41500: High-Dimensional Statistics **Winter 2020**

Isotonic Regression

Lecturer: Chao Gao *Scribe: Xinze Li*

1 Problem Formulation

Isotonic regression is also called shape-constraint estimation. The advantage of this kind of estimation is that we do not need tuning parameter to do the estimation. We assume the nonparametric model as follows

$$y_i = f(x_i) + z_i, \quad i \in [n], z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

We also assume that f is nondecreasing. In particular, we suppose

$$\begin{aligned} y_i = \theta_i + z_i, \quad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad i \in [n] \\ \theta_1 \le \theta_2 \le \cdots \le \theta_n \end{aligned} \tag{1}$$

So we could compute the following estimator

$$\widehat{\theta} = \arg \min_{\theta_1 \le \theta_2 \le \cdots \le \theta_n} \|y - \theta\|^2 \tag{2}$$

Note that the constraint is linear, and the objective is quadratic, so this is an easy quadratic programming problem. In particular, this is a convex programming problem, which is easily computable. We define the following convex set naturally

$$C = \{\theta \in \mathbb{R}^n : \theta_1 \le \theta_2 \le \cdots \le \theta_n\} \tag{3}$$

We also construct the vector $v_j = \sum_{i=1}^{j} e_j$, where $\{e_j\}_{j=1}^{n}$ is the canonical base of $\mathbb{R}^n$. Now note that $\widehat{\theta} \in C$, and that for every $\epsilon > 0$, $\widehat{\theta} - \epsilon v_j \in C$, so we could derive the basic inequality.

$$\left\|\widehat{\theta} - y\right\|^2 \le \left\|\widehat{\theta} - \epsilon v_j - y\right\|^2 \tag{4}$$

Expanding the RHS gives

$$0 \le \epsilon j - 2\left\langle \widehat{\theta} - y, v_j \right\rangle$$

Let ϵ goes to zero gives

$$\left\langle \widehat{\theta} - y, v \right\rangle \le 0$$

In other words,

$$\sum_{i=1}^{j} \widehat{\theta}_i \le \sum_{i=1}^{j} y_i, \quad \forall j \in [n] \tag{5}$$

Now notice that if $\widehat{\theta}_j < \widehat{\theta}_{j+1}$, the inequality strictly holds, then for ϵ sufficiently small, we have $\widehat{\theta} + \epsilon v_j \in C$. Following the same computation gives

$$\left\langle \widehat{\theta} - y, v_j \right\rangle \geq 0$$

Combining these reasoning, we have the following characterization of $\widehat{\theta}$

$$\begin{cases} \sum_{i=1}^{j} \widehat{\theta}_i \leq \sum_{i=1}^{j} y_i & \forall j \in [n] \\ \sum_{i=1}^{j} \widehat{\theta}_i = \sum_{i=1}^{j} y_i & \forall j \text{ s.t. } \widehat{\theta}_j < \widehat{\theta}_{j+1} \text{ or } j = n \end{cases} \tag{6}$$

We could also observe that the solution is unique (because the objective is strongly convex) and that the curve of the sum of θ_i is the largest convex function below the curve of the sum of y_i (including $(0, 0)$). Let f be an arbitrary continuous function, then we could define $\bar{f}$ as the convex

