq |\phi_{f_\epsilon}(T)| > 0$  for all  $T > T^*$  and  $|t| \leq T$ . Using that  $|\psi_F| \leq 1$  for any distribution function  $F$ , it follows from the previous display that for all  $T > T^*$  we have

$$\begin{aligned} \int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw &\leq \frac{1}{2\pi |\phi_{f_\epsilon}(T)|^2} \int_{-T}^T \frac{|\phi_{f_\epsilon}(t)|^2}{t^2} \left| \psi_{\widehat{\mathbb{L}}_n}(t) - \psi_{\mathbb{L}_n}(t) \right|^2 dt + \frac{4}{\pi T} \\ &\leq \frac{1}{2\pi |\phi_{f_\epsilon}(T)|^2} \int_{\mathbb{R}} \frac{|\phi_{f_\epsilon}(t)|^2}{t^2} \left| \psi_{\widehat{\mathbb{L}}_n}(t) - \psi_{\mathbb{L}_n}(t) \right|^2 dt + \frac{4}{\pi T}. \end{aligned}$$

Now, using again Plancherel's identity we have

$$\begin{aligned} \int_{\mathbb{R}} \frac{|\phi_{f_\epsilon}(t)|^2}{t^2} \left| \psi_{\widehat{\mathbb{L}}_n}(t) - \psi_{\mathbb{L}_n}(t) \right|^2 dt &= \int_{\mathbb{R}} |\phi_{f_\epsilon}(t)|^2 |\phi_{\widehat{\mathbb{L}}_n - \mathbb{L}_n}(t)|^2 dt \\ &= \int_{\mathbb{R}} |\phi_{f_\epsilon \star (\widehat{\mathbb{L}}_n - \mathbb{L}_n)}(t)|^2 dt \\ &= 2\pi \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right)^2 dy \end{aligned}$$

since  $\widehat{\mathbb{H}}_n = f_\epsilon \star \widehat{\mathbb{L}}_n$  and  $\mathbb{H}_n^0 = f_\epsilon \star \mathbb{L}_n$ . Here,  $(f \star g)(y) := \int_{\mathbb{R}} f(z)g(y-z)dz$ . Hence, it follows from Assumption A4 that for sufficiently large  $T$ ,

$$\int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \leq \frac{T^{2\beta}}{d_0^2} \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right)^2 dy + \frac{4}{\pi T}. \quad (54)$$

Assuming that we have

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right)^2 dy = O(n^{-1}), \quad (55)$$

it will follow that for all sufficiently large  $T$ ,

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \leq O(T^{2\beta} n^{-1}) + \frac{4}{\pi T}.$$

For  $T = T_n \sim n^{1/(2\beta+1)}$  we get

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \leq O \left( \frac{1}{n^{1/(2\beta+1)}} \right),$$

which proves Proposition 3.

Now, we will show (55). From the inequality  $(a+b)^2 \leq 2(a^2+b^2)$ , which holds for any  $a$  and  $b$  in  $\mathbb{R}$ , and the definition of  $\widehat{m}_n$  it follows that

$$\begin{aligned} \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right)^2 dy &\leq 2 \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n(y) \right)^2 dy + 2 \int_{\mathbb{R}} \left( \mathbb{H}_n^0(y) - \mathbb{H}_n(y) \right)^2 dy \\ &\leq 4 \int_{\mathbb{R}} \left( \mathbb{H}_n^0(y) - \mathbb{H}_n(y) \right)^2 dy \\ &\leq 8 \int_{\mathbb{R}} \left( \mathbb{H}_n^0(y) - H_0(y) \right)^2 dy + 8 \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - H_0(y) \right)^2 dy \end{aligned}$$

where

$$E[(\mathbb{H}_n(y) - H_0(y))^2] = n^{-1} H_0(y)(1 - H_0(y)),$$

and

$$\begin{aligned} E[(\mathbb{H}_n^0(y) - H_0(y))^2] &= \frac{1}{n} \text{Var} \Phi_\epsilon(y - m_0(X)) \\ &= \frac{1}{n} \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - H_0(y))^2 dF_0(x). \end{aligned}$$

Both the integrals

$$I_1 = \int_{\mathbb{R}} H_0(y)(1 - H_0(y)) dy \quad (56)$$

and

$$I_2 = \int_{\mathbb{R}} \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - H_0(y))^2 dF_0(x) dy \quad (57)$$

are finite, see Appendix A. This yields the result.  $\square$

**Proof of Proposition 4.** In the sequel we denote by  $\mathbb{P}_n^X$  and  $P^X$  the empirical probability measure associated with  $X_1, \dots, X_n$  and the true corresponding probability measure. Then, the two integrals in Proposition 4 are the integrated square of the empirical processes

$$\widehat{\mathbb{L}}_n(w) - \widehat{L}_n^0(w) = (\mathbb{P}_n^X - P^X) \mathbb{1}_{\{\widehat{m}_n(\cdot) \leq w\}}$$

and

$$\mathbb{L}_n(w) - L_0(w) = (\mathbb{P}_n^X - P^X) \mathbb{1}_{\{m_0(\cdot) \leq w\}}.$$

In Appendix B, we recall some of the basic tools of empirical processes that we need in this proof. In what follows, the notation  $\lesssim$  means smaller or equal modulo a universal positive multiplicative constant. For all fixed  $w \in \mathbb{R}$  and  $m \in \mathcal{M}$ , let  $k_{w,m}$  be the function defined by  $k_{w,m}(x) = \mathbb{1}_{m(x) \leq w}$  for all  $x \in [0, 1]$ . Consider the set of functions

$$\mathcal{I} := \left\{ k_{w,m}, \text{ with } m \in \mathcal{M} \text{ and } w \in [A, B] \right\}.$$

Using the same notation as in van der Vaart and Wellner (1996) (for completeness, we provide definitions in Appendix B), let us write  $\mathbb{G}_n k = \sqrt{n}(\mathbb{P}_n^X - P^X)k$  for  $k \in \mathcal{I}$ . Since  $\mathcal{I}$  is a subset of the class of monotone non-increasing functions  $f : \mathbb{R} \mapsto [0, 1]$ , it follows from van der Vaart and Wellner (1996, Theorem 2.7.5) that there exists a universal constant  $C > 0$ , such that for any  $\delta > 0$  and any probability measure  $Q$ ,

$$\log N_B(\delta, \mathcal{I}, L_2(Q)) \leq \frac{C}{\delta}$$

(where  $N_B(\cdot, \cdot, \cdot)$  is defined in Appendix B). Since  $\mathcal{I}$  admits  $F(t) = 1$  as an envelope, this and the inequality in (26) imply that

$$\begin{aligned} J(1, \mathcal{I}) &\leq \sup_Q \int_0^1 \sqrt{1 + \log N_B(2\delta, \mathcal{I}, L_2(Q))} d\delta \\ &\leq \int_0^1 \sqrt{1 + \frac{C}{2\delta}} d\delta \leq 1 + \sqrt{2C} < \infty, \end{aligned}$$

where  $J(\delta, \mathcal{F})$  is defined in (27). Since  $X_1, \dots, X_n$  are i.i.d. it follows now from van der Vaart and Wellner (1996, Theorem 2.14.1) that

$$(E [\|\mathbb{G}_n\|_{\mathcal{I}}^2])^{1/2} \lesssim J(1, \mathcal{I}). \quad (58)$$

Let us denote

$$M_n = \max \left( \int_A^B \left( \widehat{\mathbb{L}}_n(w) - \widehat{L}_n^0(w) \right)^2 dw, \int_A^B \left( \mathbb{L}_n(w) - L_0(w) \right)^2 dw \right).$$

Then,

$$0 \leq M_n \leq \frac{B-A}{n} \|\mathbb{G}_n\|_{\mathcal{I}}^2.$$

The first two claims in the proposition now follow from (58) combined to the Markov's inequality.

