Adaptive Online Forecasting of Trends
(a.k.a. towards *Online Trend Filtering*)

Yu-Xiang Wang

Based on joint work with Dheeraj Baby

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Nonparametric regression

- 50+ years of associated literature

  [Nadaraya, Watson, 1964]

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

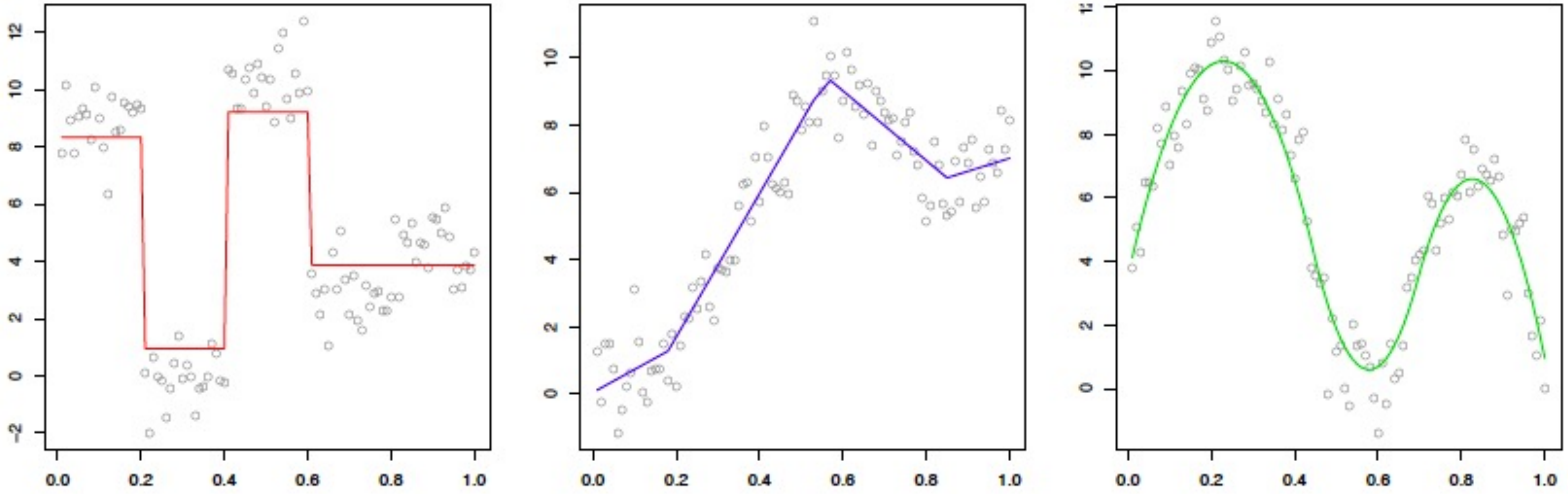
- Also known as smoothing, signal denoising /filtering in signal processing & control.

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

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



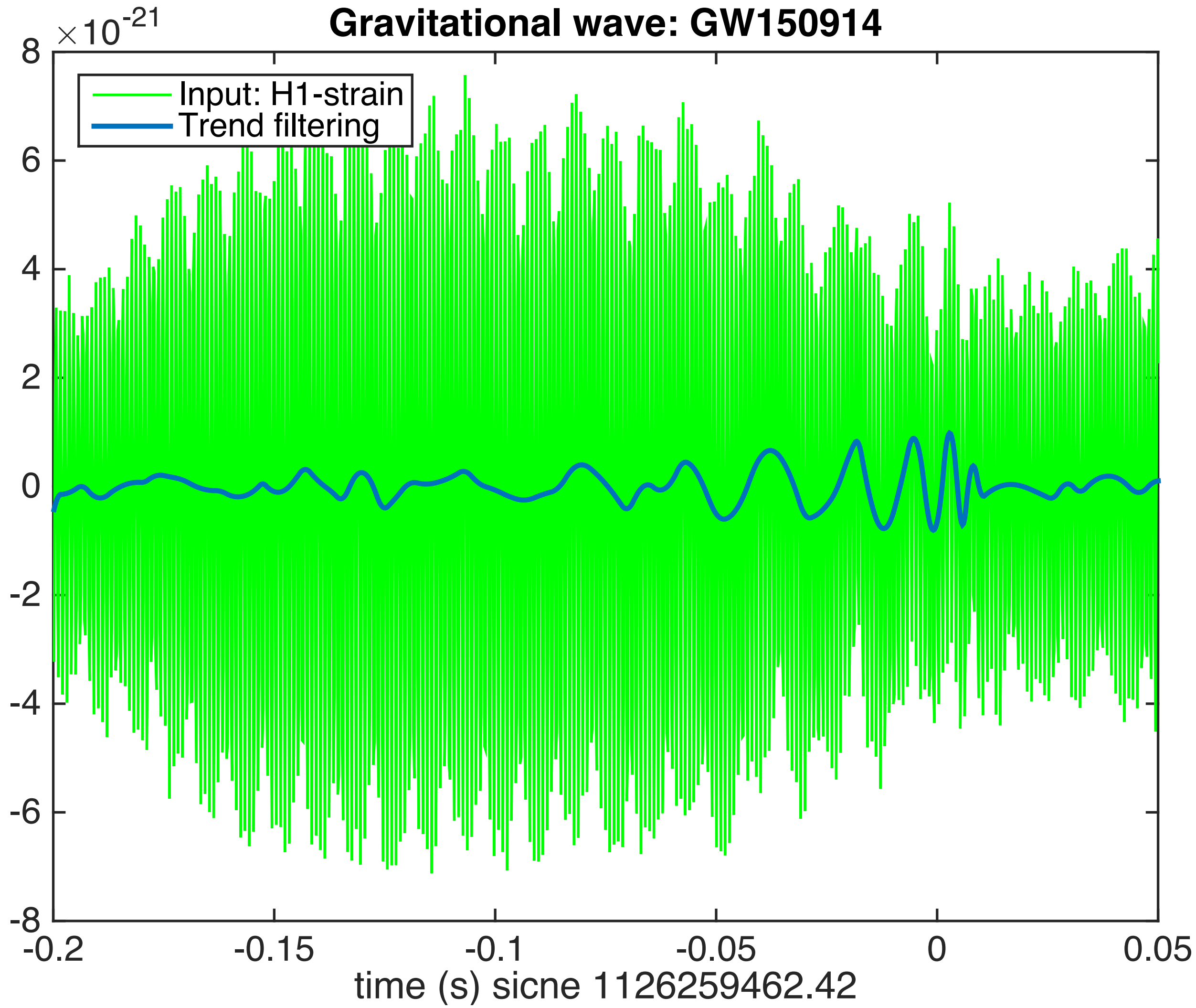
Constant, $k=0$ (Fused lasso)

