Isotonic Regression in General Dimensions

Sabyasachi Chatterjee
Department of Statistics
University of Illinois at Urbana-Champaign

29 January 2018

References

1. S.Chatterjee, A. Guntuboyina and B. Sen (2015) **On Risk Bounds in Isotonic and Other Shape Constrained Regression Problems**

2. S.Chatterjee, A. Guntuboyina and B. Sen (2016) **On Matrix Estimation under Monotonicity Constraints**

3. S.Chatterjee, Roy Han, Tengyao Wang and Richard Samworth (2017) **Isotonic Regression in General Dimensions**

Outline

1. Isotonic Regression– Definition
2. Minimax Rate Optimality of the LSE
3. Statistical Dimension of the Monotone Cone
4. Adaptivity of the LSE

Setting (Monotone Function Estimation)

We are given data

$$Y_i = f^*(x_i) + \epsilon_i \qquad \text{for } i = 1, \ldots, n$$

where ϵ_i's are i.i.d $N(0, \sigma^2)$ with σ^2 unknown.

$f^* : [0,1]^d \to \mathbb{R}$ is an unknown function which is coordinate wise monotone non decreasing. The problem is to recover f^*.

$x_1, \ldots, x_n \in [0,1]^d$ are design points which could be assumed to be fixed or chosen i.i.d at random. In this talk, we consider the fixed lattice design case.

Setting (Sequence Estimation)

Let $L_{d,n}$ be the d dimensional lattice $[1, \ldots, k]^d$ where $k = n^{1/d}$.

Natural partial ordering on $L_{d,n}$. We have $u \leq v$ iff $u_j \leq v_j$ for all $1 \leq j \leq d$.

θ^* is monotone with respect to the natural partial order.

We are given data $Y \sim N(\theta^*, \sigma^2 I_{n \times n})$ with σ^2 unknown.

We measure the performance of an estimator $\hat{\theta}$ in terms of the mean squared error:

$$R(\hat{\theta}, \theta^*) = \frac{1}{n} \mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|^2$$

where $\|\cdot\|$ is Euclidean norm.

Least Squares Estimator

The space of monotone functions on the lattice $L_{d,n}$ is a closed convex cone.

$$\mathcal{M}_{d,n} = \{\theta \in \mathbb{R}^n : \theta_u \leq \theta_v \text{ for all } u \leq v \in L_{d,n}\}.$$

LSE is simply the Euclidean projection of Y onto the set $\mathcal{M}_{d,n}$.

$$\hat{\theta} = \Pi_{\mathcal{M}_{d,n}}(Y) = \operatorname*{argmin}_{v \in \mathcal{M}_{d,n}} \|Y - v\|^2.$$

How good is the LSE as an estimator of θ^*?

Some History

The $d = 1$ case is a canonical problem in Shape Constrained Regression and has been studied by many authors. See Brunk (1955); van Eeden (1958); van de Geer (1990); Donoho (1991); Birge and Massart (1993); Zhang (2002); Chatterjee, Guntuboyina and Sen (2015); Bellec(2016).

The $d > 1$ case is much less studied. The only work we are aware of is Chatterjee, Guntuboyina and Sen (2016) who studied the $d = 2$ case.

The cases $d > 2$ have not been studied at all in the literature.

MINIMAX RATE OPTIMALITY OF THE LSE

The case of $d = 1$

The LSE achieves the cube root rate of convergence in MSE, see (Zhang(2002)).

$$R(\hat{\theta}, \theta^*) \leq C\Big(\frac{V(\theta^*)\sigma^2}{n}\Big)^{2/3} + \frac{\sigma^2 \log n}{n}.$$

when $V(\theta^*) = \theta^*_n - \theta^*_1$.

The LSE is minimax rate optimal.

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{1,n} : V(\theta) \leq 1} R(\hat{\theta}, \theta) \geq C\Big(\frac{\sigma^2}{n}\Big)^{2/3}$$

The case of $d = 2$

The LSE achieves the square root rate of convergence in MSE, see (Chatterjee, Guntuboyina and Sen(2016)).

$$R(\hat{\theta}, \theta^*) \leq C \frac{\sigma V(\theta^*)\, (\log n)^4}{\sqrt{n}} + \frac{\sigma^2 (\log n)^8}{n}.$$

The LSE is minimax rate optimal upto a polylog factor.

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{2,n} : V(\theta) \leq 1} R(\hat{\theta}, \theta^*) \geq C\big(\frac{\sigma}{\sqrt{n}}\big).$$

LSE is minimax rate optimal upto polylog factor in all dimensions

The LSE achieves a $O(n^{-1/d})$ rate of convergence in MSE for $d \geq 3$.

$$\sup_{\theta \in \mathcal{M}_{d,n} : V(\theta^*) \leq 1} R(\hat{\theta}, \theta^*) \leq C n^{-1/d} (\log n)^4.$$

The LSE is minimax rate optimal upto a log factor for $d \geq 3$.

$$\inf_{\tilde{\theta}} \sup_{\theta \in \mathcal{M}_{d,n} : V(\theta) \leq 1} R(\tilde{\theta}, \theta) \geq c_d n^{-1/d}.$$

where $V(\theta)$ is the range of θ^*.

