Nonparametric Regression meets Online Learning

Wavelets, Local Adaptivity and T^{1/3} Dynamic Regret

Yu-Xiang Wang

Based on joint work with Dheeraj Baby

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Outline

- A tour of locally adaptive nonparametric regression

- From regression to forecasting

- Open problems

Nonparametric regression

- 50+ years of associated literature

  [Nadaraya, Watson, 1964]

  - Kernels, splines, local polynomials
  - Gaussian processes and RKHS
  - CART, neural networks

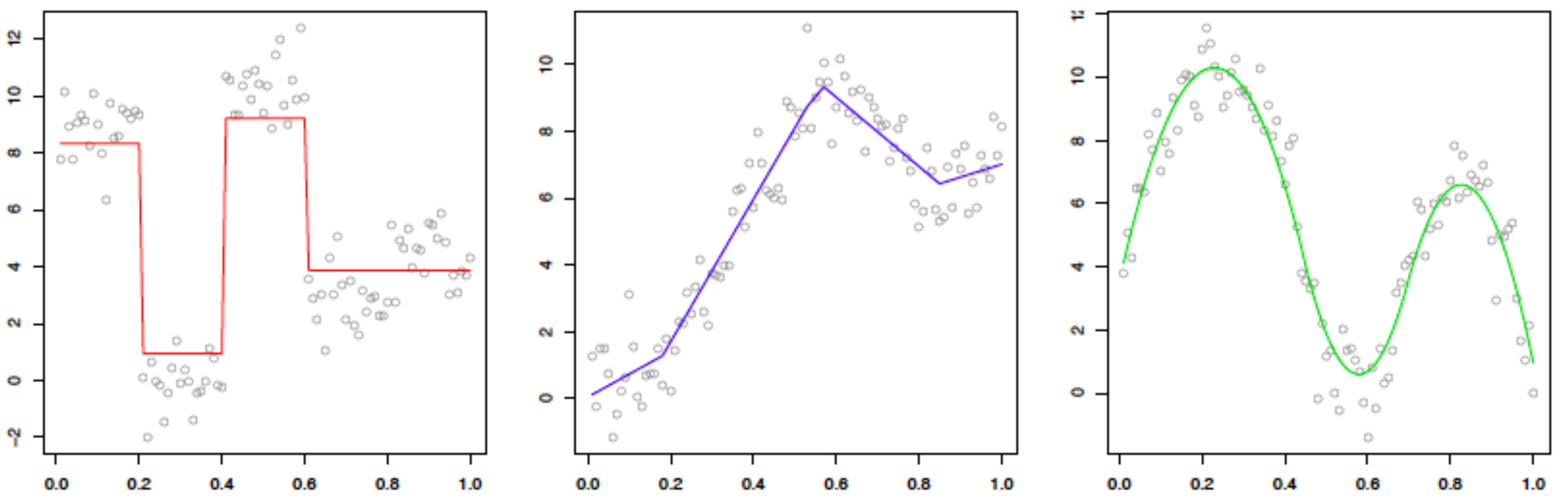
- Also known as smoothing, signal denoising /filtering in signal processing & control.

Adapting to local smoothness

- Some parts smooth, other parts wiggly.
  - Wavelets [Donoho&Johnston,1998], adaptive kernel [Lepski,1999], adaptive splines [Mammen&Van De Geer,2001]
  - a.k.a, multiscale, multi-resolution compression, used in JPEG2000.
  - New comer: Trend filtering! [Steidl,2006; Kim et. al. 2009, Tibshirani, 2013; W.,Smola, Tibshirani, 2014]

Univariate trend filtering

$$\min_{\beta \in \mathbb{R}^n} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda \|D^{(k+1)}\beta\|_1$$



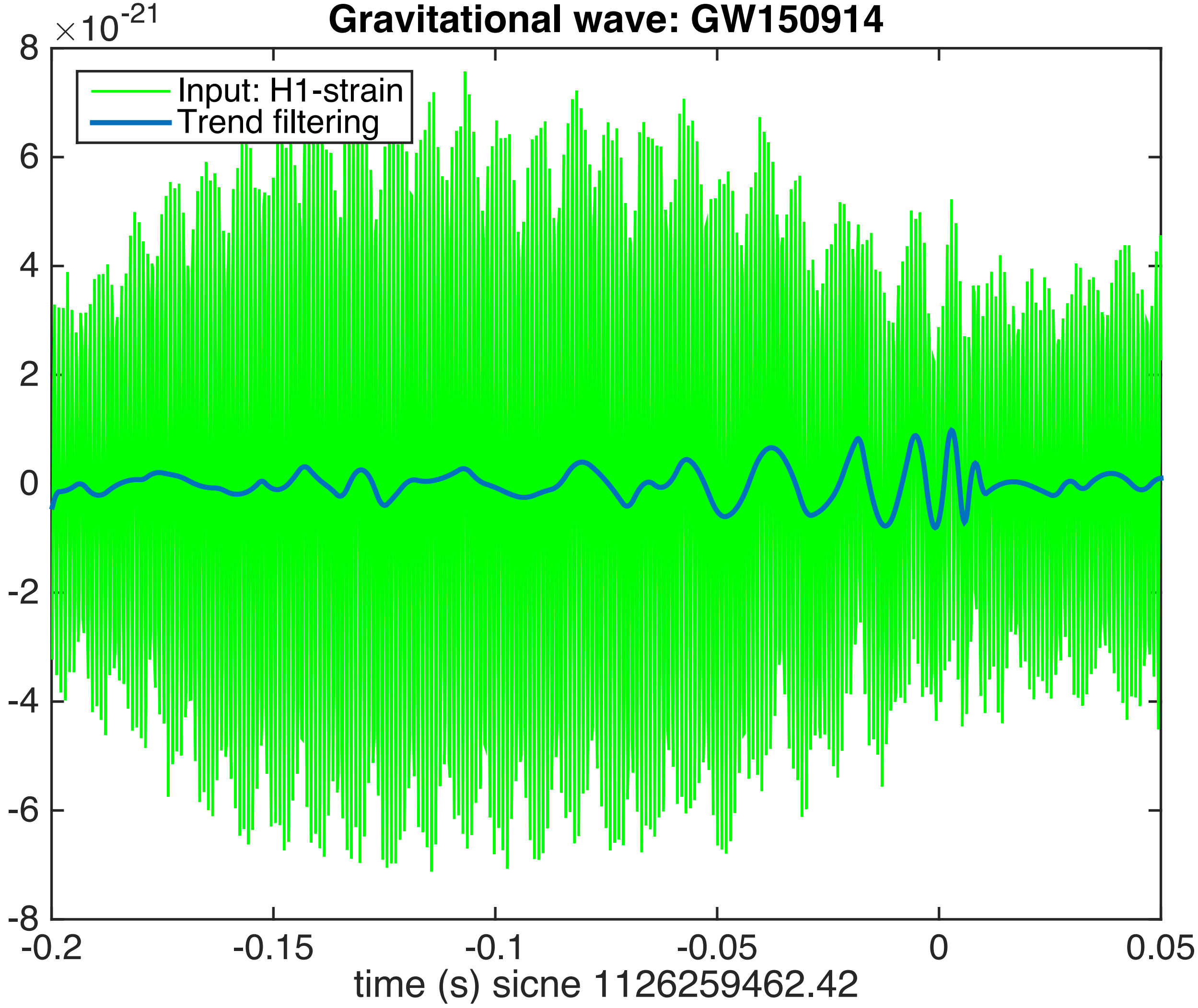
Constant, $k = 0$
(Fused lasso)

