50ème Journées de Statistique

1 Juin 2018

Adaptivity and Structure in Optimization and Estimation

John Lafferty

Department of Statistics and Data Science
Yale University

Theme

This talk is about the interplay between global and local structure in statistical learning. Two different settings:

- Stochastic convex optimization
- Shape constrained estimation

Outline

I. Adaptivity and geometry in optimization
   - How hard is it to optimize *this* function?

II. Graph structured signal denoising
   - What is behavior of shape constraints on graphs?

III. Prediction rule reshaping
   - How can shape constraints be used with ML algorithms?

I: Geometry and adaptivity in optimization

Heard around the Chicago Statistics lunch table:

> “Computer scientists are pessimists”
>
> “I don’t care about minimax”

The pessimism is about worst-case thinking, which is overly conservative.

What are alternatives?

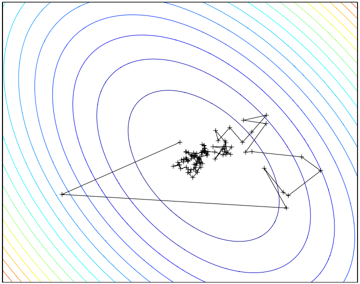
Per-Instance Complexity

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{C} \end{aligned}$$

f convex, $\mathcal{C}$ closed and convex.

Algorithms get noisy gradients of f at T query points.

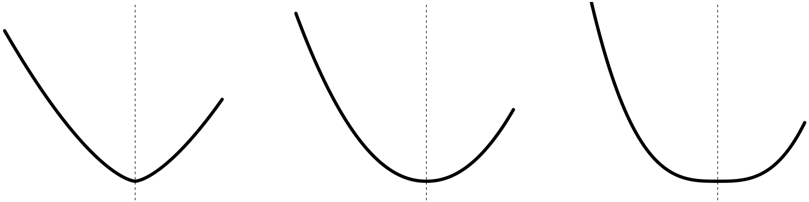
Example: logistic regression or SVMs for large-scale datasets.



Question: *How hard is it to optimize this function f?*

Per-Instance Complexity

- How hard is it to optimize *this* function?



Sabyasachi Chatterjee
UIUC

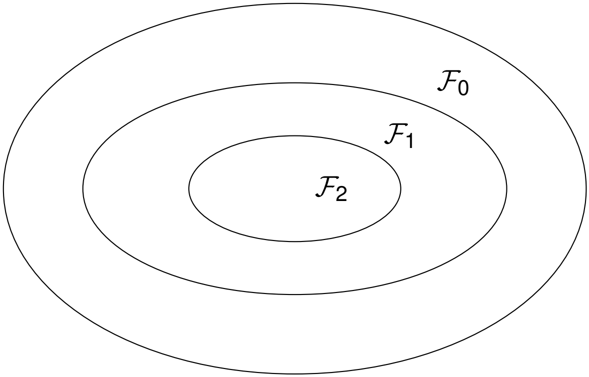
John Duchi
Stanford

Yuancheng Zhu
UPenn

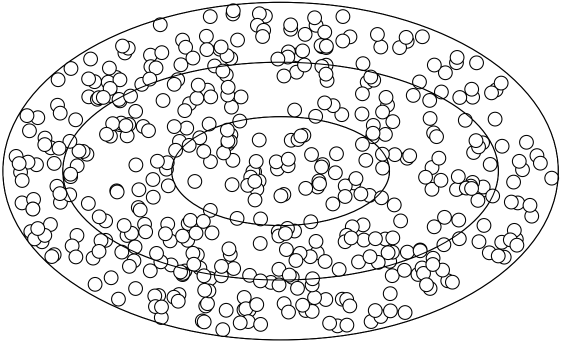
Perpective

- Stochastic gradient descent central to modern machine learning
- Our work shows that SGD is adaptive on a very fine scale
- Making sense of this requires a break with traditional formulations
    - A complexity class with a single member doesn't make sense using standard notions
- Ideas bridging optimization and statistics

Traditional adaptivity



Superadaptivity



Minimax complexity of convex optimization

- $\mathcal{F}$ *class of convex functions* on a convex set $\mathcal{C} \subset \mathbb{R}^d$.
- $\mathcal{O}$ a *stochastic first-order oracle*: query $(f, x) \in \mathcal{F} \times \mathcal{C}$, returns $Z \in \mathbb{R}^d$, with mean $f'(x) \in \partial f(x)$.
- $\mathcal{A}_T$ class of *all optimization methods* that make T queries to $\mathcal{O}$.

Minimax complexity

$$R_T(\mathcal{F}) = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}\big[\mathrm{err}(f, A)\big] = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}\Big[f(x_{T+1}) - \inf_{x \in \mathcal{C}} f(x)\Big]$$

Nemirovski and Yudin (1983)

Minimax complexity for convex optimization

Known that

$$R_T(\mathcal{F}_{\text{sc}}) \asymp 1/T \qquad \text{strongly convex functions}$$

$$R_T(\mathcal{F}_L) \asymp 1/\sqrt{T} \qquad \text{Lipschitz functions}$$

Agarwal et al. (2010) extend analysis to *d*-dimensional case, also sparse setting

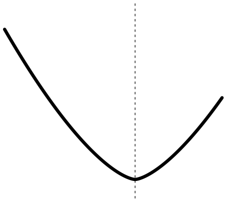
Raginsky and Rakhlin (2011) information theoretic proof technique; parallels minimax lower bounds in statistics

Shortcomings of the framework?