Some Comments

The metric entropy of bounded monotone functions (Gao, Wellner (07)) in $d \geq 3$ dimensions scales like $\frac{1}{\epsilon}^{2(d-1)}$. Hence the entropy integral diverges at a super logarithmic rate.

First(?) example of a global empirical risk minimization procedure is nearly minimax rate optimal over such a massive parameter space.

Worst case risk (upto log factors) is $n^{-\min\{2/(d+2),1/d\}}$. Transition of rate from $d = 1$ to $d \geq 3$ with $d = 2$ being the transition case.

Proof of the minimax lower bound for $d = 2$

Consider the lattice points on the anti diagonal $x + y = \sqrt{n} + 1$.

Clearly, this set of points forms an antichain; that is no two points are comparable.

If there are k points on the antichain; the problem is atleast as hard as estimating k normal means lying in $[0, 1]$; hence one obtains a minimax lower bound ck/n.

The largest antichain is the antidiagonal which has $O(\sqrt{n})$ points and hence gives a minimax lower bound of $O(n^{-1/2})$.

Proof of the minimax lower bound for general d.

Consider the set of lattice points $x_1 + \cdots + x_d = k = n^{1/d}$. Clearly this set forms an antichain.

Standard combinatorics then tells us the cardinality of this antichain is $O(k^{d-1}) = O(n^{1-1/d})$.

This immediately proves the minimax lower bound for the MSE scaling like $n^{-1/d}$.

Proof of Upper Bound for LSE

It is well known (Saurav Chatterjee(2015)) that the risk is intimately driven by the function

$$f(t) = \mathbb{E} \sup_{\theta \in \mathcal{M}_{n,d} : \|\theta - \theta^*\| \leq t} \langle Z, \theta - \theta^* \rangle.$$

Step 1: Upper bound the risk for the origin. This involves upper bounding the statistical dimension of $\mathcal{M}_{d,n}$.

Step 2: Use step 1 and Cauchy Schwarz inequality to bound $f(t)$ and derive a risk bound.

STATISTICAL DIMENSION

Statistical Dimension

Given a cone $C \subset \mathbb{R}^n$, a natural measure of its size is given by its Statistical dimension $\delta(C)$.

$$\delta(C) = \mathbb{E}\|\Pi_C(z)\|^2 \tag{1}$$

where $z \sim N(0, I_{n\times n})$.

For us, this is just the unnormalized risk at the origin.

An equivalent description is

$$\delta(C) = \mathbb{E}\Big(\sup_{\theta\in C, \|\theta\|\leq 1} \sum_{i=1}^{n} Z_i\theta_i\Big)^2.$$

Statistical Dimension of Monotone Cone

$$\delta(\mathcal{M}_{1,n}) = 1 + \frac{1}{2} + \dots \frac{1}{n}.$$

$$c(\log n)^2 \leq \delta(\mathcal{M}_{2,n}) \leq C(\log n)^8$$

$$c_d n^{1-2/d} \leq \delta(\mathcal{M}_{d,n}) \leq C n^{1-2/d}(\log n)^8$$

The statistical dimension becomes super logarithmic for $d > 2$.

Proof Ideas

To prove the upper bounds for $d \geq 3$, it is useful to view the lattice $L_{d,n}$ as a collection of $k^{d-2} = n^{(d-2)/d}$ many two dimensional lattices.

Enlarge $\mathcal{M}_{d,n}$ by removing the constraints between the lattices. Then an upper bound to $\delta(\mathcal{M}_{d,n})$ is just the sum of $\delta(\mathcal{M}_{2,n^{2/d}})$ times the number of two dimensional lattices.

To prove the lower bound we make the Gaussian supremum inner product large by setting the values to be proportional to the Gaussian vector on the antichain.

ADAPTATION OF THE LSE

Adaptive Risk Bounds for $d = 1$

When $d = 1$, Chatterjee, Guntuboyina and Sen(15); Bellec(2016) prove that

$$R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in \mathcal{M}_{1,n}} \left(\frac{\|\theta^* - \theta\|^2}{n} + \frac{\sigma^2 k(\theta)}{n} \log \frac{en}{k(\theta)} \right)$$

where $k(\theta)$ is the number of constant pieces of θ.

LSE adapts to piecewise constant functions at a parametric rate upto a log factor.

Adaptive Risk Bounds for $d = 2$

In Bivariate Isotonic Regression Chatterjee, Guntuboyina and Sen(2016) proved

$$R(\hat{\theta}, \theta^*) \leq \inf_{\theta \in \mathcal{M}_{2,n}} \left(\frac{\|\theta^* - \theta\|^2}{n} + \frac{\sigma^2 k(\theta)}{n} (\log(en))^8 \right).$$

$k(\theta)$ is the smallest k s.t there exists a rectangular block-wise partition of the $\sqrt{n} \times \sqrt{n}$ square into k blocks such that θ is constant on each block.

