Isotonic regression in deterministic and random design settings

Viktor Jönsson

January 24, 2015

Acknowledgements

I would like to thank my supervisor Dragi Anevski for his endless support, patience and encouragement during this thesis and the last two and a half years.

Abstract

This thesis will treat the subject of constrained statistical inference and will have its focus on isotonic regression, which is the problem of estimating functions that are assumed to be monotone.

A characterisation of isotonic regression and a solution to this problem will be given. The PAVA algorithm to compute the isotonic regression estimator will be introduced along with asymptotic distribution results for this. The aim is to investigate the properties of the estimator when the observation points are random dependent variables which also depend on the unknown function itself.

Contents

1 Introduction

This paper will focus on constrained statistical inference where nonparametric estimation techniques will be used. The main advantages of nonparametric estimation techniques are their robustness against misspecifications in parameter models and its applicability in general settings such as estimating monotone, unimodal or convex functions. These can be used to estimate density and probability distribution functions, hazard rates, spectral densities et cetera. This leads us to study in great detail the problem of isotonic regression, which is the regression problem where the regressor is constrained to be an nondecreasing function (or nonincreasing, referred to as antitonic regression) and the corresponding PAVA algorithm for the construction of the isotone regression function. In most cases we have an equidistant design point on the unit interval and the observations in these points are given as a true function f plus some noise, usually Gaussian. In other words, we obtain observations of the form

$$y_i \quad = \quad f(t_i) + \epsilon_i$$

where f is an unknown increasing function, $t_i = i/n$ and ϵ_i are normally distributed. The aim of this paper is to investigate the case where the design points are random dependent variables and in addition depend on the unknown function f. Thus we consider a stochastic differential equation of the form

$$dX(t) \quad = \quad f(X(t))dt + \sigma dW(t)$$

and we are interested in estimating f in the asymptotic case where σ tends to zero and $W(t)$ is a Brownian motion. This is commonly referred to as a filtering problem and has been studied by Ibragimov, Hasminski [5] and Nussbaum [7] among others. It will be shown later that σ will tend to zero at the rate $n^{-1/2}$ as dt tends to zero at the rate n^{-1}. As f determines the drift in the stochastic differential equation, it is obvious that the unknown function will affect the design points and that these are random dependent variables since they are given by the observed stochastic process.

Thus we will start by stating and giving a characterisation to the isotonic regression problem and subsequently derive the Pool Adjacent Violator Algorithm, PAVA, which computes the isotonic regressor in practice. Furthermore, it will be shown what restrictions f most follow in general in order for the stochastic differential equation to have a solution. Limit distributions for the isotonic regression estimator will be derived for the deterministic design point case and for a specific stochastic differential equation. Finally simulations will be done in order to investigate the derived results for this stochastic differential equation and also another stochastic differential equation for which there has not been derived any limit distribution results. We will conclude by discussing the limit distribution results for the isotonic regression estimator for general stochastic differential equations of the type above.

2 Isotonic regression

We will start by doing a review of isotonic regression for the sake of completeness. The problem of isotonic regression has been treated earlier and can be found in for instance [9], which is the standard reference for constrained statistical inference and isotonic regression, and also [3]. Loosely speaking, isotonic regression is the problem of finding the best estimator, subject to some criterion function, of a regression function which has the constraint of being nondecreasing. For this reason, we start by considering binary relations, $\leq$. If x, y, z are elements in the set T, a binary relation $\leq$ defined on T is said to be:

$$\begin{cases} \text{reflexive:} & x \leq x \\ \text{transitive:} & x \leq y\,, y \leq z \implies x \leq z \\ \text{antisymmetric:} & x \leq y\,, y \leq x \implies x = y \end{cases} .$$

If a binary relation is reflexive and transitive, it is said to be a quasi-order. A binary relation which is reflexive, transitive, antisymmetric and also comparable, meaning that for any two elements $x, y \in T$ either $x \leq y$ or vice versa, is said to be a simple order. The standard inequality $\leq$ defined on $\mathbb{R}$ is easily seen to be a simple order and from here on, if not stated explicitly, $\leq$ will refer to this. We are now ready to give the definition of an isotonic function, [10].

Definition (isotonic function):
A function f defined on a set T with a quasi-order $\leq$ is said to be isotonic if for $x, y \in T$, $x \leq y \implies f(x) \leq f(y)$. Similarly, a function g is said to be antitonic if for $x, y \in T$, $x \leq y \implies g(y) \leq g(x)$. We will exclusively consider the standard order $\leq$ on $\mathbb{R}$ for which an isotonic function corresponds to a monotonically increasing function and an antitonic function to a monotonically decreasing function.

Now suppose we have observations y_i which can be described as $y(t_i) = m(t_i) + \epsilon_i$ where $t_i \in T$ for some set T, m is an unknown nondecreasing function and ϵ_i are assumed to be independent error terms of the measurements with variance σ_i. The aim is to estimate m under the constraint that m is nondecreasing. For this reason, we define the isotonic regressor $\hat{m}$ as the solution to

$$\hat{m} = \underset{z\in\mathcal{F}}{\text{argmin}} \sum_{i=1}^{n} (y(t_i) - z(t_i))^2 w_i$$

where $\mathcal{F} = \{z : T \to \mathbb{R} \ , z \, \text{nondecreasing}\}$, that is, the weighted least square estimator restricted to nondecreasing functions defined on T. Therefore, let us define the norm as the weighted least squares, that is,

$$||u||^2 = \sum_{i=1}^{n} u(t_i)^2 w_i.$$

If we assume that T is a finite set, $\mathcal{F} = \{z : T \to \mathbb{R} \ , z \, \text{nondecreasing}\}$ becomes a closed, convex cone, that is,

$$\begin{cases} pz \in \mathcal{F} \quad , z \in \mathcal{F} \, , p > 0 \\ pz_1 + (1-p)z_2 \in \mathcal{F} \quad , z_1, z_2 \in \mathcal{F} \, , 0 < p < 1 \\ \{z_i\}_{i=1}^{n} \in \mathcal{F} \, , ||z_n - z|| \longrightarrow 0 \, , n \longrightarrow \infty \Longrightarrow z \in \mathcal{F}. \end{cases}$$

The first one is obvious since if $z \in \mathcal{F}$ is nondecreasing, scaling it by a constant does not change this fact. As for the second one, consider $z(s) = (pz_1 + (1-p)z_2)(s)$ for $s \in T$. Then for $s \le t \, , s, t \in T$

$$z(s) = pz_1(s) + (1-p)z_2(s) \le pz_1(t) + (1-p)z_2(t) = z(t)$$

since $z_1, z_2 \in \mathcal{F}$ and thus $z \in \mathcal{F}$. The last implication holds since for $\{z_i\} \in \mathcal{F}$ and $s \le t \in T$,

$$\begin{aligned} z_n(s) &\le z_n(t) \Longrightarrow \\ \lim_{n\longrightarrow\infty} z_n(s) &\le \lim_{n\longrightarrow\infty} z_n(t) \Longrightarrow^{\text{assumption}} \\ z(s) &\le z(t) \end{aligned}$$

where the last step requires the finiteness of T.

Thus we wish to find the isotonic regressor $\hat{m}$ such that

$$\hat{m} = \underset{z\in\mathcal{F}}{\text{argmin}} \sum_{i=1}^{n} (y(t_i) - z(t_i))^2 w_i$$

where $\mathcal{F} = \{z : T \to \mathbb{R}\ ,\ z \text{ nondecreasing}\}$ was shown to be a convex,closed cone for $T = \{t_i\}_{i=1}^n$ finite. Now define

$$\begin{aligned}\phi(z) &= ||y - z||^2 \\ &= (y - z, y - z) \\ &= \sum_{i=1}^{n} (y(t_i) - z(t_i))^2 w_i.\end{aligned}$$

In the various next steps we will give a characterisation, which is given in [3], of the isotonic regressor and the conditions it needs to fulfil.

Theorem 1:
Let T be any set, $\mathcal{F}$ any convex set of functions and y, w be arbitrary functions defined on T. Then

$$\hat{m} \in \mathcal{F} = \underset{z \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (y(t_i) - z(t_i))^2 w_i \tag{1}$$

if and only if

$$(y - \hat{m}, \hat{m} - z) = \sum_{t_i \in T} [y(t_i) - \hat{m}(t_i)][\hat{m}(t_i) - z(t_i)]w(t_i) \geq 0$$

for $z \in \mathcal{F}$. The isotonic regressor $\hat{m}$ is unique if it exists.

Proof:
($\Longrightarrow$):
Assume $\hat{m}$ is the isotonic regressor. Thus $\hat{m} \in \mathcal{F}$ and since $\mathcal{F}$ is convex, for an arbitrary $z \in \mathcal{F}$, for $0 \leq \alpha \leq 1$, $(1 - \alpha)\hat{m} + \alpha z \in \mathcal{F}$. This gives

$$\begin{aligned}\varphi(\alpha) &= ||y - ((1 - \alpha)\hat{m} + \alpha z)||^2 \\ &= \sum_{i=1}^{n} (y(t_i) - ((1 - \alpha)\hat{m}(t_i) + \alpha z(t_i)))^2 w(t_i) \Longrightarrow \\ \varphi'(\alpha) &= 2\sum_{i=1}^{n} [y(t_i) - ((1 - \alpha)\hat{m}(t_i) + \alpha z(t_i))][\hat{m}(t_i) - z(t_i)]w(t_i).\end{aligned}$$

By the assumption that $\hat{m}$ was the isotonic regressor, this implies that $\varphi(\alpha)$ takes its smallest value for $\alpha = 0$. Since φ is a quadratic function of α, we must have that $\varphi'(0) \geq 0$. As seen above,

$$\begin{aligned}
\varphi'(0) &= 2\sum_{i=1}^{n}[y(t_i) - \hat{m}(t_i)][\hat{m}(t_i) - z(t_i)]w(t_i) \\
&= (y - \hat{m}, \hat{m} - z) \geq 0.
\end{aligned}$$

($\Longleftarrow$):
Let $u, z \in \mathcal{F}$ and suppose u satisfies $(y - u, u - z) \geq 0 \, \forall z \in \mathcal{F}$. This gives

$$\begin{aligned}
||y - z||^2 &= ||y - u + u - z||^2 \\
&= ((y - u) + (u - z), (y - u) + (u - z)) \\
&= (y - u, y - u) + 2(y - u, u - z) + (u - z, u - z) \\
&= ||y - u||^2 + ||u - z||^2 + 2(y - u, u - z) \\
&\geq ||y - u||^2 + ||u - z||^2 \\
&\geq ||y - u||^2 \,,
\end{aligned}$$

where the first inequality follows by the assumption. This proves that u is the isotonic regressor.

(Uniqueness):
To prove that there only exists one solution to the isotonic regression problem, consider the case where $u_1, u_2 \in \mathcal{F}$ are assumed to be solutions to the isotonic regression problem. From above, we get

$$\begin{gathered}
(y - u_1, u_1 - u_2) \geq 0 \\
(y - u_2, u_2 - u_1) \geq 0
\end{gathered}$$

which by adding the two gives

$$\begin{aligned}
(y - u_1, u_1 - u_2) + (y - u_2, u_2 - u_1) &\geq 0 \Longleftrightarrow \\
(y - u_1, u_1 - u_2) - (y - u_2, u_1 - u_2) &\geq 0 \Longleftrightarrow \\
(u_2, u_1 - u_2) - (u_1, u_1 - u_2) &\geq 0 \Longleftrightarrow \\
-(u_1 - u_2, u_1 - u_2) &\geq 0 \Longleftrightarrow \\
||u_1 - u_2||^2 &\leq 0.
\end{aligned}$$

Since a norm is always nonnegative, we have shown that $||u_1 - u_2||^2 = 0$, that is, $u_1 = u_2$ and hence the solution is unique. $\square$

Theorem (ϕ attains its minimum value):
Let y and w be arbitrary functions defined on the finite set T. Then the solution to the isotonic regression problem exists and is attained.

Proof:
Since ϕ is a convex and thus continuous function over the closed set $\mathcal{F}$, due to the finiteness of T, the minimum of ϕ on $\mathcal{F}$ exists and is attained. Thus $\hat{m}$ exists in $\mathcal{F}$. $\square$

We will give an equivalent characterisation to the one above which will enable us to find an algorithm that computes the solution to the isotonic regression problem. The characterisation is given in another theorem.