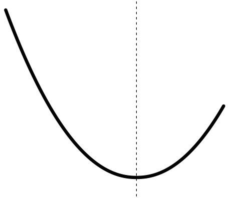
- Ignores cost of computing the gradient – $O(1)$
- Does not allow for (decreasing) bias
- Does not account for computations on past gradients, e.g., quasi-Newton algorithms
- *Too pessimistic and “global”*

Per-instance complexity

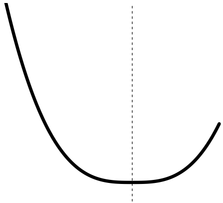
What about a particular instance?



$\ll 1/T?$



$\asymp 1/T?$



$\ll 1/\sqrt{T}?$

Local minimax complexity

Minimax complexity

$$R_T(\mathcal{F}) = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}\big[\mathrm{err}(f, A)\big]$$

Local minimax complexity

$$R_T(f; \mathcal{F}) = \sup_{g \in \mathcal{F}} \inf_{A \in \mathcal{A}_T} \max_{h \in \{f,g\}} \mathbb{E}\big[\mathrm{err}(h, A)\big]$$

T. Cai and M. Low, "A framework for estimation of convex functions," *Stat. Sinica*, 2015

Why is this interesting?

Local minimax complexity

$$R_T(f;\mathcal{F}) = \sup_{g\in\mathcal{F}} \inf_{A\in\mathcal{A}_T} \max_{h\in\{f,g\}} \mathbb{E}\big[\mathrm{err}(h,A)\big]$$

To show this is a reasonable complexity measure, need to:

1. Relate it to geometry of the function
2. Show it agrees with rates for known classes
3. Demonstrate that beating it is impossible
4. Give an algorithm that achieves it

Definitions

Set of minimizers	$\mathcal{X}_f^* = \arg\min_{x \in \mathcal{C}} f(x)$
Error function	$\operatorname{err}(x, f) = \inf_{y \in \mathcal{X}_f^*} \|x - y\|$
Minima separation	$d(f, g) = \inf_{x \in \mathcal{X}_f^*, y \in \mathcal{X}_g^*} \|x - y\|$ for $f, g \in \mathcal{F}$
Subgradient gap	$\kappa(f, g) = \sup_{x \in \mathcal{C}} \|f'(x) - g'(x)\|$

Modulus of continuity

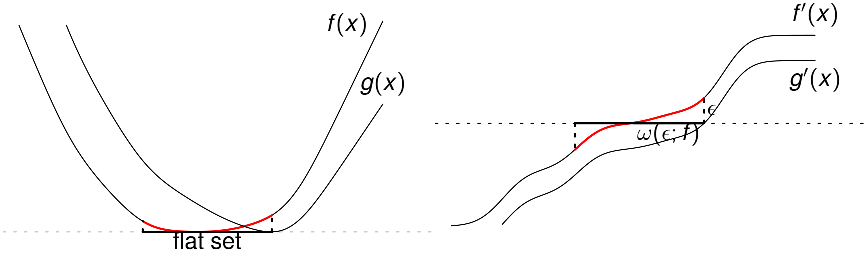
Two dissimilarity measures between f and g:

$d(f, g)$	*distance between minimizers*
$\kappa(f, g)$	*largest separation between subgradients*

Modulus of continuity of d with respect to κ at function f:

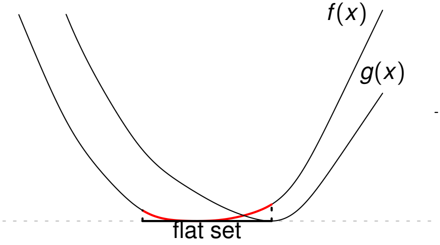
$$\omega_f(\epsilon) = \sup_{g \in \mathcal{F}} \left\{ d(f, g) \;:\; \kappa(f, g) \leq \epsilon \right\}$$

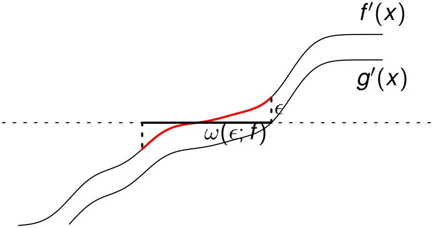
Modulus of continuity



$$\omega_f(\epsilon) = \sup \left\{ \inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{C}, |f'(y)| < \epsilon \right\}$$

