Isotonic Regression in General Dimensions

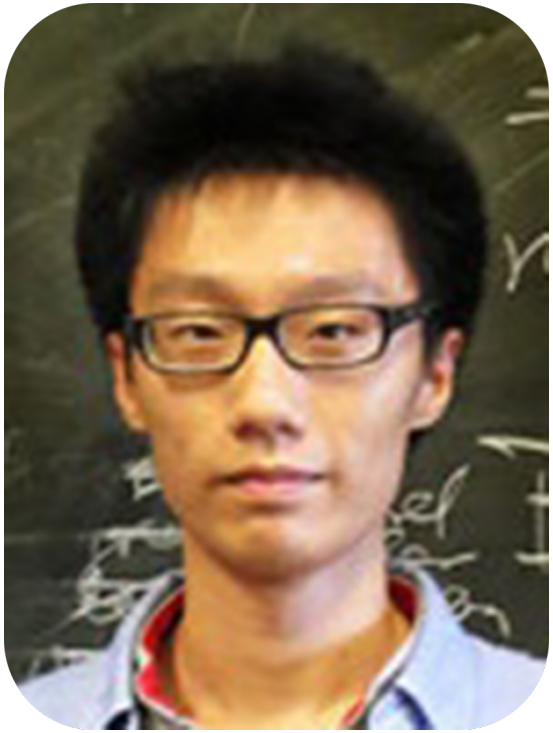
Tengyao Wang

University of Cambridge

Statistics Seminar, Université libre de Bruxelles

12 Oct 2017

Collaborators



Qiyang (Roy) Han

Sabyasachi Chatterjee

Richard Samworth

Shape-constrained estimation

Imposing shape restrictions on function classes

- Monotonicity, convexity, log-concavity, ...

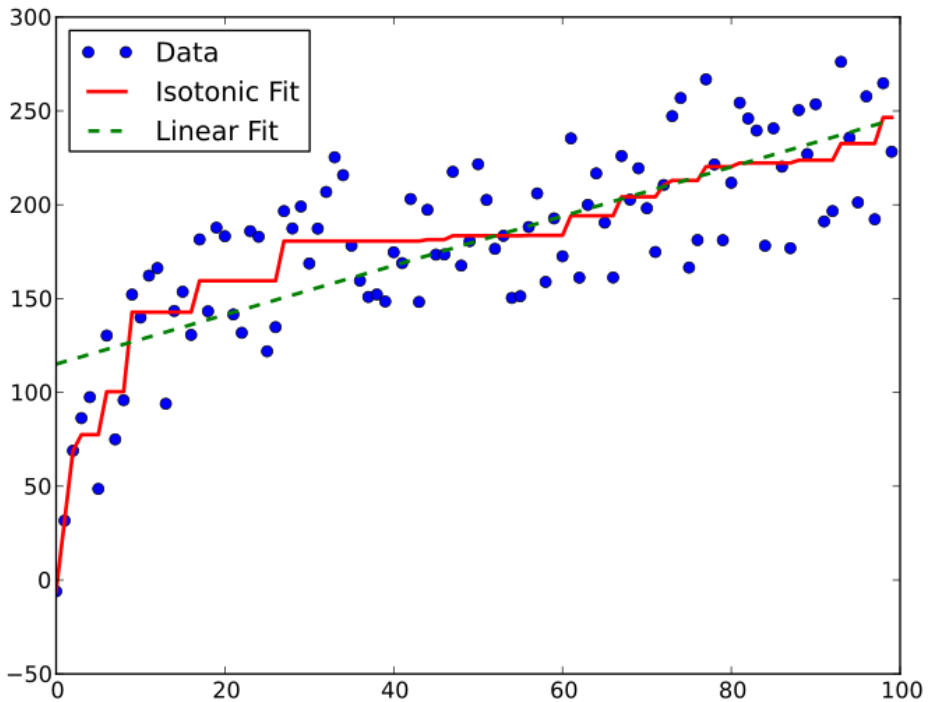
An attractive alternative to traditional non-parametric inference

- More flexible than parametric family.
- Fully automatic estimators: tuning parameter free.