Linear, $k = 1$

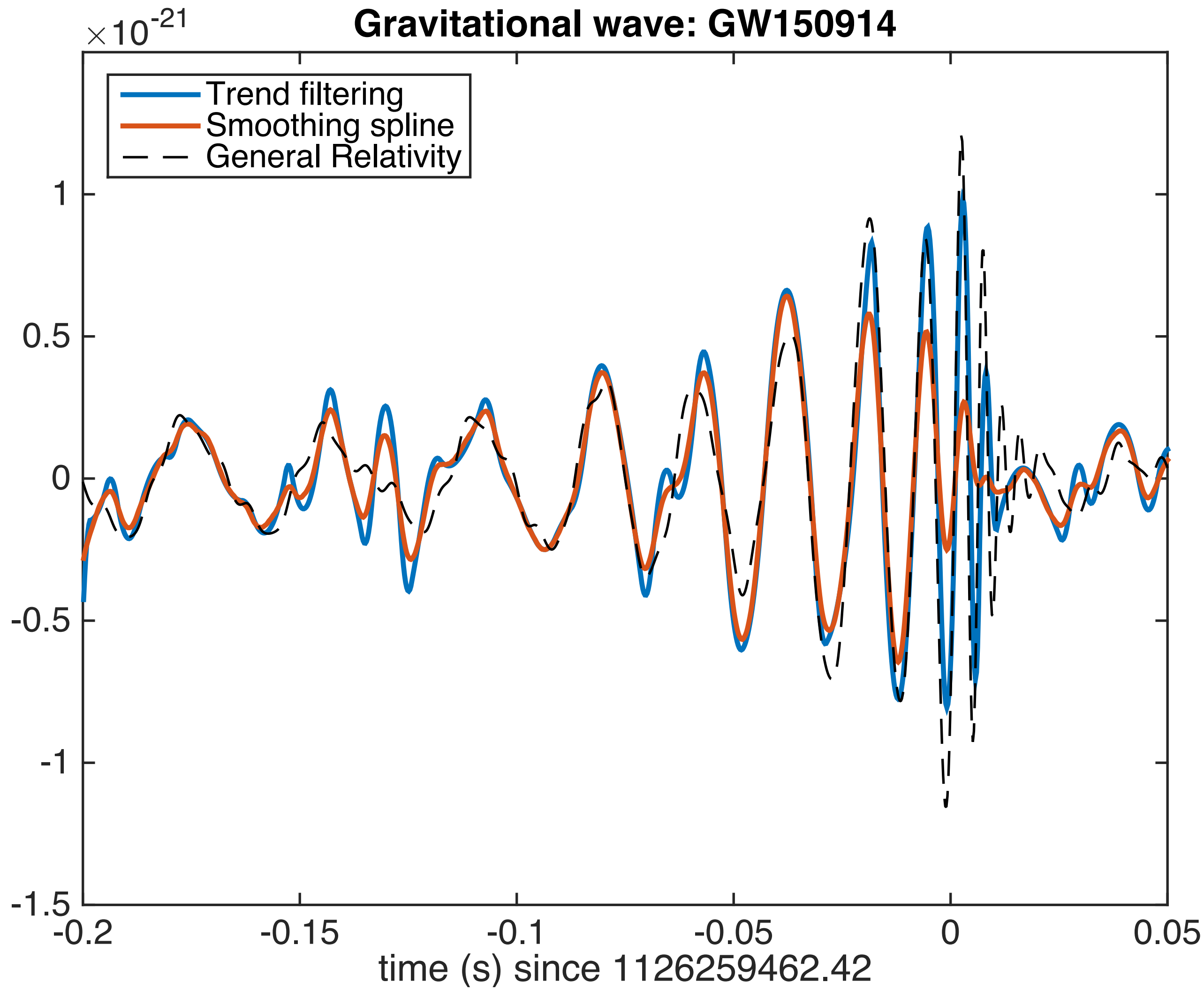
Quadratic, $k = 2$

(figure extracted from: Tibshirani (2014))

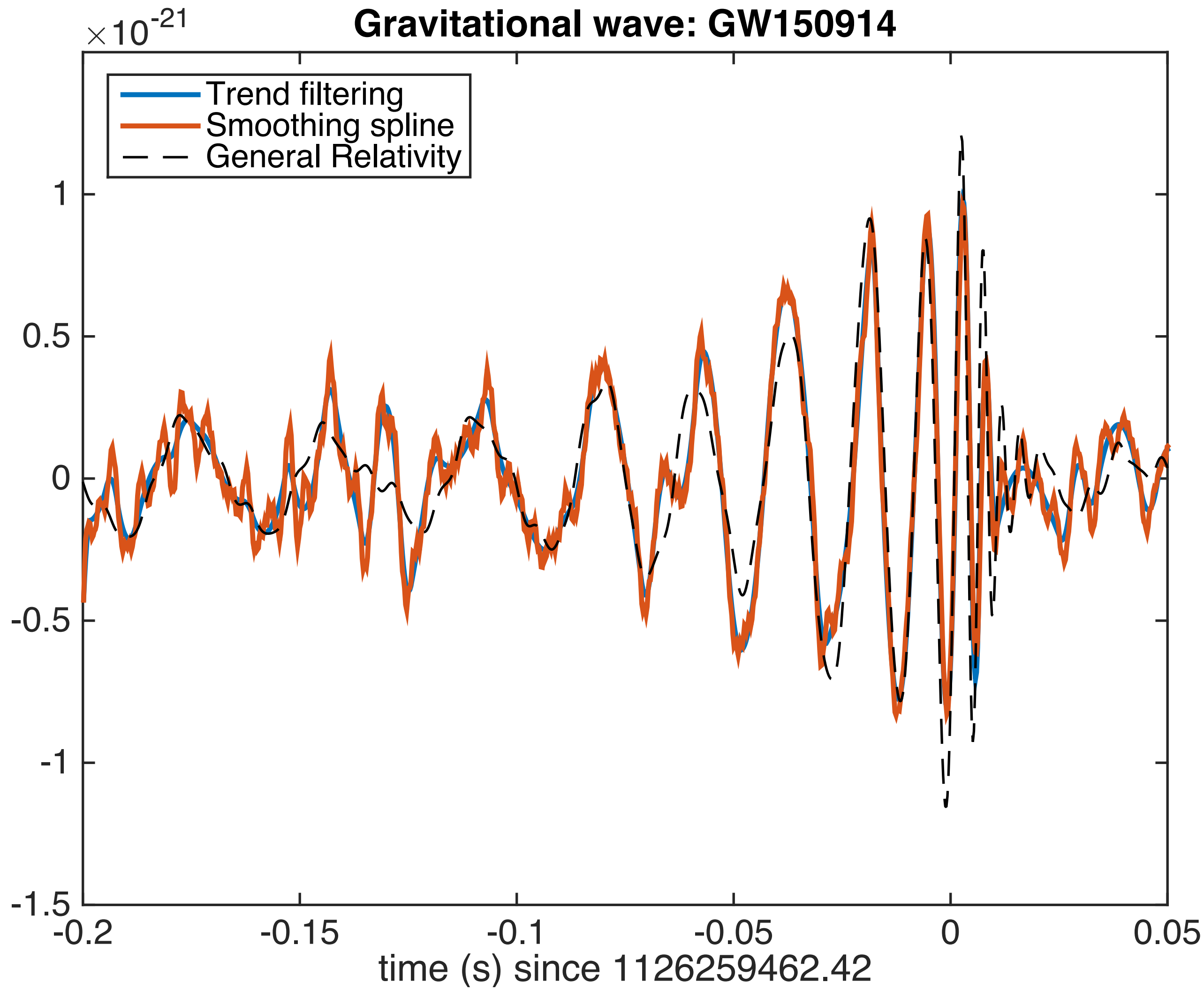
A BIG Example: merger of two black holes



A BIG Example: merger of two black holes



A BIG Example: merger of two black holes



Theory behind trend filtering

(Tibshirani, 2014, Annals of Statistics)

