Electronic Journal of Statistics
Vol. 13 (2019) 646–677
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1520

Contraction and uniform convergence of isotonic regression

Fan Yang and Rina Foygel Barber

Department of Statistics
University of Chicago
Chicago, Illinois, IL 60637
e-mail: fyang1@uchicago.edu; rina@uchicago.edu

Abstract: We consider the problem of isotonic regression, where the underlying signal x is assumed to satisfy a monotonicity constraint, that is, x lies in the cone $\{x \in \mathbb{R}^n : x_1 \leq \cdots \leq x_n\}$. We study the isotonic projection operator (projection to this cone), and find a necessary and sufficient condition characterizing all norms with respect to which this projection is contractive. This enables a simple and non-asymptotic analysis of the convergence properties of isotonic regression, yielding uniform confidence bands that adapt to the local Lipschitz properties of the signal.

MSC 2010 subject classifications: 62G08, 62G07.
Keywords and phrases: Isotonic regression, contraction, data-adaptive band, convergence rates, density estimation.

Received April 2018.

1. Introduction

Isotonic regression is a powerful nonparametric tool used for estimating a monotone signal from noisy data. Specifically, our data consists of observations $y_1, \ldots, y_n \in \mathbb{R}$, which are assumed to be noisy observations of some monotone increasing signal—for instance, we might assume that $\mathbb{E}[y_1] \leq \cdots \leq \mathbb{E}[y_n]$. Isotonic (least-squares) regression solves the optimization problem

$$\text{Minimize } \|y - x\|_2^2 \text{ subject to } x_1 \leq \cdots \leq x_n$$

in order to estimate the underlying signal.

This regression problem can be viewed as a convex projection, since $\mathcal{K}_{\mathsf{iso}} = \{x \in \mathbb{R}^n : x_1 \leq \cdots \leq x_n\}$ is a convex cone. We will write

$$\mathsf{iso}(y) := \mathcal{P}_{\mathcal{K}_{\mathsf{iso}}}(y) = \operatorname*{arg\,min}_{x \in \mathbb{R}^n} \left\{ \|y - x\|_2^2 : x_1 \leq \cdots \leq x_n \right\}$$

to denote the projection to this cone, which solves the least-squares isotonic regression problem. This projection can be computed in finite time with the Pool Adjacent Violators Algorithm (PAVA), developed by Barlow et al. [2].

In fact, mapping y to $\mathsf{iso}(y)$ is known to also solve the isotonic binary regression problem. This arises when the data is binary, that is, $y \in \{0,1\}^n$. If we assume that $y_i \sim \text{Bernoulli}(x_i)$, then the constrained maximum likelihood estimator is exactly equal to the projection $\mathsf{iso}(y)$ (Robertson et al. [21]).

In this paper, we examine the properties of the isotonic projection operator $x \mapsto \mathsf{iso}(x)$, with respect to different norms $\|\cdot\|$ on $\mathbb{R}^n$. We examine the conditions on $\|\cdot\|$ needed in order to ensure that $x \mapsto \mathsf{iso}(x)$ is contractive with respect to this norm, and in particular, we define the new "sliding window norm" which measures weighted averages over "windows" of the vector x, i.e. contiguous stretches of the form $(x_i, x_{i+1}, \ldots, x_{j-1}, x_j)$ for some indices $1 \leq i \leq j \leq n$. This sliding window norm then provides a tool for analyzing the convergence behavior of isotonic regression in a setting where our data is given by $y_i = x_i + \text{noise}$. If the underlying signal x is believed to be (approximately) monotone increasing, then $\mathsf{iso}(y)$ will provide a substantially better estimate of x than the observed vector y itself. By using our results on contractions with respect to the isotonic projection operator, we obtain clean, finite-sample bounds on the pointwise errors, $\left|x_i - \mathsf{iso}(y)_i\right|$, which are locally adaptive to the behavior of the signal x in the region around the index i and hold uniformly over the entire sequence.

1.1. Background

There is extensive literature studying convergence rates of isotonic regression, in both finite-sample and asymptotic settings. For an asymptotic formulation of the problem, since the signal $x \in \mathbb{R}^n$ must necessarily change as $n \to \infty$, a standard method for framing this as a sequence of problems indexed by n is to consider a fixed function $f : [0,1] \to \infty$, and then for each n, define $x_i = f(i/n)$ (or more generally, $x_i = f(t_i)$ for points t_i that are roughly uniformly spaced). Most models in the literature assume that $y_i = x_i + \sigma \cdot \epsilon_i$, where the noise terms ϵ_i are i.i.d. standard normal variables (or, more generally, are zero-mean variables that satisfy some moment assumptions or are subgaussian).

One class of existing results treats *global* convergence rates, where the goal is to bound the error $\|x - \mathsf{iso}(y)\|_2$, or more generally to bound $\|x - \mathsf{iso}(y)\|_p$ for some ℓ_p norm. The estimation error under the ℓ_2 norm was studied by Van de Geer [22], Wang and Chen [24], Meyer and Woodroofe [18], among others. Van de Geer [23] obtains the asymptotic risk bounds for certain 'bounded' isotonic regression under Hellinger distance, whereas Zhang [26] establishes the non-asymptotic risk bounds for general ℓ_p norm—in particular, for $p = 2$, they show that the least-squares estimator $\mathsf{iso}(y)$ of the signal x has error scaling as $\|x - \widehat{x}\|_2/\sqrt{n} \sim n^{-1/3}$. Recent work by Chatterjee et al. [9] considers non-asymptotic minimax rates for the estimation error, focusing specifically on $\|x - \widehat{x}\|_2$ for any estimator $\widehat{x}$ to obtain a minimax rate. Under a Gaussian noise model, they prove that the minimax rate scales as $\|x - \widehat{x}\|_2/\sqrt{n} \gtrsim n^{-1/3}$ over the class of monotone and Lipschitz signals x, which matches the error rate of the constrained maximum likelihood estimator (i.e. the isotonic least-squares projection, $\mathsf{iso}(y)$) established earlier. They also study minimax rates in a range of settings, including piecewise constant signals, which we will discuss later on.[1]

[1] Chatterjee et al. [9]'s results, which they describe as "local minimax" bounds, are "local" in the sense that the risk bound they provide is specific to an individual signal $x \in \mathbb{R}^n$, but the error is nonetheless measured with respect to the ℓ_2 norm, i.e. "globally" over the entire length of the signal.

A separate class of results considers *local* convergence rates, where the error at a particular index, i.e. $\left|x_i - \mathsf{iso}(y)_i\right|$ for some particular i, may scale differently in different regions of the vector. In an asymptotic setting, where $n \to \infty$ and the underlying signal comes from a function $f : [0,1] \to \mathbb{R}$, we may consider an estimator $\widehat{f} : [0,1] \to \mathbb{R}$, where $\widehat{f}(t)$ is estimated via $\mathsf{iso}(y)_i$ for $t \approx i/n$. Results in the literature for this setting study the asymptotic rate of convergence of $\left|f(t) - \widehat{f}(t)\right|$, which depends on the local properties of f near t. Brunk [6] establishes the convergence rate as well as the limiting distribution when $f'(t)$ is positive, whereas Wright [25] generalizes the result to the case of t lying in a flat region, i.e. $f'(t) = 0$. Cator [8] shows that the isotonic estimator adapts to the unknown function locally and is asymptotically minimax optimal for local behavior. Relatedly, Dümbgen [11] gives confidence bands in the related Gaussian white noise model, by taking averages over windows of the data curve, i.e. ranges of the form $[t_0, t_1]$ near the point t of interest.

In addition, many researchers have considered the related problem of monotone density estimation, where we aim to estimate a monotone decreasing density from n samples drawn from that distribution. This problem was first studied by Grenander [14], and has attracted much attention since then, see Rao [19], Groeneboom [15], Birgé [3], Birgé and Massart [4], Carolan and Dykstra [7], Balabdaoui et al. [1], Jankowski [17], among others. Birgé [3] proves a $n^{-1/3}$ minimax rate for the ℓ_2 error in estimating the true monotone density $f(t)$—the same rate as for the isotonic regression problem. The pointwise i.e. ℓ_∞ error has also been studied—Durot et al. [12] shows that, for Lipschitz and bounded densities on $[0,1]$, asymptotically the error rate for estimating $f(t)$ scales as $(n/\log(n))^{-1/3}$, uniformly over all t bounded away from the endpoints. Adaptive convergence rates are studied by Cator [8]. Later we will show that our results yield a non-asymptotic error bound for this problem as well, which matches this known rate.

Several related problems for isotonic regression have also been studied. First, assuming the model $y_i = x_i + \sigma \cdot \epsilon_i$ for standard normal error terms ϵ_i, estimating σ has been studied by Meyer and Woodroofe [18], among others. Estimators of σ for general distribution of ϵ_i are also available, see Rice [20], Gasser et al. [13]. We discuss the relevance of these tools for constructing our confidence bands in Section 4. Second, we can hope that our estimator $\mathsf{iso}(y)$ can recover x accurately only if x itself is monotone (or approximately monotone); thus, testing this hypothesis is important for knowing whether our confidence band can be expected to cover x itself or only its best monotone approximation, $\mathsf{iso}(x)$. Drton and Klivans [10] study the problem of testing the null hypothesis $x \in \mathcal{K}_{\mathsf{iso}}$ (or more generally, whether the signal x belongs to some arbitrary pre-specified cone $\mathcal{K}$), based on the volumes of lower-dimensional faces of the cone (see Drton and Klivans [10, Theorem 2 and Section 3]).

Main contributions In the context of the existing literature, our main contributions are: (1) the new analysis of the contraction properties of isotonic projection, and the specific example of the sliding window norm, and (2) clean, finite-sample estimation bounds for isotonic regression, which are locally adap-

tive to the local Lipschitz behavior of the underlying signal, and match known asymptotic convergence rates. The contraction and sliding window norm allow us to prove the isotonic regression convergence results in just a few simple lines of calculations, while the arguments in the existing literature are generally substantially more technical (for example, approximating the estimation process via a Brownian motion or Brownian bridge).

2. Contractions under isotonic projection

In this section, we examine the contractive behavior of the isotonic projection,

$$\mathsf{iso}(x) = \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \big\{ \|x - y\|_2 : y_1 \le \cdots \le y_n \big\},$$

with respect to various norms on $\mathbb{R}^n$. Since this operator projects x onto a convex set (the cone $\mathcal{K}_{\mathsf{iso}}$ of all ordered vectors), it is trivially true that

$$\|\mathsf{iso}(x) - \mathsf{iso}(y)\|_2 \le \|x - y\|_2,$$

but we may ask whether the same property holds when we consider norms other than the ℓ_2 norm.

Formally, we defined our question as follows:

Definition 1. For a seminorm $\|\cdot\|$ on $\mathbb{R}^n$, we say that isotonic projection is contractive with respect to $\|\cdot\|$ if

$$\|\mathsf{iso}(x) - \mathsf{iso}(y)\| \le \|x - y\| \text{ for all } x, y \in \mathbb{R}^n.$$

We recall that a seminorm must satisfy a scaling law, $\|c \cdot x\| = |c| \cdot \|x\|$, and the triangle inequality, $\|x + y\| \le \|x\| + \|y\|$, but may have $\|x\| = 0$ even if $x \neq 0$. From this point on, for simplicity, we will simply say "norm" to refer to any seminorm.

For which types of norms can we expect this contraction property to hold? To answer this question, we first define a simple property to help our analysis:

Definition 2. For a norm $\|\cdot\|$ on $\mathbb{R}^n$, we say that $\|\cdot\|$ is nonincreasing under neighbor averaging (NUNA) if

$$\left\| \left(x_1, \ldots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, \ldots, x_n \right) \right\| \le \|x\|$$

for all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n-1$.

Our first main result proves that the NUNA property exactly characterizes the contractive behavior of isotonic projection—NUNA is both necessary and sufficient for isotonic projection to be contractive.

Theorem 1. *For any norm $\|\cdot\|$ on $\mathbb{R}^n$, isotonic projection is contractive with respect to $\|\cdot\|$ if and only if $\|\cdot\|$ is nonincreasing under neighbor averaging (NUNA).*

(The proof of Theorem 1 will be given in Section 6.)