Modulus of continuity



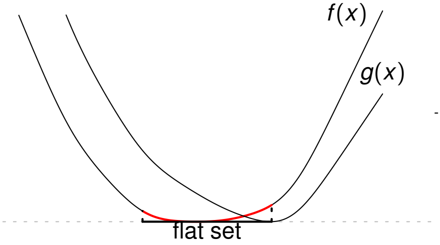


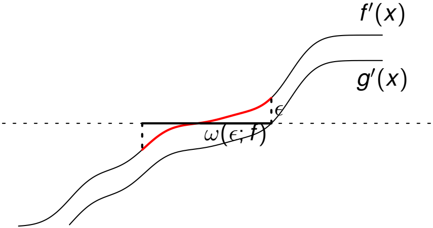
$$\omega_f(\epsilon) = \sup\left\{\inf_{x\in\mathcal{X}_f^*} |x - y| : y \in \mathcal{C}, |f'(y)| < \epsilon\right\}$$

$$f(x) = \frac{1}{\alpha}|x|^{\alpha}$$

$$\omega_f(\epsilon) = \epsilon^{1/(\alpha-1)}$$

Modulus of continuity





$$\omega_f(\epsilon) = \sup \left\{ \inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{C}, |f'(y)| < \epsilon \right\}$$

$$f(x) = \frac{1}{\alpha}|x|^\alpha \qquad f(x) = \begin{cases} \frac{1}{\alpha}|x|^\alpha & x \leq 0 \\ \frac{1}{\beta}|x|^\beta & x > 0 \end{cases}$$

$$\omega_f(\epsilon) = \epsilon^{1/(\alpha-1)} \qquad \omega_f(\epsilon) = \epsilon^{1/(\alpha \vee \beta - 1)}$$

Modulus characterizes local minimax

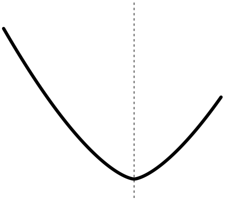
Theorem. For all sufficiently large T,

$$C_1\,\omega_f\left(\frac{\sigma}{\sqrt{T}}\right) \le R_T(f;\mathcal{F}) \le C_2\,\omega_f\left(\frac{\sigma}{\sqrt{T}}\right).$$

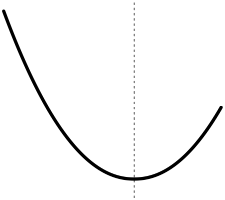
- Proof works for several distances $d(f,g)$
- For $f(x) = c|x - x^*|^\alpha$, x-error decays as $O(T^{-1/2(\alpha-1)})$, f-error decays as $O(T^{-\alpha/2(\alpha-1)})$
- Agrees with known rates for uniformly convex functions

Donoho and Liu, “Geometrizing rates of convergence: I, II, III”
Iouditski and Nesterov, “Primal-dual subgradient methods for minimizing uniformly convex functions” (2014)

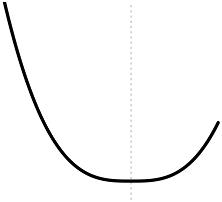
Per-instance complexity



$\asymp 1/T^{3/2}$



$\asymp 1/T$



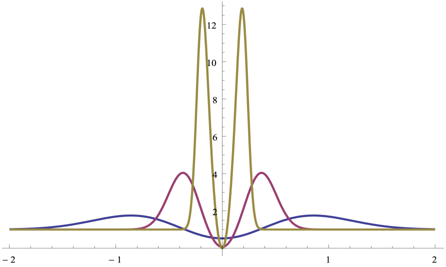
$\asymp 1/T^{2/3}$

Superefficiency

- In statistics, Fisher information is the benchmark for efficiency of (parametric) estimators

- The MLE satisfies $\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, I(\theta)^{-1})$

- LeCam-Hájek: If an estimator satisfies $\sqrt{(}\widetilde{\theta}_n - \theta) \rightsquigarrow N(0, v(\theta))$ with $v(\theta) \ll 1/I(\theta)$, then it must perform poorly at a nearby point



Superefficiency in convex optimization

Suppose an algorithm $A \in \mathcal{A}_T$ outperforms the modulus:

$$\mathbb{E}_f \operatorname{err}(\widehat{x}_A, f) \le \delta_T \, \omega_f \left(\frac{\sigma}{\sqrt{T}} \right),$$

with $\delta_T \to 0$, $e^T \delta_T \to \infty$. Then exists functions with $\kappa(f, g_T) \to 0$ and

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{g_T} \operatorname{err}(\widehat{x}_A, g_T)}{\omega_{g_T}\left(\sigma \sqrt{T^{-1} \log(1/\delta_T)} \right)} > 0.$$

Thus ω_f can be viewed as analogue of Fisher information for stochastic convex optimization

A superadaptive algorithm: Binary search

Given a budget of T queries:

- Query midpoint $T/\log T$ times
- Average the noisy gradients
- If result is positive, go left; otherwise right
- Step is correct if derivative bigger than $C/\sqrt{T}$.