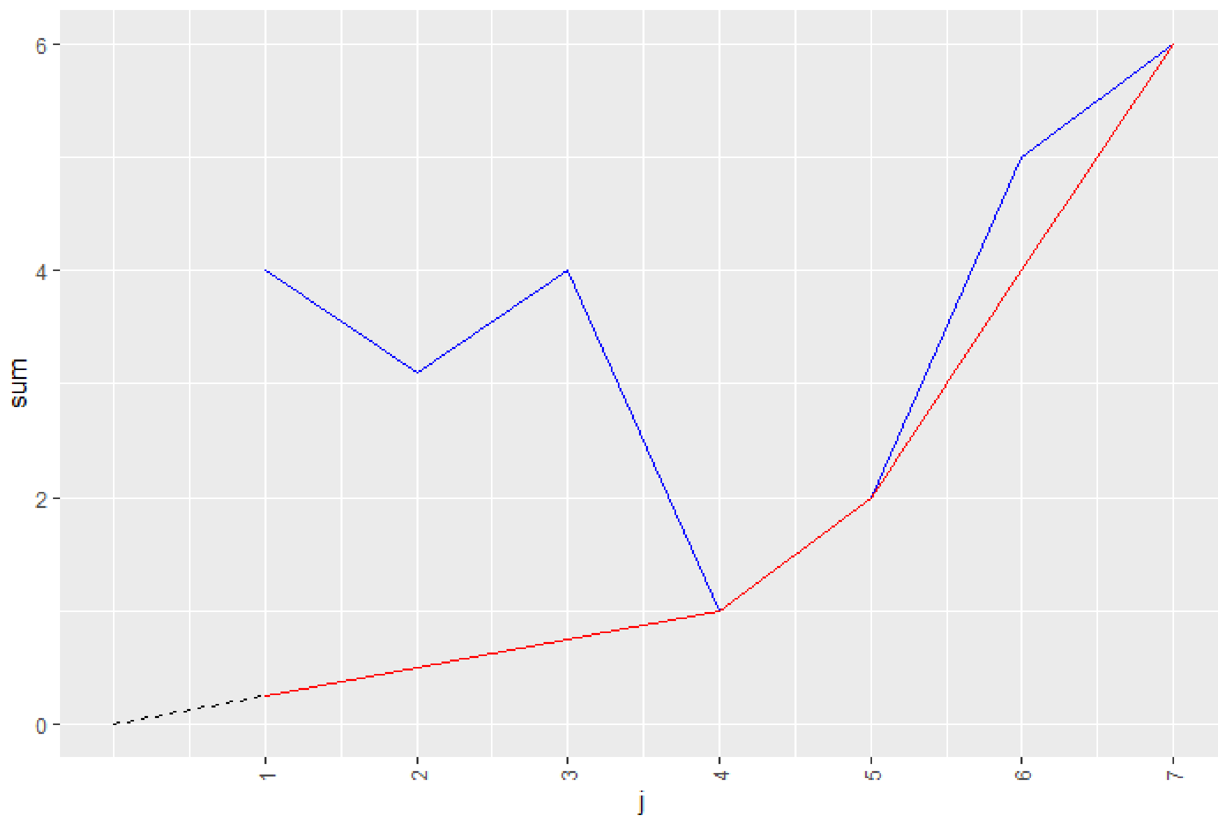

Figure 1: Simple Example of Convex Minorant

minorant of f, i.e., the largest convex function that is smaller than or equal to f everywhere. We could also define the right derivative of $\bar{f}$ as

$$
\begin{aligned}
\bar{f}^{(1,r)}(t) &= \inf_{w>t} \sup_{s\le t} \frac{f(w)-f(s)}{w-s} \\
&= \sup_{s\le t} \inf_{w>t} \frac{f(w)-f(s)}{w-s}
\end{aligned}
$$

We could also define the left derivative similarly. We now could conclude that the solution to the convex programming problem 2 is

$$
\widehat{\theta}_i = \min_{l\ge i} \max_{k\le i} \bar{y}_{[k,l]}, \quad \bar{y}_{[k,l]} = \frac{1}{l-k+1} \sum_{j=k}^{l} y_j \tag{7}
$$

We could use **PAVA** (pool-adjacent-violators-algorithm) to get the solution in linear time.

2 Algorithm

PAVA algorithm is easy to describe. It works just as its name says: to pool adjacent violators. So we first scan the array from the first element to the last element. Once there is a violator $y_i > y_{i+1}$, we pool these two nodes together to be a single node, with weight 2 and value equal to the weighted average. Because once we do these pooling procedure, the length of the array decrease by one, the time complexity of the algorithm is $O(n)$. We now verify the correctness of the algorithm.
First note that if $y_1 \le y_2 \le \cdots \le y_n$ already, then

$$
y = \arg \min_{\theta\le\cdots\le\theta_n} \sum_{i=1}^{n} w_i \left(y_i - \theta_i\right)^2
$$

Suppose $y_5 > y_6$ for simplicity. It is easy to note that

$$
\min_{\theta_1\le\cdots\le\theta_n} \sum_{i=1}^{n} w_i \left(y_i - \theta_i\right)^2 = \min_{\theta_1\le\cdots\le\theta_5=\theta_6\le\cdots\le\theta_n} \sum_{i=1}^{n} w_i \left(y_i - \theta_i\right)^2
$$