In particular, this theorem allows us to easily prove that isotonic projection is contractive with respect to the ℓ_p norm for any $p \in [1, \infty]$, and more generally as well, via the following lemma:

Lemma 1. *Suppose that $\|\cdot\|$ is a norm that is invariant to permutations of the entries of the vector, that is, for any $x \in \mathbb{R}^n$ and any permutation π on $\{1, \dots, n\}$,*

$$\|x\| = \|x_\pi\| \text{ where } x_\pi := (x_{\pi(1)}, \dots, x_{\pi(n)}).$$

(In particular, the ℓ_p norm, for any $p \in [1, \infty]$, satisfies this property.) Then $\|\cdot\|$ satisfies the NUNA property, and therefore isotonic projection is a contraction with respect to $\|\cdot\|$.

Proof of Lemma 1. Let π swap indices i and $i+1$, so that

$$x_\pi = (x_1, \dots, x_{i-1}, x_{i+1}, x_i, x_{i+2}, \dots, x_n).$$

Then

$$\left\| \left(x_1, \dots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, \dots, x_n \right) \right\| = \left\| \frac{x + x_\pi}{2} \right\| \leq \frac{1}{2} \left(\|x\| + \|x_\pi\| \right) = \|x\|,$$

where we apply the triangle inequality, and the assumption that $\|x_\pi\| = \|x\|$. This proves that $\|\cdot\|$ satisfies NUNA. By Theorem 1, this implies that isotonic projection is contractive with respect to $\|\cdot\|$. □

3. The sliding window norm

We now introduce a *sliding window* norm, which will later be a useful tool for obtaining uniform convergence guarantees for isotonic regression. For any pair of indices $1 \leq i \leq j \leq n$, we write $i : j$ to denote the stretch of $j - i + 1$ many coordinates indexed by $\{i, \dots, j\}$,

$$x = (x_1, \dots, x_{i-1}, \underbrace{x_i, \dots, x_j}_{\text{window } i:j}, x_{j+1}, \dots, x_n).$$

Fix any function

$$\psi : \{1, \dots, n\} \to \mathbb{R}_+ \text{ such that } \psi \text{ is nondecreasing and } i \mapsto i/\psi(i) \text{ is concave.} \tag{1}$$

The sliding window norm is defined as

$$\|x\|_\psi^{\mathsf{SW}} = \max_{1 \leq i \leq j \leq n} \left\{ \left| \overline{x}_{i:j} \right| \cdot \psi(j - i + 1) \right\},$$

where $\overline{x}_{i:j} = \frac{x_i + \cdots + x_j}{j - i + 1}$ denotes the average over the window $i : j$.

The following key lemma proves that our contraction theorem, Theorem 1, can be applied to this sliding window norm. (This lemma, and all lemmas following, will be proved in Appendix A.3.)

Lemma 2. *For any function ψ satisfying the conditions (1), the sliding window norm $\|\cdot\|_\psi^{\mathsf{SW}}$ satisfies the NUNA property, and therefore, isotonic projection is contractive with respect to this norm.*

This lemma is a key ingredient to our convergence analyses for isotonic regression. It will allow us to use the sliding window norm to understand the behavior of $\mathsf{iso}(y)$ as an estimator of $\mathsf{iso}(x)$, where y is a vector of noisy observations of some target signal x. In particular, we will consider the special case of subgaussian noise [2] . The following lemma can be proved with a very basic union bound argument:

Lemma 3. *Let $x \in \mathbb{R}^n$ be a fixed vector, and let $y_i = x_i + \sigma\epsilon_i$, where the ϵ_i's are independent, zero-mean, and subgaussian. Then taking $\psi(i) = \sqrt{i}$, we have*

$$\mathbb{E}\left[\|x-y\|_\psi^{\mathsf{SW}}\right] \le \sqrt{2\sigma^2\log(n^2+n)} \text{ and } \mathbb{E}\left[\left(\|x-y\|_\psi^{\mathsf{SW}}\right)^2\right] \le 8\sigma^2\log(n^2+n),$$

and for any $\delta > 0$,

$$\mathbb{P}\left\{\|x-y\|_\psi^{\mathsf{SW}} \le \sqrt{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}\right\} \ge 1-\delta.$$

As a specific example, in a Bernoulli model, if the signal is given by $x \in [0,1]^n$ and our observations are given by $y_i \sim \text{Bernoulli}(x_i)$ (each drawn independently), then this model satisfies the subgaussian noise model with $\sigma = 1$.

4. Estimation bands

In this section, we will develop a range of results bounding our estimation error when we observe a (nearly) monotone signal plus noise. These results will all use the sliding window contraction result in Lemma 2 as the main ingredient in our analysis.

We begin with a deterministic statement that is a straightforward consequence of the sliding window contraction result:

Theorem 2. *For any $x, y \in \mathbb{R}^n$, for all indices $k = 1, \dots, n$,*

$$\begin{aligned}&\max_{1\le m\le k}\left\{\overline{\mathsf{iso}(y)}_{(k-m+1):k} - \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\}\\ &\le \mathsf{iso}(x)_k \le \min_{1\le m\le n-k+1}\left\{\overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\}\end{aligned} \tag{2}$$

[2] We call a random variable X subgaussian if $\mathbb{P}\{|X - \mathbb{E}[X]| \ge t\} \le 2\exp(-t^2/2)$ for any $t > 0$.

and

$$\max_{1\le m\le k}\left\{\overline{\mathsf{iso}(x)}_{(k-m+1):k}-\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\} \le \mathsf{iso}(y)_k \le \min_{1\le m\le n-k+1}\left\{\overline{\mathsf{iso}(x)}_{k:(k+m-1)}+\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}\right\}. \tag{3}$$

Note that these two statements are symmetric; they are identical up to reversing the roles of x and y.

Proof of Theorem 2. We have $\mathsf{iso}(x)_k \ge \overline{\mathsf{iso}(x)}_{(k-m+1):k} \ge \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}$, where the first inequality uses the monotonicity of $\mathsf{iso}(x)$ while the second uses the definition of the sliding window norm along with the fact that $\|\mathsf{iso}(x)-\mathsf{iso}(y)\|_\psi^{\mathsf{SW}} \le \|x-y\|_\psi^{\mathsf{SW}}$ by Lemma 2. This proves the lower bound for (2); the upper bound, and the symmetric result (3), are proved analogously. □

This simple reformulation of our contraction result, in fact forms the backbone of all our estimation band guarantees.

These bounds bound the difference between $\mathsf{iso}(x)$ and $\mathsf{iso}(y)$, computed using either y (as in (2)) or x (as in (3)). Thus far, the two results are entirely symmetrical—they are the same if we swap the vectors x and y.

We will next study the statistical setting where we aim to estimate a signal x based on noisy observations y, in which case the vectors x and y play distinct roles, and so the two versions of the bands will carry entirely different meanings. Before proceeding, we note that the above bounds cannot give results on x itself, but only on its projection $\mathsf{iso}(x)$. If x is far from monotonic, we cannot hope that the monotonic vector $\mathsf{iso}(y)$ would give a good estimate of x. We will consider a relaxed monotonicity constraint: we say that $x \in \mathbb{R}^n$ is ϵ_{iso}-monotone if

$$x_i \le x_j + \epsilon_{\mathsf{iso}} \text{ for all } 1 \le i \le j \le n.$$

(If x is monotonic then we can simply set $\epsilon_{\mathsf{iso}} = 0$.) We find that ϵ_{iso} corresponds roughly to the ℓ_∞ distance between x and its isotonic projection $\mathsf{iso}(x)$:

Lemma 4. *For any $x \in \mathbb{R}^n$ that is ϵ_{iso}-monotone,*

$$\|x-\mathsf{iso}(x)\|_\infty \le \epsilon_{\mathsf{iso}}.$$

Conversely, any $x \in \mathbb{R}^n$ with $\|x-\mathsf{iso}(x)\|_\infty \le \epsilon$ must be (2ϵ)-monotone.

With this in place, we turn to our results for the statistical setting.

4.1. Statistical setting

We will consider a subgaussian noise model, where $x \in \mathbb{R}^n$ is a fixed signal, and the observation vector y is generated as

$$y_i = x_i + \sigma\epsilon_i, \text{ where the } \epsilon_i\text{'s are independent, zero-mean, and subgaussian.} \tag{4}$$

Lemma 3 proves that, in this case, setting $\psi(m) = \sqrt{m}$ would yield $\|x-y\|_\psi^{\mathsf{SW}} \le$

$\sqrt{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}$ with probability at least $1-\delta$. Of course, we could consider other models as well, e.g. involving correlated noise or heavy-tailed noise, but restrict our attention to this simple model for the sake of giving an intuitive illustration of our results.

In order for this bound on the sliding window to be useful in practice, we need to obtain a bound or an estimate for the noise level σ. Under the Bernoulli model $y_i \sim \text{Bernoulli}(x_i)$, we can simply set $\sigma = 1$. More generally, it may be possible to estimate σ from the data itself, for instance if the noise terms ϵ_i are i.i.d. standard normal, Meyer and Woodroofe [18] propose estimating the noise level σ with the maximum likelihood estimator (MLE), $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (y_i - \mathsf{iso}(y)_i)^2$, or the bias-corrected MLE given by

$$\widehat{\sigma}^2 = \frac{\sum_i (y_i - \mathsf{iso}(y)_i)^2}{n - c_1 \cdot \mathsf{df}\big(\mathsf{iso}(y)\big)},$$

where c_1 is a known constant while $\mathsf{df}\big(\mathsf{iso}(y)\big)$ is the number of "degrees of freedom" in the monotone vector $\mathsf{iso}(y)$, i.e. the number of distinct values in this vector.

We next consider the two different types of statistical guarantees that can be obtained, using the two symmetric formulations in Theorem 2 above.

4.2. Data-adaptive bands

We first consider the problem of providing a confidence band for the signal x in a practical setting, where we can only observe the noisy data y and do not have access to other information. In this setting, the bound (2) in Theorem 2, combined with Lemma 3's bound on $\|x-y\|_\psi^{\mathsf{SW}}$ for the subgaussian model, yields the following result:

Theorem 3. *For any signal $x \in \mathbb{R}^n$ and any $\delta > 0$, under the subgaussian noise model (4), then with probability at least $1-\delta$, for all $k = 1, \ldots, n$,*

$$\begin{aligned} &\max_{1\le m\le k} \left\{ \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} \\ &\le \mathsf{iso}(x)_k \le \min_{1\le m\le n-k+1} \left\{ \overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\}. \end{aligned} \tag{5}$$

If additionally x is ϵ_{iso}-monotone, then we also have

$$\begin{aligned} &\max_{1\le m\le k} \left\{ \overline{\mathsf{iso}(y)}_{(k-m+1):k} - \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} - \epsilon_{\mathsf{iso}} \\ &\le x_k \le \min_{1\le m\le n-k+1} \left\{ \overline{\mathsf{iso}(y)}_{k:(k+m-1)} + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} + \epsilon_{\mathsf{iso}}. \end{aligned} \tag{6}$$

We emphasize that these bounds give us a uniform confidence band for $\mathsf{iso}(x)$ (or for x itself, if it is monotone) that can be computed without assuming anything about the properties of the signal; for instance, we do not assume that the signal is Lipschitz with some known constant, or anything of this sort. We only need to know the noise level σ, which can be estimated as discussed in Section 4.1. In this sense, the bounds are data-adaptive—they are computed using the observed projection $\mathsf{iso}(y)$, and adapt to the properties of the signal (for instance, if x is locally constant near k, then the upper and lower confidence bounds will be closer together).

Comparison to existing work The flavor of our data-adaptive band is close to that given in Dümbgen [11], where the author gives confidence bands for signals in a continuous Gaussian white noise model. Although in Section 5 of Dümbgen [11] the result is applied to the discrete case, the confidence band there is only valid asymptotically as pointed our by the author, whereas our band is valid for finite samples. Moreover, the computation of the band in Dümbgen [11] involves Monte Carlo simulation to estimate several key quantiles, and hence is much heavier than the computation of our band. Another difference is that Dümbgen [11] employes kernel estimators in their bands while we use the isotonic least squares estimator in our construction.