Binary search algorithm

```
Input: $T, r$
Initialize: $(a_0, b_0)$, $E = \lfloor r \log T \rfloor$, $T_0 = \lfloor T/E \rfloor$,
for $e = 1, \ldots, E$ do:
    Query $x_e = (a_e + b_e)/2$ for $T_0$ times to get $Z_t^{(e)}$ for $t = 1, \ldots, T_0$
    Calculate the average $\bar{Z}_{T_0}^{(e)} = \frac{1}{T_0} \sum_{t=1}^{T_0} Z_t^{(e)}$
    If $\bar{Z}_{T_0}^{(e)} > 0$, set $(a_{e+1}, b_{e+1}) = (a_e, x_e)$
    If $\bar{Z}_{T_0}^{(e)} \leq 0$, set $(a_{e+1}, b_{e+1}) = (x_e, b_e)$
end
Output: $x_E$
```

Binary search achieves the benchmark

Theorem. With probability at least $1 - \delta$ and for large enough T,

$$\inf_{x \in \mathcal{X}_f^*} |x_T - x| \leq \widetilde{C} \omega_f \left(\frac{\sigma}{\sqrt{T}} \right)$$

where the term $\widetilde{C}$ hides a dependence on $\log T$ and $\log(1/\delta)$.

Superadaptation clearly seen in simulations.

Blockwise SGD is superadaptive in $d > 1$

- Can't (easily) do binary search in higher dimension
- Nesterov's blockwise SGD is extremely simple:

  1. Divide computational budget into equal chunks
  2. Run SGD in each block with fixed step size
  3. Halve the step size in each block

- Superadaptive (up to log factors)

Iouditski and Nesterov, "Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization" (2014)

Other settings Recent results of Yuancheng Zhu

	zeroth-order oracle	first-order oracle
fixed/random design	regression	1st-order regression
active design	0th-order optimization	optimization

$$R_{*,\dagger}(n, f) = \sup_{g \in \mathcal{F}} \inf_{A \in \mathcal{A}_{n,\dagger}} \max_{h \in \{f,g\}} \mathbb{E}_h \, \mathrm{err}_*(A, h)$$

See also...

J. Duchi and F. Ruan (2017), “Local asymptotics for stochastic optimization: Optimality, constraint identification, and dual averaging”

- Local minimax for optimization analogous to Hájek-Le Cam
- Based on Nesterov’s dual averaging

W. Su and Y. Zhu (2018), “Statistical inference for online learning and stochastic approximation via hierarchical incremental gradient descent”

- Inference (confidence intervals) for SGD
- Hierarchical design, computationally “free”

Part I: Summary

- Framework for assessing complexity of minimizing individual convex functions.
- Close connections to new and old statistical theory
- Challenge: More natural definition?

II: Graph structured signals

- Estimation and testing of signals that respect the structure of a network or graph in some way.
- We formulate a form of isotonic regression on graphs, and study the risk properties of the least squares estimator.

Trend filtering on graphs, Wang et al., (2014)
Normal means on graphs, Arias-Castro et al., (2008, 2014), Sharpnack (2013)
Lipschitz learning on graphs, Kyng et al., (2015)

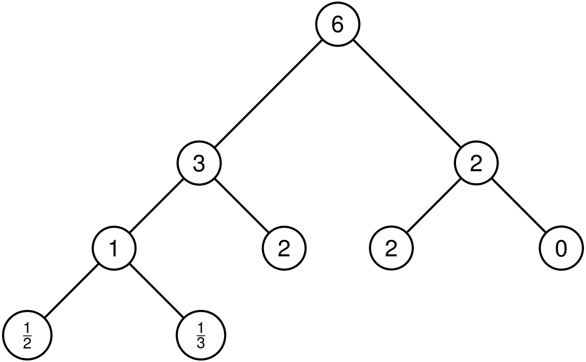
II: Graph structured signals

Sabyasachi Chatterjee

UIUC

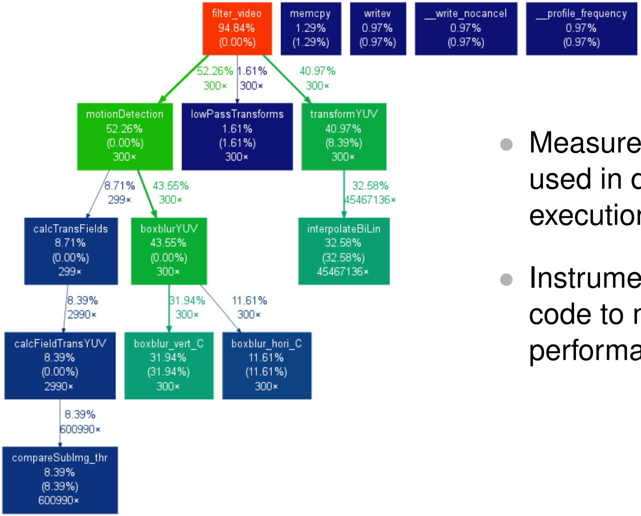
Flows on trees

Imagine a fluid flowing into a node and dividing among the children – possibly with some leakage.