Isotonic regression

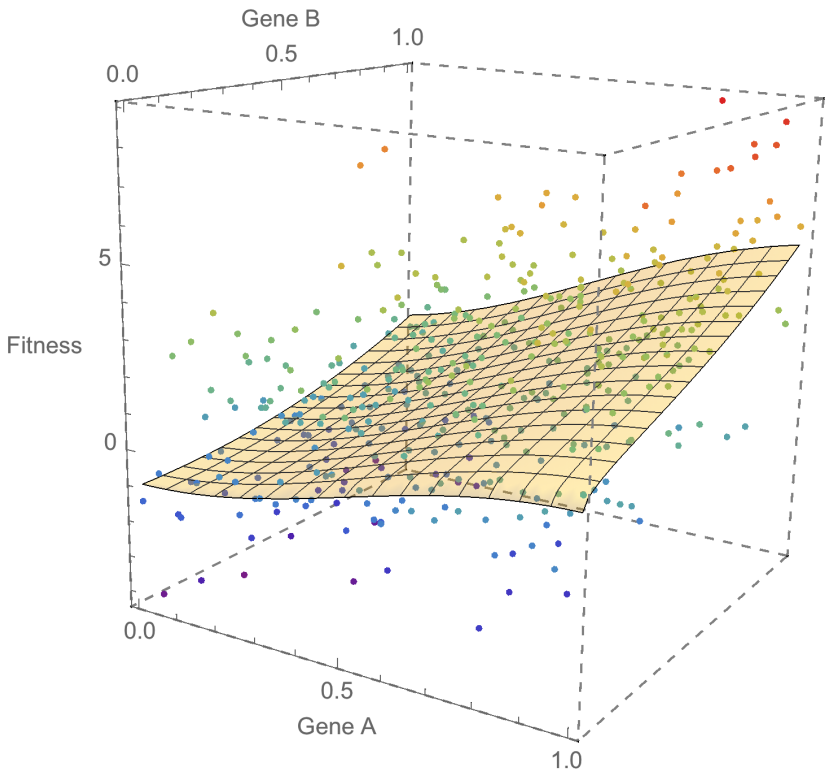
One of the oldest and most widely studied problems in shape-constrained estimation.

Estimating a monotone function f from noisy observations at design points $x_1 < x_2 < \cdots < x_n$.



Isotonic regression in general dimensions

Example: genetic effects on phenotypes such as height or fitness are monotone, but additive structure are often too restrictive (Mani *et al.*, 2007).



Isotonic regression in general dimensions

The natural partial ordering on $\mathbb{R}^d$:

$$x \preceq x' \quad \Leftrightarrow \quad x_j \leq x'_j \text{ for all } j = 1, \ldots, d.$$

Consider the class of block increasing functions:

$$\mathcal{F}_d := \left\{ f : [0,1]^d \to \mathbb{R}, f(x) \leq f(x') \text{ when } x \preceq x' \right\}.$$

Problem description

- Observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ satisfying

$$Y_i = f_0(X_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $f_0 : [0,1]^d \to \mathbb{R}$ is Borel measurable, $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0,1)$ and $X_1, \ldots, X_n$ are either fixed or random design points.

- We study the performance of the least squares estimator

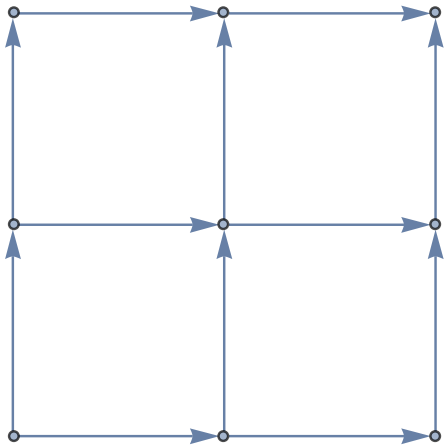
$$\hat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{F}_d} \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2$$

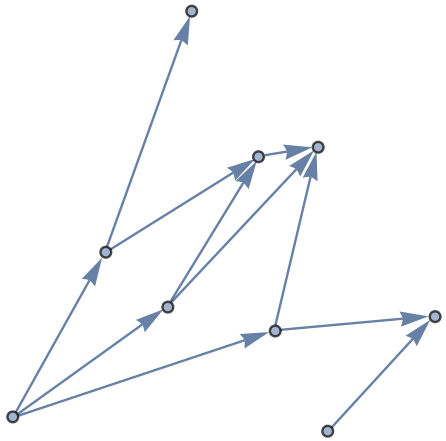
in terms of its empirical risk

$$R(\hat{f}_n, f_0) := \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \{\hat{f}_n(X_i) - f_0(X_i)\}^2\right].$$

Computation

- Design points form a directed acyclic graph G with respect to the natural partial ordering $\preceq$ on $\mathbb{R}^d$.





- Vectors respecting the partial ordering lie in the monotone cone

$$\mathcal{M}(G) := \bigcap_{X_i \preceq X_{i'}} \{y_i \leq y_{i'}\}.$$

- Least squares estimator: projection onto the monotone cone, solvable using von Neumann's algorithm or interior point methods.