This is because if $\theta_5 < \theta_6$ strictly, then we could always pull θ_5 towards θ_6 or pull θ_6 towards θ_5, to decrease the objective value, and in the same time still maintain the isotonic property. We could then write

$$
\begin{aligned}
w_5\left(y_5-\theta\right)^2 + w_6\left(y_6-\theta\right)^2 &= \left(w_5+w_6\right)\left(\frac{w_5y_5+w_6y_6}{w_5+w_6}-\theta\right)^2 + w_5\left(y_5 - \frac{w_5y_5+w_6y_6}{w_5+w_6}\right)^2 \\
&+ w_6\left(y_6 - \frac{w_5y_5+w_6y_6}{w_5+w+6}\right)^2 \\
&= \left(w_5+w_6\right)\left(\frac{w_5y_5+w_6y_6}{w_5+w_6}-\theta\right)^2 + f(y_5, y_6, w_5, w_6)
\end{aligned}
$$

Thus, we have

$$\arg \min_{\theta\leq\cdots\leq\theta_n} \sum_{i=1}^{n} w_i \left(y_i - \theta_i\right)^2 = \arg \min_{\theta_1\leq\cdots\leq\theta_4\leq\theta\leq\theta_7\leq\cdots\leq\theta_n} \left[\sum_{i=1}^{4} w_i \left(y_i - \theta_i\right)^2 \right.$$
$$\left. + \left(w_5 + w_6\right) \left(\frac{w_5 y_5 + w_6 y_6}{w_5 + w_6} - \theta\right)^2 + \sum_{i=7}^{n} w_i \left(y_i - \theta_i\right)^2\right]$$

And this is exactly what the algorithm is doing.

3 Property of PAVA Estimator

Theorem 3.1 (Meyer-Woodroofe-Zhang)**.** *If* $\theta_n - \theta_1 \leq V$*, then*

$$\mathbf{E} \left\|\widehat{\theta} - \theta\right\|^2 \lesssim \sigma^2 \left(\log(en) + n^{1/3} \left(\frac{V}{\sigma}\right)^{2/3}\right) \tag{8}$$

Remark 3.2. Suppose θ is a piecewise constant function with K blocks and that if $\theta_i \in j_{th}$ block, then $\theta_i = \mu_j$. Now note that if we let $\widehat{\theta}_i$ be the average value of all y_i in the $j_t h$ block, then we have

$$\mathbf{E} \left\|\widehat{\theta} - \theta\right\|^2 = \sum_{i=1}^{n} \left(\widehat{\theta}_i - \theta_i\right)^2 = \sum_{j=1}^{K} \left(\sum_{i\in B_j} \frac{\sigma^2}{n_j}\right) = \sigma^2 K$$

where B_j is the set of all elements in the j_{th} block. Another case is when $\frac{V}{\sigma} = O(1)$, then $E \left\|\widehat{\theta} - \theta\right\|^2 \preceq \sigma^2 n^{1/3}$. This indicates that n elements are divided into $O(n^{1/3})$ pieces.

Proof. Following the above notation, and let

$$\widehat{\theta}_i = \min_{l\geq i} \max_{k\leq i} \bar{y}_{[k,l]} \tag{9}$$

Also use the following notation

$$x_+ = \max(0, x), \quad x_- = \max(0, -x), \quad |x| = x_+ + x_-$$

Then we have

$$\mathbf{E} \left\|\widehat{\theta} - \theta\right\|^2 = \sum_{i=1}^{n} \mathbf{E} \left\|\widehat{\theta} - \theta\right\|_+^2 + \mathbf{E} \left\|\widehat{\theta} - \theta\right\|_-^2$$

Also

$$\widehat{\theta}_i = \min_{l\geq i} \max_{k\leq i} \bar{y}_{[k,l]} \leq \min_{i\leq l\leq i+m_i} \max_{k\leq i} \left(\bar{\theta}_{[k,l]} + \bar{z}_{[k,l]}\right)$$

where m_i is defined as follows

$$m_i = \max_{m} \left\{m : \bar{\theta}_{[i,i+m]} - \theta_i \leq v(m), i + m \leq n\right\}$$

$v(m)$ is a bias function to be determined. Using this definition, we have the inequality

$$\bar{\theta}_{[k,l]} \leq \bar{\theta}_{[i,l]} \leq \theta_i + v(m_i)$$

Thus,

$$\widehat{\theta}_i \leq \theta_i + v(m_i) + \min_{i \leq l \leq i+m_i} \max_{k \leq i} \bar{z}_{[k,l]}$$