Now, using the inequality  $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$  for any real numbers  $a, b$  and  $c$ , we have

$$\begin{aligned} \int_A^B \left( \widehat{L}_n^0(w) - L_0(w) \right)^2 dw &\leq 3 \int_A^B \left( \widehat{\mathbb{L}}_n(w) - \widehat{L}_n^0(w) \right)^2 dw \\ &\quad + 3 \int_A^B \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \\ &\quad + 3 \int_A^B \left( \mathbb{L}_n(w) - L_0(w) \right)^2 dw. \end{aligned}$$

It thus follows from Proposition 3 that (16) holds provided that  $B-A = O_P(n^{2\beta/(2\beta+1)})$ .  $\square$

**Proof of Proposition 5.** The first equality in Proposition 5 follows from Lemma 3 above combined with the definition of  $\widehat{L}_n^0$  and  $L_n^0$ , while the second equality follows from Proposition 4.  $\square$

**Proof of Proposition 6.** It follows from Lemma 4 below that

$$\int_a^b |\widehat{m}_n(x) - m_0(x)| dF_0(x) \leq \int_{A_n}^{B_n} |F_0 \circ \widehat{m}_n^{-1}(x) - F_0 \circ m_0^{-1}(x)| dx.$$

The proposition then follows from applying the Cauchy-Schwarz inequality.  $\square$

**Lemma 4.** Let  $f : [0, 1] \rightarrow \mathbb{R}$  and  $g : [0, 1] \rightarrow \mathbb{R}$  be right-continuous non-decreasing functions. Let  $f^{-1}$  and  $g^{-1}$  be the corresponding generalized inverses, see (6) where the infimum of an empty set is defined to be one. Let  $H : [0, 1] \rightarrow [0, 1]$  be a continuous non-decreasing function. Then, for all  $a < b$  in  $[0, 1]$  we have

$$\int_a^b |f(t) - g(t)| dH(t) \leq \int_{I(a)}^{S(b)} |H \circ g^{-1}(x) - H \circ f^{-1}(x)| dx$$

where

$$I(a) = f(a) \wedge g(a) ; S(b) = f(b) \vee g(b).$$

**Proof** [Proof of Lemma 4.]

For all real numbers  $u$ , let  $u_+ = \max(u, 0)$ . We then have

$$\int_a^b |f(t) - g(t)| dH(t) = I_1 + I_2 \quad (59)$$

where

$$I_1 = \int_a^b (f(t) - g(t))_+ dH(t) \text{ and } I_2 = \int_a^b (g(t) - f(t))_+ dH(t).$$

Let us deal first with  $I_1$ . We have

$$\begin{aligned} I_1 &= \int_a^b \int_0^\infty \mathbb{I}_{\{x \leq f(t) - g(t)\}} dx dH(t) = \int_a^b \int_{g(t)}^\infty \mathbb{I}_{\{x \leq f(t)\}} dx dH(t) \\ &= \int_a^b \int_{g(t)}^{f(b)} \mathbb{I}_{\{x \leq f(t)\}} dx dH(t), \end{aligned}$$

where we use a change of variable for the second equality and the monotonicity of  $f$  for the third one. Similar to (35), the equivalence

$$t \geq f^{-1}(x) \iff f(t) \geq x$$

holds for all  $t \in [0, 1)$  and  $x \in \mathbb{R}$ . Combining this with the Fubini theorem, we arrive at

$$\begin{aligned} I_1 &= \int_a^b \int_{g(t)}^{f(b)} \mathbb{I}_{\{t \geq f^{-1}(x)\}} dx dH(t) \\ &= \int_{g(a)}^{f(b)} \int_a^b \mathbb{I}_{\{t \geq f^{-1}(x)\}} \mathbb{I}_{\{t < g^{-1}(x)\}} dH(t) dx. \end{aligned}$$

Hence, it follows from the continuity of  $H$  that

$$\begin{aligned} I_1 &= \int_{g(a)}^{f(b)} (H(g^{-1}(x) \wedge b) - H(f^{-1}(x) \vee a))_+ dx \\ &\leq \int_{g(a)}^{f(b)} (H(g^{-1}(x)) - H(f^{-1}(x)))_+ dx, \end{aligned}$$

since  $H$  is non-decreasing. Since  $I(a) \leq g(a)$  and  $S(b) \geq f(b)$ , this implies that

$$I_1 \leq \int_{I(a)}^{S(b)} (H \circ g^{-1}(x) - H \circ f^{-1}(x))_+ dx.$$

Interchanging the roles of  $f$  and  $g$ , we obtain

$$I_2 \leq \int_{I(a)}^{S(b)} (H \circ f^{-1}(x) - H \circ g^{-1}(x))_+ dx$$

and therefore,

$$I_1 + I_2 \leq \int_{I(a)}^{S(b)} |H \circ f^{-1}(x) - H \circ g^{-1}(x)| dx.$$

Lemma 4 then follows from (59). ■

**Proof of Theorem 1.** *Proof of the first claim.* If we have (12), then  $A_n$  and  $B_n$  from Proposition 6 are both of the order  $O_P(1)$ . Hence, the first claim in Theorem 1 is an immediate consequence of Proposition 6 combined with Proposition 5.

*Proof of the second claim.* Now, we show that (12) holds true for all  $a$  and  $b$  such that  $0 < F_0(a) \leq F_0(b) < 1$ . It follows from the definition of  $\widehat{\mathbb{L}}_n$  and  $\mathbb{L}_n$  together with the Hölder inequality and Proposition 3 that

$$\begin{aligned} \int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left( 1 - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) \leq w\}} \right) dw &= \int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left| \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right| dw \\ &\leq \left( \|m_0\|_\infty \int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_n(w) - \mathbb{L}_n(w) \right)^2 dw \right)^{1/2} \\ &= o_P(1). \end{aligned}$$

On the other hand,

$$\int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left( 1 - \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) \leq w\}} \right) dw \geq \|m_0\|_\infty n^{-1} \sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) > 2\|m_0\|_\infty\}}$$

and therefore,

$$\sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) > 2\|m_0\|_\infty\}} = o_P(n).$$

By monotonicity of  $\widehat{m}_n$  this implies that

$$\mathbb{1}_{\{\widehat{m}_n(b) > 2\|m_0\|_\infty\}} \sum_{i=1}^n \mathbb{1}_{\{X_i > b\}} \leq \sum_{i=1}^n \mathbb{1}_{\{\widehat{m}_n(X_i) > 2\|m_0\|_\infty\}} = o_P(n).$$

By the law of large numbers,  $n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_i > b\}}$  converges in probability to  $1 - F_0(b) > 0$  and therefore, it follows from the previous display that

$$\mathbb{1}_{\{\widehat{m}_n(b) > 2\|m_0\|_\infty\}} = o_P(1).$$

This implies that

$$\lim_{n \rightarrow \infty} P(\widehat{m}_n(b) > 2\|m_0\|_\infty) = 0.$$

One can prove similarly that

$$\lim_{n \rightarrow \infty} P(\widehat{m}_n(a) < -2\|m_0\|_\infty) = 0.$$

This implies (12) by monotonicity of  $\widehat{m}_n$ , which completes the proof of the second claim in Theorem 1. □

**Proof of Theorem 2** As the arguments are very similar to those used in the proof of Theorem 1, we focus here on how the converge rate is obtained in the supersmooth

case. Under Assumptions A0–A3, Assumption A4' and Assumption A5, we can show that this rate of convergence is driven by

$$O\left(\exp\left(\frac{2T^\beta}{\gamma}\right)T^{-2\alpha}n^{-1}\right) + \frac{4}{\pi T}$$

for  $T = T_n \rightarrow \infty$  as  $n \rightarrow \infty$  which should be determined so that the above expression is smallest. This means that the first term should converge to 0 or equivalently that there exists a sequence  $(K_n)_n$  such that  $\log(T_n) = K_n \rightarrow \infty$  and  $K_n \leq \log n$  such that

$$\frac{2T_n^\beta}{\gamma} = \log n + (2\alpha - 1)K_n$$

or equivalently

$$T_n = c\left(\log n + (2\alpha - 1)K_n\right)^{1/\beta}, \quad \text{for } c = (\gamma/2)^{1/\beta}.$$

It is not difficult to see that the optimal choice of the sequence  $(K_n)_n$  is  $K_n = (1 - a)\log n$  for some  $a \in (0, 1)$  (the case  $a = 0$  is impossible because otherwise we would have  $T_n = c(2\alpha)^{1/\beta}(\log(T_n))^{1/\beta}$ ). This in turn yields

$$T_n = c\left(a\log n + 2\alpha\log(T_n)\right)^{1/\beta},$$

implying that  $T_n \sim ca^{1/\beta}(\log n)^{1/\beta}$  and that rate of convergence is  $(\log n)^{-1/(2\beta)}$ .  $\square$

**Proof of Theorem 3.** Without loss of generality, we can assume that the support points of  $\epsilon$  are all in  $[-1, 1]$ . From Proposition 2 we know that with probability 1 there exists a solution to (8) which is bounded in the sup-norm by  $K_0 + 2$ . In the sequel, we denote by  $\widehat{m}_n$  such a solution.