Existing theoretical results

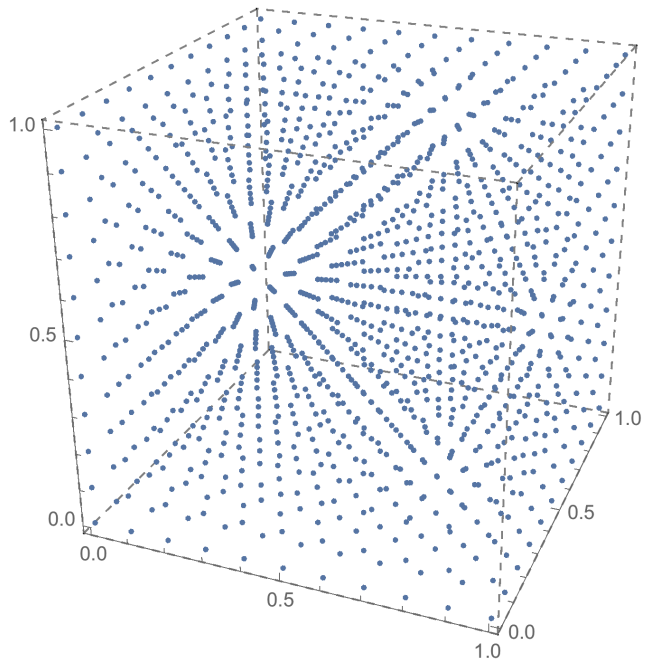
- Most theoretical results are on univariate isotonic regression (e.g. Brunk (1955); Barlow *et al.* (1972); van de Geer (1990); Birgé and Massart (1993); Meyer and Woodroofe (2000); Zhang (2002); Durot (2007, 2008); Chatterjee, Guntuboyina and Sen (2015); Bellec (2017); Yang and Barber (2017)).
- Chatterjee, Guntuboyina and Sen (2017) establishes the bivariate rate under fixed lattice designs.
- A separate line of work assumes additive structure in multivariate settings (e.g. Mammen and Yu (2007); Chen and Samworth (2016)).

Fixed lattice designs

Fixed lattice designs

Suppose $n = n_1^d$, we first consider the case of a fixed lattice design

$$\{X_1, \ldots, X_n\} = \mathbb{L}_{d,n} := \{1, \ldots, n_1\}^d \subseteq \mathbb{R}^d.$$



In this case, we write $\theta_0 := \big(f_0(X_i)\big)_{1\le i\le n}$ and $\hat{\theta}_n := \big(\hat{f}_n(X_i)\big)_{1\le i\le n}$.

Risk function $R(\hat{f}_n, f_0) = R(\hat{\theta}_n, \theta_0) := n^{-1}\|\hat{\theta}_n - \theta_0\|_2^2$.

A key ingredient

- Recall: the least squares estimator $\hat{\theta}_n$ is the projection of $(Y_1, \ldots, Y_n)^\top$ onto the monotone cone

$$\mathcal{M}(\mathbb{L}_{d,n}) := \{(y_1, \ldots, y_n) : y_i \leq y_{i'} \text{ whenever } X_i \preceq X_{i'}\}.$$

- For simplicity, we assume that $\|f_0\|_\infty \leq 1$.

- We use Sourav Chatterjee's characterisation of least squares projection onto a convex set (Chatterjee, 2014).

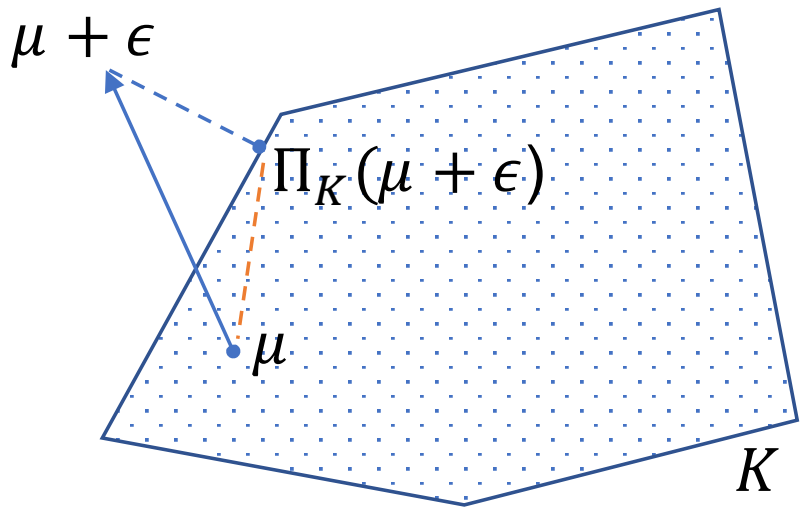
A key ingredient

Theorem (Chatterjee, 2014). Suppose μ belong be a closed convex set $K \subseteq \mathbb{R}^d$ and $\epsilon \sim N_d(0, I_d)$. If

$$t_0 := \arg\max_t \left\{ \mathbb{E} \sup_{\nu \in K, \|\nu - \mu\|_2 \le t} \langle \epsilon, \nu - \mu \rangle - \frac{t^2}{2} \right\} \ge 1,$$