Rearranging the order gives

$$\left(\widehat{\theta}_i - \theta_i\right)_+ \leq \left(v(m_i) + \min_{i \leq l \leq i+m_i} \max_{k \leq i} (z)_{[k,l]}\right)_+ \leq v(m_i) + \left(\min_{i \leq l \leq i+m_i} \max_{k \leq i} (z)_{[k,l]}\right)_+$$

So we have

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{E}\left(\widehat{\theta}_i - \theta_i\right)_+^2 &\leq 2\sum_{i=1}^{n} v(m_i)^2 + 2\sum_{i=1}^{n} \mathbf{E}\left(\min_{i \leq l \leq i+m_i} \max_{k \leq i} \bar{z}_{[k,l]}\right)_+^2 \\ \sum_{i=1}^{n} \mathbf{E}\left(\widehat{\theta}_i - \theta_i\right)_-^2 &\leq 2\sum_{i=1}^{n} \mathbf{E}\left(\min_{i \leq l \leq i+m_i} \max_{k \leq i} \bar{z}_{[k,l]}\right)_-^2 \end{aligned} \tag{10}$$

Combining these two inequalities gives

$$\mathbf{E}\left\|\widehat{\theta} - \theta\right\|^2 \leq 2\sum_{i=1}^{n} v(m_i)^2 + 2\sum_{i=1}^{n} \mathbf{E}\left(\min_{i \leq l \leq i+m_i} \max_{k \leq i} \bar{z}_{[k,l]}\right)^2$$

Before we do the next analysis, we first recall knowledge in probability theory. We say that $(\mathcal{F}_i)_{i=1}^{\infty}$ is a filtration if $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$. And that (X_t) is a (sub)martingale if the following two requirements are met:

1. X_t is an adaptive process, i.e., it is measurable w.r.t. $\mathcal{F}_t$.
2. $\mathbf{E}\left(X_t | \mathcal{F}_{t-1}\right)(\geq) = X_{t-1}$.

An obvious but interesting fact is that if ϕ is a convex function and (X_t) is a martingale w.r.t. $\mathcal{F}_t$, then $(\phi(X_t))$ is a submartingale. This is just a one-line proof:

$$\mathbf{E}\left(\phi\left(X_t\right) | \mathcal{F}_{t-1}\right) \geq \phi\left(\mathbf{E}\left(X_t | \mathcal{F}_{t-1}\right)\right) = \phi\left(X_{t-1}\right)$$

We now introduce Doob's maximal inequality

Lemma 3.3. *If $(M_n)_{n=1}^{\infty}$ is a positive submartingale, then*

$$\mathbf{E}\left(\max_{1 \leq m \leq n} M_m\right)^2 \leq 4\mathbf{E}M_n^2$$

Example 3.4. *Suppose $z_i \overset{i.i.d.}{\sim} P$, and have finite second moment: $\mathbf{E} z_i^2 < \infty$. Let S_n be the mean of the first n elements: $S_n = \frac{1}{n}\sum_{i=1}^n z_i$. Then we claim that S_n is a reverse martingale*

$$\mathbf{E}\left(S_{n-1}|S_n\right) = \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbf{E}\left(z_i|S_n\right) = \frac{1}{n-1}S_n = S_n$$

The second equality holds because the best guess of z_i when we know S_n is just S_n.

Using this, we could prove that

$$\begin{aligned}
\mathbf{E}\left(\min_{i\le l\le i+m_i}\max_{k\le i}\bar{z}_{[k,l]}\right)_+^2 &\le \mathbf{E}\left(\max_{k\le i}\bar{z}_{[k,i+m_i]}\right)_+^2 \\
&\le 4\mathbf{E}\left(\bar{z}_{[i,i+m_i]}\right)_+^2 \\
&= \frac{4}{m_i+1}\mathbf{E}\left(\mathcal{N}(0,\sigma^2)\right)_+^2 \\
&\lesssim \frac{\sigma^2}{m_i+1}
\end{aligned}$$

So we could write

$$\mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 \lesssim \sum_{i=1}^n\left(v(m_i)+\frac{\sigma^2}{m_i+1}\right)$$