Recall  $\widehat{\mathbb{H}}_n$  and  $\mathbb{H}_n^0$  from (52). Using the Cauchy-Schwarz inequality it follows that

$$\begin{aligned} \int_{\mathbb{R}} \left| \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right| dy &= \int_{-K_0-3}^{K_0+3} \left| \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right| dy \\ &\leq \left( (2K_0 + 6) \int_{\mathbb{R}} \left| \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right|^2 dy \right)^{1/2}. \end{aligned}$$

Since the equality in (55) holds under Assumptions A0 to A2 it follows that

$$\int_{\mathbb{R}} \left| \widehat{\mathbb{H}}_n(y) - \mathbb{H}_n^0(y) \right| dy = O_P(n^{-1/2}).$$

Now note that  $\widehat{\mathbb{H}}_n$  and  $\mathbb{H}_n^0$  are distribution functions with bounded support, and hence they admit a finite first moment. Therefore, denoting by  $W_1(F, G)$  the Wasserstein-distance of first order between two probability distributions with respective distribution functions  $F$  and  $G$ , it follows from Proposition 9 in Appendix C that

$$W_1\left(\widehat{\mathbb{H}}_n, \mathbb{H}_n^0\right) = O_P(n^{-1/2}).$$

With  $\widehat{\mathbb{L}}_n$  and  $\mathbb{L}_n$  taken from (14), it follows from Proposition 10 in Appendix C that there exists some constant  $C > 0$  that depends only on the distribution of  $\epsilon$  and  $K_0$  such that

$$W_1\left(\widehat{\mathbb{L}}_n, \mathbb{L}_n\right) \leq C W_1\left(\widehat{\mathbb{H}}_n, \mathbb{H}_n^0\right).$$

Thus,

$$W_1\left(\widehat{\mathbb{L}}_n, \mathbb{L}_n\right) = O_P(n^{-1/2}).$$

Using again Proposition 9 in Appendix C the latter rate yields

$$\int_{\mathbb{R}} \left| \widehat{\mathbb{L}}_n(y) - \mathbb{L}_n(y) \right| dy = O_P(n^{-1/2}).$$

Combining this with Proposition 4 and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \int_{\mathbb{R}} \left| \widehat{L}_n^0(y) - L_0(y) \right| dy &= \int_{-K_0-2}^{K_0+2} \left| \widehat{L}_n^0(y) - L_0(y) \right| dy \\ &= O_P(n^{-1/2}). \end{aligned}$$

Using the result of Lemma 4, it follows that

$$\begin{aligned} \int_0^1 |\widehat{m}_n(x) - m_0(x)| dF_0(x) &\leq \int_{m_0(0) \wedge \widehat{m}_n(0)}^{m_0(1) \vee \widehat{m}_n(1)} |F_0 \circ \widehat{m}_n^{-1}(y) - F_0 \circ m_0^{-1}(y)| dy \\ &= \int_{m_0(0) \wedge \widehat{m}_n(0)}^{m_0(1) \vee \widehat{m}_n(1)} \left| \widehat{L}_n^0(y) - L_0(y) \right| dy \\ &\leq \int_{-K_0-2}^{K_0+2} \left| \widehat{L}_n^0(y) - L_0(y) \right| dy \end{aligned}$$

implying that  $\int_0^1 |\widehat{m}_n(x) - m_0(x)| dF_0(x) = O_P(n^{-1/2})$ .  $\square$

**Proof of Theorem 4.** *Proof of Claim 1.* We start with the case where the noise has an absolutely continuous density. We restrict attention to the ordinary smooth case because the arguments are very similar in the supersmooth one. Now, since the proof in the ordinary smooth case follows the same lines of the proof of Theorem 1, details are omitted and the reader is referred to the latter proof. Here, we only point out the main existing differences. Similar to (14) and (52), we define

$$\widehat{\mathbb{L}}_{n_x, n_y}(w) := \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) \leq w\}} \quad \text{and} \quad \mathbb{L}_{n_x}(w) := \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbb{1}_{\{m_0(X_i) \leq w\}}$$

for all  $w \in \mathbb{R}$ , and

$$\widehat{\mathbb{H}}_{n_x, n_y}(y) = \frac{1}{n_x} \sum_{i=1}^{n_x} \Phi_\epsilon(y - \widehat{m}_{n_x, n_y}(X_i)), \quad \text{and} \quad \mathbb{H}_{n_x}^0(y) = \frac{1}{n_x} \sum_{i=1}^{n_x} \Phi_\epsilon(y - m_0(X_i))$$

for all  $y \in \mathbb{R}$ . With similar arguments as for the proof of (54) we obtain that for all sufficiently large  $T$ ,

$$\int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right)^2 dw \leq \frac{T^{2\beta}}{d_0^2} \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_x}^0(y) \right)^2 dy + \frac{4}{\pi T}. \quad (60)$$

Moreover, it follows from the definition of  $\widehat{m}_{n_x, n_y}$  that

$$\begin{aligned} \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_x}^0(y) \right)^2 dy &\leq 2 \int_{\mathbb{R}} \left( \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_y}(y) \right)^2 dy \\ &\quad + 2 \int_{\mathbb{R}} \left( \mathbb{H}_{n_x}^0(y) - \mathbb{H}_{n_y}(y) \right)^2 dy \end{aligned}$$

which is less than or equal to

$$4 \int_{\mathbb{R}} (\mathbb{H}_{n_x}^0(y) - \mathbb{H}_{n_y}(y))^2 dy \leq 8 \int_{\mathbb{R}} (\mathbb{H}_{n_x}^0(y) - H_0(y))^2 dy + 8 \int_{\mathbb{R}} (\mathbb{H}_{n_y}(y) - H_0(y))^2 dy$$

where

$$E[(\mathbb{H}_{n_y}(y) - H_0(y))^2] = n_y^{-1} H_0(y)(1 - H_0(y)),$$

and

$$\begin{aligned} E[(\mathbb{H}_{n_x}^0(y) - H_0(y))^2] &= \frac{1}{n_x} \text{Var} \Phi_\epsilon(y - m_0(X)) \\ &= \frac{1}{n_x} \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - H_0(y))^2 dF_0(x). \end{aligned}$$

The integrals  $I_1$  and  $I_2$  defined in (56) and (57) are finite, since it can be shown that  $\int_{-\infty}^0 H_0(y) dy < \infty$  and  $\int_0^\infty (1 - H_0(y)) dy < \infty$ ; see Appendix A. Hence, we obtain that

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_x}^0(y) \right)^2 dy = O((n_x \wedge n_y)^{-1}). \quad (61)$$

Combining this with (60) proves that for all sufficiently large  $T$ ,

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right)^2 dw \leq T^{2\beta} O((n_x \wedge n_y)^{-1}) + \frac{4}{\pi T}.$$

For  $T \sim (n_x \wedge n_y)^{1/(2\beta+1)}$  we get

$$\int_{\mathbb{R}} E \left( \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right)^2 dw \leq O \left( \frac{1}{(n_x \wedge n_y)^{1/(2\beta+1)}} \right), \quad (62)$$

which proves an analogue of Proposition 3 in the case of possibly unequal sample sizes.

Next, we consider an analogue of Proposition 4. For this task, we denote by  $\mathbb{P}_{n_x}^X$  and  $P^X$  the empirical probability measure associated with  $X_1, \dots, X_{n_x}$  and the true corresponding probability measure. We consider the empirical processes

$$\widehat{\mathbb{L}}_{n_x, n_y}(w) - \widehat{L}_{n_x, n_y}^0(w) = (\mathbb{P}_{n_x}^X - P^X) \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(\cdot) \leq w\}}$$

and

$$\mathbb{L}_{n_x}(w) - L_0(w) = (\mathbb{P}_{n_x}^X - P^X) \mathbb{1}_{\{m_0(\cdot) \leq w\}};$$

where

$$\widehat{L}_{n_x, n_y}^0(w) = \int \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(x) \leq w\}} dF_0(x)$$

and  $L_0$  is defined in (15). Then, with similar arguments as in the proof of Proposition 4, we obtain that for all random variables  $A < B$  (that may depend on  $n$ ) it holds that

$$\int_A^B \left( \widehat{\mathbb{L}}_{n_x, n_y}(w) - \widehat{L}_{n_x, n_y}^0(w) \right)^2 dw \leq (B - A) O_P(1/n_x) \leq (B - A) O_P(1/(n_x \wedge n_y)),$$

$$\int_A^B \left( \mathbb{L}_{n_x}(w) - L_0(w) \right)^2 dw \leq (B - A) O_P(1/n_x) \leq (B - A) O_P(1/(n_x \wedge n_y)),$$

where  $O_P(1/n_x)$  is uniform in  $A$  and  $B$ . Moreover, if  $B - A = O_P((n_x \wedge n_y)^{2\beta/(2\beta+1)})$ , then

$$\int_A^B \left( \widehat{L}_{n_x, n_y}^0(w) - L_0(w) \right)^2 dw = O_P((n_x \wedge n_y)^{-1/(2\beta+1)}).$$