We observe a noisy measurement $X_i = \mu_i + \epsilon_i$ at each node i

Example: Statistical code profiling (gprof)



- Measures time/storage used in different parts of execution tree
- Instruments compiled code to monitor performance

Denoising flows

- Graph structured form of isotonic regression.
  - What is the behavior of the least-squares estimator?
  - How does it depend on the structure of the tree?
  - What is the fundamental limit of flow estimation?

Isotonic regression well studied: Ayer et al. (1955), Brunk (1955), van Eeden (1958), Birgé and Massart (1993), Donoho (1991), van de Geer (1990), Zhang (2002), Chatterjee et al. (2013), ...

The isotonic case

For isotonic regression, $\mu_1 \geq \cdots \geq \mu_n$, risk of LSE scales as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mu_i - \widehat{\mu}_i)^2 \leq C\left(\frac{\sigma^2(\mu_1 - \mu_n)}{n}\right)^{2/3} = O\big(n^{-2/3}\big)$$

Matches the minimax rate.

C.H. Zhang (2002)

LSE for tree flows: Logarithmic depth

Theorem. Let T_n be a sequence of trees with n nodes and depth h_n. For any flow $\mu \in \mathcal{F}(T_n)$, the LSE has risk

$$\frac{1}{n}\mathbb{E}\|\widehat{\mu} - \mu\|^2 \leq C \left(\frac{\sigma^2 h_n (1 + \log n)^3}{n} + \frac{\sigma \mu_1 \sqrt{h_n} (1 + \log n)^{3/2}}{n} \right)$$

where μ_1 is the flow at the root.

LSE for tree flows: Logarithmic depth

Theorem. Let T_n be a sequence of trees with n nodes and depth h_n. For any flow $\mu \in \mathcal{F}(T_n)$, the LSE has risk

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|\widehat{\mu} - \mu\|^2 &\leq C\left(\frac{\sigma^2 h_n(1+\log n)^3}{n} + \frac{\sigma\mu_1\sqrt{h_n}(1+\log n)^{3/2}}{n}\right)\\ &= \widetilde{O}\left(\frac{h_n}{n}\right)\end{aligned}$$

where μ_1 is the flow at the root.

LSE for tree flows: Logarithmic depth

Theorem. Let T_n be a sequence of trees with n nodes and depth h_n. For any flow $\mu \in \mathcal{F}(T_n)$, the LSE has risk

$$\begin{aligned}\frac{1}{n}\mathbb{E}\|\widehat{\mu} - \mu\|^2 &\leq C\left(\frac{\sigma^2 h_n(1 + \log n)^3}{n} + \frac{\sigma\mu_1\sqrt{h_n}(1 + \log n)^{3/2}}{n}\right)\\ &= \widetilde{O}\left(\frac{h_n}{n} + \frac{\mu_1\sqrt{h_n}}{n}\right)\end{aligned}$$

where μ_1 is the flow at the root.

Hardest trees?

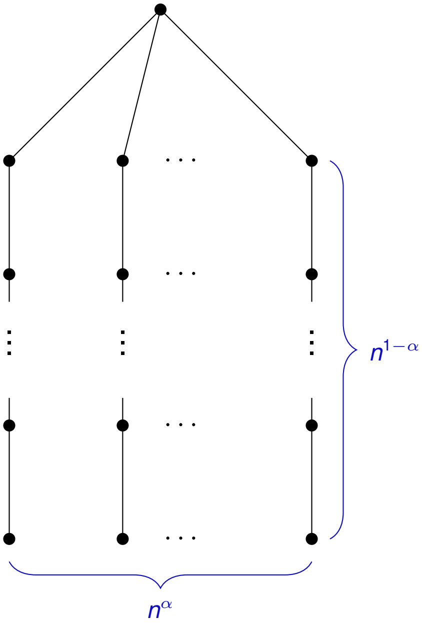
- Flow estimation is easier for stars than for paths

$$\widetilde{O}(1/n) \;\; vs. \;\; O(n^{-2/3})$$

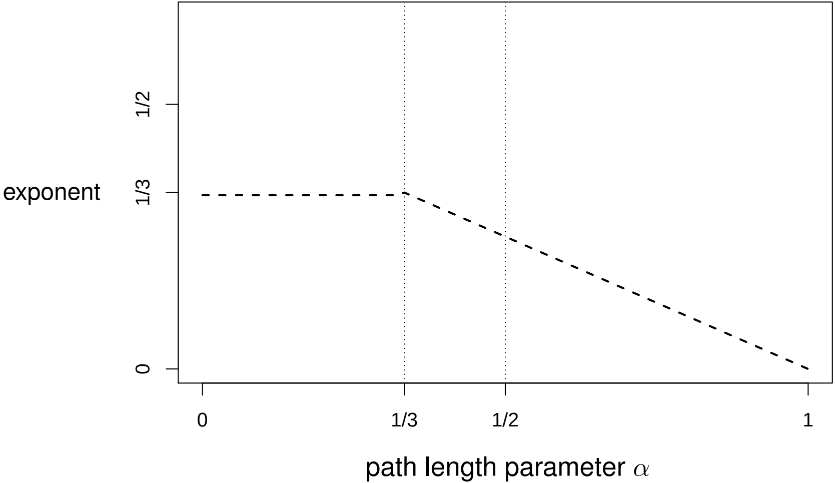
- Is the path the “hardest” flow estimation problem?