- Observations: $$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots n$$

- TV-class: $$\mathcal{F}_k = \{f : \mathrm{TV}(f^{(k)}) \leq C\}$$

- Error rate: $$O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$$

- Best achievable rate for linear smoothers

$$O_{\mathbb{P}}(n^{-(2k+1)/(2k+2)})$$

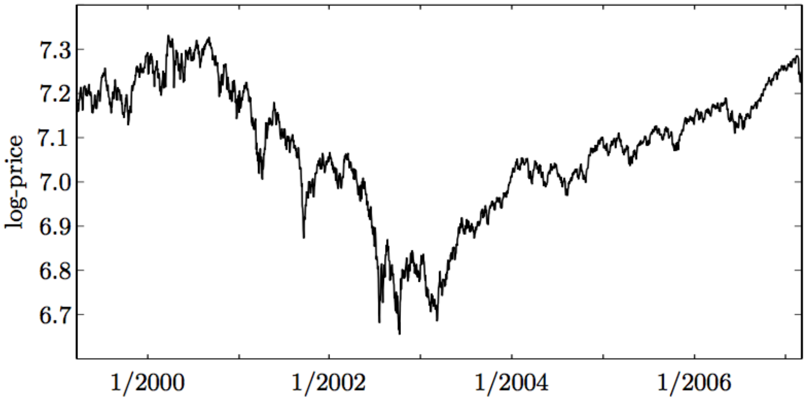
Generalizations of trend filtering

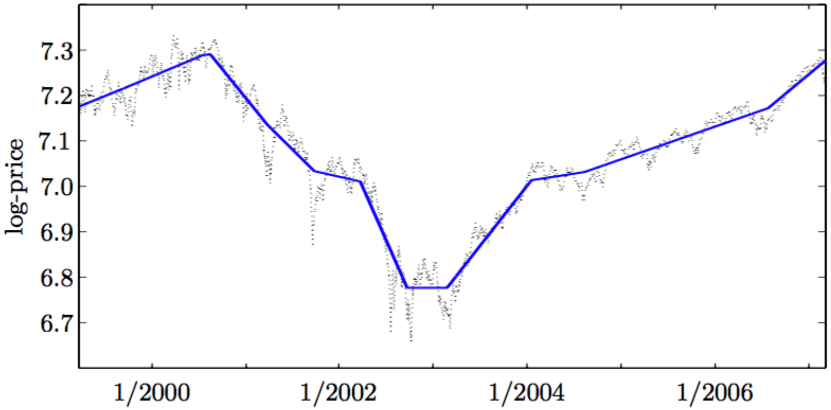
- To multi-dimensional signal observed on a lattices/grid: images/video
  - d>1, k=0 (Sadhanala, W., Tibshirani, NeurIPS 2016)
  - d=2, k>0 (Sadhanala, W., Tibshirani, NeurIPS 2017)
- To signals on a general graphs
  - (W., Sharpnack, Smola, Tibshirani, JMLR 2016)
- Type of results:
  - Minimax rate, minimax linear rate, adaptivity, phase transition phenomena
  - fast algorithms, various applications.

Story of another time.

Back to univariate trend filtering: does it solve the motivating application?

- L1-trend filtering (Kim et al, 2009)
  - Motivation: time series!
  - e.g., SnP500, CO2 emission, market demand





- Two major problems in time series:
  - Forensics: making things of what happened.
  - **Forecasting: predict the future**

This talk: Online Nonparametric Forecasting

$\boxed{\text{Individual sequence} \quad \theta_1, \ldots, \theta_n \in \mathbb{R}}$

- At each time step $t = 1, \ldots, n$
  - Prediction $\hat{\theta}_t$ is made by the forecaster
  - $y_t = \theta_t + \epsilon_t$, $\epsilon_t \sim iid \, \mathrm{subgauss}(0, \sigma^2)$ is revealed by Nature

Minimize the Total Squared Error (TSE): $R(n) = \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2]$

This talk: Online Nonparametric Forecasting

Individual sequence $\theta_1, \ldots, \theta_n \in \mathbb{R}$

- At each time step $t = 1, \ldots, n$
  - Prediction $\hat{\theta}_t$ is made by the forecaster
  - $y_t = \theta_t + \epsilon_t$, $\epsilon_t \sim iid\, \mathrm{subgauss}(0, \sigma^2)$ is revealed by Nature

Minimize the Total Squared Error (TSE): $R(n) = \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2]$

More difficult than batch problem where one observes all noisy data points before fitting the data

Weaken the adversary

- We aim to build a forecaster that has sub-linear TSE as a function of n against all possible ground truth sequences
- Impossible unless some regularity conditions are applied to the adversary's moves
- Hence need to restrict ourselves to some class of ground truth sequences

Bounded Variation Class

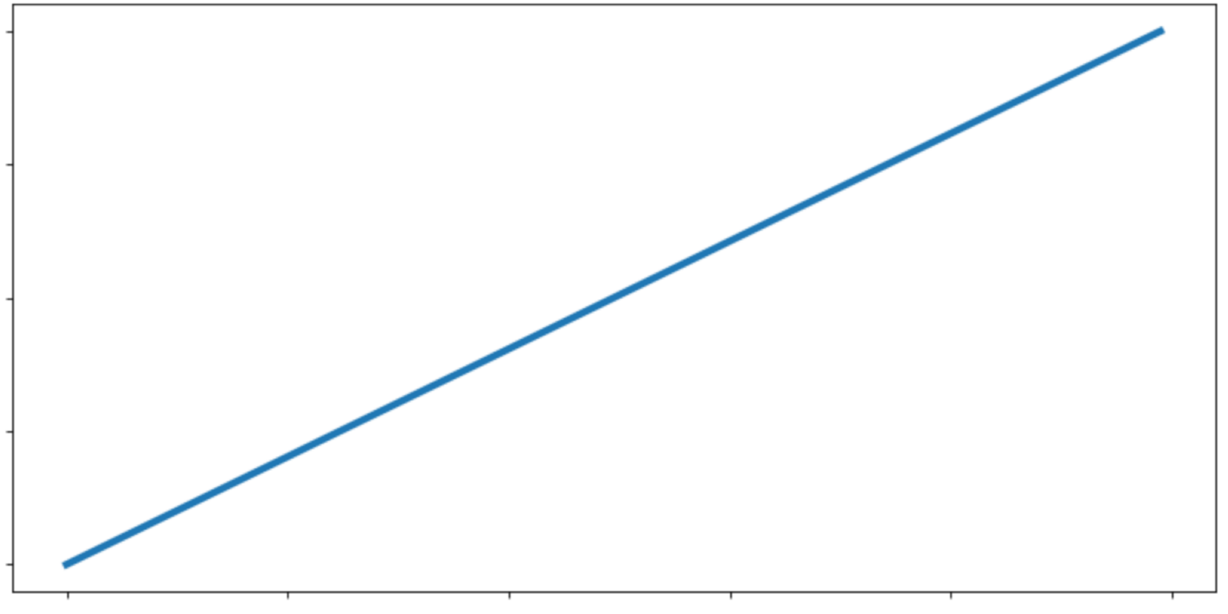
- Bounded variation sequences $\Theta = (\theta_1, \ldots, \theta_n)^T \in \mathbb{R}^n$

$$\text{where } \|D\Theta\|_1 = \sum_{t=2}^{n} |\theta_t - \theta_{t-1}| \leq C_n$$