4.3. Convergence rates

While the results of Theorem 3 give data-adaptive bounds that do not depend on properties of x, from a theoretical point of view we would also like to understand how the estimation error depends on these properties. For the data-adaptive bands, we used the result (2) relating $\mathsf{iso}(x)$ and $\mathsf{iso}(y)$, but for this question, we will use the symmetric result (3) instead, which immediately yields the following theorem.

Theorem 4. *For any signal $x \in \mathbb{R}^n$ and any $\delta > 0$, under the subgaussian noise model (4), then with probability at least $1-\delta$, for all $k = 1, \ldots, n$,*

$$\begin{aligned} &- \min_{1 \le m \le k} \left\{ \left(\mathsf{iso}(x)_k - \overline{\mathsf{iso}(x)}_{(k-m+1):k} \right) + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} \\ &\quad \le \mathsf{iso}(y)_k - \mathsf{iso}(x)_k \\ &\quad \le \min_{1 \le m \le n-k+1} \left\{ \left(\overline{\mathsf{iso}(x)}_{k:(k+m-1)} - \mathsf{iso}(x)_k \right) + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\}. \end{aligned} \tag{7}$$

If additionally x is ϵ_{iso}-monotone, then we also have

$$- \min_{1 \le m \le k} \left\{ \left(x_k - \overline{x}_{(k-m+1):k} \right) + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} - \epsilon_{\mathsf{iso}}$$

$$\leq \mathsf{iso}(y)_k - x_k \leq \min_{1\leq m\leq n-k+1} \left\{ \left(\overline{x}_{k:(k+m-1)} - x_k\right) + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\} + \epsilon_{\mathsf{iso}}. \tag{8}$$

Proof of Theorem 4. For the first bound (7), we simply subtract $\mathsf{iso}(x)_k$ from the inequalities (3). For the second bound (8) in the case that x is approximately monotone, we instead subtract x_k from (3), and also use the fact that $\|x - \mathsf{iso}(x)\|_\infty \leq \epsilon_{\mathsf{iso}}$ by Lemma 4, which implies that $\left|\overline{x}_{k:(k+m-1)} - \overline{\mathsf{iso}(x)}_{k:(k+m-1)}\right| \leq \epsilon_{\mathsf{iso}}$, and similarly $\left|\overline{x}_{(k-m+1):k} - \overline{\mathsf{iso}(x)}_{(k-m+1):k}\right| \leq \epsilon_{\mathsf{iso}}$. □

Comparison to existing work In the monotone setting (i.e. $x = \mathsf{iso}(x)$), Chatterjee et al. [9] derive related results bounding the pointwise error $\left|x_k - \mathsf{iso}(y)_k\right|$. Specifically, they use the "minmax" formulation of the isotonic projection, $\mathsf{iso}(y)_k = \min_{j\geq k} \max_{i\leq k} \overline{y}_{i:j}$, and give the following argument:

$$\begin{aligned}
\mathsf{iso}(y)_k - x_k &= \min_{1\leq m\leq n-k+1} \max_{i\leq k} \overline{y}_{i:(k+m-1)} - x_k \\
&\leq \min_{1\leq m\leq n-k+1} \left\{ \left(\max_{i\leq k} \overline{x}_{i:(k+m-1)} - x_k \right) + \max_{i\leq k} \left|\overline{x}_{i:(k+m-1)} - \overline{y}_{i:(k+m-1)}\right| \right\} \\
&\leq \min_{1\leq m\leq n-k+1} \left\{ \left(\overline{x}_{k:(k+m-1)} - x_k\right) + \underbrace{\max_{i\leq k} \left|\overline{x}_{i:(k+m-1)} - \overline{y}_{i:(k+m-1)}\right|}_{\text{(Err)}} \right\},
\end{aligned}$$

where the first step defines $m = j - k + 1$ and uses the "minmax" formulation, while the third uses the assumption that x is monotone. They then bound the error term (Err) in expectation. We can instead bound it as (Err) $\leq \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\sqrt{m}}$, which is exactly the same as the upper bound in our result (8). Their "minmax" strategy can analogously be used to obtain the corresponding lower bound as well.

4.4. Locally constant and locally Lipschitz signals

If the signal x is monotone, Chatterjee et al. [9]'s results, which are analogous to our bounds in (8), yield implications for many different classes of signals: for instance, they show that for a piecewise constant signal x taking only s many unique values, the ℓ_2 error scales as

$$\frac{1}{n}\|x - \mathsf{iso}(y)\|_2^2 \leq \frac{16 s\sigma^2}{n} \log\left(\frac{en}{s}\right).$$

We therefore see that

$$\left|x_k - \mathsf{iso}(y)_k\right| \lesssim \sqrt{\frac{log(n)}{n}} \tag{9}$$

for "most" indices k when the signal is piecewise constant.

We can instead consider a Lipschitz signal: we say that x is L-Lipschitz if $|x_i - x_{i+1}| \le L/n$ for all i. (Rescaling by n is natural as we often think of $x_i = f(i/n)$ for some underlying function f). In this setting, our results in Theorem 4 immediately yield the bound

$$\left|x_k - \mathsf{iso}(y)_k\right| \le \min_{1 \le m \le k \wedge (n-k+1)} \left\{ \frac{L(m-1)}{2n} + \sqrt{\frac{2\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{m}} \right\}, \tag{10}$$

where the term $\frac{L(m-1)}{2n}$ is a bound on $\left(\overline{x}_{k:(k+m-1)} - x_k\right)$ and $\left(x_k - \overline{x}_{(k-m+1):k}\right)$ when x is L-Lipschitz. It's easy to see that the optimal scaling is achieved by taking $m = \left\lceil \left(\frac{n\sqrt{\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}}{L} \right)^{2/3} \right\rceil$, in which case we obtain the bound

$$\left|x_k - \mathsf{iso}(y)_k\right| \le 2\sqrt[3]{\frac{L\sigma^2 \log\left(\frac{n^2+n}{\delta}\right)}{n}} \tag{11}$$

for all $m \le k \le n - m + 1$. (For indices k nearer to the endpoints, we are forced to choose a smaller m, and the scaling will be worse.)

We can also compute convergence rates in a more general setting, where the signal x is locally Lipschitz—its behavior may vary across different regions of the signal. As discussed in Section 1.1, many papers in the literature consider asymptotic local convergence rates—local in the sense of giving *pointwise* error bounds, which for a single signal $x = (x_1, \ldots, x_n)$, may be larger for indices i falling within a region of the signal that is strictly increasing, and smaller for indices i falling into a locally flat region. We would hope to see some interpolation between the $n^{-1/3}$ rate expected for a strictly increasing stretch of the signal, as in (11), versus the improved parametric rate of $n^{-1/2}$ in a locally constant region as in (9).

Our confidence bands can also be viewed as providing error bounds that are local in this sense, i.e. that adapt to the local behavior of the signal x as we move from index $i = 1$ to $i = n$. To make this more precise, we will show how our bounds scale locally with the sample size n to obtain the $n^{-1/3}$ and $n^{-1/2}$ rates described above. Consider any monotone signal x. Suppose the signal x is locally constant near k, with $x_{k-cn+1} = \cdots = x_k = \cdots = x_{k+cn-1}$ for some positive constant $c > 0$. Then our bound (8) applied with $m = cn$ yields

$$\left|x_k - \mathsf{iso}(y)_k\right| \lesssim \sqrt{\frac{\sigma^2 \log(n)}{n}}. \tag{12}$$

For other indices, however, where the signal is locally strictly increasing with a Lipschitz constant L, then taking $m \sim \left(\frac{\sigma^2 n \log(n)}{L}\right)^{2/3}$ yields the $n^{-1/3}$ scaling obtained above in (11). It is of course also possible to achieve an interpolation between the $n^{-1/2}$ and $n^{-1/3}$ rates via our results, as well.

Many works in the literature consider the local adaptivity problem in an asymptotic setting; here we will describe the results of Cator [8]. Consider an asymptotic setting where the signal $x = (x_1, \dots, x_n)$ comes from measuring (at n many points) a monotone function $f : [0,1] \to \mathbb{R}$, and we are interested in the local convergence rate at some fixed $t \in (0,1)$. Cator [8] show that if the first α derivatives of f at t satisfy $f^{(1)}(t) = \cdots = f^{(\alpha-1)}(t) = 0$ and $f^{(\alpha)}(t) > 0$, the convergence rate for estimating $f(t)$ scales as $n^{-\alpha/(2\alpha+1)}$. In particular, if $\alpha = 1$ (f is strictly increasing at t) then they obtain the $n^{-1/3}$ rate seen before, while if $\alpha = \infty$ (f is locally constant near t) then they obtain the faster parametric $n^{-1/2}$ rate. Of course, any α in between 1 and ∞ will produce some power of n between these two. Our work can be viewed as a finite-sample version of these types of results.

4.5. Convergence rates in the ℓ_2 norm

We next show that the tools developed in this paper can be used to yield a bound on the ℓ_2 error, achieving the same $n^{-1/3}$ scaling as in Chatterjee et al. [9]. While achieving an $n^{-1/3}$ elementwise requires a Lipschitz condition on the signal (as in our result (11) above), here we do not assume any Lipschitz conditions and require only a bound on the total variation,

$$V \coloneqq \mathsf{iso}(x)_n - \mathsf{iso}(x)_1.$$

Our proof uses similar techniques as Chatterjee et al. [9]'s result.

Theorem 5. *For any signal $x \in \mathbb{R}^n$, under the subgaussian noise model (4), we have*

$$\frac{1}{n}\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 \le 48\left(\frac{V\sigma^2\log(2n)}{n}\right)^{2/3} + \frac{96\sigma^2\log^2(2n)}{n}.$$

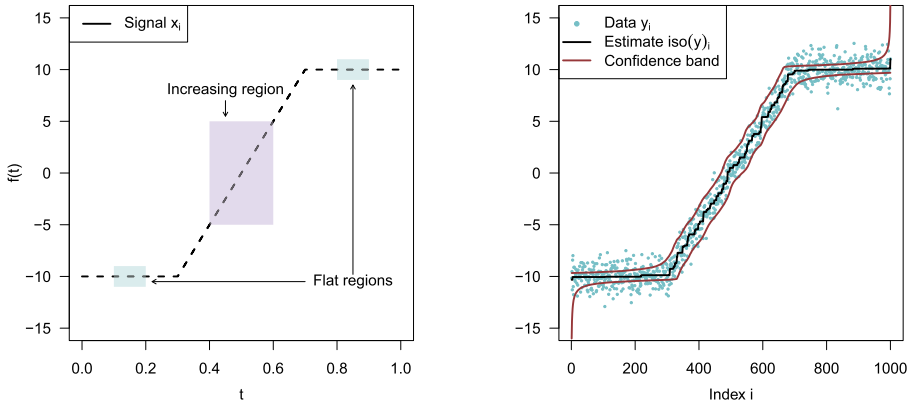
As long as $n \gg \frac{\sigma^2\log^4(2n)}{V^2}$, the first term is the dominant one, matching the result of Chatterjee et al. [9, Theorem 4.1] with a slight improvement in the log term. (The constants in this result are of course far from optimal.) This result is proved in Appendix A.1.

4.6. Simulation: local adaptivity

To demonstrate this local adaptivity in practice, we run a simple simulation. The signal is generated from an underlying function $f(t)$ defined over $t \in [0,1]$, with

$$f(t) = \begin{cases} -10, & 0 \le t \le 0.3, \\ \text{linearly increasing from } -10 \text{ to } 10, & 0.3 \le t \le 0.7, \\ 10, & 0.7 \le t \le 1, \end{cases}$$

as illustrated in Figure 1(a). For a fixed sample size n, we set $x_i = f\left(\frac{i}{n+1}\right)$ and $y_i = x_i + N(0,1)$. We then compute a data-adaptive confidence band as given



(a) Signal (b) Results for $n = 1000$

FIG 1. *(a) The function $f(t)$ used to generate signals $x \in \mathbb{R}^n$ for various n, with flat and increasing regions highlighted. (b) At sample size $n = 1000$, the observed data y, estimated signal* $\mathsf{iso}(y)$*, and data-adaptive confidence band computed as in* (5).