Linear, $k=1$

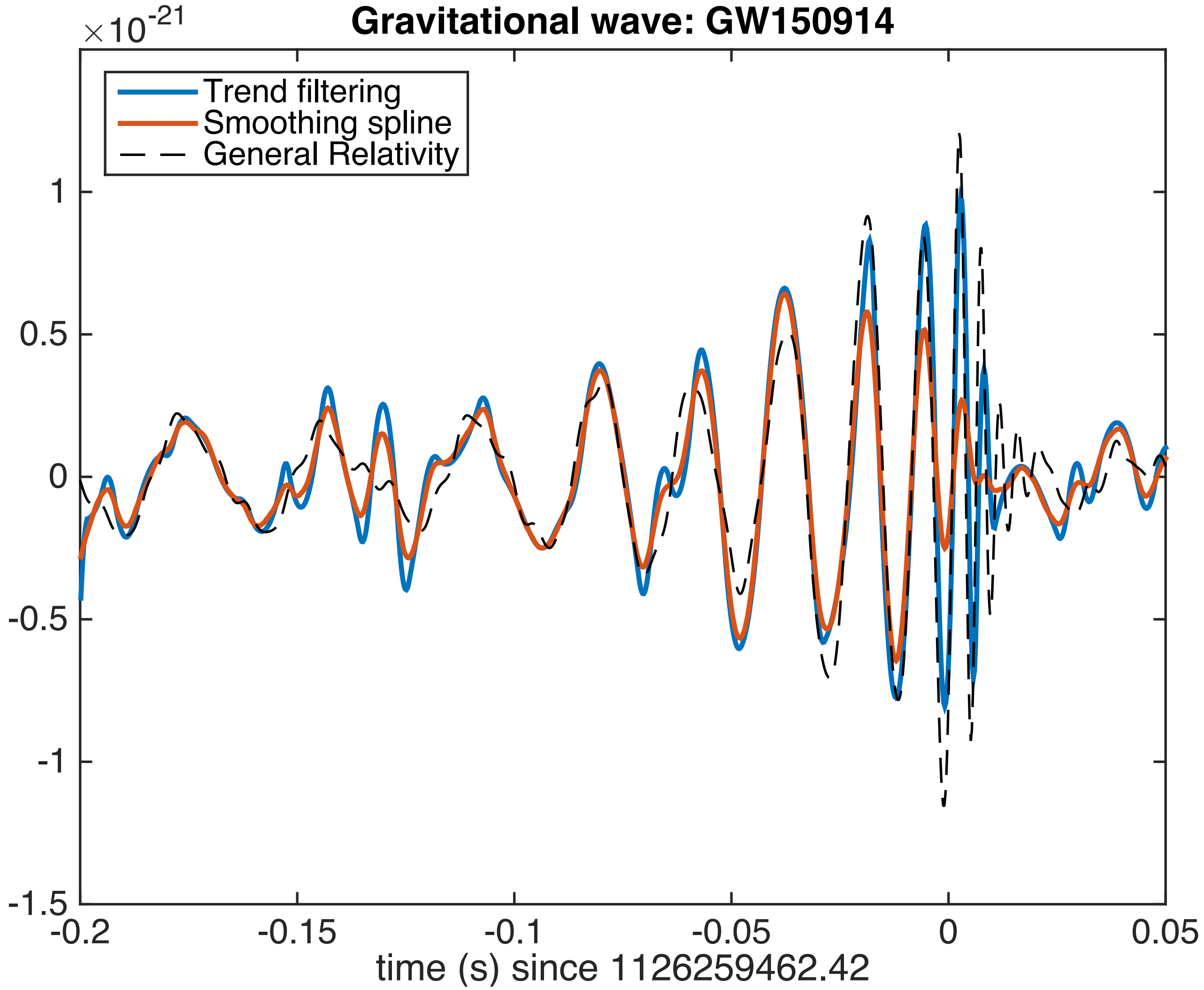
Quadratic, $k=2$

(figure extracted from: Tibshirani (2014))