Theorem 2:
Let $\mathcal{F}$ be any convex cone of functions defined on the set T and y, w arbitrary functions also with domain T. Then $\hat{m} \in \mathcal{F}$ is the isotonic regressor if and only if for $z \in \mathcal{F}$,

$$\begin{aligned} (y - \hat{m}, \hat{m}) &= 0 \qquad (2) \\ (y - \hat{m}, z) &\leq 0. \qquad (3) \end{aligned}$$

Proof:
($\Longrightarrow$):
Let $\hat{m}$ be the solution to the isotonic regression problem and $\alpha > 0$. Since $\mathcal{F}$ is a convex cone this implies that $z = \alpha\hat{m} \in \mathcal{F}$ since $\hat{m} \in \mathcal{F}$. By theorem 1 we get

$$\begin{aligned} (y - \hat{m}, \hat{m} - \alpha\hat{m}) &\geq 0 \Longleftrightarrow \\ (1 - \alpha)(y - \hat{m}, \hat{m}) &\geq 0. \end{aligned}$$

Since this must hold $\forall \alpha > 0$ we obtain for $\alpha > 1$

$$(y - \hat{m}, \hat{m}) \leq 0$$

whereas for $0 < \alpha \leq 1$ we get

$$(y - \hat{m}, \hat{m}) \geq 0.$$

This shows that $(y - \hat{m}, \hat{m}) = 0$. Using theorem 1 again,

$$\begin{aligned} (y - \hat{m}, \hat{m} - z) &\geq 0 \Longleftrightarrow \\ (y - \hat{m}, \hat{m}) - (y - \hat{m}, z) &\geq 0 \Longleftrightarrow \\ -(y - \hat{m}, z) &\geq 0\,, \end{aligned}$$

that is, (3) is satisfied as well.

($\Longleftarrow$):

By subtracting (3) from (2) we obtain $(y - \hat{m}, \hat{m} - z) \geq 0$ which by theorem 1 shows that $\hat{m}$ is the isotonic regressor. $\square$

Theorem 2 will be used to find an algorithm which computes the isotonic regressor, for which we will need the following theorem.

Theorem 3:

Let $\mathcal{F}$ be the convex cone of nondecreasing functions defined on the set T. Then for any $z \in \mathcal{F}$ we can express z in the following way:

$$z = \sum_{i=1}^{n} \alpha_i \eta_i \tag{4}$$

where α_i are nonnegative weights defined on T and $\eta_i \in \mathcal{F}$. In other words, any function in $\mathcal{F}$ can be expressed as a linear combination of base functions in $\mathcal{F}$.

Proof:

$$\begin{aligned} z(t_k) &= z(t_k) - z(t_{k-1}) + z(t_{k-1}) - z(t_{k-2}) + z(t_{k-2}) - ... - z(t_1) + z(t_1) \\ &= z(t_k) - z(t_{k-1}) + z(t_{k-1}) - z(t_{k-2}) + ... - z(t_1) + \max(z(t_1), 0) + \\ &\quad \min(z(t_1), 0) \\ &= z(t_k) - z(t_{k-1}) + z(t_{k-1}) - z(t_{k-2}) + ... - z(t_1) + \max(z(t_1), 0) - \\ &\quad (-1)\min(z(t_1), 0) \\ &= \sum_{i=2}^{n} \alpha_i(t)\eta_i(t_k) + \alpha_{11}\eta_{11}(t_k) + \alpha_{12}\eta_{12}(t_k)\,, \end{aligned}$$

where $\forall i \geq 2$

$$\begin{aligned} \alpha_i(t) &= z(t_i) - z(t_{i-1}) \\ \eta_i(t) &= \mathbb{1}\{t_i \leq t\}\,, \end{aligned}$$

and

$$\begin{aligned}
\alpha_{11} &= \max(z(t_1), 0) \\
\alpha_{12} &= \max(-z(t_1), 0) \\
\eta_{11}(t) &= \mathbb{1}\{t_1 \leq t\} \\
\eta_{12}(t) &= -\mathbb{1}\{t_1 \leq t\}.
\end{aligned}$$

The fact that $z \in \mathcal{F}$ and $\max(a, 0) \geq 0\, \forall\, a \in \mathbb{R}$ ensures that $\alpha_i \geq 0$ and, since the indicator function is increasing, $\eta_i \in \mathcal{F}$. We recall that $\eta_{12}(t) = -\mathbb{1}\{t_1 \leq t\} \equiv -1\, \forall t \in T$ and hence an increasing function. This concludes the proof. $\square$

We will now restrict our attention to the case where $\mathcal{F}$ is the set of all nondecreasing functions on T, that is, $\mathcal{F} = \{z : T \longrightarrow \mathbb{R}\,,\, z \,\text{nondecreasing}\}$. As was proved above, $\mathcal{F}$ is then a convex cone which is also closed for finite T. The continuous case where T is an interval will be treated in a later section and therefore we assume T is finite in this section. Thus $\mathcal{F}$ satisfies theorem 1 and 2 and it was also shown that every function $z \in \mathcal{F}$ could be generated by base functions in $\mathcal{F}$. This enables us to find the solution to the isotonic regression problem,

$$\begin{aligned}
\hat{m} &= \operatorname*{argmin}_{z \in \mathcal{F}} \sum_{i=1}^{n} [y(t_i) - z(t_i)]^2 w(t_i) \\
&= \operatorname*{argmin}_{z \in \mathcal{F}} \phi(z)
\end{aligned}$$

where y, w are arbitrary functions defined on T and it was shown above that there exists a unique solution and this is attained. We now use (4) to prove the following theorem.

Theorem 4:
Let η_i be the base functions which generate $\mathcal{F}$. Then $\hat{m}$ is the isotonic regressor if and only if

$$\begin{aligned}
(y - \hat{m}, \eta_i) &\leq 0\,, \forall i = 1, ..., n &\quad (5) \\
(y - \hat{m}, \eta_i) &= 0\,, \text{if } \hat{\alpha}_i > 0 &\quad (6)
\end{aligned}$$

with $\hat{\alpha}_i = \hat{m}(t_i) - \hat{m}(t_{i-1})$.

Proof:
($\Longrightarrow$):
By (2) we know that $(y-\hat{m}, z) \le 0\,, \forall z \in \mathcal{F}$ and $(y-\hat{m}, \hat{m}) = 0$ which, since $\hat{m} \in \mathcal{F}$, gives due to (2)

$$\begin{aligned} 0 &\ge (y-\hat{m}, z) \\ &= (y-\hat{m}, \sum_{i=1}^{n} \alpha_i \eta_i) \\ &= \sum_{i=1}^{n} \alpha_i (y-\hat{m}, \eta_i). \end{aligned}$$

Now, since $\alpha_i \ge 0\,\forall i$ we obtain $(y-\hat{m}, \eta_i) \le 0\,\forall i$. From (3) we get

$$\begin{aligned} 0 &= (y-\hat{m}, \hat{m}) \\ &= (y-\hat{m}, \sum_{i=1}^{n} \hat{\alpha}_i \eta_i) \\ &= \sum_{i=1}^{n} \hat{\alpha}_i (y-\hat{m}, \eta_i). \end{aligned}$$

Since $(y-\hat{m}, \eta_i) \le 0\,, \hat{\alpha}_i \ge 0\,\forall i$, all terms are nonpositive and thus when $\hat{\alpha}_i > 0$ we must have $(y-\hat{m}, \eta_i) = 0$.
($\Longleftarrow$):

$$\begin{aligned} (y-\hat{m}, z) &= (y-\hat{m}, \sum_{i=1}^{n} \alpha_i \eta_i) \\ &= \sum_{i=1}^{n} \alpha_i (y-\hat{m}, \eta_i) \le 0 \end{aligned}$$

since by the first assumption (5), this is a sum of nonpositive terms. Furthermore,

$$(y-\hat{m}, \hat{m}) = \sum_{i=1}^{n} \hat{\alpha}_i (y-\hat{m}, \eta_i) = 0$$

as this is a sum of zero terms due to the second assumption. Subtracting the first equation from the second gives $(y-\hat{m}, \hat{m}-z) \ge 0$ which by theorem 2 proves that $\hat{m}$ is the isotonic regressor. $\square$

Hence we obtain for the solution to the isotonic regression problem, $\hat{m}$,

$$\begin{aligned}
(y-\hat{m},\eta_k) &= \sum_{i=1}^{n}[y(t_i)-\hat{m}(t_i)]\eta_k(t_i)w(t_i) \\
&= \sum_{i=1}^{n}[y(t_i)-\hat{m}(t_i)]\mathbb{1}\{t_k \le t_i\}w(t_i) \\
&= \sum_{i=k}^{n}[y(t_i)-\hat{m}(t_i)]w(t_i) \\
&= \begin{cases} \le 0\,, k=1,...,n \\ = 0\,, \hat{\alpha_k} = \hat{m}(t_k)-\hat{m}(t_{k-1}) > 0 \end{cases}
\end{aligned}$$

according to (5). Moreover, since $\pm\eta_1(t) = \pm\mathbb{1}\{t_1 \le t\} \equiv \pm 1$ are isotonic functions on T, we have $(y-\hat{m},\eta_1) \le 0$ as well as $(y-\hat{m},-\eta_1) \le 0$ which implies that $(y-\hat{m},\eta_1) = 0$. In other words, we have

$$\begin{aligned}
(y-\hat{m},\eta_1) &= \sum_{i=1}^{n}[y(t_i)-\hat{m}(t_i)]\eta_1(t_i)w(t_i) \\
&= \sum_{i=1}^{n}[y(t_i)-\hat{m}(t_i)]\mathbb{1}\{t_1 \le t_i\}w(t_i) \\
&= \sum_{i=1}^{n}[y(t_i)-\hat{m}(t_i)]w(t_i)\,,
\end{aligned}$$

which in turn gives

$$\begin{aligned}
(y-\hat{m},\eta_1)-(y-\hat{m},\eta_{k+1}) &= \sum_{i=1}^{k}[y(t_i)-\hat{m}(t_i)]w(t_i) \\
&= \begin{cases} \ge 0\,, k=1,...,n \\ = 0\,, \hat{\alpha}_{k+1} = \hat{m}(t_{k+1})-\hat{m}(t_k) > 0 \end{cases} \quad \Longleftrightarrow \\
\sum_{i=1}^{k} y(t_i)w(t_i) &= \begin{cases} \ge \sum_{i=1}^{k}\hat{m}(t_i)w(t_i)\,, k=1,..,n \\ = \sum_{i=1}^{k}\hat{m}(t_i)w(t_i)\,, \hat{\alpha}_{k+1} = \hat{m}(t_{k+1})-\hat{m}(t_k) > 0 \end{cases} .
\end{aligned}$$

Define the cumulative sums

$$\begin{aligned}\hat{M}_k &= \sum_{i=1}^{k} \hat{m}(t_i)w(t_i)\\ \hat{Y}_k &= \sum_{i=1}^{k} y(t_i)w(t_i)\end{aligned}$$

which gives

$$\begin{aligned}\hat{M}_k &\leq \hat{Y}_k\,, k = 1, ..., n\\ \hat{M}_k &= \hat{Y}_k\,, \hat{\alpha}_{k+1} = \hat{m}(t_{k+1}) - \hat{m}(t_k) > 0\end{aligned}$$

where

$$\hat{\alpha}_{k+1} = \frac{\hat{M}_{k+1} - \hat{M}_k}{w(t_{k+1})} - \frac{\hat{M}_k - \hat{M}_{k-1}}{w(t_k)}.$$

Thus we have found a characterisation which gives a direct correspondence between the observed values and the isotonic regressor through the cumulative sums $\hat{Y}_k$ and $\hat{M}_k$ respectively. Introducing the cumulative sum of the weight function as $\hat{W}_k = \sum_{i=1}^{k} w(t_i)$, by plotting $\hat{p}_k = (\hat{W}_k, \hat{M}_k)$ and $p_k = (\hat{W}_k, \hat{Y}_k)$ we obtain the following three properties:
(i)
Since

$$\frac{\hat{M}_{k+1} - \hat{M}_k}{\hat{W}_{k+1} - \hat{W}_k} = \frac{\hat{m}(t_{k+1})w(t_{k+1})}{w(t_{k+1})} = \hat{m}(t_{k+1})$$

which corresponds to the slope of the function that is obtained by connecting the points $\hat{p}_k$ by straight lines, we see that this function is convex since its derivative is increasing due to $\hat{m}$ being an isotonic function. We also conclude that $\hat{m}$ corresponds to the left derivative of the function described by the cumulative sum $\hat{M}_k$.
(ii)
Obviously we have that $\hat{p}_k \leq p_k$ from the characterisation above, $\hat{M}_k \leq \hat{Y}_k$, which implies that $\hat{p}_k$ is a minorant to p_k.
(iii)
For the points $t_i \in T$ such that $\hat{m}$ is strictly increasing, that is, $\hat{m}(t_{i+1}) >$

$\hat{m}(t_i)$, we obtained $\hat{M}_k = \hat{Y}_k$. In other words, if $\hat{m}$ is strictly increasing this is equivalent to $\hat{M}$ being strictly convex which implied that $\hat{M} \equiv \hat{Y}$.

(i) and **(ii)** imply that $\hat{p}_k$ is a convex minorant to p_k. Since $\hat{p}_k = p_k$ for the points where $\hat{m}$ is strictly increasing, we cannot obtain a greater $\hat{p}_k$ which satisfies **(i)** and **(ii)**. Thus, we conclude that $\hat{p}_k$ is the greatest convex minorant of p_k.

We summarise our results of the isotonic regression problem in a theorem.