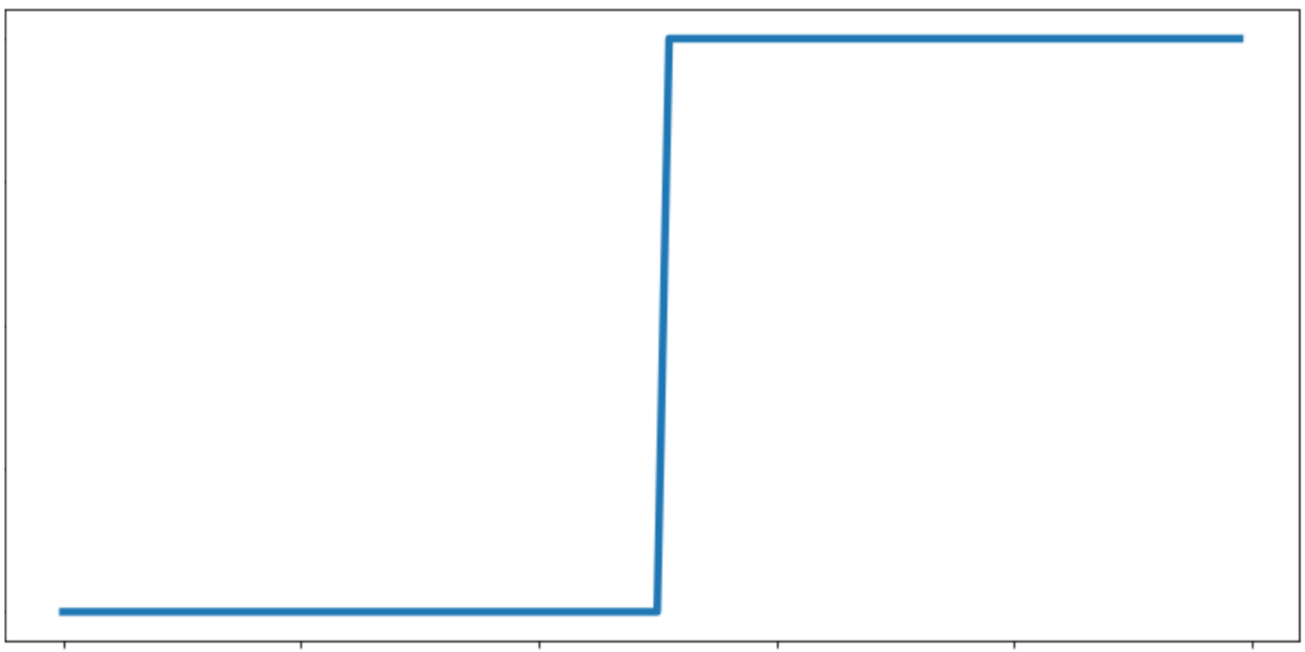
From trend filtering problems, this is the Total Variation class with k=0, d=1.

- Constrain the variation budget
- Features a rich class of sequences

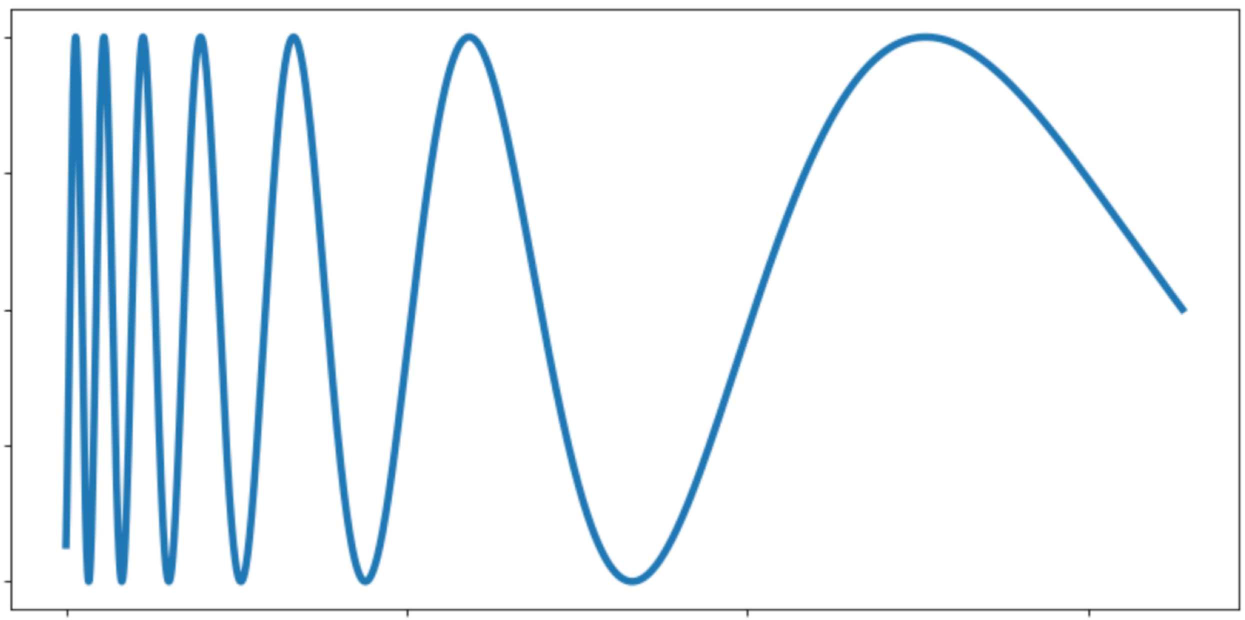
Spatially Homogeneous trends



Spatially Inhomogeneous trends



Spatially Inhomogeneous trends



Minimax TSE

$$\tilde{R}(n) = \min_{algos}\left(\max_{\Theta;\|D\Theta\|_1 \le C_n} R(n)\right)$$

- For batch setting: $\tilde{R}(n) = \Omega(n^{1/3}\sigma^{4/3}C_n^{2/3})$ → **From theory of non-parametric regression**

[6] Donoho et.al

Minimax TSE

$$\tilde{R}(n) = \min_{algos} \left(\max_{\Theta; \|D\Theta\|_1 \le C_n} R(n) \right)$$

- For batch setting: $\tilde{R}(n) = \Omega(n^{1/3}\sigma^{4/3}C_n^{2/3})$
- It can be shown that for forecasting:

$$\tilde{R}(n) = \Omega(n^{1/3}\sigma^{4/3}C_n^{2/3} + \boxed{C_n^2}) \longrightarrow$$

Forecasting is harder than smoothing

This is a very basic problem, what are existing ways of solving it?

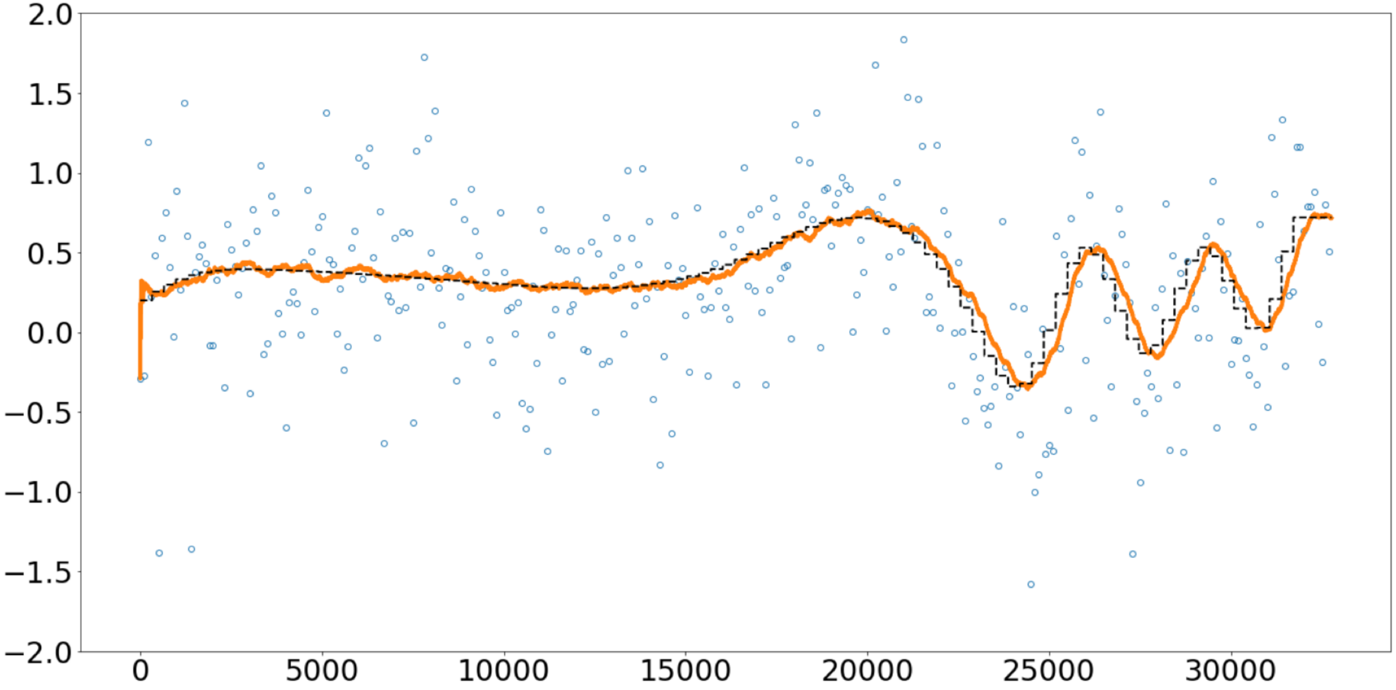
- Classical time-series forecasting
  - AR, MA, ARMA, ARiMA (Box-Jenkins style)