LSE adapts to bivariate non decreasing functions which are piecewise constant on rectangles at a parametric rate upto a log factor.

Adaptation to Intrinsic Dimensionality for $d = 2$

For $r = 0, 1, 2$, we say a vector $\theta_0 \in \mathcal{M}(\mathbb{L}_{2,n})$ is a *function of r variables*, written $\theta_0 \in \mathcal{M}_r(\mathbb{L}_{2,n})$, if θ_0 only depends on r many coordinates out of $d = 2$.

For $d \geq 2$, there exists constant $C > 0$

$$\sup_{\theta_0 \in \mathcal{M}_r(\mathbb{L}_{2,n}) \cap B_\infty(1)} R(\hat{\theta}_n, \theta_0) \leq C_d \begin{cases} n^{-1} \log^8 n & \text{if } r = 0 \\ n^{-2/3} \log^8 n & \text{if } r = 1 \\ n^{-1/2} \log^4 n & \text{if } r = 2. \end{cases}$$

Adaptive Risk Bound for $d > 2$?

The LSE has a $O(n^{-2/d})$ rate of convergence when θ^* is a constant function; parametric adaptation therefore is not possible.

However this rate is faster than the minimax rate of convergence $O(n^{-1/d})$.

It is still natural to surmise that the LSE will have faster rate of convergence than the minimax rate whenever θ^* is piecewise constant on rectangles; subsets of the lattice of the form $\prod_{i=1}^{d}[a_i, b_i]$.

Adaptive Risk Bound for $d > 2$?

It is useful to view the lattice $L_{d,n}$ as a collection of two dimensional lattices; for example by fixing $d - 2$ coordinates.

One can then apply the existing adaptation result for 2 dimensional lattices in Chatterjee, Guntuboyina, Sen(2016) on each of these lattices.

Adaptive Risk Bound for $d > 2$

For a rectangle $\prod_{i=1}^{d}[a_i, b_i]$, call it a *two dimensional sheet* if $|i : a_i = b_i| \geq d - 2$.

For any $\alpha \in \mathcal{M}_{d,n}$, define $k(\alpha)$ to be the cardinality of the minimal partition of $L_{d,n}$ into two dimensional sheets.

$$R(\hat{\theta}, \theta^*) \leq \inf_{\alpha \in \mathcal{M}_{d,n}} \left(\frac{\|\theta^* - \alpha\|^2}{n} + \frac{\sigma^2 k(\alpha)}{n} (\log(en))^8 \right).$$

The above theorem works even if the model is misspecified.

Adaptation to general rectangular level sets

Let $\mathcal{M}_{k,d,n}$ be the collection of all $\theta \in \mathcal{M}(d,n)$ such that there exists a partition $L_{d,n} = \cup_{i=1}^{k} R_i$ where $R_1, ..., R_k$ are rectangles with the property that θ is constant on each rectangle.

$$R(\hat{\theta}, \theta^*) \leq \inf_{k} \Big\{ \inf_{\alpha \in \mathcal{M}_{k,d,n}} \left(\frac{\|\theta^* - \alpha\|^2}{n} + C\big(\frac{k}{n}\big)^{2/d} (\log n)^8 \right) \Big\}.$$

If $\theta^* \in \mathcal{M}_{k,d,n}$ then we get a $\tilde{O}(k/n)^{2/d}$ rate of convergence.

Adaptation to Intrinsic Dimensionality for $d \geq 3$.

For $r = 0, 1, \ldots, d$, we say a vector $\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n})$ is a *function of r variables*, written $\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n})$, if θ_0 only depends on r many coordinates out of d.

For $d \geq 2$, there exists constant $C_d > 0$, depending only on d, such that

$$\sup_{\theta_0 \in \mathcal{M}_r(\mathbb{L}_{d,n}) \cap B_\infty(1)} R(\hat{\theta}_n, \theta_0) \leq C_d \begin{cases} n^{-2/d} \log^8 n & \text{if } r \leq d-2 \\ n^{-4/(3d)} \log^{16/3} n & \text{if } r = d-1 \\ n^{-1/d} \log^4 n & \text{if } r = d. \end{cases}$$

Summary

LSE is minimax rate optimal with $\tilde{O}(n^{-1/d})$ rate of convergence.

The Statistical Dimension of the Monotone Cone becomes super logarithmic as soon as $d > 2$.

Nearly parametric adaptation to piecewise constant functions is no longer obtained as in $d = 1$ and 2 but faster rates than the minimax rates are still obtained when θ^* has additional structure such as piecewise constant on rectangles or when intrinsic dimensionality is lower than d.

THANK YOU!