To balance variance and bias, we choose

$$v(m)^2 = \frac{\sigma^2}{m+1}$$

So we have the following analysis

$$\begin{aligned}
\mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 &\lesssim \sigma^2\sum_{i=1}^n\frac{1}{m_i+1} \\
&= \sigma^2\sum_{i=1}^n\sum_{l\ge 0}\mathbb{1}\left(2^l\le m_i<2^{l+1}\right)\frac{1}{m_i+1} + \sigma^2\sum_{i=1}^n\mathbb{1}\left(0\le m_i<1\right)\frac{1}{m_i+1} \\
&\le \sigma^2\sum_{l\ge 0}\frac{1}{2^l+1}\sum_{i=1}^n\mathbb{1}\left(2^l\le m_i<2^{l+1}\right) + \sigma^2\sum_{i=1}^n\mathbb{1}\left(0\le m_i<1\right) \\
&= \sigma^2\sum_{l\ge 0}\frac{1}{2^l+1}\left(H\left(2^{l+1}\right)-H\left(2^l\right)\right) + \sigma^2 H(1)
\end{aligned}$$

where

$$H(m) = \sum_{i=1}^n\mathbb{1}\left(m_i\le m\right)$$

So if $H(m) \le \widetilde{H}(m)$ for every $m \ge 0$, then

$$\sigma^2 \sum_{l \ge 0} \frac{1}{2^l+1} \left(H\left(2^{l+1}\right) - H\left(2^l\right)\right) + \sigma^2 H(1) \le \sigma^2 \sum_{l \ge 0} \frac{1}{2^l+1} \left(\widetilde{H}\left(2^{l+1}\right) - \widetilde{H}\left(2^l\right)\right) + \sigma^2 \widetilde{H}(1)$$

Now recall m_i's defintion

$$m_i = \max_m \left\{m : \bar{\theta}_{[i,i+m]} - \theta_i \le v(m), i+m \le n\right\}$$

thus if $m_i < m$, then $\bar{\theta}_{[i,i+m]} > v(m)$. Thus, we have

$$\begin{aligned}
H(m) &= \sum_{i=1}^{n} \mathbb{1}\left(m_i \le m\right) \\
&\le \sum_{i=1}^{n-m} \mathbb{1}\left(\bar{\theta}_{[i,i+m]} - \theta_i > v(m)\right) + m \\
&\le \sum_{i=1}^{n-m} \frac{\bar{\theta}_{[i,i+m]} - \theta_i}{v(m)} + m \\
&\le \sum_{i=1}^{n-m} \frac{\theta_{i+m} - \theta_i}{v(m)} + m \\
&\le \frac{mV}{v(m)} + m \\
&= m + \frac{mV}{\sigma}\sqrt{m+1}
\end{aligned}$$

The last inequality holds because all terms are canceled except for the first m and the last m term and they form m pairs. Then the bound is derived because of the total variation is V. So we have a bound for $H(m)$:

$$H(m) \le \min\left(n, m\sqrt{m+1} \cdot \frac{V}{\sigma}\right) + \min\left(n, m\right) = \widetilde{H}(m)$$

So now we have the analysis

$$\begin{aligned}
\mathbf{E}\left\|\widehat{\theta} - \theta\right\|^2 &\lesssim \sigma^2 \cdot \sum_{l \ge 0} \frac{1}{2^l+1} \left(\widetilde{H}\left(2^{l+1}\right) - \widetilde{H}\left(2^l\right)\right) + \sigma^2 \widetilde{H}(1) \\
&= \sigma^2 \cdot \underbrace{\sum_{l \ge 0} \frac{1}{2^l+1} \left[\min\left(n, 2^{l+1}\sqrt{2^{l+1}+1} \cdot \frac{V}{\sigma}\right) - \min\left(n, 2^l\sqrt{2^l+1} \cdot \frac{V}{\sigma}\right)\right]}_{\mathrm{I}} \\
&\quad + \sigma^2 \cdot \underbrace{\sum_{l \ge 0} \frac{1}{2^l+1} \left[\min\left(n, 2^{l+1}\right) - \min\left(n, 2^l\right)\right]}_{\mathrm{II}} + \sigma^2 \cdot \underbrace{\left(\min\left(n, \frac{V}{\sigma}\right) + 1\right)}_{\mathrm{III}}
\end{aligned}$$

For the second part II

$$\mathrm{II} \leq \sum_{2^l \leq n} \frac{1}{2^l+1} \min\left(n, 2^{l+1}\right) \leq \sum_{2^l \leq n} 2 \lesssim \log\left(en\right)$$