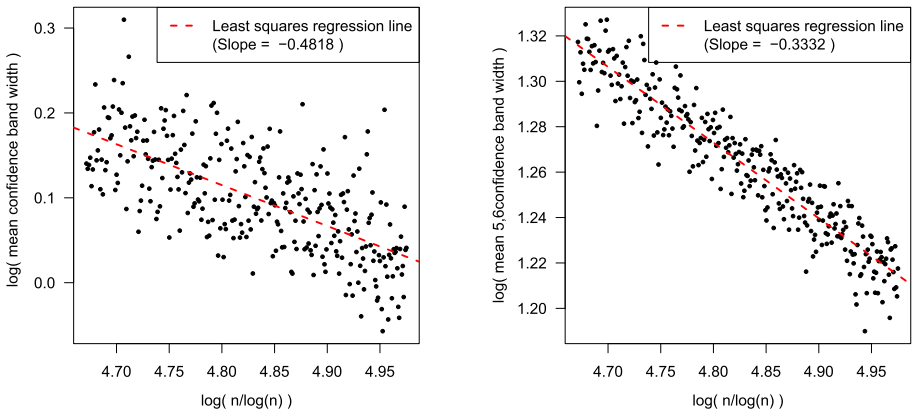
in (6), with known noise level $\sigma = 1$, with target coverage level $1 - \delta = 0.9$, and with $\epsilon_{\mathsf{iso}} = 0$ as the signal x is known to be monotone. For sample size $n = 1000$, the resulting estimate $\mathsf{iso}(y)$ and confidence band are illustrated in Figure 1(b).

We then repeat this experiment at sample sizes $n = 700, 701, 702, \ldots, 1000$. For each sample size n, we take the mean width of the confidence band averaged over (a) the locally constant ("flat") regions of the signal, defined by all indices i corresponding to values $t \in [0.1, 0.2] \cup [0.8, 0.9]$, and (b) a strictly increasing region, defined by indices i corresponding to $t \in [0.4, 0.6]$. (These regions are illustrated in Figure 1(a).)

Our theory predicts that the mean confidence band width scales as $\sim \sqrt{\frac{\log(n)}{n}}$ in the flat regions, and $\sim \sqrt[3]{\frac{\log(n)}{n}}$ in the increasing region. To test this, we take a linear regression of the log of the mean confidence band width against $\log\left(\frac{n}{\log(n)}\right)$, and find a slope $\approx -1/2$ in the flat regions and $\approx -1/3$ in the increasing region, confirming our theory. These results are illustrated in Figure 2.

Note that our data-adaptive estimation bands given by Theorem 3 are calculated without using prior knowledge of the signal's local behavior (locally constant / locally Lipschitz)—the confidence bands computed in Theorem 3 are able to adapt to this unknown structure automatically.

We next check the empirical coverage level of these confidence bands. Ideally we would want to see that, over repeated simulations, the true monotone sequence $x = (x_1, \ldots, x_n)$ lies entirely in the band roughly $1-\delta = 90\%$ of the time. While our theory guarantees that coverage will hold with probability *at least* 90%, our bounds are of course somewhat conservative. We observe empirically that the coverage is in fact too high—it is essentially 100%—but nonetheless, the width of the confidence band is not too conservative. In particular, shrink-



(a) Results for flat regions (b) Results for increasing region

FIG 2. *For each sample size* $700 \le n \le 1000$*, log mean width of the confidence band over a region. (a) Flat region:* $t \in [0.1, 0.2] \cup [0.8, 0.9]$*, where slope* $\approx -1/2$*, i.e. pointwise error scales as* $(n/\log(n))^{-1/2}$*, as predicted in* (12)*. (b) Increasing region:* $t \in [0.4, 0.6]$*, where slope* $\approx -1/3$*, as predicted in* (11)*.*

ing the width of the confidence band by a factor of ≈ 0.855 empirically leads to achieving the target 90% coverage level; in other words, our confidence bands are around 17% too wide. (Of course, this ratio is specific to our choice of data distribution, and is likely to vary in different settings.)

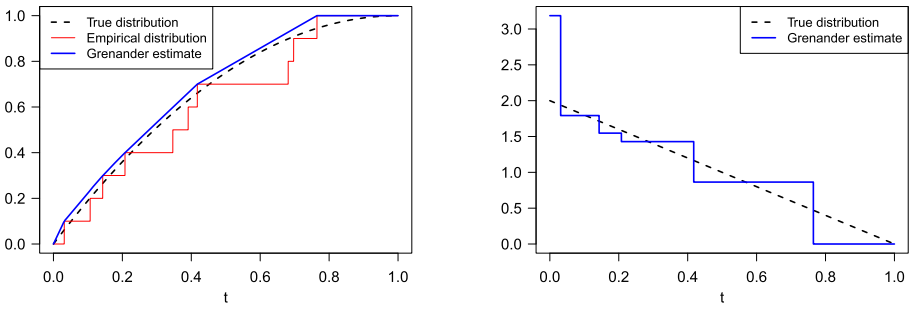
5. Density estimation

As a second application of the tools developed for the sliding window norm, we consider the problem of estimating a monotone nonincreasing density g on the interval $[0, 1]$, using n samples drawn from this density.

Let $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} g$ be n samples drawn from the target density f, sorted into an ordered list $Z_{(1)} \le \cdots \le Z_{(n)}$. The Grenander estimator for the monotone density g is defined as follows. Let $\widehat{G}_n$ be the empirical cumulative distribution function for this sample,

$$\widehat{G}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{Z_i \le t\},$$

and let $\widehat{G}_{\mathsf{Gren}}$ be the minimal concave upper bound on $\widehat{G}_n$. Finally, define the Grenander estimator of the density, denoted by $\widehat{g}_{\mathsf{Gren}}$, as the left-continuous piecewise constant first derivative of $\widehat{G}_{\mathsf{Gren}}$. This process is illustrated in Figure 3. It is known (Robertson et al. [21]) that $\widehat{g}_{\mathsf{Gren}}$ can be computed with a simple isotonic projection of a sequence. Namely, for $i = 1, \dots, n$, let $y_i = n(Z_{(i)} - Z_{(i-1)})$ where we set $Z_{(0)} := 0$, and calculate the isotonic projection $\mathsf{iso}(y)$.



(a) Cumulative distribution function (b) Density

FIG 3. *Illustration of the Grenander estimator for a monotone decreasing density.*

Then the Grenander estimator is given by

$$\widehat{g}_{\mathsf{Gren}} = \begin{cases} 1/\mathsf{iso}(y)_1, & 0 \le t \le Z_{(1)}, \\ 1/\mathsf{iso}(y)_2, & Z_{(1)} < t \le Z_{(2)}, \\ \dots & \\ 1/\mathsf{iso}(y)_n, & Z_{(n-1)} < t \le Z_{(n)}, \\ 0, & Z_{(n)} < t \le 1. \end{cases} \tag{13}$$

If we assume that f is Lipschitz and lower-bounded, then our error bounds for isotonic regression transfer easily into this setting, yielding the following theorem (proved in Appendix A.2):

Theorem 6. *Let $g : [0,1] \to [c, \infty)$ be a nonincreasing L-Lipschitz density, let $Z_1, \dots, Z_n$ be i.i.d. draws from g, and define the Grenander estimator $\widehat{g}_{\mathsf{Gren}}$ as in (13). Then for any $\delta > 0$, if*

$$\Delta := 9\left(\frac{1}{c} + \frac{L}{2c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \le \frac{1}{c+L},$$

then

$$\mathbb{P}\left\{\sup_{\Delta \le t \le 1-\Delta} \left|g(t) - \widehat{g}_{\mathsf{Gren}}(t)\right| \le \frac{\Delta}{\frac{1}{c+L} \cdot \left(\frac{1}{c+L} - \Delta\right)}\right\} \ge 1 - \delta.$$

This result is similar to that of Durot et al. [12], which also obtains a $(n/\log(n))^{-1/3}$ convergence rate uniformly over t (although in their work, t is allowed to be slightly closer to the endpoints, by a log factor). Their results are asymptotic, while our work gives a finite-sample guarantee. As mentioned in Section 1.1, Cator [8] also derives locally adaptive error bounds whose scaling depends on the local Lipschitz behavior or local derivatives of f. Our locally adaptive results for sequences may also be applied here to obtain a locally adaptive confidence band on the density g, but we do not give details here.

6. Proof for contractive isotonic projection

In this section, we prove our main result Theorem 1 showing that, for any seminorm, the nonincreasing-under-neighbor-averaging (NUNA) property is necessary and sufficient for isotonic projection to be contractive under this semi-norm.

Before proving the theorem, we introduce a few definitions. First, for any index $i = 1, \dots, n-1$, we define the matrix

$$A_i = \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & \mathbf{I}_{n-i-1} \end{pmatrix} \in \mathbb{R}^{n \times n}, \tag{14}$$

which averages entries i and $i+1$. That is,

$$A_i x = \left(x_1, \dots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, \dots, x_n \right).$$

We also define an algorithm for isotonic projection that differs from PAVA, and in fact does not converge in finite time, but is useful for the purpose of theoretical analysis. For any $x \in \mathbb{R}^n$ and any index $i = 1, \dots, n-1$, define

$$\mathsf{iso}_i(x) = \begin{cases} x, & \text{if } x_i \le x_{i+1}, \\ A_i x, & \text{if } x_i > x_{i+1}. \end{cases}$$

In other words, if neighboring entries i and $i+1$ violate the monotonicity constraint, then we average them. The following lemma shows that, by iterating this step infinitely many times (while cycling through the indices $i = 1, \dots, n-1$), we converge to the isotonic projection of x.

Lemma 5. *Fix any $x = x^{(0)} \in \mathbb{R}^n$, and define*

$$x^{(t)} = \mathsf{iso}_{i_t}(x^{(t-1)}) \text{ where } i_t = 1 + \mathrm{mod}(t-1, n-1) \text{ for } t = 1, 2, 3, \dots. \tag{15}$$

Then

$$\lim_{t \to \infty} x^{(t)} = \mathsf{iso}(x).$$

With this slow projection algorithm in place, we turn to the proof of our theorem.

Proof of Theorem 1. First suppose that $\|\cdot\|$ satisfies the NUNA property. We will prove that, for any $x, y \in \mathbb{R}^n$ and any index $i = 1, \dots, n-1$,

$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| \le \|x - y\|. \tag{16}$$

If this is true, then by Lemma 5, this is sufficient to see that isotonic projection is contractive with respect to $\|\cdot\|$, since the map $x \mapsto \mathsf{iso}(x)$ is just a composition of (infinitely many) steps of the form $x \mapsto \mathsf{iso}_i(x)$. More concretely, defining $x^{(t)}$ and $y^{(t)}$ as in Lemma 5, (16) proves that $\|x^{(t)} - y^{(t)}\| \le \|x^{(t-1)} - y^{(t-1)}\|$ for

each $t \geq 1$. Applying this inductively proves that $\|x^{(t)} - y^{(t)}\| \leq \|x - y\|$ for all $t \geq 1$, and then taking the limit as $t \to \infty$, we obtain $\|\mathsf{iso}(x) - \mathsf{iso}(y)\| \leq \|x - y\|$.

Now we turn to proving (16). We will split into four cases.