- Modern online learning and dynamic regret
  - Incur an online sequence of square losses.
  - Receive noisy gradients as feedback
  - TSE = Dynamic Regret

$$\sum_{t=1}^{n} \ell_t(\hat{\theta}_t) - \sum_{t=1}^{n} \min_{\theta_t} \ell_t(\theta_t)$$

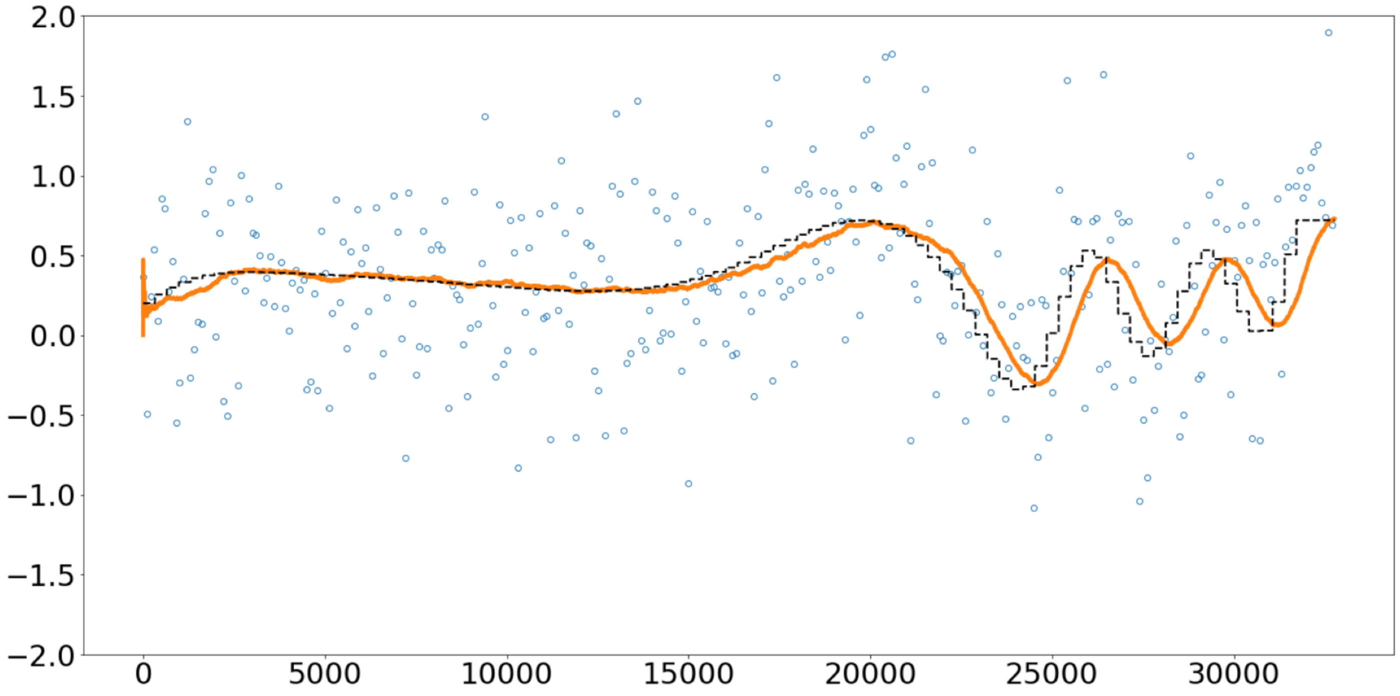
- What? Pointwise optimal comparators?
  - Constrain how quickly loss functions can change (Besbes et al, 2013)
  - Alternative view: constrain the comparator sequence (Zinkevich, 2003)

Why Moving Averages won't work?



MA win-size = 500

Why Moving Averages won't work?



MA win-size = 1000

Linear Forecasters are sub-optimal

- MA is a Linear Forecaster: a policy that predicts a fixed linear function of past observations
- It can be shown that:

$$\tilde{R}_{lin}(n) = \min_{algos} \left(\max_{\Theta; \|D\Theta\|_1 \leq C_n} R(n) = \Omega(n^{1/2}) \right)$$

Restrict to linear algorithms

[6] Donoho et.al

ARMA, Exponential Smoother, Online Gradient Descent etc

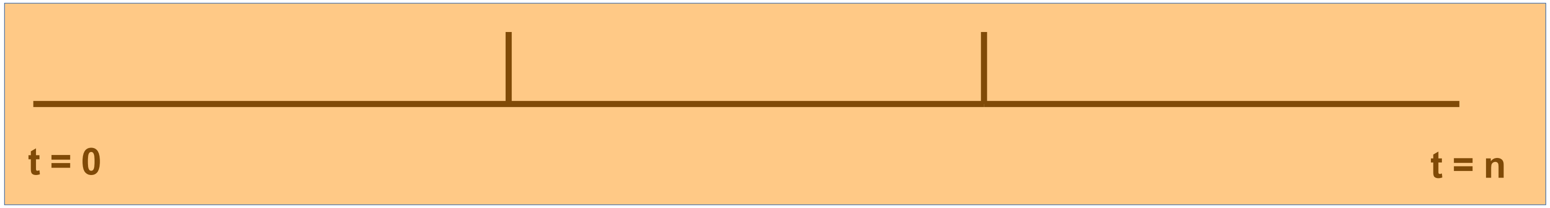
Existing policies are suboptimal

Policy	TSE	Lower bound
Restarting OGD [1,2]	$O(n^{1/2})$	$\Omega(n^{1/3})$
AOMD [3]	$O(n^{1/2})$	$\Omega(n^{1/3})$