Next, similar to Proposition 5 we obtain that for all random variables  $A < B$  such that  $B - A = O_P(1)$  it holds that

$$\begin{aligned} \int_A^B \left( F_0 \circ \widehat{m}_{n_x, n_y}^{-1}(w) - F_0 \circ m_0^{-1}(w) \right)^2 dw &= \int_A^B \left( \widehat{L}_n^0(w) - L_0(w) \right)^2 dw \\ &= O_P((n_x \wedge n_y)^{-1/(2\beta+1)}). \end{aligned}$$

Proposition 6 still holds in the case of possibly different sample sizes with  $\widehat{m}_n$  replaced by  $\widehat{m}_{n_x, n_y}$ . If (12) also holds, then  $A_n$  and  $B_n$  from Proposition 6 are both of the order  $O_P(1)$ . In that case, the second assertion in Theorem 4 is an immediate consequence of Proposition 6 combined with the preceding display. Hence, it remains to prove that (12) holds. It follows from the definition of  $\widehat{\mathbb{L}}_{n_x, n_y}$  and  $\mathbb{L}_{n_x}$  together with the Cauchy-Schwarz inequality and (62) that

$$\int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left( 1 - \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) \leq w\}} \right) dw = \int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left| \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right| dw$$

which is bounded above by

$$\sqrt{\left( \|m_0\|_\infty \int_{\mathbb{R}} \left( \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right)^2 dw \right)} = o_P(1).$$

On the other hand,

$$\int_{\|m_0\|_\infty}^{2\|m_0\|_\infty} \left( 1 - \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) \leq w\}} \right) dw \geq \|m_0\|_\infty n_x^{-1} \sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) > 2\|m_0\|_\infty\}}$$

and therefore,

$$\sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) > 2\|m_0\|_\infty\}} = o_P(n_x).$$

By monotonicity of  $\widehat{m}_{n_x, n_y}$  this implies that

$$\mathbb{1}_{\{\widehat{m}_{n_x, n_y}(b) > 2\|m_0\|_\infty\}} \sum_{i=1}^{n_x} \mathbb{1}_{\{X_i > b\}} \leq \sum_{i=1}^{n_x} \mathbb{1}_{\{\widehat{m}_{n_x, n_y}(X_i) > 2\|m_0\|_\infty\}} = o_P(n_x).$$

By the law of large numbers,  $n_x^{-1} \sum_{i=1}^{n_x} \mathbb{1}_{\{X_i > b\}}$  converges in probability to  $1 - F_0(b) > 0$  and therefore, it follows from the previous display that

$$\mathbb{1}_{\{\widehat{m}_{n_x, n_y}(b) > 2\|m_0\|_\infty\}} = o_P(1).$$

This implies that

$$\lim_{n \rightarrow \infty} P(\widehat{m}_{n_x, n_y}(b) > 2\|m_0\|_\infty) = 0.$$

One can prove similarly that

$$\lim_{n \rightarrow \infty} P(\widehat{m}_{n_x, n_y}(a) < -2\|m_0\|_\infty) = 0.$$

This implies (12) by monotonicity of  $\widehat{m}_n$ .

*Proof of Claim 2.* Now, we turn to the case where  $\epsilon$  is supported on a finite set of points. Without loss of generality we assume that  $\|\epsilon\|_\infty \leq 1$  with probability 1. The same proof of Claim 4 in Proposition 2 can be again used to show existence with probability 1 of an estimator  $\widehat{m}_{n_x, n_y}$  which is bounded in the sup-norm by  $K_0 + 2$ . Now, the rate obtained above in (61) and the Cauchy-Schwarz inequality allow us to write that

$$\begin{aligned} \int_{\mathbb{R}} \left| \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_x}^0(y) \right| dy &= \int_{-K_0-3}^{K_0+3} \left| \widehat{\mathbb{H}}_{n_x, n_y}(y) - \mathbb{H}_{n_x}^0(y) \right| dy \\ &= O_P((n_x \wedge n_y)^{-1/2}). \end{aligned}$$

Using the same arguments as in the proof of Theorem 3 (in particular Proposition 9 in Appendix C) this implies that  $W_1(\widehat{\mathbb{H}}_{n_x, n_y}, \mathbb{H}_{n_x}^0) = O_P((n_x \wedge n_y)^{-1/2})$ . Since  $\widehat{\mathbb{H}}_{n_x, n_y}(y) = \int_{\mathbb{R}} \widehat{\mathbb{L}}_{n_x, n_y}(y-t) d\Phi_\epsilon(t)$  and  $\mathbb{H}_{n_x}^0(y) = \int_{\mathbb{R}} \mathbb{L}_{n_x}(y-t) d\Phi_\epsilon(t)$ , we can use again Proposition 10 in Appendix C to find a constant  $D > 0$  depending only on the distribution of  $\epsilon$  and  $K_0$  such that

$$W_1(\widehat{\mathbb{L}}_{n_x, n_y}, \mathbb{L}_{n_x}) \leq D W_1(\widehat{\mathbb{H}}_{n_x, n_y}, \mathbb{H}_{n_x}^0).$$

implying that

$$\int_{\mathbb{R}} \left| \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right| dw = \int_{-K_0-2}^{K_0+2} \left| \widehat{\mathbb{L}}_{n_x, n_y}(w) - \mathbb{L}_{n_x}(w) \right| dw = O_P((n_x \wedge n_y)^{-1/2})$$

using again Proposition 2 of Meis and Mammen (2020). Therefore,

$$\begin{aligned} \int_{-K_0-2}^{K_0+2} \left| \widehat{L}_{n_x, n_y}^0(w) - L_0(w) \right| dw &= \int_{-K_0-2}^{K_0+2} |F_0 \circ \widehat{m}_{n_x, n_y}^{-1}(w) - F_0 \circ m_0^{-1}(w)| dw \\ &= O_P((n_x \wedge n_y)^{-1/2}) \end{aligned}$$

and hence

$$\begin{aligned} \int_0^1 |\widehat{m}_{n_x, n_y}(x) - m_0(x)| dF_0(x) &\leq \int_{\widehat{m}_{n_x, n_y}(0) \wedge m_0(0)}^{\widehat{m}_{n_x, n_y}(1) \vee m_0(1)} |F_0 \circ \widehat{m}_{n_x, n_y}^{-1}(w) - F_0 \circ m_0^{-1}(w)| dw \\ &\leq \int_{-K_0-2}^{K_0+2} |F_0 \circ \widehat{m}_{n_x, n_y}^{-1}(w) - F_0 \circ m_0^{-1}(w)| dw \\ &= O_P((n_x \wedge n_y)^{-1/2}), \end{aligned}$$

which completes the proof of Theorem 4.  $\square$

**Proof of Theorem 5.** Theorem 1, Theorem 2, or Theorem 3 imply that along any subsequence of  $\{n\}_{n=1}^\infty$  we can find a further subsequence  $\{n_i\}_{i=1}^\infty$  such that with probability 1

$$\lim_{i \rightarrow \infty} \int_a^b |\widehat{m}_{n_i}(x) - m_0(x)| dF_0(x) = 0.$$

Call this event  $E_1$ . For notational ease let  $\widehat{m}_{n_i} \equiv \widehat{m}_i$ . Further, by Corollary 2.3 of Stein and Shakarchi (2005) (stated for Lebesgue measure but the proof does not rely on the Lebesgue measure at all and the result holds for a general measure space), there exists another subsubsequence (which we call again  $\{n_i\}_{i=1}^\infty$  for convenience) such that  $\widehat{m}_i$  converges  $F_0$ -a.e. to  $m_0$ .

Recall  $C$  is a compact in the interior of  $[a, b] \cap \mathcal{S}_0$ . Then since  $m_0$  is continuous on  $C$ ,  $\widehat{m}_i$  converges on a dense subset of  $[a, b] \cap \mathcal{S}_0$  to  $m_0$  (for any points  $\alpha, \beta \in [a, b] \cap \mathcal{S}_0$ , the  $F_0$  measure of  $(\alpha, \beta)$  is given by  $F_0(\beta) - F_0(\alpha)$ , so if  $F_0(\beta) - F_0(\alpha) > 0$  then there must be a point of convergence in  $(\alpha, \beta)$ , since convergence is  $F_0$ -a.e.), and both  $\widehat{m}_i$  and  $m_0$  are monotone, it follows that  $\widehat{m}_i$  converges pointwise on all of  $C$  to  $m_0$  (one can sandwich any point in  $C$ , including its boundary points, by sequences of points above and below at which  $\widehat{m}_i$  converges to  $m_0$  and appeal to monotonicity).