And the first part I is just the same analysis

$$\begin{aligned}
\mathrm{I} &\leq \sum_{2^l\sqrt{2^l+1}V/\sigma \leq n} \frac{1}{2^l+1} \min\left(n, 2^{l+1}\sqrt{2^{l+1}+1} \cdot \frac{V}{\sigma}\right) \\
&\lesssim \sum_{2^{3l/2} \leq n\sigma/V} 2^{\frac{l}{2}} \frac{V}{\sigma} \\
&\lesssim \left(n\frac{\sigma}{V}\right)^{\frac{2}{3}\times\frac{1}{2}} \cdot \frac{V}{\sigma} \\
&= n^{\frac{1}{3}} \cdot \left(\frac{V}{\sigma}\right)^{\frac{2}{3}}
\end{aligned}$$

And the third part III

$$\mathrm{III} \lesssim n^{\frac{1}{3}} \cdot \left(\frac{V}{\sigma}\right)^{\frac{2}{3}}$$

Thus, we conclude that

$$\mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 \lesssim \sigma^2\left(\log(en) + n^{1/3}\left(\frac{V}{\sigma}\right)^{2/3}\right)$$

$\square$

Actually, we could prove that

Theorem 3.5 ([Zhang, 2002])**.**

$$\sum_{i=n_1}^{n_2} \mathbf{E}\left(\widehat{\theta}_i - \theta_i\right)^2 \lesssim \sigma^2\left[\log\left(e(n_2-n_1)\right) + (n_2-n_1)^{\frac{1}{3}} \cdot \left(\frac{\theta_{n_2}-\theta_{n_1}}{\sigma}^{\frac{2}{3}}\right)\right] \tag{11}$$

For more details, see homework 9. If we assume that

$$\begin{aligned}
&\Theta_k^{\uparrow} = \Big\{\theta \in \mathbb{R}^n : \text{ there exist } \{a_j\}_{j=0}^k \text{ and } \{\mu_j\}_{j=1}^k \text{ such that} \\
&0 = a_0 \leq a_1 \leq \cdots \leq a_k = n \\
&\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k, \text{ and } \theta_i = \mu_j \text{ for all } i \in (a_{j-1} : a_j]\}
\end{aligned} \tag{12}$$

the parameter space of nondecreasing vectors with at most k pieces. Then we have that if $\theta \in \Theta_k^{\uparrow}$, then

$$\mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 \lesssim \sigma^2 \sum_{j=1}^{k} \log\left(en_j\right)$$

where n_j is the length of each pieces. Combining these two results, we know that if θ has at most k pieces, monotone and have total variation at most V, then

$$\mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 \lesssim \sigma^2 \min\left\{\log(en) + n^{1/3}\left(\frac{V}{\sigma}\right)^{2/3}, k\log\left(\frac{en}{k}\right)\right\}$$

And from [Gao et al., 2017], we know the minimax rate

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta_k^{\uparrow}} \mathbb{E}\left\|\widehat{\theta}-\theta^*\right\|^2 \geq \begin{cases} c\sigma^2, & k=1 \\ c\sigma^2 k \log\log(16n/k), & k \geq 2 \end{cases} \tag{13}$$

We could achieve the minimax rate using the following estimator

$$\widehat{\theta} = \arg\min_{\theta \in \Theta_k^{\uparrow}} \|y-\theta\|^2$$

The $\log\log$ term appears because of the law of iterated logarithm

$$\max_{1 \leq m \leq n} \left|\frac{1}{\sqrt{m}} \sum_{i=1}^{m} z_i\right| \asymp \sigma\sqrt{\log\log n}, \quad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$$

Also if we define

$$\Theta^{\uparrow}(V) = \{\theta \in \mathbb{R}^n, \theta_1 \leq \cdots \leq \theta_n, \theta_n - \theta_1 \leq V\}$$

then we have

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta^{\uparrow}} \mathbf{E}\left\|\widehat{\theta}-\theta\right\|^2 \asymp \sigma^2 n^{1/3}\left(\frac{V}{\sigma}\right)^{2/3}$$

This is a result from Polyak, Nemirovski and Tsybakov.

References

[Gao et al., 2017] Gao, C., Han, F., and Zhang, C.-H. (2017). On estimation of isotonic piecewise constant signals.

[Zhang, 2002] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555.