Theorem 5:
Assume T is a finite set of real numbers with the usual inequality $\leq$ on $\mathbb{R}$, $\mathcal{F}$ is the set of nondecreasing functions on T and y, w arbitrary functions defined on T. Then the isotonic regressor

$$\hat{m} \quad = \quad \underset{z \in \mathcal{F}}{\operatorname{argmin}} \sum_{t_i \in T} [y(t_i) - z(t_i)]^2 w(t_i) \tag{7}$$

is given by the left derivative of the greatest convex minorant of the cumulative sum diagram $(\hat{W}_k, \hat{Y}_k)$.

2.1 The pool adjacent violator algorithm (PAVA)

By (7) the solution to the isotonic regression was obtained by taking the left hand slope of the greatest convex minorant to the cumulative sum diagram $p_k = (W_k, Y_k)$. Thus, in order to get the isotonic regressor we plot p_k and draw straight lines between the points. If the obtained function is convex, the greatest convex minorant $\hat{M}_k$ simply becomes Y_k which would correspond to the case where the observations y_i are increasing and obviously the solution to the isotonic regression problem becomes $\hat{m}(t_i) = y(t_i)$. Otherwise, there exists a point p_k for some $k = 2, ..., n$ which violates the convexity and the straight lines between p_{k-1}, p_k and p_k, p_{k+1} constitute a concave function. The greatest function which is convex and has starting point p_{k-1} and end point p_{k+1} is the straight line between p_{k-1} and p_{k+1}, which is then a minorant to the cumulative sum diagram from p_{k-1}, p_{k+1}. Every time a violator is encountered, this process is repeated until a convex function has been obtained. Since we obtained the greatest minorant by replacing the cumulative sum diagram with the straight lines, we have in fact obtained the greatest convex minorant to the cumulative sum diagram, $p_k = (W_k, Y_k)$. The solution to the isotonic regression problem is now obtained by taking the left hand slope of this function.

However, the left hand slope of the cumulative sum diagram at the point $p_k = (W_k, Y_k)$ is simply for $k = 2, ..., n$

$$\begin{aligned}\frac{Y_k - Y_{k-1}}{W_k - W_{k-1}} &= \frac{\sum_{i=1}^{k} w_i y_i - \sum_{i=1}^{k-1} w_i y_i}{\sum_{i=1}^{k} w_i - \sum_{i=1}^{k-1} w_i} \\ &= \frac{w_k y_k}{w_k} = y_k.\end{aligned}$$

Furthermore, if we had a violator at the point p_k we simply replaced the cumulative sum diagram with the straight line from p_{k-1} to p_{k+1}. This latter straight line has slope

$$\frac{Y_{k+1} - Y_{k-1}}{W_{k+1} - W_{k-1}} = \frac{\sum_{i=k}^{k+1} w_i y_i}{\sum_{i=k}^{k+1} w_i}.$$

This is in fact an average and for notational convenience we define for $t_i \leq t_j \in T$

$$\mathrm{Av}(t_i, ..., t_j) = \frac{\sum_{k=i}^{j} w_k y_k}{\sum_{k=i}^{j} w_k}.$$

We note that $\mathrm{Av}(t_i, t_i) = y_i$. Thus the solution to the isotonic regression problem, $\hat{m}$, can be obtained by the PAVA (Pool Adjacent Violator Algorithm) in the following way:

(i)

If $y(t_1) \leq y(t_2) \leq ... \leq y(t_n)$, the isotonic regressor $\hat{m}(t_i) = y(t_i), i = 1, .., n$, which corresponds to $\mathrm{Av}(t_1, t_1) \leq \mathrm{Av}(t_2, t_2) \leq ... \mathrm{Av}(t_n, t_n)$.

(ii)

If not, there exists at least one $k = 2, ..., n$ such that $\mathrm{Av}(t_k, t_k) > \mathrm{Av}(t_{k+1}, t_{k+1})$. As described above, we replace the violator of the cumulative sum diagram with the straight line between the adjacent points which yielded a slope between p_{k-1}, p_k and p_k, p_{k+1} which was equal to $\frac{\sum_{i=k}^{k+1} w_i y_i}{\sum_{i=k}^{k+1} w_i} = \mathrm{Av}(t_k, t_{k+1})$. Thus we replace $\hat{m}(t_k)$ and $\hat{m}(t_{k+1})$ with this slope, or equivalently $\mathrm{Av}(t_k, t_k)$ and $\mathrm{Av}(t_{k+1}, t_{k+1})$ are replaced by $\mathrm{Av}(t_k, t_{k+1})$. Hence, t_k and t_{k+1} belong to the same block, b_k, with equal slopes. If $\mathrm{Av}(t_1, t_1) \leq \mathrm{Av}(t_2, t_2) \leq ... \leq \mathrm{Av}(t_{k-1}, t_{k-1}) \leq \mathrm{Av}(t_k, t_{k+1}) \leq ... \leq \mathrm{Av}(t_n, t_n)$, we are done.

(iii)

Otherwise, repeat this process until there are no more violators and the slope in block $i \leq$ slope in block j for $i < j$. Then $\hat{m}(t_i)$ is given by the slope in the block which contains t_i.

3 Existence of solution of stochastic differential equation

This section will also be given for completeness and has been treated by various authors before. We will follow the characterisation given in [8]. As mentioned above, we are interested in stochastic differential equations of the form $dX(t) = f(t, X(t))dt + \sigma dW(t)$ where f is a function, on which we will impose some restrictions/conditions in subsequent chapters, and $W(t)$ is a standard Brownian motion. Before we give a strict definition and meaning to the stochastic differential equation of the form above, we recall some definitions such as σ-algebras, adapted processes et cetera.

Definition σ-algebra:
If Ω is a given set of elements, a σ-algebra $\mathcal{F}$ on Ω is a family $\mathcal{F}$ of subsets of Ω such that
(i): $\emptyset \in \mathcal{F}$
(ii): $A \in \mathcal{F} \Longrightarrow A^C \in \mathcal{F}\,, A^C = \Omega \setminus A$
(iii): $A_1, A_2, ... \in \mathcal{F} \Longrightarrow A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
$(\Omega, \mathcal{F})$ is called a measureable space. $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space if $\mathbb{P}$ is a measure such that $\mathbb{P} : \mathcal{F} \longrightarrow [0, 1]$ with the following properties:
(i): $\mathbb{P}(\emptyset) = 0 \quad , \mathbb{P}(\Omega) = 1$
(ii): If $A_1, A_2, ... \in \mathcal{F}$ are disjoint sets, that is $A_i \cap A_j = \emptyset\,, i \neq j$, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Definition (measureability):
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \longrightarrow \mathbb{R}^n$. Then X is said to be $\mathcal{F}$-measureable if

$$X^{-1}(U) \quad = \quad \{w \in \Omega : X(w) \in U\} \in \mathcal{F}$$

for any open set $U \in \mathbb{R}^n$.

Definition (Adapted process):
Let $\{\mathcal{G}_t\}_{t \geq 0}$ be an increasing family of σ-algebras of subsets of Ω, $\mathcal{G}_t \subset \mathcal{G}_s\,, t \leq s$. Then the process $g(t, w) : [0, \infty) \times \Omega \longrightarrow \mathbb{R}^n$ is said to be adapted with respect to $\mathcal{G}_t$ if $\forall t \geq 0$, the function $w \longrightarrow g(t, w)$ is $\mathcal{G}_t$-measureable for each fixed t.

We are interested in stochastic differential equations which in general take the form

$$dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t)$$

where $W(t)$ is a standard Brownian motion, that is,

$$\begin{aligned}
&(i) W(0) = 0\,, W(t) \sim N(0, t)\,, t \geq 0 \\
&(ii) W(t) - W(s) \,\text{and}\, W(s) \,\text{are independent for}\, \forall\, 0 \leq s \leq t\,, \\
&W(t) - W(s) \sim N(0, t - s).
\end{aligned}$$

The stochastic differential equation is simply the differential form for

$$X(t) = X(0) + \int_0^t \mu(s, X(s))ds + \int_0^t \sigma(s, X(s))dW(s)$$

where the last integral is a so called Ito integral, which is defined as follows, see [8].

Definition (Ito integral):
Let $\nu = \nu(S, T)$ be the class of functions $f(t, \omega) : [0, \infty) \times \Omega \longrightarrow \mathbb{R}$ such that
(i) $(t, \omega) \longrightarrow f(t, \omega)$ is $\mathcal{B}, \mathcal{F}$-measureable, where $\mathcal{B}$ denotes the Borel σ-algebra on $[0, \infty)$.
(ii) $f(t, \omega)$ is $\mathcal{F}_t$-adapted, where $\mathcal{F}_t$ is the σ-algebra generated by the random variables $W(s)\,, s \leq t$. ($\mathcal{F}_t$ is the smallest σ-algebra containing all the sets of the form $\{\omega : W(t_1) \in F_1, ..., W(t_k) \in F_k\}$ for $k = 1, 2, ...,$ $t_j \leq t$ and Borel sets F_j.
(iii) $E[\int_S^T f(t, \omega)^2 dt] < \infty$.
Then the Ito integral is to be interpreted as, for $f \in \nu(S, T)$,

$$\int_S^T f(t, \omega)dW(t) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} f(t_i, \omega)[W(t_{i+1}) - W(t_i)]$$

for $S = t_0 \leq t_1 \leq ... \leq t_n = T$. In other words, the function f is evaluated at the left point of the interval $[t_i, t_{i+1})$ and is multiplied with the $forward$ increment of the Brownian motion W.

Two important properties of the Ito integral as well as a theorem are listed below. The proofs are given in [8].
(i) $E[\int_S^T f(t, \omega)dW(t)] = 0$
(ii) $E[(\int_S^T f(t, \omega)dW(t))^2] = E(\int_S^T f(t, \omega)^2 dt)$, Ito isometry.

Theorem 6 (Ito formula):
Let W_t be a Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$ and let X_t be the Ito process

$$X_t = X_0 + \int_0^t \mu(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s$$

or in differential form $dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t$. Then if $g(t, x) \in C^2([0, \infty) \times \mathbb{R})$, $Y_t = g(t, X_t)$ is an Ito process with differential form given by

$$dY_t = \frac{\partial g}{\partial t}(t, X_t)dt + \frac{\partial g}{\partial x}(t, X_t)dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)(dX_t)^2$$

where the infinitesimal operators are to be calculated according to $dt \cdot dt = dW_t \cdot dt = dt \cdot dW_t = 0$ and $dW_t \cdot dW_t = t$.

Theorem 7 (Existence and uniqueness theorem for stochastic differential equations):
Let $T > 0$ and $\mu(\cdot, \cdot) : [0, T] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $\sigma(\cdot, \cdot) : [0, T] \times \mathbb{R}^n \longrightarrow \mathbb{R}^{n+m}$ be measurable functions satisfying

$$|\mu(t, x)| + |\sigma(t, x)| \leq C(1 + |x|)\,,\; x \in \mathbb{R}^n\,, t \in [0, T]$$

for some constant C and $|\sigma|^2 = \sum |\sigma_{i,j}|^2$ and such that

$$|\mu(t, x) - \mu(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D|x - y|\,,\; x, y \in \mathbb{R}^n\,, t \in [0, T]$$

for some constant D. Let Z be a random variable which is independent of the σ-algebra $\mathcal{F}_\infty$ generated by $W_s\,, s \geq 0$ and such that $E(|Z|^2) < \infty$. Then the stochastic differential equation

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t\,,\; 0 \leq t \leq T\,, X_0 = Z$$

has a unique t-continuous solution X_t, each component of which belongs to $\nu[0, T]$.

Proof:
The interested reader is referred to [8] for the proof.

4 Isotonic regression in a random design point setting

The aim of this paper is to estimate the function f in the model $dX(t) = f(X(t))dt + \sigma dW(t)$ where W is a Wiener process and under the restriction that f is a nondecreasing function. We recall that the stochastic differential equation above is to be interpreted in the Ito sense and thus $dW(t)$ is a forward increment. We start by considering the discrete case, that is a practical situation where the stochastic process $X(t)$ has been observed at the discrete time points t_i , $i = 1, ..., n \in T$ for some set $T \in \mathbb{R}$. Although the characterisation of isotonic regression given above allowed for arbitrary t_i, we will restrict our attention to the case with equidistant time points and without loss of generality assume that $T = [0, 1]$. Thus, we consider $t_i = i/n$, $i = 0, ..., n-1$. In the discrete setting we will thus have the model as

$$\Delta X(t_i) \quad = \quad f(X(t_i))\Delta t_i + \sigma \Delta W(t_i)$$

with $\Delta X(t_i) = X(t_{i+1}) - X(t_i)$, $\Delta t_i = t_{i+1} - t_i = 1/n$ and $\Delta W(t_i) = W(t_{i+1}) - W(t_i)$. (This is merely an approximation of the stochastic differential equation and although one might be interested in better approximations and the order of the error term of the difference of the approximation and the true value, we refrain from investigating this further.)