And, we can strengthen the convergence to uniform convergence on  $C$ , since  $m_0$  and  $\widehat{m}_i$  are monotone and  $m_0$  is continuous on  $C$ , again, for any  $\omega \in E_1$ . The elementary proof is as follows. Fix  $\epsilon > 0$ . By uniform continuity of  $m_0$  on (the compact)  $[a, b]$ , there exists  $\delta > 0$  such that  $|m_0(x) - m_0(y)| \leq \epsilon$  for all  $x, y$  such that  $|x - y| \leq \delta$ . Cover  $C$  with the set of open (in  $C$ 's subspace topology) sets  $A(x, \delta) := \{y \in C : |x - y| < \delta/2\}$  for all  $x \in C$ , and extract by compactness a finite subcover of these open sets,  $A(x_i, \delta)$ , for  $i = 1, \dots, N$ . Let  $x_{i1} := \inf A(x_i, \delta)$  and  $x_{i2} := \sup A(x_i, \delta)$ . Since  $C$  is closed,  $x_{ij} \in C$ ,  $j = 1, 2$ . By pointwise convergence  $\widehat{m}_n(x_{ij})$  is within  $\epsilon$  of  $m_0(x_{ij})$ ,  $j = 1, 2$ , for all  $i$  and  $n$  large enough. Now, take any  $x \in C$ ; let  $j$  be such that  $x_{j1} \leq x \leq x_{j2}$ . Using monotonicity of  $\widehat{m}_n$  and of  $m_0$ , we have for  $n$  large enough that  $\widehat{m}_n(x) \leq \widehat{m}_n(x_{j2}) \leq m_0(x_{j2}) + \epsilon \leq m_0(x) + 2\epsilon$ . Similarly, using  $x_{j1}$ , we have  $\widehat{m}_n(x) \geq m_0(x) - 2\epsilon$ , which proves the uniform convergence on  $C$ .

Hence, for all  $\omega \in E_1$ , for all subsequences of  $\{n_i\}_{i=1}^\infty$ , we can find a further subsequence (depending on  $\omega$ ) along which  $\sup_{x \in C} |\widehat{m}_i(x) - m_0(x)|$  converges to zero. Hence, for all  $\omega \in E_1$ , this supremum distance converges to zero along the subsequence  $\{n_i\}_{i=1}^\infty$ . Therefore,

$$P \left( \limsup_{i \rightarrow \infty} \sup_{x \in C} |\widehat{m}_i(x) - m_0(x)| = 0 \right) = 1.$$

Since for any subsequence of  $\{n\}_{n=1}^\infty$ , we have almost sure convergence along a subsubsequence, it follows that along the original sequence  $\{n\}$

$$\sup_{x \in C} |\widehat{m}_n(x) - m_0(x)| \rightarrow 0, \text{ in probability}$$

which completes the proof.  $\square$

#### E.4 Proofs for Section 4

We begin with an auxiliary lemma. Below, we use the same notation  $\widehat{\mathbb{H}}_n$  and  $\mathbb{H}_n^0$  as in (52). We recall that  $H_0$  denotes the distribution function of  $Y$ . We use the same notation  $P_n^X$  and  $P^X$  as in the proof of Proposition 4 and denote, moreover, by  $E^X$  the expectation corresponding to  $P^X$ .

**Lemma 5.** *Let Assumptions A0 and A1 hold. Assume that  $\epsilon$  is supported on  $[-1; 1]$  and independent of  $X$ . For any solution  $\widehat{m}_n \in \mathcal{M}$  to (8) that is bounded in the sup-*

norm by  $K_0 + 2$ , we then have

$$\begin{aligned} \int_{\mathbb{R}} \left\{ H_0(y) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy &\leq 2 \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y) \right\}^2 dy + O_P(n^{-1}) \\ &= O_P(n^{-1}). \end{aligned} \quad (63)$$

Moreover,

$$\left\{ \int \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dy dF_0(x) \right\}^2 = O_P(n^{-1}). \quad (64)$$

**Proof** [Proof of Lemma 5.] We can write

$$n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] = (\mathbb{P}_n^X - P^X) \Phi_\epsilon(y - \widehat{m}_n(\cdot)).$$

For a fixed  $y$ , consider the class of non-decreasing functions

$$\left\{ x \mapsto -\Phi(y - m(x)), x \in [0, 1], m \text{ monotone non-decreasing and } \|m\|_\infty \leq K_0 + 2 \right\}.$$

Using entropy arguments as in the proof of Proposition 4 by replacing the class  $\mathcal{I}$  with the one defined above we can show that for all random variables  $A < B$  such that  $B - A = O_P(1)$ , it holds that

$$\int_A^B \left\{ n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy = O_P(n^{-1}).$$

Since  $\epsilon$  is supported on  $[-1, 1]$ , and  $\widehat{m}_n$  is assumed to be bounded in the sup-norm by  $K_0 + 2$ , the above integral over  $[A, B]$  with  $A = -K_0 - 3$  and  $B = K_0 + 3$ , is equal to the same integral over the whole real line  $\mathbb{R}$ . Hence, we get

$$\int_{\mathbb{R}} \left\{ n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy = O_P(n^{-1}). \quad (65)$$

On the other hand, with  $H_0$  the distribution function of  $Y$  we have

$$\begin{aligned} E \int_{\mathbb{R}} (\mathbb{H}_n(y) - H_0(y))^2 dy &= \int_{\mathbb{R}} E(\mathbb{H}_n(y) - H_0(y))^2 dy \\ &= n^{-1} \int_{\mathbb{R}} H_0(y)(1 - H_0(y)) dy \end{aligned}$$

since  $n\mathbb{H}_n(y)$  is a binomial random variable with parameter  $n$  and probability of success  $H_0(y)$ . The integral on the right hand side is finite since  $Y$  has bounded support (included in  $[-K_0 - 1, K_0 + 1]$ ) and therefore,

$$\int_{\mathbb{R}} (\mathbb{H}_n(y) - H_0(y))^2 dy = O_P(n^{-1}).$$

Combining this with (65) together with the fact that for all real numbers  $a$  and  $b$ , we have  $(a + b)^2 \leq 2a^2 + 2b^2$ , we conclude that

$$\begin{aligned} &\int_{\mathbb{R}} \left\{ H_0(y) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy \\ &\leq 2 \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right\}^2 dy + O_P(n^{-1}). \end{aligned}$$

The first inequality follows by definition of  $\widehat{\mathbb{H}}_n$ .

Now, it follows from the definition of  $\widehat{m}_n$  that

$$\begin{aligned} \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y) \right\}^2 dy &\leq \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \mathbb{H}_n^0(y) \right\}^2 dy \\ &\leq 2 \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - H_0(y) \right\}^2 dy + 2 \int_{\mathbb{R}} \left\{ \mathbb{H}_n^0(y) - H_0(y) \right\}^2 dy. \end{aligned}$$

Since  $\epsilon$  is independent of  $X$ , both  $n\mathbb{H}_n(y)$  and  $n\mathbb{H}_n^0(y)$  are the average of  $n$  i.i.d. bounded random variables with mean  $H_0(y)$  and therefore,

$$E \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y) \right\}^2 dy \leq 2n^{-1} \int_{\mathbb{R}} \text{Var}(1_{\{Y \leq y\}}) dy + 2n^{-1} \int_{\mathbb{R}} \text{Var}(\Phi_\epsilon(y - m_0(X))) dy.$$

Both integrals on the right-hand side are finite since the integrands are bounded and equal to zero for all  $y \leq -K_0 - 1$  and all  $y \geq K_0 + 1$ . Hence,

$$\int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \widehat{\mathbb{H}}_n(y) \right\}^2 dy = O_P(n^{-1}).$$

This completes the proof of (63).