- Case 1: $x_i \leq x_{i+1}$ and $y_i \leq y_{i+1}$. In this case, $\mathsf{iso}_i(x) = x$ and $\mathsf{iso}_i(y) = y$, and so trivially,
$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| = \|x - y\|.$$
- Case 2: $x_i > x_{i+1}$ and $y_i > y_{i+1}$. In this case, we have
$$\big[\mathsf{iso}_i(x)\big]_i = \big[\mathsf{iso}_i(x)\big]_{i+1} = \frac{x_i + x_{i+1}}{2}$$
and
$$\big[\mathsf{iso}_i(y)\big]_i = \big[\mathsf{iso}_i(y)\big]_{i+1} = \frac{y_i + y_{i+1}}{2},$$
while all entries $j \notin \{i, i+1\}$ are unchanged. Therefore, we can write
$$\mathsf{iso}_i(x) - \mathsf{iso}_i(y) = A_i \cdot (x - y).$$
Since $\|\cdot\|$ satisfies the NUNA property, therefore,
$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| = \|A_i \cdot (x - y)\| \leq \|x - y\|.$$
- Case 3: $x_i \leq x_{i+1}$ and $y_i > y_{i+1}$. Let
$$t = \frac{y_i - y_{i+1}}{x_{i+1} - x_i + y_i - y_{i+1}}.$$
Note that $t \in [0, 1]$ by the definition of this case. A trivial calculation shows that
$$\big[\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\big]_i = x_i - \frac{y_i + y_{i+1}}{2} = (1 - t/2) \cdot (x_i - y_i) + t/2 \cdot (x_{i+1} - y_{i+1})$$
and
$$\begin{aligned}\big[\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\big]_{i+1} &= x_{i+1} - \frac{y_i + y_{i+1}}{2} \\ &= t/2 \cdot (x_i - y_i) + (1 - t/2) \cdot (x_{i+1} - y_{i+1})\end{aligned}$$
This means that we have
$$\mathsf{iso}_i(x) - \mathsf{iso}_i(y) = (1 - t) \cdot (x - y) + t \cdot A_i \cdot (x - y),$$
and so
$$\|\mathsf{iso}_i(x) - \mathsf{iso}_i(y)\| \leq (1 - t) \cdot \|x - y\| + t \cdot \|A_i \cdot (x - y)\| \leq \|x - y\|,$$
since $\|\cdot\|$ satisfies NUNA.

- Case 4: $x_i > x_{i+1}$ and $y_i \le y_{i+1}$. By symmetry, this is equivalent to Case 3.

This proves (16), and therefore, is sufficient to show that isotonic projection is a contraction with respect to $\|\cdot\|$.

Now we prove the converse. Suppose that $\|\cdot\|$ does not satisfy NUNA. Then we can find some x and some i such that

$$\|A_i x\| > \|x\|.$$

Without loss of generality we can assume $x_i \le x_{i+1}$ (otherwise simply replace x with $-x$—since $\|\cdot\|$ is a norm, we will have $\|-A_i x\| = \|A_i x\| > \|x\| = \|-x\|$).

Let $B = \max_{1\le j\le n-1} |x_j - x_{j+1}|$, and let $\Delta = x_{i+1} - x_i \in [0, B]$.

Now define

$$y = (\Delta-(i-1)B, \Delta-(i-2)B, \dots, \Delta-B, \underbrace{\Delta}_{\text{entry } i}, \underbrace{0}_{\text{entry } i+1}, B, 2B, \dots, (n-i-1)B),$$

and $z = y + x$. We can check that $\mathsf{iso}(z) = z$, since

$$z_{j+1} - z_j = \begin{cases} x_{j+1} - x_j + B \ge 0, & \text{if } j \ne i, \\ x_{i+1} - x_i - \Delta = 0, & \text{if } j = i. \end{cases}$$

On the other hand, using the fact that $0 \le \Delta \le B$, we have

$$\mathsf{iso}(y) = \Big(\Delta - (i-1)B, \Delta - (i-2)B, \dots, \Delta - B, \frac{\Delta}{2}, \frac{\Delta}{2}, B, 2B, \dots, (n-i-1)B\Big),$$

and so

$$\begin{aligned} \mathsf{iso}(z) - \mathsf{iso}(y) &= (z-y) + (y - \mathsf{iso}(y)) = x + \Big(0, \dots, 0, \frac{\Delta}{2}, -\frac{\Delta}{2}, 0, \dots, 0\Big) \\ &= \Big(x_1, x_2, \dots, x_{i-1}, \frac{x_i + x_{i+1}}{2}, \frac{x_i + x_{i+1}}{2}, x_{i+2}, x_{i+3}, \dots, x_n\Big) = A_i x. \end{aligned}$$

Therefore, $\|\mathsf{iso}(z) - \mathsf{iso}(y)\| > \|z - y\|$, proving that isotonic projection is not contractive with respect to $\|\cdot\|$. □

7. Discussion

We study contraction properties of isotonic regression with an application on a novel sliding window norm. We then use these tools to construct data-adaptive estimation bands and obtain non-asymptotic uniform estimation bound of isotonic estimator. Our results are adaptive to the local behavior of the unknown

signal, and can be used in a related density estimation problem. The analysis tools we developed are potentially useful for other shape-restricted problems.

We expect to apply our results on the high dimensional inference or calibration problems, where isotonic regression could serve as an important tool. Furthermore, ordering constraints, and more generally shape constraints, can take many different forms in various applications—for instance, two commonly studied constraints are isotonicity over a two-dimensional grid (in contrast to the one-dimensional ordering studied here), and convexity or concavity. It may be possible to extend the contraction results to a more general shape-constrained regression setting. We leave these problems for future work.

Appendix A: Additional proofs

A.1. Proof of ℓ_2 error rate (Theorem 5)

First we define the cube $A = [\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]^n \subset \mathbb{R}^n$, and let $z = \mathcal{P}_A(\mathsf{iso}(y))$ be the projection of $\mathsf{iso}(y)$ to this cube, which is computed by truncating each entry $\mathsf{iso}(y)_i$ to the range $[\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]$. Note that $\mathsf{iso}(x) + z$ is now a monotone vector with range given by $(\mathsf{iso}(x) + z)_n - (\mathsf{iso}(x) + z)_1 \le 2V$. Now, fix any integer $M \ge 1$ and find integers

$$0 = k_0 < \cdots < k_M = n$$

such that

$$\left|(\mathsf{iso}(x) + z)_{k_m} - (\mathsf{iso}(x) + z)_{k_{m-1}+1}\right| \le \frac{2V}{M} \text{ for all } m = 1, \ldots, M,$$

which we can find since the total variation of the vector $\mathsf{iso}(x) + z$ is bounded by $2V$ (Lemma 11.1 in Chatterjee et al. [9]).

For each $m = 1, \ldots, M$, let $I_m = \{k_{m-1} + 1, \ldots, k_m\}$ be the set of indices in the mth bin. We can calculate

$$\begin{aligned}\max_{i \in I_m}(z - \mathsf{iso}(x))_i - \min_{i \in I_m}(z - \mathsf{iso}(x))_i &\le \max_{i \in I_m} z_i + \max_{i \in I_m} \mathsf{iso}(x)_i - \min_{i \in I_m} z_i - \min_{i \in I_m} \mathsf{iso}(x)_i \\ &= (\mathsf{iso}(x) + z)_{k_m} - (\mathsf{iso}(x) + z)_{k_{m-1}+1} \le \frac{2V}{M}.\end{aligned}$$

This implies that

$$\|(z - \mathsf{iso}(x))_{I_m} - \overline{(z - \mathsf{iso}(x))}_{I_m} \cdot \mathbf{1}_{I_m}\|_2 \le \frac{2V}{M} \cdot \sqrt{k_m - k_{m-1}}.$$

We also have

$$\left|\overline{z}_{I_m} - \overline{\mathsf{iso}(x)}_{I_m}\right| \le \frac{\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}}}{\sqrt{k_m - k_{m-1}}}$$

by our choice of the sliding window norm. Next, by the triangle inequality we have

$$\begin{aligned}
&\|z_{I_m} - \mathsf{iso}(x)_{I_m}\|_2^2 \\
&\le \left(\|\overline{z}_{I_m} \cdot \mathbf{1}_{I_m} - \overline{\mathsf{iso}(x)}_{I_m} \cdot \mathbf{1}_{I_m}\|_2 + \|(z - \mathsf{iso}(x))_{I_m} - \overline{(z - \mathsf{iso}(x))}_{I_m} \cdot \mathbf{1}_{I_m}\|_2\right)^2 \\
&= \left(\left|\overline{z}_{I_m} - \overline{\mathsf{iso}(x)}_{I_m}\right| \cdot \sqrt{k_m - k_{m-1}} + \|(z - \mathsf{iso}(x))_{I_m} - \overline{(z - \mathsf{iso}(x))}_{I_m} \cdot \mathbf{1}_{I_m}\|_2\right)^2 \\
&\qquad \le \left(\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}} + \frac{2V}{M} \cdot \sqrt{k_m - k_{m-1}}\right)^2 \le 2\big(\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2 \\
&\qquad\qquad + \frac{8V^2}{M^2}(k_m - k_{m-1}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|z - \mathsf{iso}(x)\|_2^2 &= \sum_{m=1}^{M} \|z_{I_m} - \mathsf{iso}(x)_{I_m}\|_2^2 \\
&\le 2M\big(\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2 + \frac{8V^2}{M^2}\sum_{m=1}^{M}(k_m - k_{m-1}).
\end{aligned}$$

Since $\sum_{m=1}^{M}(k_m - k_{m-1}) = n$ trivially, we can simplify this to

$$\|z - \mathsf{iso}(x)\|_2^2 \le 2M\big(\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}}\big)^2 + \frac{8V^2 n}{M^2}.$$

Now, since z is the projection of $\mathsf{iso}(y)$ to the range $[\mathsf{iso}(x)_1, \mathsf{iso}(x)_n]$, it follows trivially that $\|z - \mathsf{iso}(x)\|_\psi^{\mathsf{SW}} \le \|\mathsf{iso}(x) - \mathsf{iso}(y)\|_\psi^{\mathsf{SW}}$ and therefore is bounded by $\|x - y\|_\psi^{\mathsf{SW}}$ by our contraction result. Therefore,

$$\|z - \mathsf{iso}(x)\|_2^2 \le 2M\big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{8V^2 n}{M^2}.$$

Next, we have

$$\begin{aligned}
\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 &\le 2\|z - \mathsf{iso}(x)\|_2^2 + 2\|\mathsf{iso}(y) - z\|_2^2 \\
&= 2\|z - \mathsf{iso}(x)\|_2^2 + 2\sum_{i=1}^{n}(z_i - \mathsf{iso}(y)_i)_+^2 + 2\sum_{i=1}^{n}(\mathsf{iso}(y)_i - z_i)_+^2 \\
\le 4M\big(\|x - y\|_\psi^{\mathsf{SW}}\big)^2 &+ \frac{16V^2 n}{M^2} + 2\sum_{i=1}^{n}(\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2 + 2\sum_{i=1}^{n}(\mathsf{iso}(y)_i - \mathsf{iso}(x)_n)_+^2.
\end{aligned}$$

It remains to bound these last two terms. First we bound $\sum_{i=1}^{n}(\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2$. If $\mathsf{iso}(y)_1 \ge \mathsf{iso}(x)_1$ then this term is zero, so now we focus on the case that $\mathsf{iso}(y)_1 < \mathsf{iso}(x)_1$. For any $1 \le j \le i$, we have

$$\mathsf{iso}(x)_j - \mathsf{iso}(y)_j \ge \mathsf{iso}(x)_1 - \mathsf{iso}(y)_j$$

and so

$$\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i \le \overline{\mathsf{iso}(x)}_{1:i} - \overline{\mathsf{iso}(y)}_{1:i} \le \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\sqrt{i}}.$$

Therefore,

$$\sum_{i=1}^n (\mathsf{iso}(x)_1 - \mathsf{iso}(y)_i)_+^2 \le \sum_{i=1}^n \left(\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\sqrt{i}}\right)^2 = \big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 \cdot 2\log(2n),$$

by bounding the harmonic series $1 + \frac{1}{2} + \cdots + \frac{1}{n}$. We also have $\sum_{i=1}^n (\mathsf{iso}(y)_i - \mathsf{iso}(x)_n)_+^2 \le \big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 \cdot 2\log(2n)$ by an identical argument. Combining everything, then,

$$\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 \le 4M\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{16V^2 n}{M^2} + 8\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 \log(2n).$$

Setting

$$M = \lceil M_0 \rceil \text{ where } M_0 = \frac{2V^{2/3} n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}},$$

we obtain

$$\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 \le 4\left(\frac{2V^{2/3}n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}} + 1\right)\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 + \frac{16V^2 n}{\left(\frac{2V^{2/3}n^{1/3}}{\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{2/3}}\right)^2} + 8\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2 \log(2n).$$

After simplifying (and assuming $n \ge 2$ to avoid triviality), this bound becomes

$$\frac{1}{n}\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2 \le \frac{12V^{2/3}\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{4/3}}{n^{2/3}} + \frac{12\log(2n)}{n} \cdot \big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2.$$