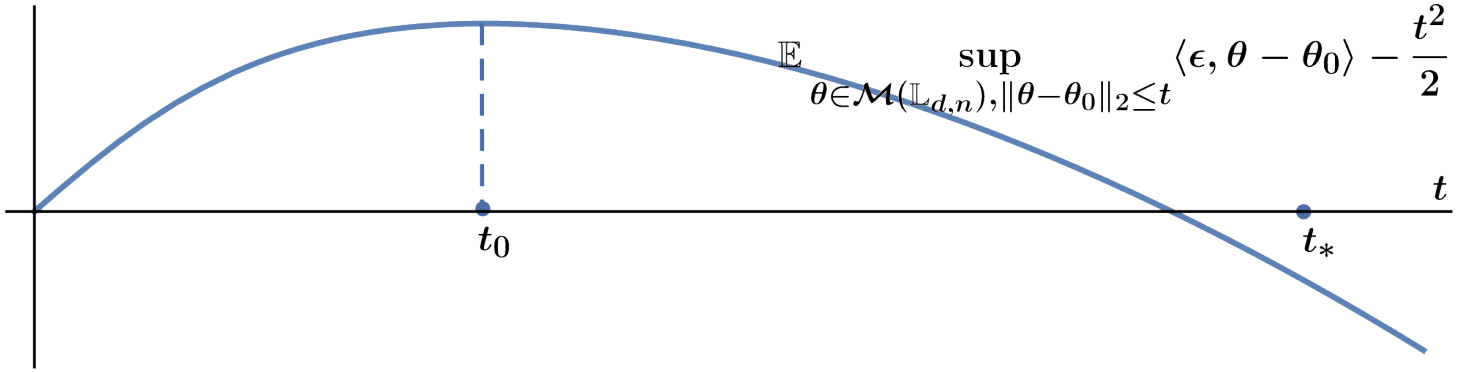
then

$$\mathbb{E}\|\Pi_K(\mu + \epsilon) - \mu\|_2^2 \asymp t_0^2.$$



Towards an upper bound

The function $t \mapsto \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta - \theta_0\|_2 \le t} \langle \epsilon, \theta - \theta_0 \rangle - t^2/2$ is strictly concave and tends to $-\infty$.



Hence $R(\hat{\theta}_n, \theta_0) \le n^{-1} t_*$ for any t_* satisfying

$$\mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta - \theta_0\|_2 \le t_*} \langle \epsilon, \theta - \theta_0 \rangle \le \frac{t_*^2}{2}.$$

Towards an upper bound

We manipulate the left hand side to obtain

$$\begin{aligned}
\text{LHS} &= \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta - \theta_0\|_2 \le t_*} \Big\{ \langle \epsilon, \theta - \bar{\theta}_0 \rangle + \langle \epsilon, \bar{\theta}_0 - \theta_0 \rangle \Big\} \\
&\le \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}), \|\theta - \bar{\theta}_0\|_2 \le t_* + n^{1/2}} \langle \epsilon, \theta - \bar{\theta}_0 \rangle \\
&= \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B(0, t_* + n^{1/2})} \langle \epsilon, \theta \rangle \\
&= \{ t_* + n^{1/2} \} \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B(0,1)} \langle \epsilon, \theta \rangle \le \{ t_* + n^{1/2} \} \delta^{1/2} \big(\mathcal{M}(\mathbb{L}_{d,n}) \big),
\end{aligned}$$

where $\delta(\mathcal{M}(\mathbb{L}_{d,n})) := \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B(0,1)} \langle \epsilon, \theta \rangle^2$ is the statistical dimension of the cone $\mathcal{M}(\mathbb{L}_{d,n})$ (cf. Amelunxen *et al.* (2014)). Therefore, it suffices to choose

$$t_* = \delta^{1/2} \big(\mathcal{M}(\mathbb{L}_{d,n}) \big) + \big\{ \delta \big(\mathcal{M}(\mathbb{L}_{d,n}) \big) + 2 n^{1/2} \delta^{1/2} \big(\mathcal{M}(\mathbb{L}_{d,n}) \big) \big\}^{1/2}.$$

The statistical dimension

Inductively reduce the problem to computing the stat. dim. of the bivariate monotone cone:

$$\delta\big(\mathcal{M}(\mathbb{L}_{d,n})\big) \le n_1 \delta\big(\mathcal{M}(\mathbb{L}_{d-1,n_1^{d-1}})\big) \le \cdots \le n_1^{d-2}\delta\big(\mathcal{M}(\mathbb{L}_{2,n_1^2})\big).$$

But $\delta\big(\mathcal{M}(\mathbb{L}_{2,n_1^2})\big)$ is approximately the square of the Gaussian complexity of $\mathcal{M}(\mathbb{L}_{2,n_1^2}) \cap B(0,1)$, which can be controlled via Dudley's entropy integral.