Now, with  $F_0$  the distribution function of  $X$ , it follows from the assumption that  $\widehat{m}_n$  (as well as  $m_0$ ) is bounded in sup-norm by  $K_0 + 2$  that

$$\begin{aligned} &\int_{\mathbb{R}} \left\{ H_0(y) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy \\ &= \int_{\mathbb{R}} \left\{ \int (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dF_0(x) \right\}^2 dy \\ &= \int_{-K_0-3}^{K_0+3} \left\{ \int (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dF_0(x) \right\}^2 dy \end{aligned}$$

so it follows from the Jensen inequality and the Fubini theorem that

$$\begin{aligned} &\int_{\mathbb{R}} \left\{ H_0(y) - E^X [\Phi_\epsilon(y - \widehat{m}_n(X))] \right\}^2 dy \\ &\geq \frac{1}{2K_0 + 6} \left\{ \int_{\mathbb{R}} \int (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dF_0(x) dy \right\}^2 \\ &= \frac{1}{2K_0 + 6} \left\{ \int \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dy dF_0(x) \right\}^2. \end{aligned}$$

Combining this with (63) yields (64) and completes the proof of the lemma.  $\blacksquare$

### Proof of Theorem 6.

*Case of error with finite support:* Suppose that  $\epsilon$  is supported on a finite set such that all the points of the support belong to  $[-1, 1]$  and let  $k \geq 2$ . For any solution  $\widehat{m}_n$  to (8) which is bounded by  $K_0 + 2$  (which exists with probability 1 in view of Proposition 2), it follows from Theorem 3 and the identity  $a^k - b^k = (a - b)(a^{k-1} + a^{k-2}b + \dots + b^{k-1})$  that

$$\begin{aligned} \left| \int_0^1 (\widehat{m}_n^k(x) - m_0^k(x)) dF_0(x) \right| &\leq k(K_0 + 2)^{k-1} \left( \int_0^1 |\widehat{m}_n(x) - m_0(x)| dF_0(x) \right) \\ &= O_P(1/\sqrt{n}). \end{aligned}$$

To show (20), it is enough to show that

$$\int_0^1 \widehat{m}_n^k(x) d(\mathbb{F}_n(x) - F_0(x)) \equiv \frac{1}{\sqrt{n}} \mathbb{G}_n \widehat{m}_n^k \quad (66)$$

is  $O_P(1/\sqrt{n})$ . Define

$$\mathcal{M}_c = \{m : m \text{ non-decreasing on } [0, 1] \text{ and } \|m\|_\infty \leq c\},$$

and

$$\mathcal{G}_{k,c} = \{m^k : m \in \mathcal{M}_c\},$$

for  $c > 0$ . If  $k = 1$ , then  $\widehat{m}_n \in \mathcal{M}_{K_0+2}$  and it follows from similar arguments as in the proof of Proposition 4 that there exists  $M > 0$  depending only on  $K_0$  such that  $\|\mathbb{G}_n\|_{\mathcal{M}_{K_0+2}} = O_P(1)$  which implies from (66) that (20) holds true. Now, suppose that  $k \geq 2$ . Using the decomposition  $m^k = m^k \mathbb{1}_{m \geq 0} + m^k \mathbb{1}_{m < 0}$  we see that for any  $m \in \mathcal{M}_{K_0+2}$ ,  $m^k$  is either the sum or the difference (depending on whether  $k$  is odd or even) of two functions in  $\mathcal{M}_{(K_0+2)^k}$ . Using Proposition 8 with  $(\lambda_1, \lambda_2) = (1, -1)$  or  $(\lambda_1, \lambda_2) = (1, 1)$ , it follows that for any discrete measure  $Q$  and  $\delta > 0$

$$\log N(\delta, \mathcal{G}_{k, K_0+2}, L_2(Q)) \leq 2 \log N(\delta/2, \mathcal{M}_{(K_0+2)^k}, L_2(Q)).$$

Using (26) and similar arguments as in the proof of Proposition 4, we conclude again that  $\|\mathbb{G}_n\|_{\mathcal{G}_{k, K_0+2}} = O_P(1)$ . Together with (66), it follows that (20) holds true.

*Case of uniform error:* Now, we turn to the case where  $\epsilon \sim \mathcal{U}[-1, 1]$ . To compute the integral on the left-hand side of (64), we distinguish between several cases. We recall that, because  $\Phi_\epsilon$  is the distribution function of a uniformly distributed random variable over  $[-1, 1]$ ,  $\Phi_\epsilon(t)$  is equal to 0 if  $t < -1$ , to 1 if  $t > 1$ , and to  $(t + 1)/2$  otherwise. For all  $x$  such that  $m_0(x) \leq \widehat{m}_n(x) \leq m_0(x) + 2$  we have

$$\begin{aligned} & \Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x)) \\ &= \begin{cases} (y - m_0(x) + 1)/2 & \text{if } m_0(x) - 1 \leq y \leq \widehat{m}_n(x) - 1 \\ (\widehat{m}_n(x) - m_0(x))/2 & \text{if } \widehat{m}_n(x) - 1 \leq y \leq m_0(x) + 1 \\ 1 - (y - \widehat{m}_n(x) + 1)/2 & \text{if } m_0(x) + 1 \leq y \leq \widehat{m}_n(x) + 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, for all  $x$  such that  $m_0(x) \leq \widehat{m}_n(x) \leq m_0(x) + 2$  we have

$$\begin{aligned} & \int_{\mathbb{R}} (\Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x))) dy \\ &= \frac{1}{2} (\widehat{m}_n(x) - m_0(x))^2 + \frac{1}{2} (\widehat{m}_n(x) - m_0(x))(m_0(x) - \widehat{m}_n(x) + 2) \\ &= (\widehat{m}_n(x) - m_0(x)). \end{aligned}$$

Similarly, for all  $x$  such that  $m_0(x) + 2 < \widehat{m}_n(x)$  we have

$$\begin{aligned} & \Phi_\epsilon(y - m_0(x)) - \Phi_\epsilon(y - \widehat{m}_n(x)) \\ &= \begin{cases} (y - m_0(x) + 1)/2 & \text{if } m_0(x) - 1 \leq y \leq m_0(x) + 1 \\ 1 & \text{if } m_0(x) + 1 \leq y \leq \widehat{m}_n(x) - 1 \\ 1 - (y - \widehat{m}_n(x) + 1)/2 & \text{if } \widehat{m}_n(x) - 1 \leq y \leq \widehat{m}_n(x) + 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

which implies that

$$\int_{\mathbb{R}} (\Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - \hat{m}_n(x))) dy = (\hat{m}_n(x) - m_0(x)).$$

Hence,

$$\begin{aligned} & \int 1_{\{m_0(x) \leq \hat{m}_n(x)\}} \int_{\mathbb{R}} (\Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - \hat{m}_n(x))) dy dF_0(x) \\ &= \int 1_{\{m_0(x) \leq \hat{m}_n(x)\}} (\hat{m}_n(x) - m_0(x)) dF_0(x) \end{aligned}$$

Similarly,

$$\begin{aligned} & \int 1_{\{\hat{m}_n(x) \leq m_0(x)\}} \int_{\mathbb{R}} (\Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - \hat{m}_n(x))) dy dF_0(x) \\ &= \int 1_{\{\hat{m}_n(x) \leq m_0(x)\}} (\hat{m}_n(x) - m_0(x)) dF_0(x) \end{aligned}$$

Combining the two previous displays yields

$$\begin{aligned} & \int \int_{\mathbb{R}} (\Phi_{\epsilon}(y - m_0(x)) - \Phi_{\epsilon}(y - \hat{m}_n(x))) dy dF_0(x) \\ &= \int (\hat{m}_n(x) - m_0(x)) dF_0(x). \end{aligned}$$

Now, from (64) it follows that

$$\left| \int_0^1 (\hat{m}_n(x) - m_0(x)) dF_0(x) \right| = O_P(1/\sqrt{n}).$$

As we already know that  $\int_0^1 \hat{m}_n(x) d(\mathbb{F}_n(x) - F_0(x)) = O_P(1/\sqrt{n})$ , the second claim of the proposition now follows.  $\square$

## Appendix F. Proof of Proposition 7

To make the notation less cumbersome, we write in the following  $Z_i$  for the  $i$ -th order statistic  $X_{(i)}$ . Suppose that  $\hat{m}_n$  takes at least two distinct values and let  $1 \leq j < j' \leq n$  be such that  $\hat{m}_n$  is constant on  $[Z_j, Z_{j'})$ , where  $Z_j < Z_{j'}$  are two successive jump points of  $\hat{m}_n$ . Consider the function  $m_{\delta}$  which is right-continuous, constant between the order statistics  $Z_1, \dots, Z_n$ , and

$$m_{\delta}(Z_i) = \begin{cases} \hat{m}_n(Z_i) + \delta, & i \in \{j, \dots, j' - 1\} \\ \hat{m}_n(Z_i), & \text{otherwise.} \end{cases}$$