Under the constraint that f is increasing, given observations $\Delta X(t_i)$ we encounter a situation similar to

$$y(x_i) \quad = \quad m(x_i) + \epsilon_i$$

where m is an increasing function, ϵ_i are independent error terms and the design points $x_1 \leq x_2 \leq ... \leq x_n$ satisfy the simple order given by the standard inequality on $\mathbb{R}$. Since the Wiener process has independent, stationary increments for disjoint intervals, together with the assumption that $t_i = i/n$, $\Delta W(t_i)$ has the same distribution $\forall\, i = 0, 1, ..., n-1$. However, in our case the design points are $X(t_i)$ which obviously are stochastic. To obtain the isotonic regression model, we consider the following steps.

(i)

Take the order statistic of the observation points $X(t_i)$ and denote it by $X_{(i)}$

such that

$$\begin{aligned} X_{(1)} &\leq X_{(2)} \leq ... \leq X_{(n)} \\ X_{(k)} &= k\text{th smallest value of} \{X(t_i), i = 1, ..., n\}. \end{aligned}$$

(ii)
Rearrange the corresponding responses $\Delta X(t_i)$ accordingly to get observations of the form $(X_{(i)}, Y_{(i)})$ where $Y_{(i)}$ is given by

$$Y_{(i)} = f(X_{(i)})\tilde{\Delta} t_i + \Delta\widetilde{W}(t_i)$$

where

$$\begin{cases} \tilde{\Delta} t_i = \Delta t_j \\ \Delta\widetilde{W}(t_i) = \Delta W(t_j) \end{cases}$$

if $X_{(i)} = X(t_j)$.

Taking the order statistic of the observed process corresponds to a permutation

$$\chi : \{1, ..., n\} \to \{\chi(1), ..., \chi(n)\} \in \{1, ..., n\}$$

such that

$$\begin{aligned} X_{\chi(1)} &\leq ... \leq X_{\chi(n)} \\ \Delta\widetilde{W}(t_i) &= \Delta W(t_{\chi(i)}). \end{aligned}$$

The observation points X_i obviously depend on each other and on $W(t_i)$ and thus so will the permutation, χ. If we denote by P the permutation matrix corresponding to the order induced by χ we obtain $\Delta\widetilde{W}(t_i), t_i = i/n$ as

$$\Delta\widetilde{W}(t_i) = PW.$$

Now if the permutation matrix P were independent of W, we would after the permutation still get errors that are independent, Gaussian random variables with the same variance as before due to the following. Given the assumption of independence, $\Delta\widetilde{W}$ is a linear combination of Gaussian random variables and hence Gaussian with expectation and variance as follows.

$$\Delta\widetilde{W} = PW \Longrightarrow \begin{cases} E[\widetilde{W}] = PE(W) = 0 \\ Cov(\Delta\widetilde{W}) = PCov(W)P^T = P\Delta t P^T = \Delta t PP^T = \Delta t I. \end{cases}$$

which would imply that when considering the regression model after the rearranging of the terms, we would not have affected the distribution of the error terms. However, since P and W in general are dependent, $\Delta\widetilde{W}(t_{\chi(i)})$ are neither independent nor Gaussian in general. A guess would be that the rearranged errors would be weakly dependent but due to time limitations, the distribution of the error terms in general settings was not investigated. The fact that the error terms are not Gaussian does not affect the regression problem, however, we must be careful of the assumptions we make regarding the dependence of the error terms as we later will try to derive limit distribution results for the estimator.

Hence we have obtained observations of the form $Y_{(i)} = f(X_{(i)})\Delta t_i + \Delta\widetilde{W}(t_i)$ which is an isotonic regression model with random design points. As we now have an isotonic regression model, we know how to find

$$\hat{m} \;=\; \underset{z\in\mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} [y_{(i)} - z(x_{(i)})]^2 w_i$$

by taking the greatest convex minorant of the cumulative sum diagram, $\sum_{i=1}^{n} y(x_{(i)})w_i$.

If f and σ satisfy the assumptions of Theorem 7 we know that the solution $X(t)$ is unique and continuous as well as $E[(\int_0^T f(t)dW(t))^2] = \int_0^T E[f^2(s)]ds < \infty$ so that in particular, X is bounded on $T = [0, 1]$. Thus the stochastic design point interval we obtain from X is well defined for the isotonic regression model. To summarise, if we want to do isotonic regression on the model

$$\Delta X(t_i) \;=\; f(X(t_i))\Delta t_i + \sigma\Delta W(t_i)$$

we take the order statistic of X and the corresponding values of Y and perform isotonic regression on f on the interval

$$[X_{(1)}, X_{(n)}] = [\min X(t), \max X(t)]\,, t = \frac{i}{n}\,, i = 1, ..., n.$$

5 Isotonic regression in the continuous case

In this section we will generalise the results in the discrete case to the continuous case. This will be done by reviewing the work in [1] and slightly clarifying this. It was shown above that in the discrete case, the solution to the isotonic regression problem

$$\hat{m} \;=\; \operatorname*{argmin}_{z \in \mathcal{F}} \sum_{i=1}^{n} (y_i - z(t_i))^2 w_i$$

with $\mathcal{F}$ the set of nondecreasing functions, was given by the left hand slope of the greatest convex minorant of the cumulative sum, $\sum_{i=1}^{n} y_i w_i$. We will show that this also holds in the continuous case.

Thus consider the case where we have observed $g \in L^2[a,b]$, that is

$$\int_a^b g^2(u)du < \infty$$

so that g is square integrable on $[a,b]$. Analogously to the discrete case, we wish to find the solution to

$$\hat{m} \;=\; \operatorname*{argmin}_{z \in \mathcal{F}} \int_a^b (g(u) - z(u))^2 du$$

where $\mathcal{F} = \{z : [a,b] \longrightarrow \mathbb{R}\,, z \,\text{nondecreasing}\}$. We make the following definitions in the continuous case, see [1].

Definition (greatest convex minorant):
The greatest convex minorant, T, of a function $y : [a,b] \longrightarrow \mathbb{R}$ is defined as

$$T(y) \;=\; \sup\{z : [a,b] \longrightarrow \mathbb{R}\,, z \leq y\,, z\,\text{convex}\}.$$

Its derivative is defined as

$$T(y)'(t) \;=\; \min_{u \leq t} \max_{v \geq t} \frac{y(v) - y(u)}{v - u}.$$

It is clear that T is continuous since it is a convex function and by definition is also satisfies $T(y)(t) \leq y(t)$. Moreover, it is also obvious that the end points of T coincide with the end points of y, that is, $T(y)(a) = y(a)$ and $T(y)(b) = y(b)$. Furthermore, T is convex so that it has a left and a right derivative. It can be shown that the definition of $T(y)'$ above coincides with the left derivative of $T(y)$ and in addition,

$$T(y)(t) \;=\; \int_0^t T(y)'(s)ds.$$

For a proof of the statement, we refer to [1].

We will show that the solution to

$$\hat{m} \;=\; \underset{z \in \mathcal{F}}{\operatorname{argmin}} \int_a^b (g(u) - z(u))^2 du$$

is obtained by taking the integral of g

$$\bar{g}(t) \;=\; \int_a^t g(s) ds$$

and subsequently taking the left derivative of the greatest convex minorant over $[a, b]$ of $\bar{g}$, that is

$$\hat{m}(t) \;=\; T(\bar{g})'(t).$$

In order to show this, we first prove a theorem regarding the support of $dT(y)'$, which is given in [1], where $dT(y)'$ is the differential of the derivative of the greatest convex minorant of y on $[a, b]$.

Theorem 8 (support of $dT(y)'$)
Let y be a continuous function defined on $[a, b] \in \mathbb{R}$ and let $T(y)$ be the greatest convex minorant of y. Then $\operatorname{supp}\{dT(y)'\} \subset \{T(y) = y\}$.

Proof:
By definition, $T(y)$ is continuous and satisfies $T(y)(a) = y(a)$ and $T(y)(b) = y(b)$. Since y is also continuous, the set $\{x : T(y)(x) < y(x)\,, x \in [a, b]\}$ is open. Since every open set in $\mathbb{R}$ is a union of open intervals in $\mathbb{R}$, the set $\{x : T(y)(x) < y(x)\,, x \in [a, b]\}$ is a union of open intervals. On such an interval we have $T(y) < y$ which implies that T is linear on such an interval for the following reason. If T were strictly convex, that is not linear, on such an interval we could always find T^*, T^* strictly convex, such that $T^* > T$ on that interval. This is a contradiction which means that T must be linear in order to be the greatest convex minorant. But if $T(y)$ is linear, this implies that its derivative $T(y)'$ is a constant on such an interval which in turn implies that $dT(y)'$ is zero on every open interval where $T(y) < y$, that is $\{T(y) < y\} \subset \{dT(y)' = 0\}$. Taking complements yields $\{dT(y)' > 0\} \;\subset \{T(y) = y\}$ since the derivative of a convex function is always nondecreasing ($dT(y)' \geq 0$) and T is a minorant of y ($T(y) \leq y$). This concludes the proof. $\square$

(For a geometrical interpretation, it is often helpful to think of the greatest convex minorant of a function as stretching a string from below the graph of the function. For intervals where the greatest convex minorant is strictly less than the function, making it linear instead of strictly convex corresponds to tightening the string instead of it hanging loosely.)

As we want to minimise $\int_a^b (g(s) - f(s))^2 ds$ with $f \in \mathcal{F}$, the set of non-decreasing functions on $[a, b]$, define

$$G(f) \quad = \quad \int_a^b (f(s) - g(s))^2 ds$$

so that we wish to find $\operatorname{argmin}_{f \in \mathcal{F}} G(f)$. In other words we minimise the L^2-norm, $||f - g||_2$ for $f \in \mathcal{F}$. For this reason we consider the Gateaux derivative at the point $f \in \mathcal{F}$ in the direction h such that for small t, $f + th \in \mathcal{F}$, defined as

$$G_f(h) \quad = \quad \lim_{t \longrightarrow 0} \frac{G(f + th) - G(f)}{t}.$$

Now we show that $\hat{f} = T(\bar{g})'$ indeed minimises $G(f)\,, f \in \mathcal{F}$. The proof is an extended version or slighty more clarified than the one given in [1].

Theorem 9 (isotonic regressor - continuous case):
Let $g \in L^2[a, b]$ be a continuous function and define $\bar{g}(t) = \int_a^t g(s) ds$. Then

$$\hat{f} \quad = \quad \underset{f \in \mathcal{F}}{\operatorname{argmin}} \int_a^b (f(s) - g(s))^2$$

is given by $T(\bar{g})'(t)$ where T denotes the greatest convex minorant with its derivative defined as above.

Proof:
Let $\hat{f} = T(\bar{g})'$ and consider the Gateaux derivative at the point $\hat{f}$ in the

direction h:

$$\begin{aligned}
G_{\hat{f}}(h) &= \lim_{t\longrightarrow 0} \frac{G(\hat{f}+th)-G(\hat{f})}{t} \\
&= \lim_{t\longrightarrow 0} \frac{\int_a^b [th(s)+(\hat{f}(s)-g(s))]^2 ds - \int_a^b (\hat{f}(s)-g(s))^2 ds}{t} \\
&= \lim_{t\longrightarrow 0} \frac{\int_a^b [t^2h(s)^2+2th(s)(\hat{f}(s)-g(s))]ds}{t} \\
&= \lim_{t\longrightarrow 0} t\int_a^b h^2(s)ds + 2\int_a^b h(s)[\hat{f}(s)-g(s)]ds \\
&= 2\int_a^b [\hat{f}(s)-g(s)]h(s)ds.
\end{aligned}$$

Now using that $\hat{f} = T(\bar{g})'$ and that $\bar{g}(t) = \int_a^t g(s)ds$ we obtain by partial integration

$$\begin{aligned}
\frac{1}{2}G_{\hat{f}}(h) &= \int_a^b [\hat{f}(s)-g(s)]h(s)ds \\
&= [(T(\bar{g})(x)-\bar{g}(x))h(x)]_{x=a}^b - \int_a^b (T(\bar{g})(s)-\bar{g}(s))\frac{dh(s)}{ds}ds \\
&= [0\cdot h(b) - 0\cdot h(a)] - \int_a^b (T(\bar{g})(s)-\bar{g}(s))dh(s) \\
&= -\int_a^b [T(\bar{g})(s)-\bar{g}(s)]dh(s)
\end{aligned}$$

since the greatest convex minorant of y, $T(y)$, coincides with the function at the end points. Now if we let $h = \hat{f} = T(\bar{g})'$ we get

$$G_{\hat{f}}(\hat{f}) = -2\int_a^b [T(\bar{g})(s)-\bar{g}(s)]dT(\bar{g})'.$$

Since T is the greatest convex minorant, we recall that $T(\bar{g})(s) - \bar{g}(s) \leq 0$ and also that its derivative is nondecreasing, which is equivalent to $dT(\cdot)'$ being nonnegative. Additionally it was proved above that the support of $dT(\cdot)'$ was included in $T(\bar{g}) = \bar{g}$. Thus we do not get any contribution from the integral above since if $T(\bar{g}) < \bar{g}$ we get $dT(\cdot) = 0$. Hence $G_{\hat{f}}(\hat{f}) = 0$. For an arbitrary $f \in \mathcal{F}$ we have

$$G_{\hat{f}}(f) = -2\int_a^b [T(\bar{g})(s)-\bar{g}(s)]dT(f)'$$

which by the properties of T is nonnegative, which implies that

$$G_{\hat{f}}(f-\hat{f}) \;=\; -2\int_a^b [T(\bar{g})(s)-\bar{g}(s)]dT(f)' \geq 0.$$