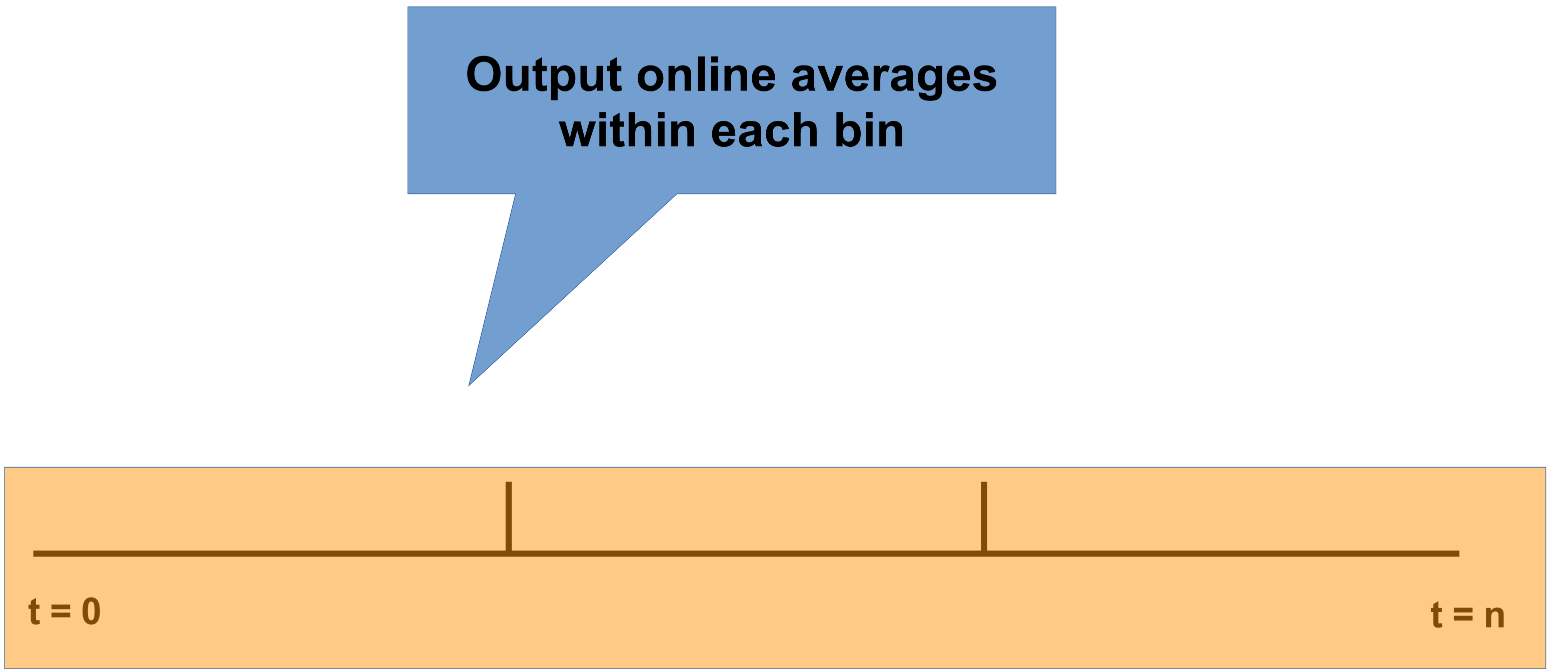
Restarting OGD in our setting

Partition the time horizon into <u>fixed size</u> bins

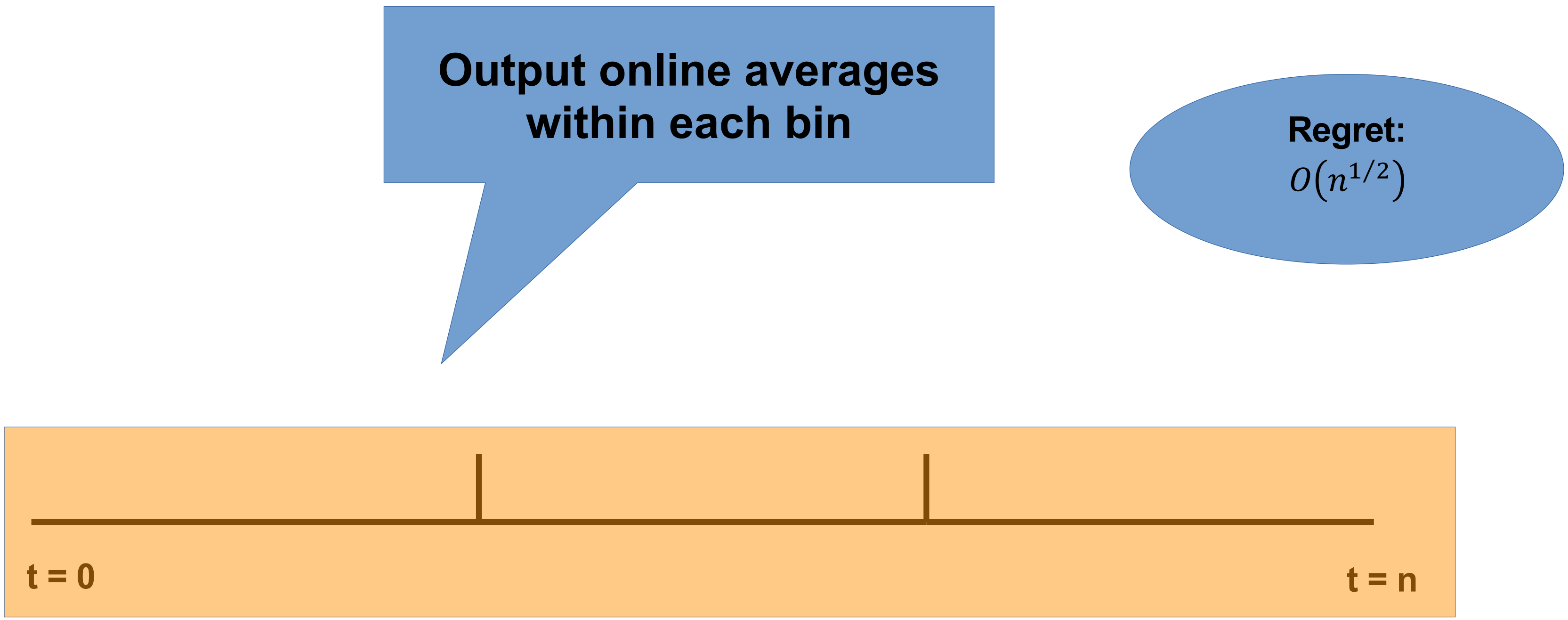
Size of bins determined by σ **n, and C**



Restarting OGD in our setting



Restarting OGD in our setting



A method to design optimal policy

- Restarting online averages
- Key Idea: Adaptively choose the restarting schedule
- Restart only when enough Total Variation is detected
- Adaptively partition the time horizon into various bins

1. A TV lower bound => Bound # of times we restart
2. A TV upper bound => Upper bound the error of a fixed baseline comparator