Then, the function  $m_{\delta}$  as defined above belongs to  $\mathcal{M}$ , provided that  $|\delta|$  is small enough. It follows from the definition (8) that  $\mathbb{M}_n(m_{\delta}) \geq \mathbb{M}_n(\hat{m}_n)$ . Using Taylor expansion of  $\Phi_{\epsilon}$  with the integral remainder term we can write that for  $i \in \{j, \dots, j' - 1\}$

$$\Phi_{\epsilon}(y - \hat{m}_n(Z_i) - \delta) = \Phi_{\epsilon}(y - \hat{m}_n(Z_i)) - \delta f_{\epsilon}(y - \hat{m}_n(Z_i)) + R_{\delta,i}(y)$$

where the remainder term  $R_{\delta,i}$  is given below. Hence,

$$\begin{aligned}
 0 &\leq \mathbb{M}_n(m_\delta) - \mathbb{M}_n(\widehat{m}_n) \tag{67} \\
 &= \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i \notin \{j, \dots, j'-1\}} \Phi_\epsilon(y - \widehat{m}_n(Z_i)) - \frac{1}{n} \sum_{i=j}^{j'-1} \Phi_\epsilon(y - \widehat{m}_n(Z_i) - \delta) \right\}^2 dy \\
 &\quad - \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right\}^2 dy \\
 &= \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i \notin \{j, \dots, j'-1\}} \Phi_\epsilon(y - \widehat{m}_n(Z_i)) \right. \\
 &\quad \left. - \frac{1}{n} \sum_{i=j}^{j'-1} \Phi_\epsilon(y - \widehat{m}_n(Z_i)) + \delta \frac{1}{n} \sum_{i=j}^{j'-1} f_\epsilon(y - \widehat{m}_n(Z_i)) - \frac{1}{n} \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right\}^2 dy \\
 &\quad - \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right\}^2 dy
 \end{aligned}$$

which equals

$$\begin{aligned}
 &\int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) + \delta \frac{1}{n} \sum_{i=j}^{j'-1} f_\epsilon(y - \widehat{m}_n(Z_i)) - \frac{1}{n} \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right\}^2 dy \\
 &\quad - \int_{\mathbb{R}} \left\{ \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right\}^2 dy
 \end{aligned}$$

which equals

$$\begin{aligned}
 &\frac{2}{n} \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) \left( \delta \sum_{i=j}^{j'-1} f_\epsilon(y - \widehat{m}_n(Z_i)) - \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right) dy \\
 &\quad + \frac{1}{n^2} \int_{\mathbb{R}} \left( \delta \sum_{i=j}^{j'-1} f_\epsilon(y - \widehat{m}_n(Z_i)) - \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right)^2 dy
 \end{aligned}$$

which equals

$$\begin{aligned}
 &\frac{2}{n} \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) \left( \delta(j' - j) f_\epsilon(y - \widehat{m}_n(Z_j)) - \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right) dy \\
 &\quad + \frac{1}{n^2} \int_{\mathbb{R}} \left( \delta(j' - j) f_\epsilon(y - \widehat{m}_n(Z_j)) - \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right)^2 dy
 \end{aligned}$$

where for  $i = j, \dots, j' - 1$

$$\begin{aligned}
 R_{\delta,i}(y) &= \int_{y - \widehat{m}_n(Z_i)}^{y - \widehat{m}_n(Z_i) - \delta} f'_\epsilon(t) \cdot (y - \widehat{m}_n(Z_i) - \delta - t) dt \\
 &= - \int_0^\delta f'_\epsilon(y - \widehat{m}_n(Z_i) - u) \cdot (u - \delta) du, \text{ letting } u = y - \widehat{m}_n(Z_i) - t \\
 &= -\delta^2 \int_0^1 f'_\epsilon(y - \widehat{m}_n(Z_i) - \delta v) (v - 1) dv, \text{ letting } v = u/\delta.
 \end{aligned}$$

Thus,  $(\mathbb{M}_n(m_\delta) - \mathbb{M}_n(\widehat{m}_n))/\delta$  equals

$$\begin{aligned} & \frac{2}{n}(j' - j) \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) f_\epsilon(y - \widehat{m}_n(Z_j)) dy \\ & - \frac{2}{n} \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) \frac{1}{\delta} \sum_{i=j}^{j'-1} R_{\delta,i}(y) dy \\ & + \frac{1}{n^2} \int_{\mathbb{R}} \left\{ \delta(j' - j)^2 f_\epsilon(y - \widehat{m}_n(Z_j))^2 - 2(j' - j) f_\epsilon(y - \widehat{m}_n(Z_j)) \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right. \\ & \quad \left. + \frac{\left( \sum_{i=j}^{j'-1} R_{\delta,i}(y) \right)^2}{\delta} \right\} dy \end{aligned}$$

which equals

$$\begin{aligned} & \frac{2}{n}(j' - j) \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) f_\epsilon(y - \widehat{m}_n(Z_j)) dy \\ & + \frac{2\delta}{n} \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) \times \\ & \quad \left( \sum_{i=j}^{j'-1} \int_0^1 f'_\epsilon(y - \widehat{m}_n(Z_i) - \delta v) (v - 1) dv \right) dy \\ & + \frac{\delta(j' - j)^2}{n^2} \int_{\mathbb{R}} f_\epsilon(y - \widehat{m}_n(Z_j))^2 dy \\ & + \frac{2\delta(j' - j)}{n^2} \int_{\mathbb{R}} f_\epsilon(y - \widehat{m}_n(Z_j)) \sum_{i=j}^{j'-1} \int_0^1 f'_\epsilon(y - \widehat{m}_n(Z_i) - \delta v) (v - 1) dv dy \\ & + \frac{\delta^3}{n^2} \int_{\mathbb{R}} \left( \int_0^1 \sum_{i=j}^{j'-1} f'_\epsilon(y - \widehat{m}_n(Z_i) - \delta v) (v - 1) dv \right)^2 dy. \end{aligned}$$

We show below that each term on the right hand side that depends on  $\delta$  takes the form of  $\delta^i, i = 1, 2, 3$  times a finite integral, so that it tends to zero as  $\delta \rightarrow 0$ . From Assumption A6, it follows that

$$\begin{aligned} & \left| \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right) \cdot \left( \sum_{i=j}^{j'-1} \int_0^1 f'_\epsilon(y - \widehat{m}_n(Z_i) - \delta v) (v - 1) dv \right) \right| \\ & \leq D(j' - j) \left| \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_\epsilon(y - \widehat{m}_n(X_i)) \right| \end{aligned}$$

which can be shown to be integrable on  $\mathbb{R}$  using the property of  $\Phi_\epsilon$  in (43). Also, Assumption A6 implies that there exists  $D' > 0$  such that  $\sup_{t \in \mathbb{R}} f_\epsilon(t) \leq D'$ . Then,

$$\int_{\mathbb{R}} f_\epsilon(y - \widehat{m}_n(Z_j))^2 dy \leq D' \int_{\mathbb{R}} f_\epsilon(y - \widehat{m}_n(Z_j)) dy = D',$$

and by Fubini's Theorem

$$\begin{aligned}
& \int_{\mathbb{R}} \sum_{i=j}^{j'-1} \int_0^1 f_{\epsilon}(y - \widehat{m}_n(Z_j)) |f'_{\epsilon}(y - \widehat{m}_n(Z_i) - \delta v)| (v-1) \, dv dy \\
&= \sum_{i=j}^{j'-1} \int_0^1 \left( \int_{\mathbb{R}} f_{\epsilon}(y - \widehat{m}_n(Z_j)) |f'_{\epsilon}(y - \widehat{m}_n(Z_i) - \delta v)| \, dy \right) (v-1) \, dv \\
&\leq D \sum_{i=j}^{j'-1} \int_0^1 (v-1) \, dv = \frac{D(j'-j)}{2}
\end{aligned}$$

using Assumption A6 and the fact that  $f_{\epsilon}$  is a density. Finally, using again Assumption A6 and Fubini's Theorem we have

$$\begin{aligned}
& \int_{\mathbb{R}} \left( \int_0^1 \sum_{i=j}^{j'-1} f'_{\epsilon}(y - \widehat{m}_n(Z_i) - \delta v) (v-1) \, dv \right)^2 \, dy \\
&\leq \frac{D(j'-j)}{2} \int_{\mathbb{R}} \int_0^1 \sum_{i=j}^{j'-1} |f'_{\epsilon}(y - \widehat{m}_n(Z_i) - \delta v)| (1-v) \, dv dy
\end{aligned}$$