Proposition. For $d \ge 2$, we have

$$n^{1-2/d} \lesssim \delta\big(\mathcal{M}(\mathbb{L}_{d,n})\big) \lesssim n^{1-2/d}\log^8 n.$$

Worst case bounds

Putting it together, we get the following worst case upper bound.

Theorem. Let $d \geq 2$. There exists a universal constant $C > 0$ such that

$$\sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R(\hat{\theta}_n, \theta_0) \leq C n^{-1/d} \log^4 n.$$

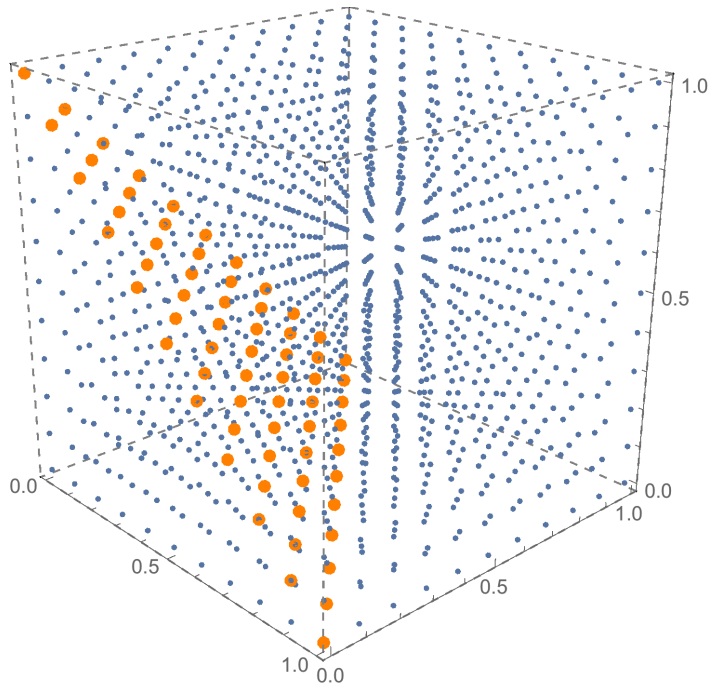
We also establish a matching minimax lower bound up to poly-log factors.

Proposition. There exists a constant $c_d > 0$, depending only on d, such that for $d \geq 2$,

$$\inf_{\tilde{\theta}_n} \sup_{\theta_0 \in \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(1)} R(\tilde{\theta}_n, \theta_0) \geq c_d n^{-1/d}.$$

Lower bound construction

- $\mathbb{L}_{d,n}$ has a large antichain A of cardinality at least $c_d n^{1-1/d}$.
- Consider all binary vectors $\theta \in \mathcal{M}(\mathbb{L}_{d,n})$ that are 1 above A and 0 below A.
- There are $\approx e^{c|A|}$ such θs whose pairwise Hamming distance $\geq |A|/2$.
- Apply Fano's/Assouad's Lemma.



Minimax optimality

- Compare to known results for $d \leq 2$:

dimension	1	2	3	4	$\cdots$
rate	$\Theta(n^{-2/3})$	$\tilde{\Theta}(n^{-1/2})$	$\tilde{\Theta}(n^{-1/3})$	$\tilde{\Theta}(n^{-1/4})$	$\cdots$
	low dimensions: $n^{-2/(d+2)}$		high dimensions: $n^{-1/d}$		

- Two competing factors drive the rate: block monotonic functions are 'no smoother' than bounded Lipschitz functions, and that existence of a large antichain prevents efficient estimation.
- This is the first example showing that empirical risk minimisation can be minimax optimal (up to poly-log factors) in such classes where the ε-entropy grows faster than $\varepsilon^{-2-\delta}$ (cf. Birgé and Massart (1993)).

Sharp oracle inequalities

Let $k(\theta)$ be the minimal number of hyperrectangles to partition $\mathbb{L}_{d,n}$ such that θ is piecewise constant w.r.t. the partition.