Define for $t \in [0,1]$

$$u(t) \;=\; G(\hat{f}+t(f-\hat{f}))$$

which has derivative

$$\begin{aligned}
u'(t) &= \lim_{h\longrightarrow 0}\frac{G(\hat{f}+(t+h)(f-\hat{f}))-G(\hat{f}+t(f-\hat{f}))}{h} \\
&= \lim_{h\longrightarrow 0}\frac{\int_a^b[\hat{f}+t(f-\hat{f})-g+h(f-\hat{f})]^2ds}{h} - \\
&\quad \frac{\int_a^b[\hat{f}+t(f-\hat{f})-g]^2ds}{h} \\
&= \lim_{h\longrightarrow 0} 2\int_a^b [f(s)-\hat{f}(s)][\hat{f}(s)+t(f(s)-\hat{f}(s))-g(s)]ds + \\
&\quad h\int_a^b [f(s)-\hat{f}(s)]^2ds \\
&= 2\int_a^b [f(s)-\hat{f}(s)][\hat{f}(s)+t(f(s)-\hat{f}(s))-g(s)]ds.
\end{aligned}$$

Evaluated at 0 we obtain

$$\begin{aligned}
u'(0) &= 2\int_a^b (f(s)-\hat{f}(s))(\hat{f}(s)-g(s))ds \\
&= G_{\hat{f}}(f-\hat{f}) \geq 0.
\end{aligned}$$

But u is the composition of a convex function G (since G is an integral of the square function it is a linear combination of convex functions and thus convex) and a linear function, which in turn is convex. Thus u is convex. As u is convex, it has a nondecreasing derivative and since $u'(0) \geq 0$ we get that $u'(t) \geq 0\,\forall t \in [0,1]$ and thus $u(t) \geq u(0)\,\forall t \in [0,1]$. In particular, $u(1) \geq u(0)$ which is equivalent to $G(\hat{f}) \leq G(f)$ for any $f \in \mathcal{F}$. To summarise,

$$\begin{aligned}
\hat{f} &= T(\bar{g})' \\
&= \operatorname*{argmin}_{f\in\mathcal{F}} \int_a^b (f(s)-g(s))^2ds. \qquad \square
\end{aligned}$$

6 Pointwise limit distributions of greatest convex minorant and its derivative

We now know how to find the isotonic regressor in both the discrete and continuous case by taking the greatest convex minorant of the cumulative sum in the discrete case and the integral in the continuous case. The aim is to find the limit distribution of our estimator as the number of observations tend to infinity. We will start by considering the case of an equidistant design point and subsequently try to generalise this to the situations where the design point setup is given by independent random variables as well as the stochastic differential equation setup described earlier.

Thus consider the isotonic regression problem in the equidistant design point, that is, we have observations

$$y(t_i) \quad = \quad m(t_i) + \epsilon_i$$

where $t_i = i/n$ and ϵ_i are independent random variables. This means we wish to find

$$\hat{m} \quad = \quad \operatorname*{argmin}_{m \in \mathcal{F}} \sum_{i=1}^{n} (y(t_i) - m(t_i))^2$$

which corresponds to finding

$$\hat{m} \quad = \quad \operatorname*{argmin}_{m \in \mathcal{F}} \int_0^1 (y(s) - m(s))^2 ds$$

which by the previous section was given by

$$\hat{m}(t) \quad = \quad T(\bar{y})'(t)$$

where $\bar{y}(t) = \int_0^t y(s)ds$. Many authors have treated the subject of the asympototic distribution of the greatest convex minorant (or least concave majorant) of stochastic processes and in this paper we will follow the scheme given in [2], where the observed stochastic process is split into a deterministic part and a stochastic one.

Anevski and Hössjer deal with sequences of stochastic processes, $\{x_n\}_{n=1}^{\infty}$, on the space $D(J)$ where J is an interval in $\mathbb{R}$ allowed to be infinite and D is the space of right-continuous functions with left hand limits. These stochastic processes can be split into a deterministic part and a stochastic part where the latter also lies in $D(J)$. Thus

$$x_n(t) \;=\; x_{b,n}(t) + v_n(t)$$

where $x_{b,n}$ is deterministic and v_n is a stochastic process defined on $D(J)$. This paper will not deal with sequences of stochastic processes and due to this restriction we instead write

$$x(t) \;=\; x_b(t) + v(t).$$

Let d_n be a sequence that tends to 0 and define the rescaled version of v at the point t_0 as

$$\tilde{v}_n(s;t_0) = d_n^{-p}(v(t_0 + sd_n) - v(t_0))$$

for $s \in d_n^{-1}(J - t_0)$ and for $1 < p < \infty$ a fixed constant. It can be shown that if certain assumptions are fulfilled,

$$d_n^{-p}(T_J(x)(t_0) - x(t_0)) \xrightarrow{\mathcal{L}} T(|s|^p + \tilde{v}(s))(0)$$

as d_n tends to zero and where T denotes the greatest convex minorant and

$$\tilde{v}(s) \;=\; \lim_{d_n \to 0} \tilde{v}_n(s).$$

Moreover, for the derivative of T it holds that

$$d_n^{-p+1}(T_J(x)'(t) - x_b'(t)) \xrightarrow{\mathcal{L}} T(|s|^p + \tilde{v}(s))'(0).$$

The various assumptions that are required to be satisfied first regard the possibility to rescale the stochastic and determinstistic parts of the process which satisfy certain conditions. Further, these rescaled processes must satisfy some growth conditions in relation to each other. Thus we start with the first assumptions regarding the rescaling of the processes.

Assumption 1:
The process

$$\tilde{v}_n(s;t_0) \;=\; d_n^{-p}(v(t_0 + sd_n) - v(t_0)) \xrightarrow{\mathcal{L}} \tilde{v}(s)$$

for some process $\tilde{v} \in D(J)$ as d_n tends to zero.

Assumption 2:
Define $g_n(s)$ at the point t_0 as

$$g_n(s; t_0) \quad = \quad d_n^{-p}(x_b(t_0 + sd_n) - l_n(s))$$

where

$$l_n(s) \quad = \quad x_b(t_0) + x_b'(t_0)sd_n$$

that is, the first two terms in the Taylor expansion of x_b at the point t_0. Moreover, $\exists A > 0$ such that for every $c > 0$

$$\sup_{s \in [-c,c]} |g_n(s) - A|s|^p| \to 0$$

as $n \to \infty$ or equivalently $d_n \to 0$.

The function $x_b(t)$ must be a convex function since it is assumed that its derivative is a nondecreasing function. In most applications $x_b(t)$ also satisfies

$$x_b(t) \quad = \quad x_b(t_0) + x_b'(t_0)(t - t_0) + A|t - t_0|^p + o(|t - t_0|^p)$$

where, in particular, $A = \frac{1}{2}x_b''(t_0)$ if $p = 2$, which we will encounter in a later section. Thus, define the rescaled function as

$$y_n(s; t_0) \quad = \quad g_n(s; t_0) + \tilde{v}_n(s; t_0).$$

Assumption 3:
For every $\delta > 0$ there are finite $0 < \tau = \tau(\delta)$ and $0 < \kappa = \kappa(\delta)$ such that

$$\lim_{n\to\infty} \inf P[\inf_{|s|>\tau} (y_n(s) - \kappa|s| > 0)] \quad > \quad 1 - \delta.$$

Assumption 4:
Given $\epsilon, \delta, \tau > 0$,

$$\lim_{n\to\infty} \sup P[\inf_{\tau \le s \le c} \frac{y_n(s)}{s} - \inf_{\tau \le s} \frac{y_n(s)}{s} > \epsilon] < \delta$$

and

$$\lim_{n\to\infty} \sup P[\inf_{-c\leq s\leq -\tau} \frac{y_n(s)}{s} - \inf_{s\leq -\tau} \frac{y_n(s)}{s} < -\epsilon] < \delta$$

for large enough $c > 0$.

Assumption 3 and 4 say that, for large enough s, the rescaled process $y_n(s)$ lies above a constant times s and in fact outgrows the identity function as we are far enough from the origin. For the proofs of the following propositions and theorems, the interested reader is referred to [2]. It can be shown that if assumption 2 holds and that for every $\epsilon, \delta > 0$, there exists a finite $\tau = \tau(\epsilon, \delta)$ such that

$$\lim_{n\to\infty} \sup P[\sup_{|s|\geq\tau} |\frac{\tilde{v}_n(s)}{g_n(s)}| > \epsilon] < \delta\,,$$

then assumption 3 and 4 hold.

We need a final assumption before we can state the first theorem of the limit distribution for the greatest convex minorant.

Assumption 5:
For every $\epsilon, \delta > 0$ there exists a $\tau = \tau(\epsilon, \delta) > 0$ such that

$$P[\sup_{|s|\geq\tau} |\frac{\tilde{v}(s;t_0)}{A|s|^p}| > \epsilon] < \delta$$

We can state the first theorem.

Pointwise limit distribution for greatest convex minorant
Let t_0 be fixed and suppose assumptions 1-5 hold. Then

$$d_n^{-p}[T_J(x)(t_0) - x(t_0)] \xrightarrow{\mathcal{L}} T[A|s|^p + \tilde{v}(s;t_0)](0)$$

as $n \to \infty$ with A the positive constant given in assumption 2.

As for the limit distribution of the derivative of the we need one more assumption.

Assumption 6:
We recall the definition of $y_n(s) = g_n(s) + \tilde{v}_n(s)$ and $y(s) = A|s|^p + \tilde{v}(s)$. Then it holds that

$$T_c(y_n)'(0) \xrightarrow{\mathcal{L}} T_c(y)'(0) \tag{8}$$

as $n \to \infty$ for each $c > 0$, that is the greatest convex minorant converges on compact intervals.

Assumption 6, namely that the greatest convex minorant converges on compact intervals, holds basically whenever the process $\tilde{v}(s)$ in

$$y(s) \quad = \quad A|s|^p + \tilde{v}(s)$$

is a Brownian motion.

Furthermore, as a consequence of assumptions 3 and 4, it can be shown ([2]) that

$$\lim_{c\to\infty} \lim_{n\to\infty} |T_c(y_n)'(0) - T_{\mathbb{R}}(y_n)'(0)| \xrightarrow{P} 0 \tag{9}$$

and

$$\lim_{c\to\infty} |T_c(y)'(0) - T_{\mathbb{R}}(y)'(0)| \xrightarrow{P} 0. \tag{10}$$

Then from (8), (9) and (10), by Slutsky's theorem, one can show that

$$d_n^{-p+1}[T(x)'(t_0) - x_b'(t_0)] \xrightarrow{\mathcal{L}} T(A|s|^p + \tilde{v}(s;t_0))'(0)$$

as $n \to \infty$. We state this as a theorem.

Limit distribution for the derivative of the greatest convex minorant:

Suppose that assumptions 1-6 above hold and let t_0 be a fixed point. Then

$$d_n^{-p+1}[T(x)'(t_0) - x_b'(t_0)] \xrightarrow{\mathcal{L}} T(A|s|^p + \tilde{v}(s;t_0))'(0)$$

as $n \to \infty$.

7 Limit distribution for estimator of trend function in dX(t)=X(t)dt+σ dW(t)

Consider the concrete example of the stochastic differential equation given by

$$dX(t) = X(t)dt + \sigma dW(t).$$

We are interested in the situation when σ tends to zero. This is known as a filtering problem and has been studied by for example Ibragimov and Hasminski [5] as well as Nussbaum [7]. Prakasa Rao has studied kernel estimators of the trend function in stochastic differential equations driven by fractional Brownian motions as the volatility tends to zero [6]. Since we are interested in performing isotonic regression on the drift function, in this case being the identity function, we are treating the increments given by the stochastic differential equations as our observations. Thus we get observations of the form

$$y_i \quad = \quad m(t_i) + \xi_i$$

with

$$\begin{cases} m(t_i) = f(X(t_i))\Delta t_i = X(t_i)\Delta t_i \\ \xi_i = \sigma_n \epsilon_i \\ \sigma_n \to 0 \, , n \to \infty \\ \epsilon_i \sim N(0, 1/n) \, , \epsilon_i \text{ ind. of } \epsilon_j \, , i \neq j \end{cases}$$

In a practical situation we get observations on the unit interval $[0, 1]$ with an equidistant step size of length $1/n$. In other words we obtain $\Delta t_i = 1/n$ and $dW(t) \sim N(0, 1/n)$ is normally distributed with variance $1/n$. As we are interested in the estimation of f, we can multiply our observations by n since

$$\underset{z \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - z(t_i))^2 \quad = \quad \underset{z \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} [n(y_i - z(t_i))]^2$$

and thus this does not change the constrained least squares estimator.