Applying Lemma 3, we have $\mathbb{E}\left[\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^2\right] \le 8\sigma^2\log(2n)$ which implies that $\mathbb{E}\left[\big(\|x-y\|_\psi^{\mathsf{SW}}\big)^{4/3}\right] \le (8\sigma^2\log(2n))^{2/3} = 4\sigma^{4/3}\log^{2/3}(2n)$. Plugging this in, we then have

$$\mathbb{E}\left[\frac{1}{n}\|\mathsf{iso}(y) - \mathsf{iso}(x)\|_2^2\right] \le 48\left(\frac{V\sigma^2\log(2n)}{n}\right)^{2/3} + \frac{96\sigma^2\log^2(2n)}{n}.$$

A.2. Proof of density estimation result (Theorem 6)

Let $G(t) = \int_{s=0}^t g(s)\,\mathsf{d}s$ be the cumulative distribution function for the density g. Since $g(t) \ge c$ everywhere, this means that $G(t)$ is strictly increasing, and is

therefore invertible. Using this lower bound on g, and the assumption that g is L-Lipschitz, we can furthermore calculate

$$0 \le (G^{-1})'(t) = \frac{1}{g(G^{-1}(t))} \le \frac{1}{c} \quad \text{and} \quad \big|(G^{-1})''(t)\big| = \left|\frac{-g'(G^{-1}(t))}{(g(G^{-1}(t)))^3}\right| \le \frac{L}{c^3}. \tag{17}$$

Let

$$x_i = n\left(G^{-1}(i/n) - G^{-1}\big((i-1)/n\big)\right) \text{ and } y_i = n\big(Z_{(i)} - Z_{(i-1)}\big)$$

for $i = 1, \dots, n$, where $Z_{(0)} \coloneqq 0$. Note that x gives the difference in quantiles of the distribution, while y estimates these gaps empirically.

The following lemma (proved in Appendix A.3) gives a concentration result on the $Z_{(i)}$'s:

Lemma 6. *Let $Z_{(1)} \le \cdots \le Z_{(n)}$ be the order statistics of $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} g$, where the density $g : [0,1] \to [c, \infty)$ is L-Lipschitz. Then for any $\delta > 0$, with probability at least $1 - \delta$,*

$$\big|Z_{(i)} - G^{-1}(i/n)\big| \le \frac{4}{c} \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}} \tag{18}$$

for all $1 \le i \le n$, and

$$\begin{aligned}&\big|\big(Z_{(i)} - Z_{(j)}\big) - \big(G^{-1}(i/n) - G^{-1}(j/n)\big)\big| \\ &\le \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{cn} + \frac{4L|i-j|\sqrt{\log((n^2+n)/\delta)}}{c^3 n^{3/2}}\end{aligned} \tag{19}$$

for all $1 \le i < j \le n$.

From now on, we assume that these bounds (18) and (19) both hold. Plugging our definitions of x and y into these two bounds, this proves that

$$\begin{aligned}\big|\overline{x}_{i:j} - \overline{y}_{i:j}\big| &\le \frac{\sqrt{3(j-i+1)\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{c \cdot (j-i+1)} \\ &\quad + \frac{4L}{c^3} \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}}\end{aligned}$$

for all $1 \le i \le j \le n$. (If $i = 1$ then we use the bound (18) while if $i > 1$ then we use the bound (19).)

Now, defining

$$\psi(i) = \frac{i}{\frac{1}{c} \cdot \left(\sqrt{3i\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)\right) + \frac{4L}{c^3} \cdot i \cdot \sqrt{\frac{\log((n^2+n)/\delta)}{n}}},$$

we see that

$$\|x - y\|_\psi^{\mathsf{SW}} = \max_{1 \le i \le j \le n} \big|\overline{x}_{i:j} - \overline{y}_{i:j}\big| \cdot \psi(j-i+1) \le 1.$$

(Note that ψ is nondecreasing and $i \mapsto i/\psi(i)$ is concave, as required by (1).)

Next we check that x is a Lipschitz sequence. We have

$$\begin{aligned} n^{-1}(x_{i+1}-x_i) &= \left(G^{-1}\left(\frac{i+1}{n}\right) - G^{-1}(i/n)\right) + \left(G^{-1}((i-1)/n) - G^{-1}(i/n)\right) \\ &= (G^{-1})'(i/n)\cdot\frac{1}{n} + \frac{1}{2}(G^{-1})''\left(\frac{i+s}{n}\right)\cdot\frac{1}{n^2} \\ &\quad + (G^{-1})'(i/n)\cdot -\frac{1}{n} + \frac{1}{2}(G^{-1})''\left(\frac{i-1+t}{n}\right)\cdot\frac{1}{n^2} \end{aligned}$$

for some $s, t \in [0,1]$, by Taylor's theorem. The first-order terms cancel, and we know by (17) that $(G^{-1})''$ is bounded by $\frac{L}{c^3}$. Therefore, x is $\frac{L}{c^3}$-Lipschitz. Finally, x is monotone nondecreasing since g is a monotone nonincreasing density.

We then apply the calculations (10) for the Lipschitz signal setting (with $\frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)}$ taking the place of $\sqrt{\frac{2\sigma^2\log\left(\frac{n^2+n}{\delta}\right)}{m}}$, which was specific to the subgaussian noise model setting). We see that for any index $m \geq 1$,

$$\begin{aligned} &\max_{m\leq k\leq n-m+1} \left|x_k - \mathsf{iso}(y)_k\right| \leq \frac{\|x-y\|_\psi^{\mathsf{SW}}}{\psi(m)} + \frac{L(m-1)}{2nc^3} \\ &\leq \frac{\frac{1}{c}\cdot\left(\sqrt{3m\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)\right) + \frac{4L}{c^3}\cdot m\cdot\sqrt{\frac{\log((n^2+n)/\delta)}{n}}}{m} \\ &\quad + \frac{L(m-1)}{2nc^3}. \end{aligned}$$

Set $m = \left\lceil \left(n\sqrt{\log((n^2+n)/\delta)}\right)^{2/3} \right\rceil$. Since $\Delta < \frac{1}{c+L} \leq 1$ we know $\log((n^2+n)/\delta) \leq n$, so we can simplify the above bound to

$$\max_{m\leq k\leq n-m+1} \left|x_k - \mathsf{iso}(y)_k\right| \leq \left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}.$$

Now we show how this uniform bound on the difference $x-y$, translates to an error bound on the Grenander density estimator $\widehat{g}_{\mathsf{Gren}}$. First, we check that $Z_{(m)} \leq \Delta$ and $Z_{(n-m+1)} \geq 1-\Delta$. We have

$$1 \geq \|x-y\|_\psi^{\mathsf{SW}} \geq \psi(m)\cdot\left|\overline{x}_{1:m} - \overline{y}_{1:m}\right| = \frac{\psi(m)}{m}\cdot n\cdot\left|G^{-1}(m/n) - Z_{(m)}\right|.$$

We also know that $G^{-1}(m/n) \leq \frac{m}{cn}$ since G^{-1} is $(1/c)$-Lipschitz as calculated in (17), and so

$$\begin{aligned} &Z_{(m)} \leq G^{-1}(m/n) + \frac{m}{n\psi(m)} \leq \\ &\frac{m}{cn} + \frac{\sqrt{3m\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{cn} + \frac{4Lm\sqrt{\log((n^2+n)/\delta)}}{c^3n^{3/2}} \end{aligned}$$

$$\leq \left(\frac{5}{c} + \frac{4L}{c^3}\right) \sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \leq \Delta,$$

using the fact that $\log((n^2+n)/\delta) \leq n$ as before. Similarly $Z_{(n-m+1)} \geq 1-\Delta$.

Next, for any t with $\Delta \leq t \leq 1-\Delta$, find index k such that

$$Z_{(k-1)} < t \leq Z_{(k)}.$$

By the work above we will have $m \leq k \leq n-m+1$. Then $\widehat{g}_{\mathsf{Gren}}(t) = \frac{1}{\mathsf{iso}(y)_k}$ by definition of the Grenander estimator. Therefore, we have

$$\left|\widehat{g}_{\mathsf{Gren}}(t) - g(t)\right| = \left|\frac{1}{\mathsf{iso}(y)_k} - g(t)\right| \leq \left|\frac{1}{\mathsf{iso}(y)_k} - \frac{1}{x_k}\right| + \left|\frac{1}{x_k} - g(t)\right|.$$

By Lemma 6, we know

$$Z_{(k)} \leq G^{-1}\left(\frac{k}{n}\right) + \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

and

$$Z_{(k-1)} \geq G^{-1}\left(\frac{k-1}{n}\right) - \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

so we have

$$G^{-1}\left(\frac{k-1}{n}\right) - \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} < t \leq G^{-1}\left(\frac{k}{n}\right) + \frac{4}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}$$

We calculate

$$\begin{aligned} x_k &= n\left(G^{-1}\left(\frac{k}{n}\right) - G^{-1}\left(\frac{k-1}{n}\right)\right) = n(G^{-1})'\left(\frac{k-1+s}{n}\right)\cdot\frac{1}{n} \\ &= \frac{1}{g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right)} \end{aligned}$$

by Taylor's theorem for some $s \in [0,1]$, and so

$$\begin{aligned} \left|\frac{1}{x_k} - g(t)\right| &= \left|g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right) - g(t)\right| \leq L\cdot\left|G^{-1}\left(\frac{k-1+s}{n}\right) - t\right| \\ &\leq L\left|G^{-1}\left(\frac{k}{n}\right) - G^{-1}\left(\frac{k-1}{n}\right)\right| + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \\ &\leq \frac{L}{cn} + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}, \end{aligned}$$

since g is L-Lipschitz and G^{-1} is $(1/c)$-Lipschitz, as proved before. Finally,

$$\begin{aligned} &\left|\frac{1}{\mathsf{iso}(y)_k} - \frac{1}{x_k}\right| = \frac{|x_k - \mathsf{iso}(y)_k|}{x_k\cdot\mathsf{iso}(y)_k} \\ &\leq \frac{\left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{x_k\cdot\mathsf{iso}(y)_k} \leq \frac{\left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{x_k\cdot\left(x_k - \left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)}, \end{aligned}$$

from the bound on $\left|x_k - \mathsf{iso}(y)_k\right|$ above. And, we know that $x_k = \frac{1}{g\left(G^{-1}\left(\frac{k-1+s}{n}\right)\right)}$ for some $s \in [0,1]$ as above, so $x_k \geq \frac{1}{\max_{s\in[0,1]} g(s)}$. Now, since g is lower-bounded by c and is L-Lipschitz, we see that $g(s) \leq c + L$, and so $x_k \geq \frac{1}{c+L}$. Combining everything,

$$\begin{aligned}
&\left|\widehat{g}_{\mathsf{Gren}}(t) - g(t)\right| \leq \\
&\frac{\left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{\frac{1}{c+L}\cdot\left(\frac{1}{c+L} - \left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)} + \frac{L}{cn} + \frac{4L}{c}\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}} \\
&\qquad \leq \frac{\left(\frac{1}{c}\left(4 + \frac{5L}{c+L}\right) + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}}{\frac{1}{c+L}\cdot\left(\frac{1}{c+L} - \left(\frac{4}{c} + \frac{4.5L}{c^3}\right)\sqrt[3]{\frac{\log((n^2+n)/\delta)}{n}}\right)} \leq \frac{\Delta}{\frac{1}{c+L}\left(\frac{1}{c+L} - \Delta\right)}.
\end{aligned}$$

A.3. Proofs of lemmas

Proof of Lemma 2. By Theorem 1, we only need to prove that $\|\cdot\|_\psi^{\mathsf{SW}}$ satisfies NUNA. Fix any $x \in \mathbb{R}^n$ and any index $k = 1, \ldots, n-1$. Let $y = A_k x$. Note that we have

$$y_i = \begin{cases} x_i, & \text{if } i < k \text{ or } i > k+1, \\ \frac{x_k + x_{k+1}}{2}, & \text{if } i = k \text{ or } i = k+1. \end{cases}$$

Take any indices $1 \leq i \leq j \leq n$. We need to prove that $|\overline{y}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_\psi^{\mathsf{SW}}$.