which equals

$$\begin{aligned}
& \frac{D(j'-j)}{2} \int_0^1 \left( \sum_{i=j}^{j'-1} \int_{\mathbb{R}} |f'_{\epsilon}(y - \widehat{m}_n(Z_i) - \delta v)| \, dy \right) (1-v) \, dv \\
&= \frac{D(j'-j)^2}{4} \int_{\mathbb{R}} |f'_{\epsilon}(t)| \, dt < \infty,
\end{aligned}$$

by Assumption A6. By using (67) and distinguishing between the cases of positive and negative values of  $\delta$  it follows that

$$\begin{aligned}
0 &= \lim_{\delta \rightarrow 0} \frac{\mathbb{M}_n(m_{\delta}) - \mathbb{M}_n(\widehat{m}_n)}{\delta} \\
&= \frac{2}{n} (j'-j) \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - \frac{1}{n} \sum_{i=1}^n \Phi_{\epsilon}(y - \widehat{m}_n(X_i)) \right) f_{\epsilon}(y - \widehat{m}_n(Z_j)) \, dy
\end{aligned}$$

and therefore,

$$0 = \int_{\mathbb{R}} \left( \mathbb{H}_n(y) - n^{-1} \sum_{i=1}^n \Phi_{\epsilon}(y - \widehat{m}_n(X_i)) \right) f_{\epsilon}(y - \widehat{m}_k) \, dy$$

where  $\widehat{m}_k = \widehat{m}_n(Z_j) = \dots = \widehat{m}_n(Z_{j'-1})$ . This is precisely the condition given in (21).

In the case  $\widehat{m}_n$  takes a unique value, a similar reasoning give the same result, characterizing  $\widehat{m}_k$  for  $k = 1$ .

Now, the alternative expression in (22) follows from the fact that for any  $a \in \mathbb{R}$

$$\int_{\mathbb{R}} \mathbb{H}_n(y) f_{\epsilon}(y - a) \, dy = \frac{1}{n} \sum_{i=1}^n \int_{Y_i}^{\infty} f_{\epsilon}(y - a) \, dy = 1 - \frac{1}{n} \sum_{i=1}^n \Phi_{\epsilon}(Y_i - a)$$

which completes the proof.  $\square$

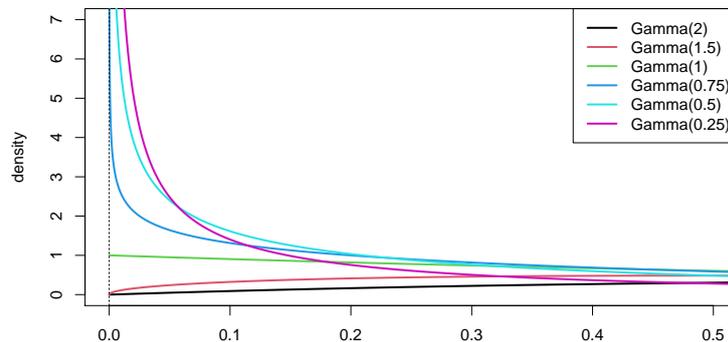


Figure 3: Plots of gamma densities with varying shape parameter.

## References

- Abubakar Abid, Ada Poon, and James Zou. Linear regression with shuffled labels. *arXiv*, May 2017.
- R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference under Order Restrictions. The Theory and Application of Isotonic Regression*. John Wiley, London-New York-Sydney, 1972.
- Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261, number 1259. American Mathematical Society, 2019.
- H. D. Brunk. Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, pages 177–197. Cambridge Univ. Press, London, 1970.
- Chris Carolan and Richard Dykstra. Asymptotic behavior of the Grenander estimator at density flat regions. *The Canadian Journal of Statistics*, 27(3):557–566, 1999. ISSN 0319-5724.
- Alexandra Carpentier and Teresa Schlüter. Learning relationships between data obtained independently. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 658–666, 2016.
- Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement error in nonlinear models*, volume 105 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2006. ISBN 978-1-58488-633-4; 1-58488-633-1. doi: 10.1201/9781420010138. URL <https://doi.org/10.1201/9781420010138>. A modern perspective.
- Eric Cator. Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17(2):714–735, 2011. ISSN 1350-7265. doi: 10.3150/10-BEJ289. URL <https://doi.org/10.3150/10-BEJ289>.

- Sabyasachi Chatterjee, Adityanand Guntuboyina, and Bodhisattva Sen. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015.
- Itai Dattner, Alexander Goldenshluger, and Anatoli Juditsky. On deconvolution of distribution functions. *The Annals of Statistics*, 39(5):2477–2501, 2011.
- Itai Dattner, Markus Reiß, Mathias Trabs, et al. Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli*, 22(1):143–192, 2016.
- Morris H DeGroot and Prem K Goel. Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, 8(2):264–278, March 1980.
- A Delaigle and I Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Ann. Inst. Statist. Math.*, 56(1):19–47, 2004.
- Cécile Durot. Sharp asymptotics for isotonic regression. *Probab. Theory Related Fields*, 122(2):222–240, 2002. ISSN 0178-8051. doi: 10.1007/s004400100171. URL <https://doi.org/10.1007/s004400100171>.
- Cécile Durot. On the  $L_p$ -error of monotonicity constrained estimators. *Ann. Statist.*, 35(3):1080–1104, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001497. URL <https://doi.org/10.1214/009053606000001497>.
- Cécile Durot and Hendrik P. Lopuhaä. A Kiefer-Wolfowitz type of result in a general setting, with an application to smooth monotone estimation. *Electron. J. Stat.*, 8(2):2479–2513, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS958. URL <https://doi.org/10.1214/14-EJS958>.
- Cécile Durot, Vladimir N. Kulikov, and Hendrik P. Lopuhaä. The limit distribution of the  $L_\infty$ -error of Grenander-type estimators. *Ann. Statist.*, 40(3):1578–1608, 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1015. URL <https://doi.org/10.1214/12-AOS1015>.
- David Edelman. Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *Ann. Statist.*, 16(4):1609–1622, 1988. ISSN 0090-5364. doi: 10.1214/aos/1176351056. URL <https://doi.org/10.1214/aos/1176351056>.
- Bob E Ellison. Two theorems for inferences about the normal distribution with applications in acceptance sampling. *Journal of the American Statistical Association*, 59(305):89–95, 1964. doi: 10.2307/2282860.
- Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991. ISSN 0090-5364. doi: 10.1214/aos/1176348248. URL <https://doi.org/10.1214/aos/1176348248>.
- R Fletcher. *Practical methods of optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, second edition, 1987.
- P. Groeneboom. Estimating a monotone density. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 539–555. Wadsworth, Belmont, CA, 1985.

- Piet Groeneboom. The concave majorant of Brownian motion. *Ann. Probab.*, 11(4): 1016–1027, 1983.
- Piet Groeneboom. Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields*, 81(1):79–109, 1989. ISSN 0178-8051. doi: 10.1007/BF00343738. URL <https://doi.org/10.1007/BF00343738>.
- Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, Cambridge, 2014.
- Peter Hall and Soumendra N Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics*, 36(5):2110–2134, 2008.
- Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Springer Science & Business Media, 2007.
- Enno Mammen. Estimating a smooth monotone regression function. *Ann. Statist.*, 19(2):724–740, 1991. ISSN 0090-5364. doi: 10.1214/aos/1176348117. URL <https://doi.org/10.1214/aos/1176348117>.
- Jan Meis and Enno Mammen. Uncoupled isotonic regression with discrete errors. *Personal communication*, 2020.
- Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, New York, New York, 1999.
- Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with an unknown permutation: Statistical and computational limits. In *Proceedings of the 2016 54th Annual Allerton Conference on Communication, Control, and Computing*, pages 417–424. IEEE, 2016.
- Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Denoising linear models with permuted data. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 446–450. IEEE, 2017.
- Ashwin Pananjady, Martin J Wainwright, and Thomas A Courtade. Linear regression with shuffled data: statistical and computational limits of permutation recovery. *IEEE Trans. Inform. Theory*, 64(5):3286–3300, 2018.
- Philippe Rigollet and Jonathan Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. *Information and Inference (to appear)*, 2019.
- Tim Robertson, F T Wright, and R L Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1988.
- Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. *IPUMS, USA: Version 10.0 [dataset]*. Minneapolis, MN: IPUMS, 2020.
- Elias M Stein and Rami Shakarchi. *Real analysis*, volume 3 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, NJ, 2005.

Jayakrishnan Unnikrishnan, Saeid Haghghatshoar, and Martin Vetterli. Unlabeled sensing with random linear measurements. *IEEE Trans. Inform. Theory*, 64(5): 3237–3253, 2018.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, 1996.

Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2): 528–555, 2002.