Reminiscent of Tan, Anandkumar and Willsky, “Learning Gaussian tree models” (2010)

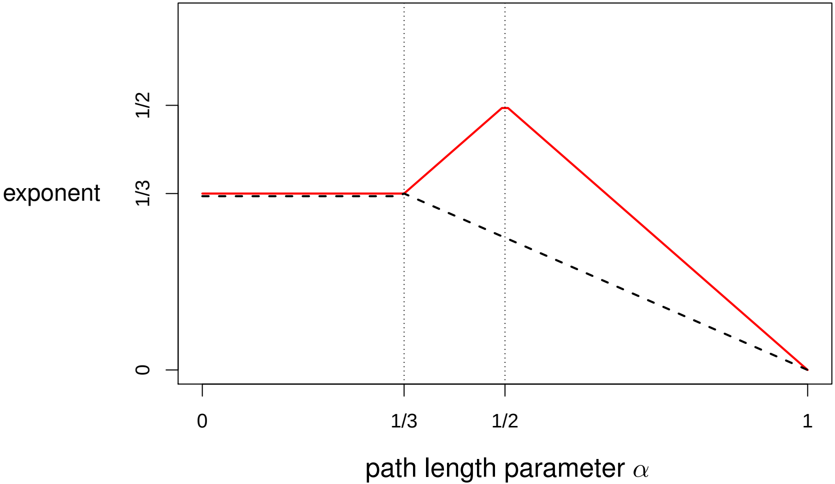
Many long paths $\mathcal{T}_{n,\alpha}$



Lower bounds for $\mathcal{T}_{n,\alpha}$



Lower bounds for $\mathcal{T}_{n,\alpha}$ and LSE



Proof techniques

Result	*Technique*
Upper bound for LSE; shallow trees	Gaussian supremum
Minimax lower bound; shallow trees	Fano's lemma
Lower bound for LSE; shallow trees	Gaussian widths
Isotonic upper bound for LSE; deep trees	Statistical dimension
Simplex upper bound for LSE; deep trees	Chaining
Minimax lower bound; monotone seqs	Assouad's lemma
Minimax lower bound; deep trees, $\alpha \leq \frac{1}{3}$	Minimax for isotonic
Minimax lower bound; deep trees, $\alpha \geq \frac{1}{3}$	Fano's lemma
Tightness of lower bound, $\alpha \geq \frac{1}{3}$	LSE on net
Tightness of LSE upper bound, $\alpha \geq \frac{1}{3}$	Gaussian widths

Proof techniques: GP suprema

For a fixed flow μ we define the Gaussian supremum function

$$f_\mu(t) := \mathbb{E}\Big(\sup_{\nu \in \mathcal{F} : \|\nu - \mu\| \leq t} \langle Z, \nu - \mu \rangle \Big) - \frac{t^2}{2}.$$

If $t^* > 0$ satisfies $f_\mu(t^*) \leq 0$ then

$$R(\widehat{\mu}, \mu) \leq \frac{C}{n} \max\left(t^{*2}, \sigma^2 \right).$$

We bound $f_\mu(t)$ by Dudley's entropy integrals (chaining)

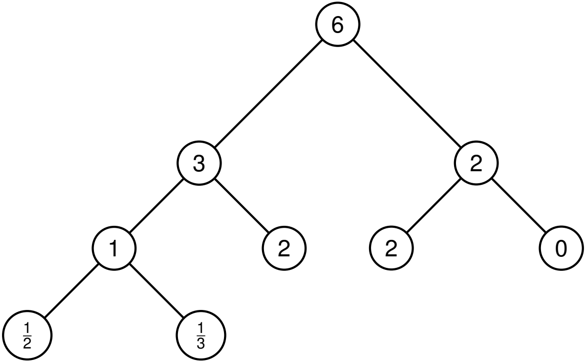
$$\mathbb{E}\Big(\sup_{\nu \in \mathcal{F} : \|\nu - \mu\| \leq t} \langle Z, \nu - \mu \rangle \Big) \leq C \int_0^{2t} \sqrt{\log N(\epsilon, B_t(\mu))}\, d\epsilon$$

which requires good upper bounds on log covering numbers.

Sourav Chatterjee (2014)

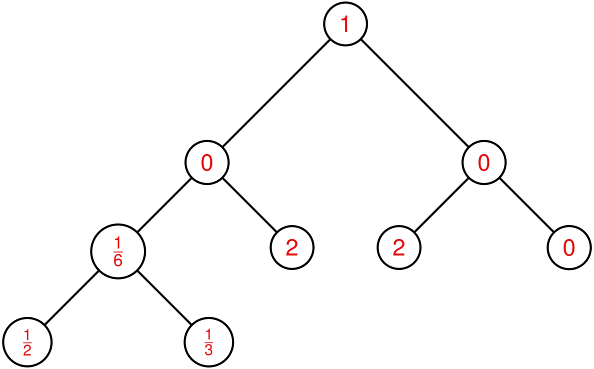
Covering number sketch: Flows and leaks

A flow is determined by its leaks. Let $\mu_1 \leq V$.