By multiplying with n we get observations of the form

$$\tilde{y}_i \quad = \quad f(X(t_i)) + n\sigma_n \Delta W(t_i)$$

with $\Delta W(t_i) \sim N(0, 1/n)$. By properties of the Gaussian process, we know that $n\sigma_n \Delta W(t_i)$ is distributed as

$$n\sigma_n \Delta W(t_i) \quad \sim \quad N(0, n\sigma_n^2)$$

and hence, for the problem to have a non-trivial solution, σ_n has to decay at the rate of $n^{-1/2}$. With $\sigma_n = \sigma/\sqrt{n}$ we obtain observations of the form

$$\tilde{y}_i \quad = \quad f(X(t_i)) + \zeta_i \quad , \zeta_i \sim N(0, \sigma^2)$$

which we recognise as the usual isotonic regression model with a random design point.

Thus, consider the case of $dX(t) = X(t)dt + \sigma dW(t)$, with $\sigma_n = \sigma\sqrt{1/n}$. The problem is now to estimate the drift function under the assumption that this is increasing, which is done by ordering the observed stochastic process and then order the observations $\Delta X(t_i)$ in correspondance with the ordering of the observed stochastic process.

Before considering the limit distribution of the unknown function f in the observed process points, we make a Taylor expansion so that we can split the process parts completely into a deterministic part and a random one, as done in [2]. Thus in the process

$$dX(t) \quad = \quad f(X(t))dt + \sigma_n dW(t)$$

we expand the stochastic process around its expected value and since σ_n tends to zero, we can expect $X(t)$ to be close to its expected value.

$$\begin{aligned} f(X(t)) &= f(E(X(t)) + X(t) - E(X(t)) \\ &= f(E(X(t)) + f'(E(X(t))(X(t) - E(X(t))) + \\ &+ f''(E(X(t)))/2(X(t) - E(X(t)))^2 + O([X(t) - E(X(t)]^3). \end{aligned}$$

Denote $E(X(t)) = \mu(t)$ and let $X(t)$ be the integral of the observations/integrated process up to time t which gives

$$\begin{aligned} X(t) &= \int_{s=0}^{t} f(X(s))ds + \sigma W(t) \qquad (11) \\ &= \int_0^1 [f(\mu(s)) + f'(\mu(s))(X(s) - \mu(s)) + O((X(s) - \mu(s))^2)]ds + \sigma W(t). \end{aligned}$$

In the case of $dX(t) = X(t)dt + \sigma_n dW(t)$ it is possible to obtain a closed form solution which is obtained in the following way.

$$\begin{aligned} dX(t) &= X(t)dt + \sigma_n dW(t) \Longleftrightarrow \\ dX(t) - X(t)dt &= \sigma_n dW(t) \Longleftrightarrow \\ d(e^{-t}X(t)) &= e^{-t}\sigma_n dW(t) \Longleftrightarrow \\ e^{-t}X(t) &= x_0 + \sigma_n \int_0^t e^{-s}dW(s) \Longleftrightarrow \\ X(t) &= x_0 e^t + \sigma_n \int_0^t e^{t-s}dW(s). \end{aligned}$$

Now since $dW(s)$ are Gaussian and independent and the integral is a linear combination, we get that $X(t)$ is a linear combination of Gaussian random varibles and hence Gaussian itself. Using the properties of the Ito integral we obtain

$$\begin{aligned} E(X(t)) &= x_0 e^t + \sigma_n E(\int_0^t e^{t-s} dW(s)) \\ &= x_0 e^t \end{aligned}$$

since the expectation of an Ito integral is zero. By the Ito isometry we also obtain the variance

$$\begin{aligned} Var(X(t)) &= Var(x_0 e^t + \sigma_n \int_0^t e^{t-s} dW(s)) \\ &= \sigma_n^2 Var(\int_0^t e^{t-s} dW(s)) \\ &= \sigma_n^2 E((\int_0^t e^{t-s} dW(s))^2) \\ &= \sigma_n^2 \int_0^t e^{2(t-s)} ds \\ &= \frac{\sigma_n^2 (e^{2t} - 1)}{2}. \end{aligned}$$

Thus

$$X(t) \sim N(x_0 e^t, \frac{\sigma_n^2 (e^{2t} - 1)}{2})$$

which implies that

$$X(t) - \mu(t) = X(t) - x_0 e^t \sim N(0, \frac{\sigma_n^2 (e^{2t} - 1)}{2}).$$

We will now prove the following theorem about the limit distribution of the greatest convex minorant of the integral of the observations given by the stochastic differential equation given by

$$dX(t) = X(t) dt + \sigma_n dW(t).$$

Theorem:
Suppose observations of the form $dX(t) = X(t)dt + \sigma_n dW(t)$ have been obtained and that we wish to estimate the drift function under the constraint that this is nondecreasing. Then it holds that

$$n^{2/3}[T_{[0,1]}(X)(t_0) - X(t_0)] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + B(s)](0)$$

where $B(s)$ is a Brownian motion, $T_{[0,1]}(x)(t_0)$ is the greatest convex minorant on $[0,1]$ of the integrated process $X(t) = \int_0^t dX(s)$ evaluated in the point t_0. Moreover, we get

$$n^{1/3}[T_{[0,1]}(X)'(t_0) - x_b'(t_0)] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + B(s)]'(0)$$

that is,

$$n^{1/3}[T_{[0,1]}(X_{\text{int}})'(t_0) - f(E[X_{\text{true}}(t_0)])] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + B(s)]'(0).$$

Proof:
By considering (11) we get the integral of our observations in this case as

$$X(t) = \int_0^t f(\mu(s))ds + \int_0^t [f'(\mu(s))(X(s) - \mu(s)) + O((X(s) - \mu(s))^2)]ds + \sigma_n W(t).$$

Denote

$$\begin{cases} x_b(t) = \int_0^t f(\mu(s))ds \\ v(t) = \int_0^t [f'(\mu(s))(X(s) - \mu(s)) + O((X(s) - \mu(s))^2)]ds + \sigma_n W(t) \end{cases}$$

It is now possible to rescale these as done in Anevski and Hössjer with $X(t) = x_b(t) + v(t)$

$$\begin{aligned} v(t+s\delta) - v(t) &= \int_t^{t+s\delta} [f'(\mu(s)(X(t) - \mu(s)) + \frac{f''(\mu(s))}{2}(X(s) - \mu(s))^2 ds + \\ &\quad \int_t^{t+s\delta} O((X(s) - \mu(s))^3)ds + \sigma_n(W(t+s\delta) - W(t)) \\ &\sim s\delta[f'(\mu(t))(X(t) - \mu(t)) + \frac{f''(\mu(t))}{2}(X(t) - \mu(t))^2 + O((X(t) - \mu(t))^3)] + \\ &\quad \sigma_n \delta^{1/2} \tilde{W}(s) \end{aligned}$$

with $\tilde{W}(s)$ being a Wiener process as well. Since

$$\begin{aligned} X(t) - \mu(t) &\sim \sigma_n N(0, (e^{2t} - 1)/2) \\ &= \sigma_n Z \end{aligned}$$

this yields

$$\delta^{-2}(v(t + s\delta) - v(t)) \sim \delta^{-2}[\delta\sigma_n s f'(\mu(t))Z + O(\sigma_n^2)sZ^2 + \sigma_n \delta^{1/2}\tilde{W}(s)].$$

We see that if δ is chosen as $\sigma_n^{2/3}$ which implies that $\sigma_n = \delta^{3/2}$, we get

$$\begin{aligned} \delta^{-2}(v(t + s\delta) - v(t)) &\sim \delta^{-2}(\delta^{5/2} s f'(\mu(t))Z + sO(\delta^3) + \delta^2\tilde{W}(s)) \\ &= \delta^{1/2} s f'(\mu(t))Z + O(\delta)Z^2 s + \tilde{W}(s) \\ &= \tilde{v}_n(s). \end{aligned}$$

Thus

$$\begin{aligned} \tilde{v}(s) &= \lim_{n\to\infty} \tilde{v}_n(s) \\ &= \tilde{W}(s) \end{aligned}$$

as $\delta \to 0$. Since $\sigma_n = n^{-1/2}$ we get that δ tends to zero at the rate $\delta = \sigma_n^{2/3} = n^{-1/3}$.

Now consider with $\mu(s) = E(X(s)) = x_0 e^s$ and with $f(t) = t$ the identity function,

$$\begin{aligned} x_b(t) &= \int_0^t f(\mu(s)ds \\ &= \int_0^t f(E(X(s)))ds \\ &= \int_0^t x_0 e^s ds. \end{aligned}$$

This yields $x_b'(t) = f(\mu(t))$ and $x_b''(t) = f'(\mu(t))\frac{d}{dt}(\mu(t)) = 1 \cdot x_0 e^t$ which are well defined. Thus

$$x_b(t + s\delta) = x_b(t) + x_b'(t)s\delta + \frac{x_b''(t)}{2}s^2\delta^2 + O(\delta^3)$$

which implies that

$$\begin{aligned}\delta^{-2}(x_b(t+s\delta)-x_b(t)-x_b'(t)s\delta) \quad &= \quad \frac{x_b''(t)s^2}{2}+O(\delta)\\ &\longrightarrow_{\delta\to 0} \quad \frac{x_b''(t)s^2}{2}=\frac{x_0e^ts^2}{2}\end{aligned}$$

Thus assumption 1 and 2 are well defined with $\tilde{v}(s)\sim N(0,s)$ and $A=x_0e^t/2$. We now need to check the remaining assumptions to able to state the limit distribution.

In order to check assumptions 3 and 4, it suffices to check that for $\epsilon,\delta>0$ there exists $\tau(\epsilon,\delta)>0$ such that

$$\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}|\frac{\tilde{v}_n(s)}{g_n(s)}|>\epsilon] \quad < \quad \delta.$$

Now we have got with $\sigma_n=n^{-1/2}$

$$\begin{cases}\tilde{v}_n(s)=\sigma_n^{1/3}sf'(\mu(t))Z+O(\sigma_n^{2/3})Z^2s+\tilde{W}(s)\\ g_n(s)=\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})\end{cases}$$

which gives

$$\begin{aligned}&\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}|\frac{\tilde{v}_n(s)}{g_n(s)}>\epsilon]=\\ =\;&\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}|\frac{\sigma_n^{1/3}sf'(\mu(t))Z+O(\sigma_n^{2/3})Z^2s+\tilde{W}(s)}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}|>\epsilon]\\ \leq\;&\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}\frac{|\sigma_n^{1/3}sf'(\mu(t))Z|+|O(\sigma_n^{2/3})Z^2s|+|\tilde{W}(s)|}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}>\epsilon]\\ \leq\;&\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}\frac{|\sigma_n^{1/3}sf'(\mu(t))Z|}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}]+P[\sup_{|s|\geq\tau}\frac{|O(\sigma_n^{2/3})Z^2s|}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}]+\\ &P[\sup_{|s|\geq\tau}\frac{|\tilde{W}(s)|}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}]\\ =\;&\lim_{n\to\infty}\sup P[\sup_{|s|\geq\tau}\frac{|\sigma_n^{1/3}f'(\mu(t))Z|}{\frac{x_0e^ts}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}]+P[\sup_{|s|\geq\tau}\frac{|O(\sigma_n^{2/3})Z^2|}{\frac{x_0e^ts}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}]+\\ &P[\sup_{|s|\geq\tau}\frac{|\tilde{W}(s)|}{\frac{x_0e^ts^2}{2}+O(\sigma_n^{2/3})}>\frac{\epsilon}{3}].\end{aligned}\tag{12}$$

Because of s in the denominator of the two first terms in (12) there exist $\tau_1(\epsilon,\delta) > 0$ and $\tau_2(\epsilon,\delta) > 0$ such that these two terms are smaller than $\delta/3$. In fact the limit superior of these terms tend to zero for any $\epsilon, \delta, \tau > 0$ since $\sigma_n \to 0$. As for the third term, due to the law of the iterated logarithm which states that

$$\lim_{t\to\infty} \sup \frac{|B(t)|}{\sqrt{2t \log\log t}} \quad = \quad 1\, a.s$$

where B is a Brownian motion, there exists $\tau_3(\epsilon,\delta) > 0$ such that this third term is also smaller than $\delta/3$. By choosing $\hat{\tau} = \max(\tau_1, \tau_2, \tau_3)$, we get that (12) fulfils

$$\begin{aligned}
&\lim_{n\to\infty} \sup P[\sup_{|s|\geq\hat{\tau}} \frac{|\sigma_n^{1/3} f'(\mu(t))Z|}{\frac{x_0 e^t s}{2} + O(\sigma_n^{2/3})} > \frac{\epsilon}{3}] + P[\sup_{|s|\geq\hat{\tau}} \frac{|O(\sigma_n^{2/3})Z^2|}{\frac{x_0 e^t s}{2} + O(\sigma_n^{2/3})} > \frac{\epsilon}{3}] + \\
&P[\sup_{|s|\geq\hat{\tau}} \frac{|\tilde{W}(s)|}{\frac{x_0 e^t s^2}{2} + O(\sigma_n^{2/3})} > \frac{\epsilon}{3}] \\
&< \delta/3 + \delta/3 + \delta/3 = \delta
\end{aligned}$$

and thus assumptions 3 and 4 are satisfied.