Theorem. Let $d \geq 2$. There exists a universal constant $C > 0$ such that for every $\theta_0 \in \mathbb{R}^{\mathbb{L}_{d,n}}$,

$$R(\hat{\theta}_n, \theta_0) \leq \inf_{\theta \in \mathcal{M}(\mathbb{L}_{d,n})} \left\{ \frac{\|\theta - \theta_0\|_2^2}{n} + C \left(\frac{k(\theta)}{n} \right)^{2/d} \log_+^8 \left(\frac{n}{k} \right) \right\}.$$

The adaptation rate of $n^{-2/d}$ cannot be improved.

Proposition. There exists $c_d > 0$, depending only on d, such that

$$R(\hat{\theta}_n, 0) \geq c_d \begin{cases} n^{-1} \log^2 n & \text{if } d = 2 \\ n^{-2/d} & \text{if } d \geq 3. \end{cases}$$

Sharp oracle inequalities

- Compare to known results for $d \leq 2$:

dimension	1	2	3	4	$\cdots$
adaptation rate	$\tilde{\Theta}(n^{-1})$	$\tilde{\Theta}(n^{-1})$	$\tilde{\Theta}(n^{-2/3})$	$\tilde{\Theta}(n^{-1/2})$	$\cdots$

low dimensions: parametric

high dimensions: $n^{-2/d}$

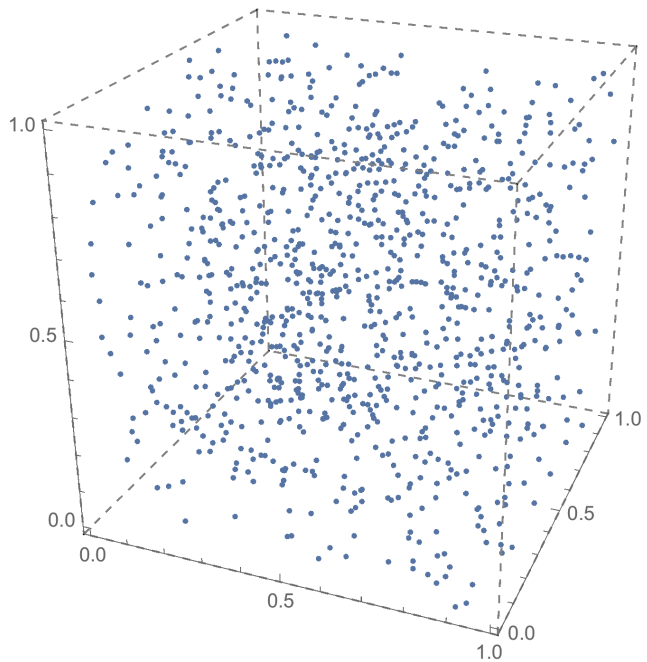
- First example of adaptation at a non-parametric rate.
- Minimax optimal worst case rate, but minimax suboptimal adaptation.

Random designs

Random designs

Uniform random design in $[0,1]^d$:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}([0,1]^d).$$



Risk function $R(\hat{f}_n, f_0) = \mathbb{E}\|\hat{f}_n - f_0\|^2_{L_2(\mathbb{P}_n)}$, where $\mathbb{P}_n := n^{-1}\sum_{i=1}^d \delta_{X_i}$.

Worst case and adaptation bounds

Theorem. Let $d \geq 2$. There exists $C_d > 0$ depending only on d such that

$$\sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} R(\hat{f}_n, f_0) \leq C_d n^{-1/d} \log^{\gamma_d} n,$$

where $\gamma_2 = 9/2$ and $\gamma_d = (d^2 + d + 1)/2$ for $d \geq 3$.

Theorem. Let $d \geq 2$. There exists a constant $C_d > 0$ depending only on d such that for any measurable function $f_0 : [0,1]^d \to \mathbb{R}$, we have

$$R(\hat{f}_n, f_0) \leq \inf_{f \in \mathcal{F}_d} \left\{ \|f - f_0\|^2_{L_2(P)} + C_d \left(\frac{k(f)}{n} \right)^{2/d} \log_+^{2\gamma_d} \left(\frac{n}{k} \right) \right\}.$$

Technical challenges

- Even the result for $d = 2$ was not known.
- For $d \geq 3$, the size of $\mathcal{F}_d$ has ε-entropy $\varepsilon^{-2(d-1)}$. Standard chaining argument via entropy integrals only gives a rate of $n^{-1/(2(d-1))}$ for the risk.
- In the fixed design case, we circumvent the problem by decomposing the d-dimensional lattice into unions of lower-dimensional lattices. Such geometric structure is not available in the random design.