- Case 1: if $j < k$ or if $i > k+1$, then neither of the indices $k, k+1$ are included in the window $i:j$, and therefore $x_{i:j} = y_{i:j}$ (i.e. all entries in the stretch of indices $i:j$ are equal). So,

$$|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{x}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_\psi^{\mathsf{SW}}.$$

- Case 2: If $i \leq k$ and $j \geq k+1$, then both indices $k, k+1$ are included in the window $i:j$. Since $y_k + y_{k+1} = x_k + x_{k+1}$ and all other entries of x and y coincide, we can trivially see that

$$|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{x}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_\psi^{\mathsf{SW}}.$$

- Case 3: if $i < k$ and $j = k$, then

$$\begin{aligned}
&|\overline{y}_{i:j}| \cdot \psi(j-i+1) = |\overline{y}_{i:k}| \cdot \psi(k-i+1) \\
&= \frac{\left|\sum_{\ell=i}^{k-1} x_\ell + \frac{x_k + x_{k+1}}{2}\right| \cdot \psi(k-i+1)}{k-i+1} \\
&= \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k-1} x_\ell + \frac{1}{2}\sum_{\ell=i}^{k+1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k-1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1} + \frac{\left|\frac{1}{2}\sum_{\ell=i}^{k+1} x_\ell\right| \cdot \psi(k-i+1)}{k-i+1} \\
&= \frac{\psi(k-i+1)}{k-i+1} \cdot \left(\frac{1}{2}|\overline{x}_{i:(k-1)}| \cdot \psi(k-i) \cdot \frac{k-i}{\psi(k-i)}\right. \\
&\left. + \frac{1}{2}|\overline{x}_{i:(k+1)}| \cdot \psi(k-i+2) \cdot \frac{k-i+2}{\psi(k-i+2)}\right) \\
&\leq \|x\|_\psi^{\mathsf{SW}} \cdot \frac{1}{2}\left[\frac{k-i}{\psi(k-i)} + \frac{k-i+2}{\psi(k-i+2)}\right] \cdot \frac{\psi(k-i+1)}{k-i+1} \\
&\leq \|x\|_\psi^{\mathsf{SW}} \cdot \frac{k-i+1}{\psi(k-i+1)} \cdot \frac{\psi(k-i+1)}{k-i+1} = \|x\|_\psi^{\mathsf{SW}},
\end{aligned}
$$

where the last inequality holds since $i \mapsto i/\psi(i)$ is concave by assumption on ψ.

- Case 4: if $i = k+1$ and $j > k+1$, by symmetry this case is analogous to Case 3.
- Case 5: if $i = j = k$, then

$$
\begin{aligned}
|\overline{y}_{i:j}| \cdot \psi(j-i+1) &= |y_k| \cdot \psi(1) = |x_k + x_{k+1}| \cdot \psi(1)/2 \\
&= |\overline{x}_{k:(k+1)}| \cdot \psi(1) \leq |\overline{x}_{k:(k+1)}| \cdot \psi(2) \leq \|x\|_\psi^{\mathsf{SW}},
\end{aligned}
$$

since $\psi(1) \leq \psi(2)$ due to the assumption that ψ is nondecreasing.
- Case 6: if $i = j = k+1$, then by symmetry this case is analogous to Case 5.

Therefore, $|\overline{y}_{i:j}| \cdot \psi(j-i+1) \leq \|x\|_\psi^{\mathsf{SW}}$ for all indices $1 \leq i \leq j \leq n$, and so $\|y\|_\psi^{\mathsf{SW}} \leq \|x\|_\psi^{\mathsf{SW}}$, as desired. □

Proof of Lemma 3. For any indices $1 \leq i \leq j \leq n$,

$$
\overline{y}_{i:j} - \overline{x}_{i:j} = \sigma \overline{\epsilon}_{i:j},
$$

and we know that $\sqrt{j-i+1} \cdot \overline{\epsilon}_{i:j}$ is subgaussian, that is,

$$
\mathbb{P}\left\{\sqrt{j-i+1} \cdot |\overline{\epsilon}_{i:j}| > t\right\} \leq 2e^{-t^2/2}
$$

for any $t \geq 0$. Now we set $t = \sqrt{2\log\left(\frac{n^2+n}{\delta}\right)}$, and take a union bound over all $n + \binom{n}{2} = \frac{n^2+n}{2}$ possible pairs of indices $i \leq j$. We then have

$$
\mathbb{P}\left\{\max_{1 \leq i \leq j \leq n} \sqrt{j-i+1} \cdot |\overline{\epsilon}_{i:j}| \leq \sqrt{2\log\left(\frac{n^2+n}{\delta}\right)}\right\} \geq 1-\delta.
$$

Setting $\psi(t) = \sqrt{t}$ proves that, on this event, $\|x-y\|_\psi^{\mathsf{SW}} \leq \sigma\sqrt{2\log\left(\frac{n^2+n}{\delta}\right)}$, as desired. For the bound in expectation, we have a similar calculation: it is known that $\mathbb{E}\left[\max_{k=1,\ldots,N} |Z_k|\right] \leq \sqrt{2\log(2N)}$ and $\mathbb{E}\left[\max_{k=1,\ldots,N} |Z_k|^2\right] \leq 8\log(2N)$

for any (not necessarily independent) subgaussian random variables Z_k. Setting $Z_k = \sqrt{j-i+1} \cdot \overline{\epsilon}_{i:j}$ for each of the $N = \frac{n^2+n}{2}$ possible pairs i, j, we obtain

$$\mathbb{E}\left[\max_{1 \le i \le j \le n} \sqrt{j-i+1} \cdot \left|\overline{\epsilon}_{i:j}\right|\right] \le \sqrt{2\log(n^2+n)}$$

and

$$\mathbb{E}\left[\left(\max_{1 \le i \le j \le n} \sqrt{j-i+1} \cdot \left|\overline{\epsilon}_{i:j}\right|\right)^2\right] \le 8\log(n^2+n). \qquad \square$$

Proof of Lemma 4. Assume that x is ϵ_{iso}-monotone, and fix any index $1 \le i \le n$. Let $j = \max\{k \le n : \mathsf{iso}(x)_k = \mathsf{iso}(x)_i\}$. Then $i \le j \le n$, $\mathsf{iso}(x)_i = \mathsf{iso}(x)_j$, and either $j = n$ or $\mathsf{iso}(x)_j < \mathsf{iso}(x)_{j+1}$. Therefore, we must have $x_j \le \mathsf{iso}(x)_j$ by properties of the isotonic projection. (This is because, if $x_j > \mathsf{iso}(x)_j$, then writing $\mathbf{e}_j$ for the jth basis vector and taking some sufficiently small $\epsilon > 0$, the vector $\mathsf{iso}(x) + \epsilon \cdot \mathbf{e}_j$ is an isotonic vector that is strictly closer to x than $\mathsf{iso}(x)$, which is a contradiction.) Therefore, $x_i \le x_j + \epsilon_{\mathsf{iso}} \le \mathsf{iso}(x)_j + \epsilon_{\mathsf{iso}} = \mathsf{iso}(x)_i + \epsilon_{\mathsf{iso}}$. The reverse bound is proved similarly.

Now we turn to the converse. For any $1 \le i \le j \le n$, we have $x_i \le \mathsf{iso}(x)_i + \epsilon \le \mathsf{iso}(x)_j + \epsilon \le x_j + 2\epsilon$, where the first and third inequalities use the bound $\|x - \mathsf{iso}(x)\|_\infty \le \epsilon$, while the second uses the fact that $\mathsf{iso}(x)$ is monotone. $\square$

Proof of Lemma 5. For $i = 1, \ldots, n-1$, let $\mathcal{K}_i = \{x \in \mathbb{R}^n : x_i \le x_{i+1}\}$, which is a closed convex cone in $\mathbb{R}^n$. We have $\mathcal{K}_{\mathsf{iso}} = \cap_{i=1}^{n-1} \mathcal{K}_i$ and it's easy to see that $\mathsf{iso}_i(x) = \mathcal{P}_{\mathcal{K}_i}(x)$. Hence the slow projection algorithm defined in (15) is actually a cyclic projection algorithm, that is, the iterates are given by

$$x^{(1)} = \mathcal{P}_{\mathcal{K}_1}(x^{(0)}), \quad x^{(2)} = \mathcal{P}_{\mathcal{K}_2}(x^{(1)}), \quad \ldots, \quad x^{(n)} = \mathcal{P}_{\mathcal{K}_1}(x^{(n-1)}), \quad \ldots.$$

In general, it is known that a cyclic projection algorithm starting at some point $x = x^{(0)}$ is guaranteed to converge to some point in the intersection of the respective convex sets, i.e. $\lim_{t\to\infty} x^{(t)} = x^* \in \cap_{i=1}^{n-1} \mathcal{K}_i = \mathcal{K}_{\mathsf{iso}}$, but without any assumptions on the nature of the convex sets $\mathcal{K}_i$, this point may not necessarily be the projection of x onto the intersection of the sets (Bregman [5], Han [16]). Therefore, we need to check that for our specific choice of the sets $\mathcal{K}_i$, the cyclic projection algorithm (15) in fact converges to $\mathsf{iso}(x)$ as claimed in the lemma.

We first claim that

$$\mathsf{iso}\big(\mathsf{iso}_i(x)\big) = \mathsf{iso}(x) \tag{20}$$

for all $x \in \mathbb{R}^n$ and all $i = 1, \ldots, n-1$. Assume for now that this is true. Since $\mathsf{iso}(\cdot)$ is contractive with respect to the ℓ_2 norm, the convergence $x^{(t)} \to x^*$ implies that $\mathsf{iso}(x^{(t)}) \to \mathsf{iso}(x^*)$. Applying (20) inductively, we know that $\mathsf{iso}(x^{(t)}) = \mathsf{iso}(x^{(0)}) = \mathsf{iso}(x)$ for all $t \ge 1$. On the other hand, since $x^* \in \mathcal{K}_{\mathsf{iso}}$, this means that $x^* = \mathsf{iso}(x^*)$. Combining everything, then, we obtain

$$\lim_{t\to\infty} x^{(t)} = x^* = \mathsf{iso}(x^*) = \lim_{t\to\infty} \mathsf{iso}(x^{(t)}) = \mathsf{iso}(x).$$

Finally, we need to prove (20). Fix any index i and any $x \in \mathbb{R}^n$. If $x_i \le x_{i+1}$, then $\mathsf{iso}_i(x) = x$ and the statement holds trivially. If not, then $x_i > x_{i+1}$ and

we have $\mathsf{iso}_i(x) = A_i x$ (recalling the definition of A_i in (14) earlier). Now let $y = \mathsf{iso}(x)$ and $z = \mathsf{iso}(A_i x)$. It is trivially true that, since $x_i > x_{i+1}$, we must have $y_i = y_{i+1}$. Also, $\langle x - y, z - y \rangle \le 0$ by properties of projection to the convex set $\mathcal{K}_{\mathsf{iso}}$, so we can calculate

$$\begin{aligned} \langle A_i x - y, z - y \rangle &= \langle x - y, z - y \rangle + \frac{x_{i+1} - x_i}{2} \cdot (z_i - y_i - z_{i+1} + y_{i+1}) \\ &\le \frac{x_{i+1} - x_i}{2} \cdot (z_i - z_{i+1}) \le 0, \end{aligned} \tag{21}$$

where the last step holds since $z_i \le z_{i+1}$ due to $z \in \mathcal{K}_{\mathsf{iso}}$ and $x_i \ge x_{i+1}$ by assumption. We also have $\|A_i x - z\|_2^2 \le \|A_i x - y\|_2^2$ since $z = \mathsf{iso}(A_i x)$, which combined with (21) proves that $y = z$. Thus (20) holds, as desired. □

Proof of Lemma 6. Let $U_1, \dots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, and let $G(t) = \int_{s=0}^{t} g(s)\, \mathsf{d}s$ be the cumulative distribution function for the density g. Since $g \ge c > 0$, $G : [0,1] \to [0,1]$ is strictly increasing, and is therefore invertible. It is known that setting $Z_{(i)} = G^{-1}(U_{(i)})$ recovers the desired distribution for the ordered sample points $Z_{(1)} \le \cdots \le Z_{(n)}$.