We also need to check assumption 5, that is, for every $\epsilon, \delta > 0$ there exists a $\tau = \tau(\epsilon,\delta) > 0$ such that

$$P[\sup_{|s|\geq\tau} |\frac{\tilde{v}(s)}{As^2}| > \epsilon] \quad < \quad \delta\, .$$

As above this holds by applying the law of the iterated logarithm.

Assumption 6 is satisfied since the Wiener process is almost surely continuous and this is enough for assumption 6 to hold.

Thus we obtain the following results for our estimator.

$$n^{2/3}[T_{[0,1]}(X)(t_0) - X(t_0)] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + W(s)](0)$$

where $T_{[0,1]}(x)(t_0)$ is the greatest convex minorant on $[0,1]$ of the integrated process $X(t) = \int_0^t dX(s)$ evaluated in the point t_0. Moreover, we get

$$n^{1/3}[T_{[0,1]}(X)'(t_0) - x_b'(t_0)] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + W(s)]'(0)$$

that is,

$$n^{1/3}[T_{[0,1]}(X_{\text{int}})'(t_0) - f(E[X_{\text{true}}(t_0)])] \xrightarrow{\mathcal{L}} T[\frac{x_0 e^{t_0}}{2}|s|^2 + W(s)]'(0).$$

Hence we have obtained an estimator of our unknown, nondecreasing function f such that this estimator converges of the order $n^{1/3}$ to the slope at the origin of the greatest convex minorant of a Wiener process plus a square function.

8 Simulations

We will now simulate the process

$$dX(t) = X(t)dt + \sigma_n dW(t)$$

and estimate the drift function for $n = 100, 500, 1000, 10000$ observations. From our results above we know that this estimator should converge of the order $n^{1/3}$ and thus we will investigate the mean of the absolute errors to see if this seems to be the case. Moreover, we will also investigate the behaviour of the mean square integrated error.

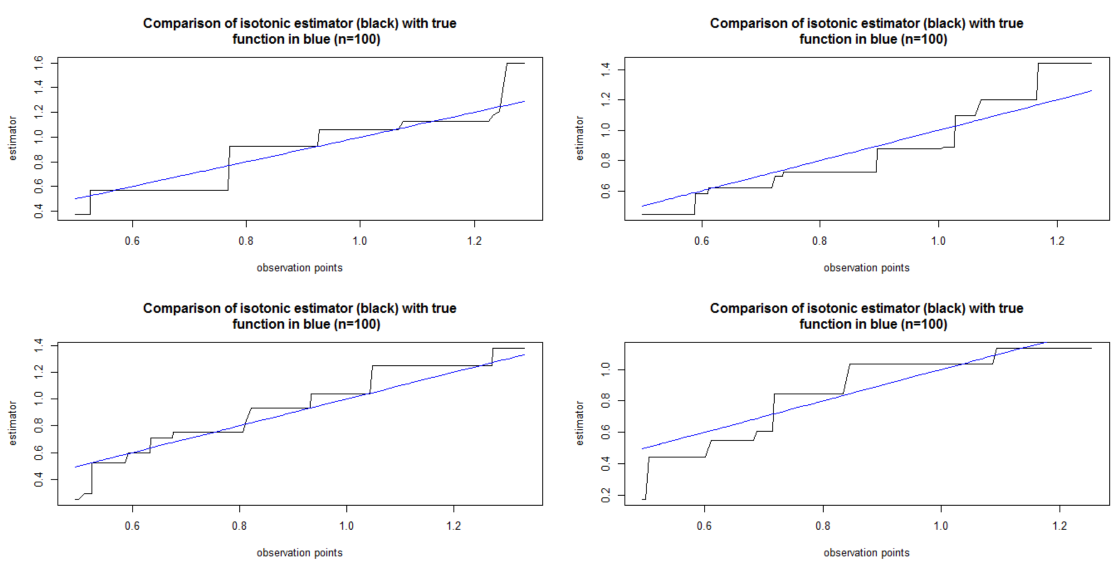

Figure 1: The isotonic estimator in black compared to the true function, the identity function, in blue for $n = 100$

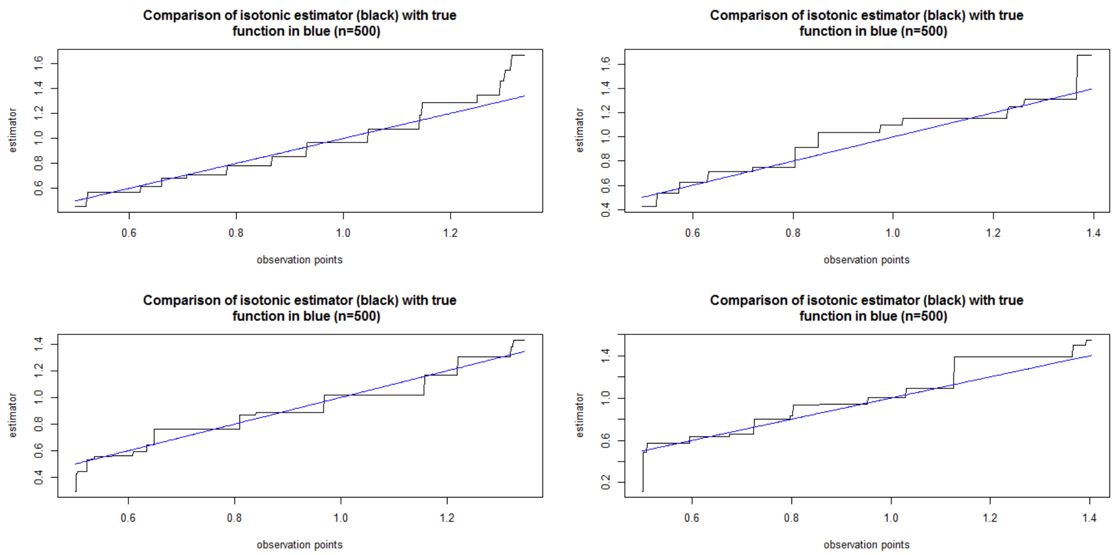

Figure 2: The isotonic estimator in black compared to the true function, the identity function, in blue for $n = 500$

We will investigate the behaviour of the mean square integrated error, $MSIE$, the mean absolute error for the observed process points, MAE_{obs}, as well as the absolute error for the true expected process points, MAE_{real}, which are given by

$$MSIE = \int_0^1 [\hat{f}(E[X(t)]) - f(E[X(t)])]dt,$$

$$MAE_{obs} = \frac{1}{n}\sum_{i=1}^{n} |\hat{f}(X(t_i)) - f(X(t_i))|$$

and

$$MAE_{real} = \frac{1}{n}\sum_{i=1}^{n} |\hat{f}(E[X(t_i)]) - f(E[X(t_i)])|$$

for the estimator and the true estimand based on n observations. We get values of approximate order

$$\begin{bmatrix} n & 100 & 500 & 1000 & 10000 \\ MSIE & 0.02 & 0.005 & 0.003 & 0.0007 \\ MAE_{obs} & 0.10 & 0.050 & 0.040 & 0.020 \\ MAE_{real} & 0.11 & 0.055 & 0.045 & 0.020 \end{bmatrix}$$

and both the MAE_{obs} and MAE_{real} seem to decrease at the order $n^{-1/3}$, which is in concordance with the theoretical rate of convergence.

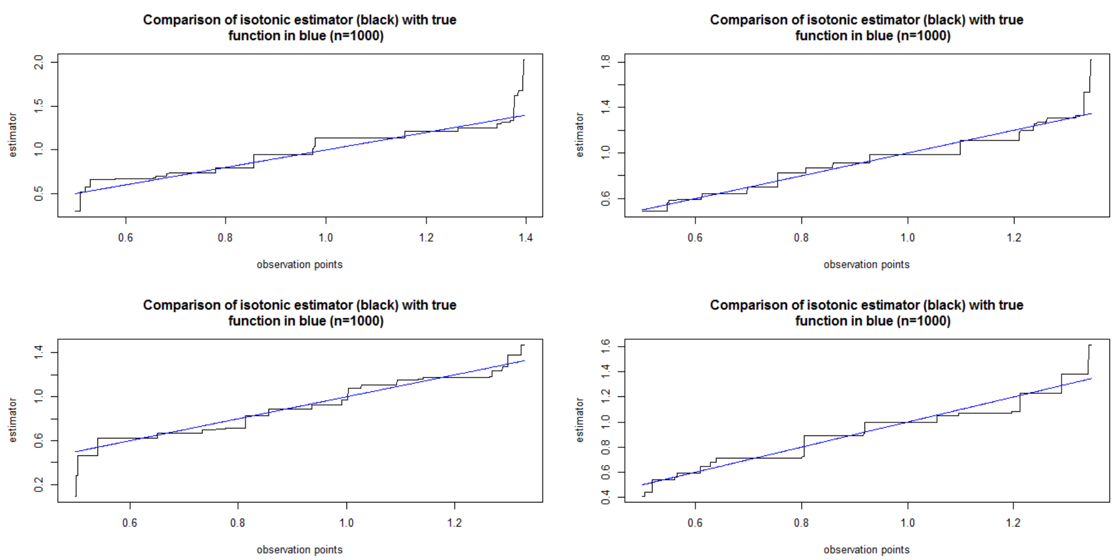

Figure 3: The isotonic estimator in black compared to the true function, the identity function, in blue for 1000 observations.

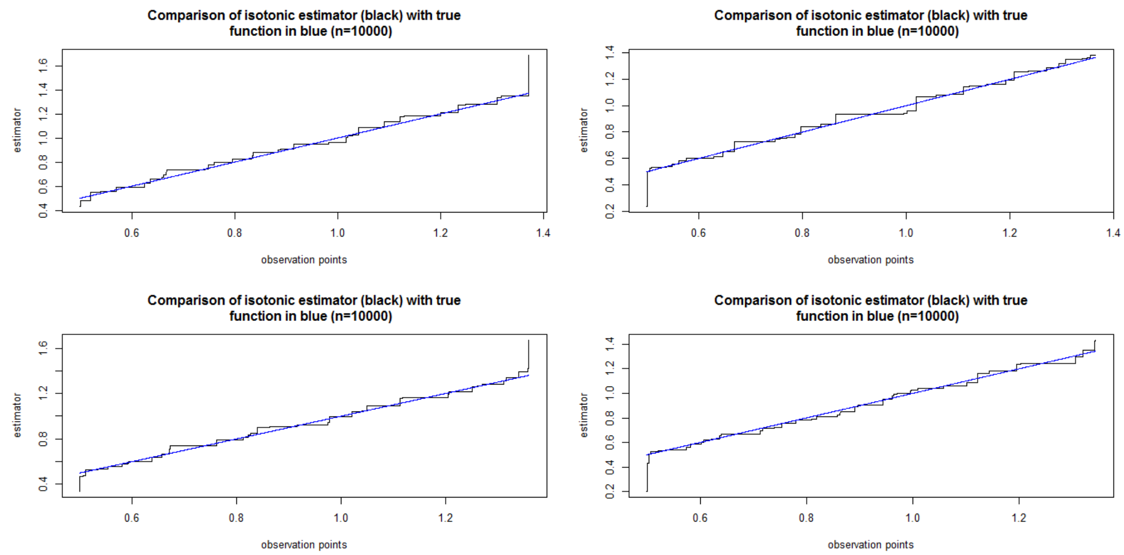

Figure 4: The isotonic estimator in black compared to the true function, the identity function, in blue for 10 000 observations.

What is worth commenting on is that there might be a tendency to have spikes at the boundary which increase with the number of observations as we get closer and closer to the boundary. However, this seems to be a general problem in nonparametric estimation and thus not explicit to our problem. Another thing is that one might be interested in obtaining a more smooth function which could be achieved by applying a kernel-smoother to the isotonic regressor. If one chooses a symmetric kernel, such as for instance a Gaussian kernel, appyling the kernel smoother to the isotonic regressor does not change this from being isotonic. Applying a kernel smoother to obtain an isotonised kernel estimator was investigated and although the obtained estimator was much smoother and also seemed to suffer less from the boundary value problem from which the isotonic regressor suffered, there was also evidence suggesting that the quality of the isotonised kernel estimator depends quite much on the bandwidth chosen. However, it would be interesting to investigate further if it is possible to find an optimal bandwidth which would enable us to find an isotonic regression estimator which is smooth and does not suffer from boundary problems to the same extent.

We will also run some simulations for another stochastic differential equation/filtering problem where

$$dX(t) = \log(X(t))dt + \sigma_n dW(t).$$

In this case the unknown function f is the log function which is increasing. However, it does not satisfy the Lipschitz condition on its domain as a whole. Since it does fulfil the Lipschitz condition for values of $X(t) \geq 1$, we choose an initial value large enough so that the process stays above 1 $\forall t \in [0, 1]$. This is due to our main interest being in investigating how well the isotonic estimator performs in this situation.