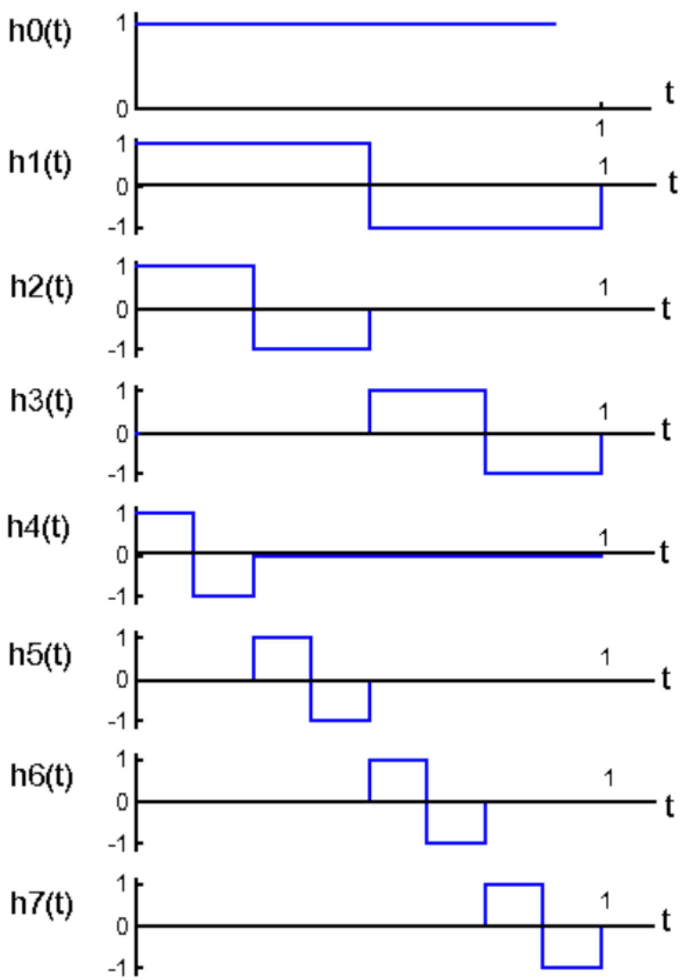
Wavelets, and wavelet smoothing

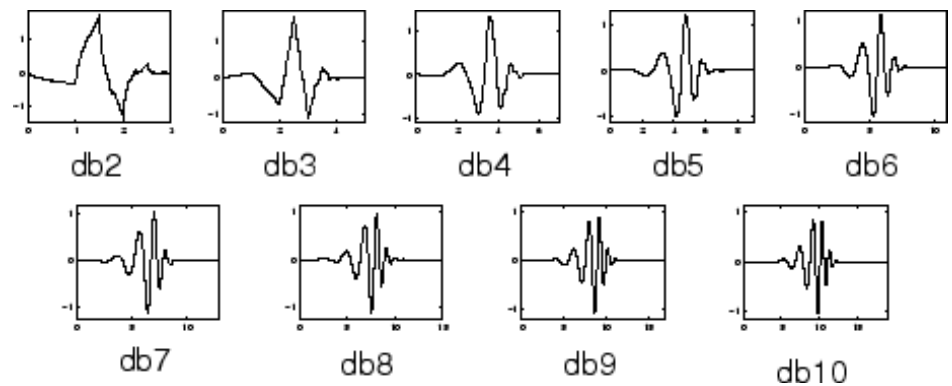
- Classical (Haar, 1909) (Ricker, 1953)

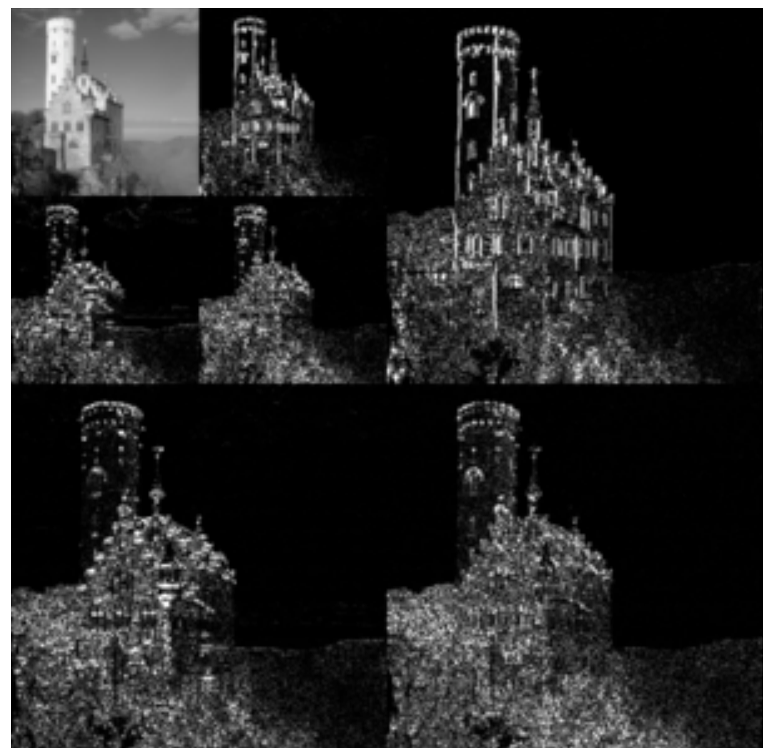
(Alfréd Haar, 1885 - 1933)
PhD student of David Hilbert

- A lot of developments in 1980s and 1990s
  - e.g., Daubechies, Coifmans et al (1980s)
- Use in statistics / statistical signal processing
  - Donoho and Johnstone (1998) et al
- Implementation: e.g., JPEG2000, DjVu, Multi-resolution analysis

Examples of wavelets and wavelet transforms







Wavelet smoothing in one slide

- Model: $y = \theta + \text{noise}$

- Wavelet smoothing algorithm
  1. Wavelet Transform: $\alpha = Hy$
  2. Thresholding: $\hat{\alpha} = \text{Soft-Threshold}_{\lambda}(\alpha)$
  3. Reconstruction: $\hat{\theta} = H^{-1}\hat{\alpha}$

Remarkable adaptivity of wavelet smoothing

- Choose $$\lambda = \sigma\sqrt{2\log n}$$
  - (or use SUREShrink as an adaptive choice)

- Where are the functions coming from:
  - Holder classes, Sobolev classes, Total Variation classes
  - Besov class(p,q, R)

- Donoho (1995), Donoho & Johnstone (1998):
  - Wavelet smoothing is **simultaneously minimax** (up to a log n term) for all **p, q, R > 0** in the Besov class

ARROWS: **A**daptive **R**estarting **R**ule in **O**nline averaging with **W**avelet **S**hrinkage

ARROWS: inputs - observed y values, time horizon n, $\delta \in (0, 1]$, total variation bound C_n, a hyper-parameter $\beta > 6$

1. Initialize $t_h = 1$, $newBin = 1$, $y_0 = 0$
2. For $t = 1$ to n:
   (a) if $newBin == 1$, predict $x_t^{t_h} = y_{t-1}$, else predict $x_t^{t_h} = \bar{y}_{t_h:t-1}$
   (b) set $newBin = 0$, observe y_t and suffer loss $(x_t^{t_h} - \theta_t)^2$
   (c) Let $\hat{y} = pad_0(y_{t_h}, ..., y_t)$ and k be the padded length.
   (d) Let $\hat{\alpha}(t_h : t) = T(H\hat{y})$
   (e) **Restart Rule:** If $\frac{1}{\sqrt{k}} \sum_{l=0}^{\log_2(k)-1} 2^{l/2} \|\hat{\alpha}(t_h : t)[l]\|_1 > n^{-1/3} C_n^{1/3} \sigma^{2/3}$ then
      i. set $newBin = 1$
      ii. set $t_h = t + 1$

Our Main Results

- By using wavelet soft-thresholding as the child smoother, our policy achieves the minimax regret:

$$\tilde{R}(n) = \tilde{O}(n^{1/3}\sigma^{4/3}C_n^{2/3} + \|D\Theta\|_2^2)$$

- With nearly linear run-time of $O(n \log n)$

- The additional factor is why forecasting is harder than smoothing.

Blackbox Recipe to turn smoothers into forecasters

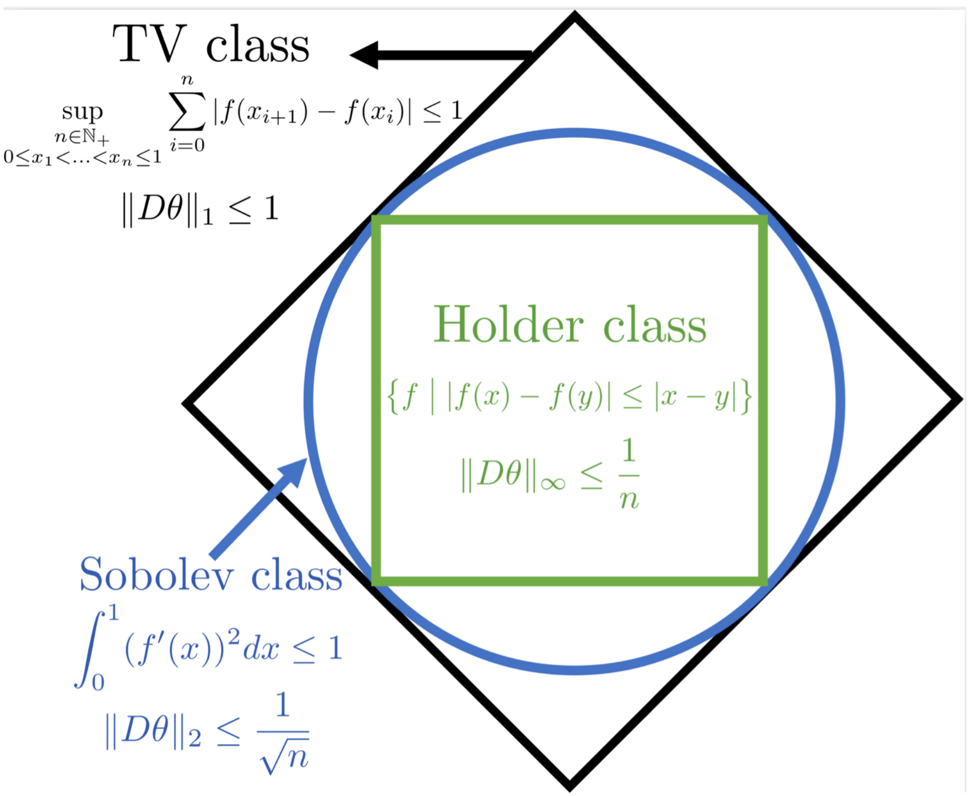
- Two ingradients:

- 1. Smoother that is adaptively minimax and produces estimates as smooth as the original with high probability.

- 2. Online Learner with logarithmic regret

- Any blackboxes that satisfy these oracle properties will work.

Beyond TV bounded sequences



ARROWS calibrated according to radius of a TV class is adaptively minimax over the Holder and Sobolev class inscribed within

Adaptivity to parameters of the model: C_n, n, σ

- Turns out that we don't have to know C_n

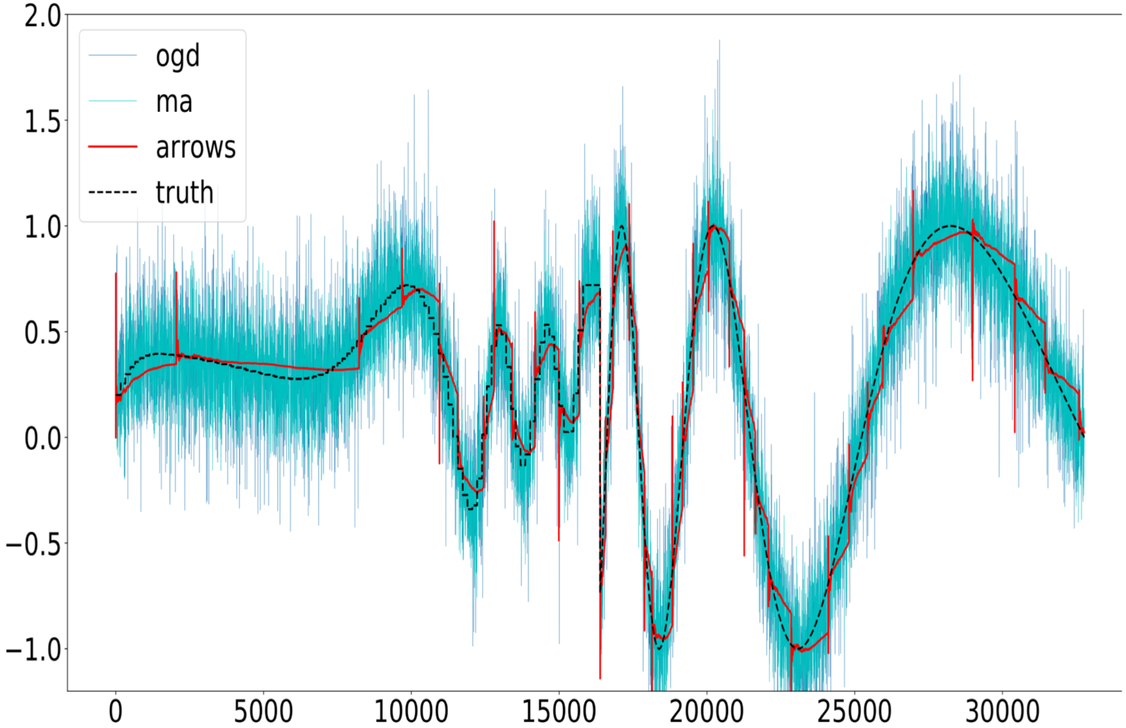
(e) **Restart Rule:** If $\frac{1}{\sqrt{k}} \sum_{l=0}^{\log_2(k)-1} 2^{l/2} \|\hat{\alpha}(t_h : t)[l]\|_1 > n^{-1/3} C_n^{1/3} \sigma^{2/3}$

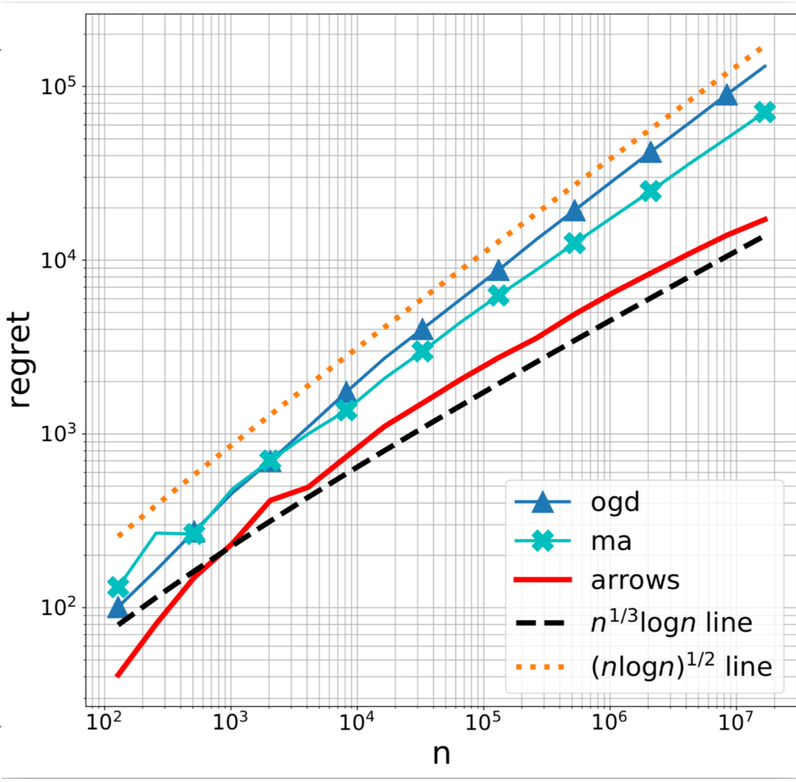
  - Replace the threshold with: $\dfrac{\sigma}{\sqrt{k}}$

- n: (only in log factors) Standard doubling trick

- σ: Easy under the Gaussian noise model.

Experimental Results





Summary

- Optimal forecasting algorithm for any sequences within a total variation class.
- Adaptive to (almost) all parameters of the problem.
- Forecasting is harder than smoothing
- Unprecedented O(n^{1/3}) dynamic regret. We hit a n^{1/2}) lower bound in almost all problems in that setting.

Open problems:

1. Get rid of the iid noise assumption
   - Regret against to the prediction of the best function in the TV-class.
   - Zinkevich style “dynamic regret” / “tracking regret”.
   - Non-constructive argument (Rakhlin and Sridharan, 2014)

2. Beyond quadratic loss functions
   - In nonstationary stochastic optimization: we have a lower bound of $\sqrt{nC_n}$ for strongly convex losses.
   - Faster rate possible for quadratic loss functions
   - What’s in between quadratic loss and strongly convex losses?

Thank you for your attention!

- Paper available at: https://arxiv.org/pdf/1906.03364.pdf
- To appear at NeurIPS 2019.
- Student author: Dheeraj Baby
- Acknowledgment: Yining Wang, Xi Chen

(Chen, Wang and W., Operations Research, 2019)

- Ryan Tibshirani, anonymous reviewers.