Sketch of proofs

- Both theorems rely on proving the following sharp bound of the adaptation risk when $f_0 = 0$:

$$R(\hat{f}_n, 0) = \mathbb{E}\|\hat{f}_n\|^2_{L_2(\mathbb{P}_n)} \le C_d n^{-2/d} \log^{2\gamma_d} n.$$

- If we assume $\|\hat{f}_n\|_\infty \le C\sqrt{\log n}$, then $\mathbb{E}\|\hat{f}_n\|^2_{L_2(\mathbb{P}_n)} \approx \mathbb{E}\|\hat{f}_n\|^2_{L_2(P)}$.
- To bound $\|\hat{f}_n\|^2_{L_2(P)} = P\hat{f}_n^2$, we start from the basic inequality

$$\begin{aligned}\|\epsilon - \hat{f}_n\|^2_{L_2(\mathbb{P}_n)} \le \|\epsilon\|^2_{L_2(\mathbb{P}_n)} &\Rightarrow 2\langle \epsilon, \hat{f}_n\rangle_{L_2(\mathbb{P}_n)} - \|\hat{f}_n\|^2_{L_2(\mathbb{P}_n)} \ge 0 \\ &\Rightarrow \left|2\langle \epsilon, \hat{f}_n\rangle_{L_2(\mathbb{P}_n)}\right| + \left|\mathbb{P}_n\hat{f}_n^2 - P\hat{f}_n^2\right| \ge P\hat{f}_n^2.\end{aligned}$$

- Write $\mathcal{M} := \mathcal{M}(\mathbb{L}_{d,n}) \cap B_\infty(0,1)$. Thus, $P\hat{f}_n^2 \in (r^2, 4r^2)$ implies

$$\sup_{f\in\mathcal{M}\cap B(0,2r)} \left|\langle \epsilon, f\rangle_{L_2(\mathbb{P}_n)}\right| \ge r^2/4 \quad \text{or} \quad \sup_{f\in\mathcal{M}\cap B(0,2r)} \left|\mathbb{G}_n f^2\right| \ge n^{1/2} r^2/2.$$

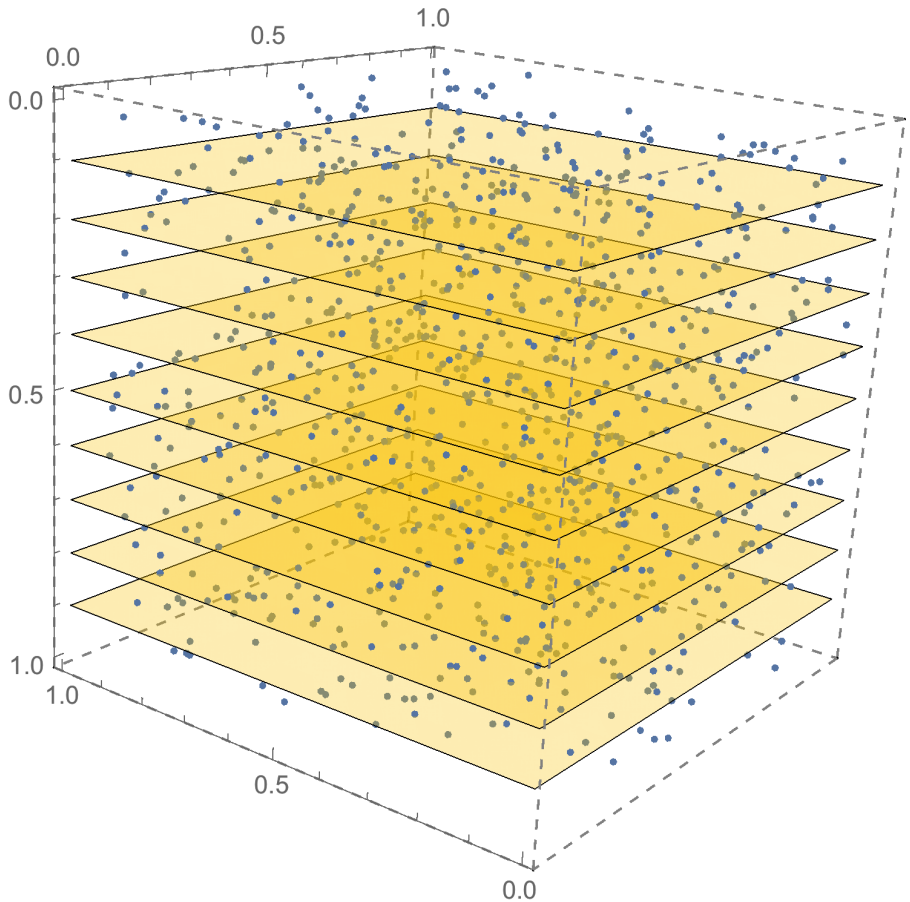
can be controlled by the first process!