Doing the same calculations for the logarithm function as the unknown function, instead of the identity function, yields errors of a magnitude as well as a rate of convergence similar to the previous case. Increasing σ in $\sigma_n = \sigma/\sqrt{n}$ naturally increases the magnitude of the error but there has not been enough time to investigate how the estimator is affected by changing the characteristics of the estimand or changing the initial values and how these might be connected et cetera. We show some plots of the estimator in the case where f is the logarithm for the same number of observations as the previous case, that is, for $n = 100, 500, 1000, 10000$.

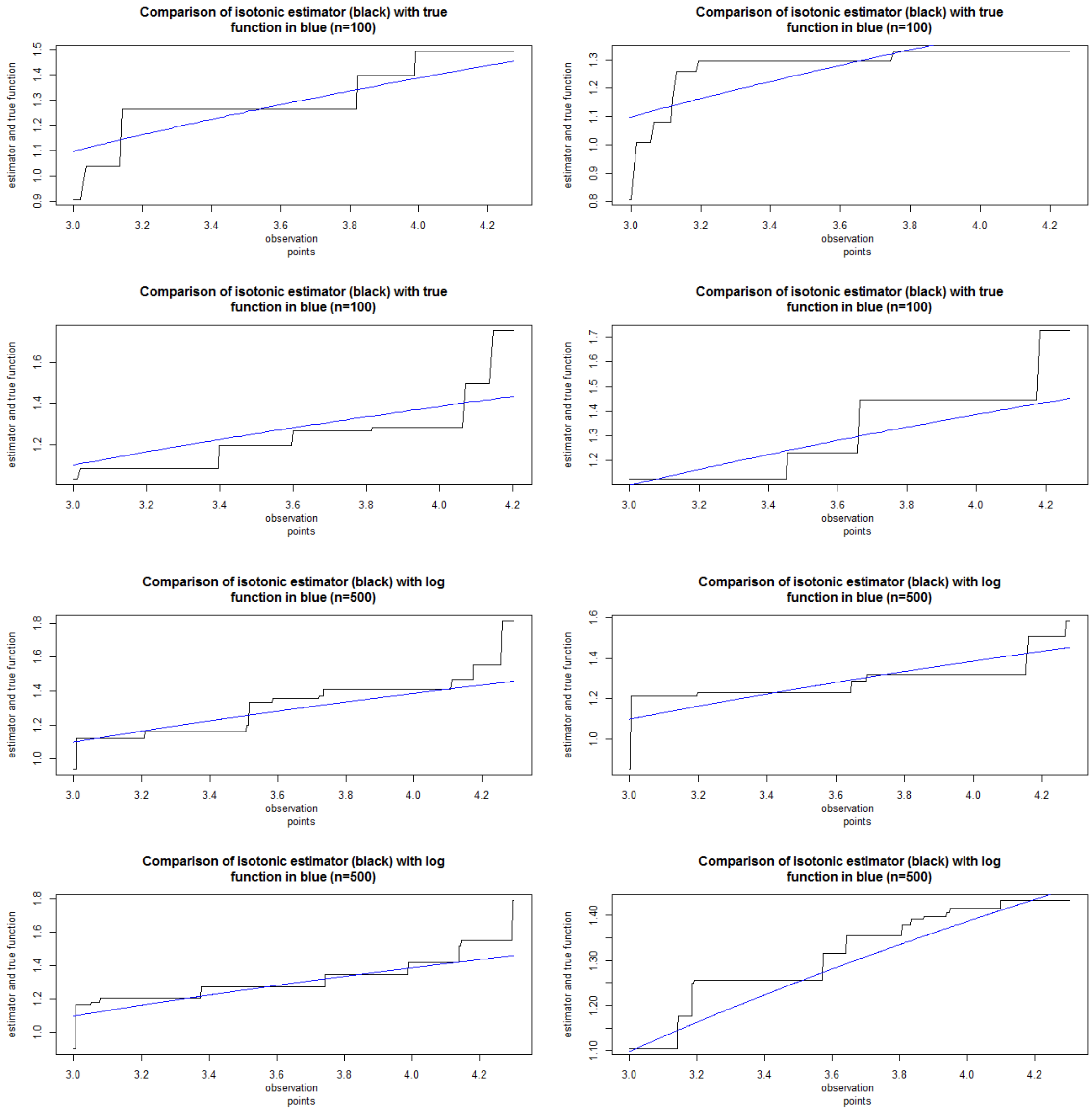

Figure 5: Isotonic regressor (black) of the logarithm function (blue) for 100 and 500 observations.

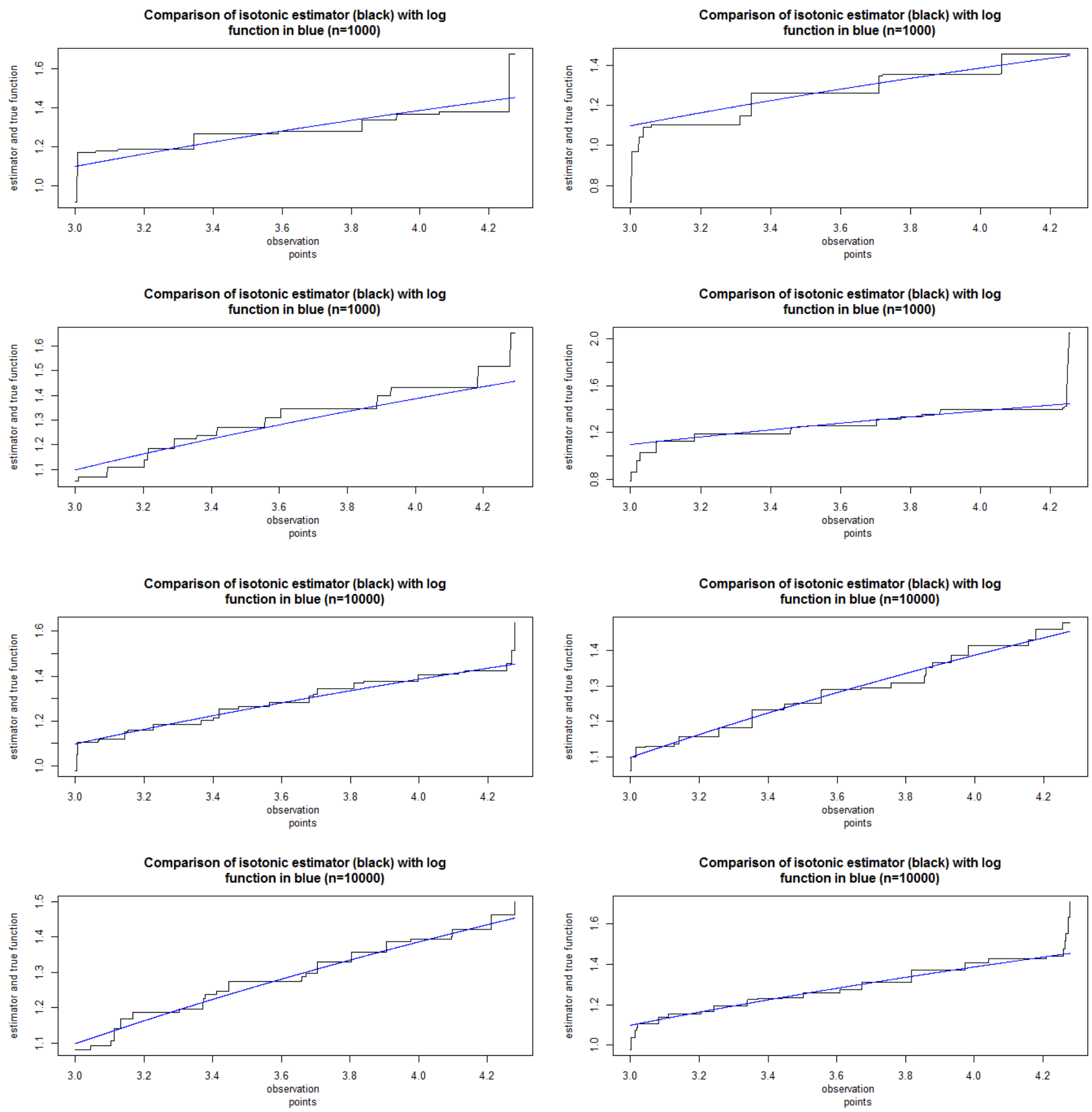

Figure 6: Isotonic regressor (black) of the logarithm function (blue) for 1000 and 10 000 observations.

9 Conclusions and discussion

We can conclude after having given a characterisation of isotonic regression in the discrete and continuous case that it is possible to apply isotonic regression to the filtering problem and that it seems that a rate of convergence of (at least) order $n^{-1/3}$ can be obtained. However, on some occasions there are spikes at the boundaries of the observed process where the isotonic regressor does not perform as desired. Although, it seems that a kernel smoother of the isotonic regressor might remedy this to some extent which, along with the increased smoothness of the estimator it provides, makes it of interest to investigate whether an optimal bandwidth can be obtained for any number of observations, since the bandwidth has a strong influence on the performance of the estimator.

However, something that was not managed to be shown which is of interest, is to be able to find limit distribution results in the actual observation points for a general unknown, monotone function f. For the standard isotonic regression problem with

$$y_i = m(t_i) + \epsilon_i \,, t_i = i/n \,,$$

where m is nondecreasing and ϵ_i are independent, identically distributed random varibles with mean zero and finite variance σ^2, it has been shown, see [2], that by taking the greatest convex minorant on $[0,1]$ of the partial sum process

$$x_n(t) = \frac{1}{n}\sum_{i=1}^{\tilde{n}} y_i + \frac{(nt-1/2)-\tilde{n}}{n} y_{\tilde{n}+1}$$

where $\tilde{n} = \lfloor nt - 1/2 \rfloor$, it holds that

$$n^{1/3}(\hat{m}(t_0) - m(t_0)) \xrightarrow{L} [4m'(t_0)\sigma^2]^{1/3} T(s^2 + B(s))'(0),$$

with $t_0 \in (0,1)$, $m'(t_0) > 0$, B a Brownian motion and $\hat{m}(t_0) = T_{[0,1]}(x_n)'(t_0)$.

In [4] it has instead been considered the regression model where we have independent observations of the bivariate random variables

$$\begin{cases} (X_i, Y_i) \,, i = 1, ..., n \\ (X, Y) \in [0,1] \times \mathbb{R} \\ Y_i = m(X_i) + \epsilon_i \,, i = 1, ..., n \\ E(\epsilon_i | X_i = 0) \end{cases}$$

with X_i continuous random variables and m assumed to be nondecreasing, for instance $m(x) = E(Y|X = x)$. If we denote $\sigma^2(t) = E(\epsilon^2|X_i = t)$, the density of X_i by f and let $t_0 \in (0, 1)$ be a fixed point where m is differentiable with $m'(t_0) > 0$, it holds that

$$n^{1/3}(\hat{m}_n(t) - m(t)) \xrightarrow{L} [\frac{4\sigma^2(t)m'(t)}{f(t)}]^{1/3} X(0)$$

where $\hat{m}_n(t)$ is the greatest convex minorant of the cumulative sum diagram of y_i as described in an earlier section and $2X(0)$ is distributed as the slope at zero of the greatest convex minorant of $\{s^2 + B(s)\,,\, s \in \mathbb{R}\}$ with B a Brownian motion.

It seems reasonable that the limit distribution for the isotonic regressor in a general filtering problem is similar to the limit distributions described above. The remaining problem is to find exactly how the dependence of the observations affects the limit distribution of the estimator and the magnitude of this dependence. In both [2] and [4] the domains are given by $[0, 1]$ or in general $[a, b]$ where a, b are fixed constants. Another aspect to the filtering problem that would be of interest to incorporate is the initial value of the process, since this influences the range of the process. Thus one would also like to include the distribution of the starting value should this be random.

References

[1] Anevski D,Soulier P. Monotone spectral density estimation. *The Annals of statistics*, 39:418–438, 2011.

[2] Anevski,D ,Hössjer O. A general asymptotic scheme for inference under order restrictions. *The Annals of Statistics*, 34:1874–1930, 2006.

[3] Anevski,Dragi. Inference under monotononicity assumptions. pages 1–24, 2006.

[4] Durot Cecile. Monotone nonparametric regression with random design. *Mathematical methods of statistics*, 17:327–341, 2008.

[5] Ibragimov I.A, Hasminski R.Z. *Statistical estimation: Asymptotic theory*. Springer, 1981.

[6] Mishra, M.N , Prakasa Rao, B.L.S. Nonparametric estimation of trend for stochastic differential equations driven by fractional brownian motions. *Stat Inference Stoch Process*, 14:101–109, 2011.

[7] Nussbaum, Michael. Spline smoothing in regression models and asymptotic efficiency in l_2. *The Annals of Statistics*, 13:984–997, 1985.

[8] Oksendahl,Bernt. *Stochastic differential equations: an introduction with applications*. Springer, 1995.

[9] Robertson T ,Wright F.T, Dykstra R. *Order restricted statistical inference*. Wiley series in probability and mathematical statistics. Wiley, 1988.

[10] Silvapulle, Mervyn J., Sen, Pranab K. *Constrained statistical inference: inequality,order and shape restrictions*. Wiley series in probability and statistics. Wiley, 2005.