Covering number sketch: Flows and leaks

A flow is determined by its leaks. Let $\mu_1 \leq V$.



Random construction

Given a flow μ, define a random flow F by its leaks:

$$\ell(F) = \begin{cases} Ve_i & \text{with probability } \dfrac{\ell_i(\mu)}{V} \\ 0 & \text{with probability } 1 - \dfrac{1}{V}\displaystyle\sum_i \ell_i(\mu). \end{cases}$$

Take convex combination of m random flows:

$$\overline{\mu}^m = \frac{1}{m}(F_1 + \cdots + F_m)$$

Recursion

Since the node flows are negatively correlated, can show

$$\mathsf{Var}(\overline{\mu}_v^m) \leq \sum_{u \in \mathsf{Subtree}(v)} \mathsf{Var}(\bar{\ell}_u^m)$$

$$\begin{aligned} \mathbb{E}\|\overline{\mu}^m - \mu\|^2 = \sum_i \mathbb{E}(\overline{\mu}_i^m - \mu_i)^2 &= \sum_i \mathsf{Var}(\overline{\mu}_i^m) \\ &\leq \sum_i d_i \mathsf{Var}(\bar{\ell}_i^m) \\ &\leq h \sum_i \mathsf{Var}(\bar{\ell}_i^m) \\ &\leq \frac{V^2 h}{m} \end{aligned}$$

Covering number

Random construction thus gives a $V^2 h/m$ covering. By simple combinatorics,

$$\log N(\epsilon, \mathcal{F}_V) \leq \frac{V^2 h}{\epsilon^2}\left(1 + \log\left(1 + \frac{n\epsilon^2}{V^2 h}\right)\right)$$

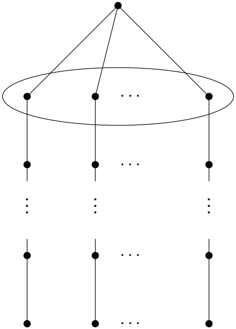
“Maurey’s argument” (van der Vaart and Wellner, 1996)

Intuition for the gap

General minimax lower bound for isotonic regression:

$$\inf_{\widetilde{\mu}} \sup_{\mu:\mu_1 \le V} \mathbb{E}\|\widetilde{\mu} - \mu\|^2 \ge C \min\left\{\sigma^2, V^2, \left(\frac{\sigma^2 V}{n}\right)^{2/3}\right\}$$

- When V is small, LSE not minimax—trivial estimator $\widehat{\mu} = 0$ achieves lower bound
- In "narrow" tree regime $\alpha \approx 1/3$, some of the root flows will be small
- A lasso-style estimator to zero entire paths?



II: Summary

- Flows: Tree structured generalization of isotonic regression
- Surprise: LSE is not minimax rate optimal
- Shape constraints in graph/network settings largely unexplored

S. Chatterjee and JL, “Denoising flows on trees,” *IEEE Trans. Info. Theory*, 2018

III: Prediction rule reshaping

- Imposing shape constraints on “black box” prediction rules
- Reshaping random forests

Matt Bonakdarpour
UChicago/Yale

Rina Foygel Barber
UChicago

Related work: Chernozhukov et al., 2010, González et al., 2015, Gupta et al., 2016, You et al., 2017, Amos et al., 2017

Motivation

- Shape constraints are natural in applications
  - House price assumed to be decreasing function of crime rate (all other predictors held constant)
- Not easily incorporated into popular machine algorithms
- We study different approaches to *reshaping* prediction rules

Prediction rule reshaping

Suppose that $\widehat{f} : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a prediction rule estimated from data (e.g., regression or classification)

Let $\mathcal{S}$ be a class of functions that satisfy a set of shape contraints. For example

$$\begin{aligned} \frac{\partial f}{\partial x_j} &\geq 0 \quad \text{(monotonic)}, \ j \in \mathcal{R}_{\uparrow} \\ \frac{\partial^2 f}{\partial x_j^2} &\leq 0 \quad \text{(concave)}, \ j \in \mathcal{R}_{\cap} \end{aligned}$$

Reshaping is the infinite dimensional optimization

$$\min_{f \in \mathcal{S}} \|f - \widehat{f}\|$$

A finite optimization problem

Let $\mathcal{S} = (\mathcal{S}_v)_{v \in \mathcal{R}}$ be candidate shape constraints, $\mathcal{R} \subset [d]$

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be set of test points

For $v \in \mathcal{R}$, define the $n \times n$ prediction matrix $\widehat{F}^v = \left[\widehat{F}^v_{i',i}\right]$

$$\begin{aligned} \widehat{F}^v_{i',i} &= \widehat{f}(x_{i,1}, x_{i,2}, \ldots, x_{i,v-1}, x_{i',v}, x_{i,v+1}, \ldots, x_{i,d}) \\ \widehat{F}^v_i(\cdot) &\equiv \widehat{f}(x_{i,1}, x_{i,2}, \ldots, x_{i,v-1}, \cdot, x_{i,v+1}, \ldots, x_{i,d}). \end{aligned}$$

where $x_{i',v}$ ranges over all n values of v-th predictor.