Next, by Lemma 8 below, with probability at least $1 - \delta$, for all indices $0 \le i < j \le n$,

$$\left| U_{(i)} - U_{(j)} - \frac{i-j}{n} \right| \le \frac{\sqrt{3|i-j| \log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n}. \tag{22}$$

From this point on, assume that this bound holds. In particular, by taking $i = 0$, this implies that

$$\left| U_{(j)} - \frac{j}{n} \right| \le \frac{\sqrt{3j \log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n} \le 4\sqrt{\frac{\log((n^2+n)/\delta)}{n}}, \tag{23}$$

for all $j = 1, \dots, n$, by assuming that $\log((n^2+n)/\delta) \le n$ (if not, then this bound holds trivially since $U_{(j)}$ and j/n both lie in $[0,1]$).

Then, since g is L-Lipschitz, for $1 \le i < j \le n$ we compute

$$\begin{aligned} &\left| \left(Z_{(i)} - Z_{(j)}\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right) \right| \\ &\quad = \left| \left(G^{-1}(U_{(i)}) - G^{-1}(U_{(j)})\right) - \left(G^{-1}(i/n) - G^{-1}(j/n)\right) \right| \\ &= \left| (U_{(i)} - U_{(j)}) \cdot (G^{-1})' \left(sU_{(i)} + (1-s)U_{(j)}\right) - \frac{i-j}{n} \cdot (G^{-1})' \left(\frac{si + (1-s)j}{n} \right) \right|, \end{aligned}$$

where the last step holds by Taylor's theorem applied to the function

$$s \mapsto \left(G^{-1}\left(sU_{(i)} + (1-s)U_{(j)}\right) - G^{-1}\left(\frac{si + (1-s)j}{n} \right) \right).$$

We can rewrite this as

$$\begin{aligned}&\left|\left(Z_{(i)}-Z_{(j)}\right)-\left(G^{-1}(i/n)-G^{-1}(j/n)\right)\right| \\ &\quad\le\left|\left(U_{(i)}-U_{(j)}\right)-\frac{i-j}{n}\right|\cdot\left(G^{-1}\right)'\left(sU_{(i)}+(1-s)U_{(j)}\right)+ \\ &\quad\frac{j-i}{n}\cdot\left|\left(G^{-1}\right)'\left(sU_{(i)}+(1-s)U_{(j)}\right)-\left(G^{-1}\right)'\left(\frac{si+(1-s)j}{n}\right)\right|,\end{aligned}$$

Since G^{-1} has bounded first and second derivatives as in (17), we then have

$$\begin{aligned}&\left|\left(Z_{(i)}-Z_{(j)}\right)-\left(G^{-1}(i/n)-G^{-1}(j/n)\right)\right| \\ &\le\left|\left(U_{(i)}-U_{(j)}\right)-\frac{i-j}{n}\right|\cdot\frac{1}{c}+\frac{j-i}{n}\cdot\frac{L}{c^3}\cdot\left|\left(sU_{(i)}+(1-s)U_{(j)}\right)-\frac{si+(1-s)j}{n}\right| \\ &\le\frac{\sqrt{3(j-i)\log((n^2+n)/\delta)}+2\log((n^2+n)/\delta)}{n}\cdot\frac{1}{c}+\frac{j-i}{n}\cdot\frac{4L}{c^3} \\ &\quad\cdot\sqrt{\frac{\log((n^2+n)/\delta)}{n}},\end{aligned}$$

applying the bounds obtained above in (22) and (23). This proves the bound (19) in the lemma. To prove the simpler bound (18), we calculate

$$\begin{aligned}\left|Z_{(i)}-G^{-1}(i/n)\right|=\left|G^{-1}(U_{(i)})-G^{-1}(i/n)\right|&\le\frac{1}{c}\left|U_{(i)}-i/n\right| \\ &\le\frac{4}{c}\sqrt{\frac{\log((n^2+n)/\delta)}{n}},\end{aligned}$$

since G^{-1} is $(1/c)$-Lipschitz and we can apply (23). □

Lemma 7. *Let $U_{(1)}\le\cdots\le U_{(n)}$ be the order statistics of $U_1,\ldots,U_n\overset{\text{iid}}{\sim}\text{Unif}[0,1]$. For any $\delta>0$,*

$$\mathbb{P}\left\{\left|U_{(i)}-\frac{i}{n}\right|\le\frac{\sqrt{3i\log(2n/\delta)}+2\log(2n/\delta)}{n}\ \textit{for all}\ i=1,\ldots,n\right\}\ge1-\delta.$$

Proof of Lemma 7. Fix any index i. If $i<3\log(2n/\delta)$, then

$$\frac{i}{n}-\frac{\sqrt{3i\log(2n/\delta)}}{n}\le0$$

and so trivially we have $U_{(i)}\ge\frac{i}{n}-\frac{\sqrt{3i\log(2n/\delta)}}{n}$. If instead $i\ge3\log(2n/\delta)$, then suppose that $U_{(i)}\le\frac{i}{n}-\frac{\sqrt{3i\log(2n/\delta)}}{n}=:p$. This means that at least i many of the U_k's lie in the interval $[0,p]$. Then

$$\mathbb{P}\left\{U_{(i)}\le\frac{i}{n}-\frac{\sqrt{3i\log(2n/\delta)}}{n}\right\}=\mathbb{P}\left\{U_{(i)}\le p\right\}=\mathbb{P}\left\{\text{Binomial}(n,p)\ge i\right\}$$

$$= \mathbb{P}\left\{\text{Binomial}(n,p) \geq np \cdot \left(1 + \frac{\sqrt{3i\log(2n/\delta)}}{i - \sqrt{3i\log(2n/\delta)}}\right)\right\}$$

$$\leq \exp\left\{-\frac{1}{3}np\left(\frac{\sqrt{3i\log(2n/\delta)}}{i - \sqrt{3i\log(2n/\delta)}}\right)^2\right\} = \exp\left\{-\frac{1}{3}\frac{\left(\sqrt{3i\log(2n/\delta)}\right)^2}{i - \sqrt{3i\log(2n/\delta)}}\right\} \leq \frac{\delta}{2n},$$

where the inequality uses the multiplicative Chernoff bound. Next, suppose that instead we have

$$U_{(i)} \geq \frac{i}{n} + \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n} =: p'.$$

This means that at most $i-1$ of the U_k's lie in the interval $[0, p']$. Then

$$\begin{aligned}
&\mathbb{P}\left\{U_{(i)} \geq \frac{i}{n} + \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n}\right\} = \mathbb{P}\left\{U_{(i)} \geq p'\right\} \\
\leq\ &\mathbb{P}\left\{\text{Binomial}(n,p') \leq i\right\} \\
=\ &\mathbb{P}\left\{\text{Binomial}(n,p') \leq np' \cdot \left(1 - \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{i + \sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right)\right\} \\
\leq\ &\exp\left\{-\frac{1}{2}np'\left(\frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{i + \sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right)^2\right\} \\
=\ &\exp\left\{-\frac{1}{2}\frac{\left(\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)\right)^2}{i + \sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}\right\} \leq \frac{\delta}{2n},
\end{aligned}$$

where again the first inequality uses the multiplicative Chernoff bound. Combining these two calculations,

$$\mathbb{P}\left\{\left|U_{(i)} - \frac{i}{n}\right| \leq \frac{\sqrt{3i\log(2n/\delta)} + 2\log(2n/\delta)}{n}\right\} \geq 1 - \delta/n.$$

Finally, taking a union bound over all i, we have proved the lemma. □

Lemma 8. *Let $U_{(1)} \leq \cdots \leq U_{(n)}$ be the order statistics of $U_1, \ldots, U_n \overset{\text{iid}}{\sim} \text{Unif}[0,1]$, and let $U_{(0)} = 0$. For any $\delta > 0$,*

$$\mathbb{P}\Bigg\{\left|U_{(i)} - U_{(j)} - \frac{i-j}{n}\right| \leq \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n}$$
$$\text{for all } 0 \leq i < j \leq n\Bigg\} \geq 1 - \delta.$$

Proof of Lemma 8. First, it is known that $U_{(j)} - U_{(i)} \sim \text{Beta}\big((j-i), (n+1)-(j-i)\big)$ for all $0 \le i < j \le n$. In particular, $U_{(j)} - U_{(i)}$ has the same distribution as $U_{(j-i)}$, and so by Lemma 7 applied with $k = j - i$ and with $2\delta/(n^2+n)$ in place of δ/n,

$$\mathbb{P}\left\{\left|U_{(j)} - U_{(i)} - \frac{j-i}{n}\right| > \frac{\sqrt{3|i-j|\log((n^2+n)/\delta)} + 2\log((n^2+n)/\delta)}{n}\right\} \le \frac{2\delta}{n^2+n}.$$

Taking a union bound over all $\binom{n+1}{2} = \frac{n^2+n}{2}$ pairs of indices i, j, then, we obtain the desired bound. □

Acknowledgements

R.F.B. was partially supported by an Alfred P. Sloan Fellowship. The authors are grateful to Sabyasachi Chatterjee for helpful feedback on earlier drafts of this paper.

References

[1] Fadoua Balabdaoui, Hanna Jankowski, Marios Pavlides, Arseni Seregin, and Jon Wellner. On the grenander estimator at zero. *Statistica Sinica*, 21(2):873, 2011. MR2829859

[2] Richard E Barlow, David J Bartholomew, JM Bremner, and H Daniel Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York, 1972. MR0326887

[3] Lucien Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, pages 995–1012, 1987. MR0902241

[4] Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1):113–150, 1993. MR1240719

[5] Lev M Bregman. The method of successive projection for finding a common point of convex sets(theorems for determining common point of convex sets by method of successive projection). *Soviet Mathematics*, 6:688–692, 1965. MR0198341

[6] Hugh D Brunk. *Estimation of isotonic regression*. University of Missouri-Columbia, 1969.

[7] Chris Carolan and Richard Dykstra. Asymptotic behavior of the grenander estimator at density flat regions. *Canadian Journal of Statistics*, 27(3):557–566, 1999. MR1745821

[8] Eric Cator. Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17(2):714–735, 2011. MR2787612

[9] Sabyasachi Chatterjee, Adityanand Guntuboyina, Bodhisattva Sen, et al. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4):1774–1800, 2015. MR3357878

[10] Mathias Drton and Caroline Klivans. A geometric interpretation of the characteristic polynomial of reflection arrangements. *Proceedings of the American Mathematical Society*, 138(8):2873–2887, 2010. MR2644900

[11] Lutz Dümbgen. Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449, 2003. MR1997491

[12] Cécile Durot, Vladimir N Kulikov, and Hendrik P Lopuhaä. The limit distribution of the l_∞-error of grenander-type estimators. *The Annals of Statistics*, pages 1578–1608, 2012. MR3015036

[13] Theo Gasser, Lothar Sroka, and Christine Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, pages 625–633, 1986. MR0897854

[14] Ulf Grenander. On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956. MR0093415

[15] Piet Groeneboom. Estimating a monotone density. *Department of Mathematical Statistics*, (R 8403):1–14, 1984. MR0822052

[16] Shih-Ping Han. A successive projection method. *Mathematical Programming*, 40(1-3):1–14, 1988. MR0923692

[17] Hanna Jankowski. Convergence of linear functionals of the grenander estimator under misspecification. *The Annals of Statistics*, 42(2):625–653, 2014. MR3210981

[18] Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, pages 1083–1104, 2000. MR1810920

[19] BLS Prakasa Rao. Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 23–36, 1969. MR0267677

[20] John Rice. Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230, 1984. MR0760684

[21] Tim Robertson, Farrol T Wright, and Richard L Dykstra. Order restricted statistical inference, 1988. MR0961262

[22] Sara Van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990. MR1056343

[23] Sara Van de Geer. Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics*, pages 14–44, 1993. MR1212164

[24] Yazhen Wang and KS Chen. The l2risk of an isotonic estimate. *Communications in Statistics-Theory and Methods*, 25(2):281–294, 1996. MR1379445

[25] Farrol T Wright. The asymptotic behavior of monotone regression estimates. *The Annals of Statistics*, 9(2):443–448, 1981. MR0606630

[26] Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002. MR1902898