Sketch of proofs

- By a peeling argument, it suffices to control

$$\sup_{f \in \mathcal{M} \cap B(0,r)} \langle \epsilon, f \rangle_{L_2(\mathbb{P}_n)} = \sup_{f \in \mathcal{M} \cap B(0,r)} \sum_{i=1}^{n} \epsilon_i f(X_i).$$

- Partition $[0,1]^d$ into slices of the form $[0,1]^{d-1} \times [\frac{\ell-1}{n_1}, \frac{\ell}{n_1}]$.
- The envelope function within the ℓ^{th} slice has $L_2(P)$ norm $O(r\ell^{-1/2})$.
- Contribution from each slice is bounded via chaining.



Lower bounds

Minimax lower bound:

Proposition. Let $d \geq 2$. There exists a constant $c_d > 0$, depending only on d, such that,

$$\inf_{\tilde{f}_n} \sup_{f_0 \in \mathcal{F}_d \cap B_\infty(1)} R(\tilde{f}_n, f_0) \geq c_d n^{-1/d}.$$

Remark. curious connection with the Erdős–Szekeres theorem.

Adaptive lower bound:

Proposition. Let $d \geq 2$. There exists a constant $c_d > 0$, depending only on d, such that

$$R(\hat{f}_n, 0) \geq c_d n^{-2/d}.$$

Remark. A version of Assouad's lemma for random loss functions was developed.

Summary

In this work

- Isotonic regression behaves differently for $d \leq 2$ and $d \geq 3$.
- Worst case rate $\tilde{\Theta}\big(n^{\frac{2}{d+2} \wedge \frac{1}{d}}\big)$, adaptative rate $\tilde{\Theta}\big((k/n)^{1 \wedge \frac{2}{d}}\big)$.
- Same behaviour under the random design.

Future directions

- Isotonic regression on a general DAG (e.g . fixed, non-lattice designs)?
- Other global loss functions (e.g. $L_2(P)$ loss)?
- Pointwise rates and adaptation results?
- Other shape constraints in general dimensions?
- Any other problems where ERM achieves the minimax rate?

More details in Han, W., Chatterjee and Samworth (2017).

References

- Amelunxen, D., Lotz, M., McCoy, M. B. and Tropp, J. A. (2014) Living on the edge: phase transition in convex programs with random data. *Inf. Inference*, **3**, 224–294.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) *Statistical Inference under Order Restrictions*. Wiley, New York.
- Bellec, P. C. (2017) Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.*, to appear.
- Birgé, L. and Massart, P. (1993) Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, **97**, 113–150.
- Brunk, H. D. (1955) Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, **26**, 607–616.
- Chatterjee, S. (2014) A new perspective on least squares under convex constraint. *Ann. Statist.*, **42**, 2340–2381.
- Chatterjee, S., Guntuboyina, A. and Sen, B. (2015) On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.*, **43**, 1774–1800.
- Chatterjee, S., Guntuboyina, A. and Sen, B. (2017) On matrix estimation under monotonicity constraints. *Bernoulli*, to appear.
- Chen, Y. and Samworth, R. J. (2016) Generalized additive and index models with shape constraints. *J. Roy. Statist. Soc., Ser. B*, **78**, 729–754.

References

- Durot, C. (2007) On the L_p-error of monotonicity constrained estimators. *Ann. Statist.*, **35**, 1080–1104.
- Durot, C. (2008) Monotone nonparametric regression with random design. *Math. Methods Statist.*, **17**, 327–341.
- Han, Q., Wang, T., Chatterjee, S. and Samworth, R. J. (2017) Isotonic regression in general dimensions. *arXiv preprint*, arxiv:1708.09468.
- Mammen, E. and Yu, K. (2007) Additive isotonic regression. In Cator, E. A. et al. (Eds.), *Asymptotics: Particles, Processes and Inverse Problems*, 179–195. Institute of Mathematical Statistics, Beachwood.
- Mani, R., Onge, R. P. S., Hartman, J. L., Giaever, G. and Roth, F. P. (2007) Defining genetic interaction. *Proc. Nat. Acad. Sci. USA*, **105**, 3461–3466.
- Meyer, M. and Woodroofe, M. (2000) On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, **28**, 1083–1104.
- van de Geer, S. A. (1990) Estimating a regression function. *Ann. Statist.* **18**, 907–924.
- Yang, F. and Barber, R. F. (2017) Uniform convergence of isotonic regression. *arXiv preprint*, arxiv:1706.01852.
- Zhang, C.-H. (2002) Risk bounds in isotonic regression. *Ann. Statist.*, **30**, 528–555.

Thank you!