A finite optimization problem

Reshaped predictions $\check{f}_{\mathcal{S}}$ *obtained by projecting matrix of predicted values onto shape constraints*:

$$\begin{aligned}\check{F}_{\mathcal{S}} = \underset{F=(F^v)_{v\in\mathcal{R}}}{\arg\min} & \sum_{v\in\mathcal{R}} \|F^v - \widehat{F}^v\|_F^2 \\ \text{such that } & F_i^v \in \mathcal{S}_v, \text{ for each } v \in \mathcal{R} \\ & \operatorname{diag}(F^v) = \operatorname{diag}(F^w), \text{ for all } v, w \in \mathcal{R}\end{aligned}$$

Reshaped predictions:

$$\check{f}_{\mathcal{S}} = \operatorname{diag}(\check{F}_{\mathcal{S}}^v)$$

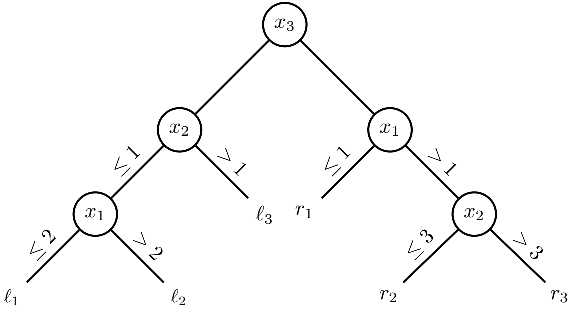
Intersecting isotonic regressions

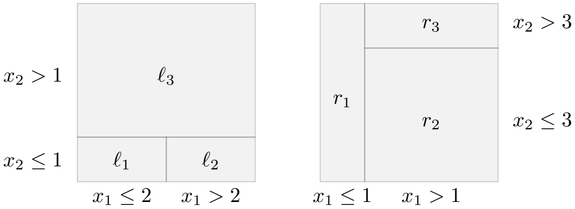
- Suppose constraints are monotonicity contraints
- Reshaping leads to “intersecting isotonic regressions” due to consistency constraints $\mathrm{diag}(F^v) = \mathrm{diag}(F^w)$
- A generalization of PAVA solves this efficiently

Reshaping random forests

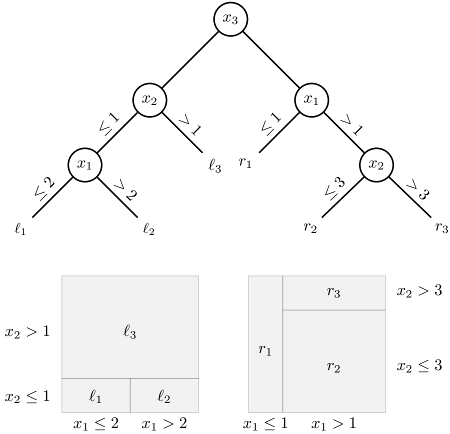
1. Grow the tree in the usual way
2. Reshape the leaf values to enforce monotonicity

Illustration





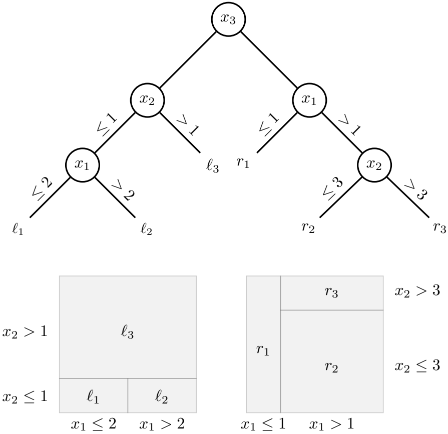
Illustration



Exact estimator, 6 constraints:

$$\mu_{\ell_2} \leq \mu_{r_2},\ \mu_{\ell_1} \leq \mu_{r_1},\ \ \mu_{\ell_1} \leq \mu_{r_2},\ \mu_{\ell_3} \leq \mu_{r_1},\ \ \mu_{\ell_3} \leq \mu_{r_2},\ \mu_{\ell_3} \leq \mu_{r_3}$$

Illustration



Over-constrained estimator, all 9 pairwise constraints

$$\mu_{\ell_1} \leq \mu_{r_1},\ \mu_{\ell_1} \leq \mu_{r_2},\ \mu_{\ell_1} \leq \mu_{r_3}, \ldots$$

Reference

“Prediction rule reshaping,” Matt Bonakdarpour, Sabyasachi Chatterjee, Rina Foygel Barber, and JL, arXiv:1805.06439

(to be presented at ICML this summer)

Summary

Interplay between global and local structure:

- Stochastic convex optimization
- Signal denoising on trees
- Reshaping prediction rules

Merci beaucoup d'avoir m'écouté!