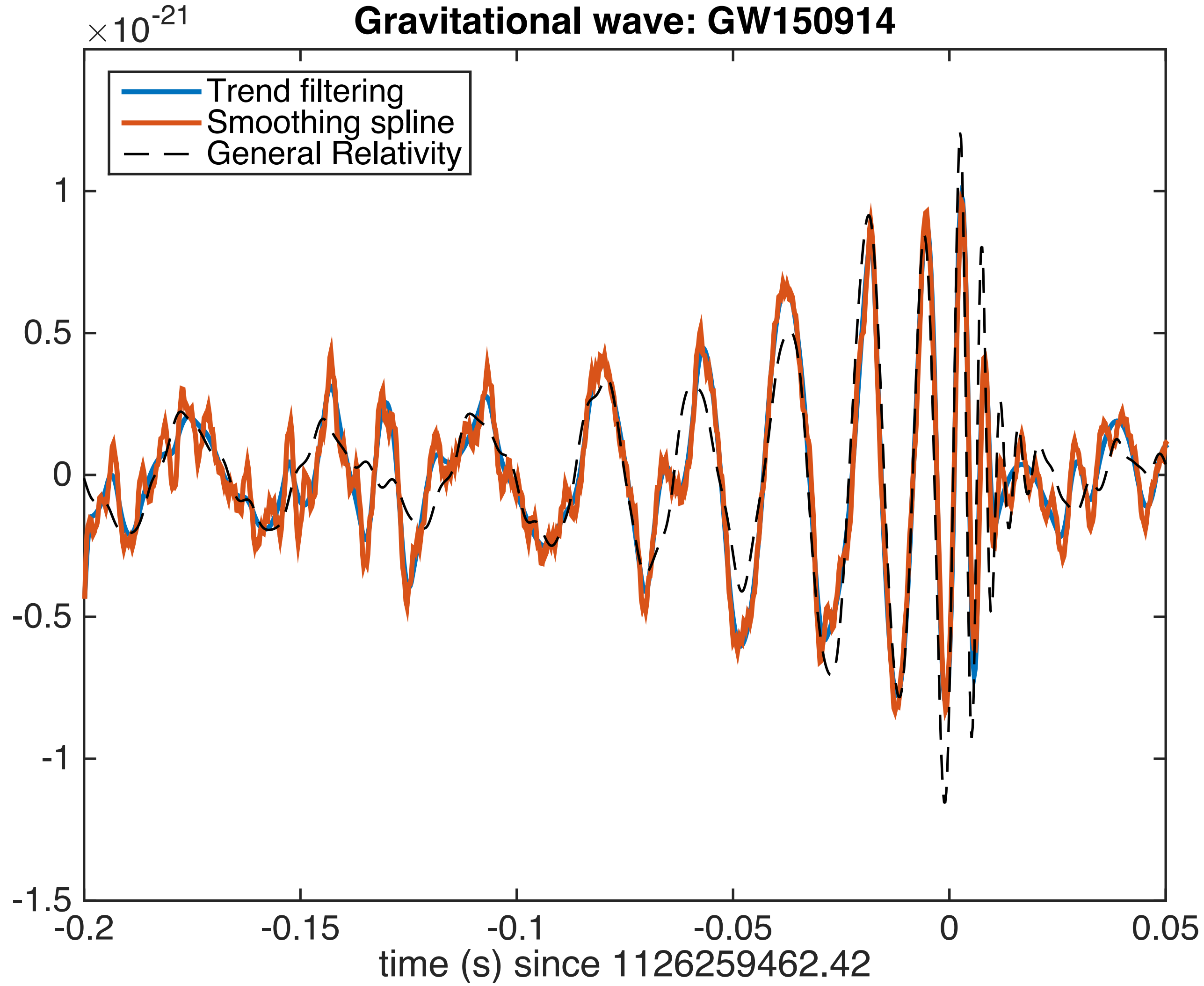
A BIG Example: merger of two black holes



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:

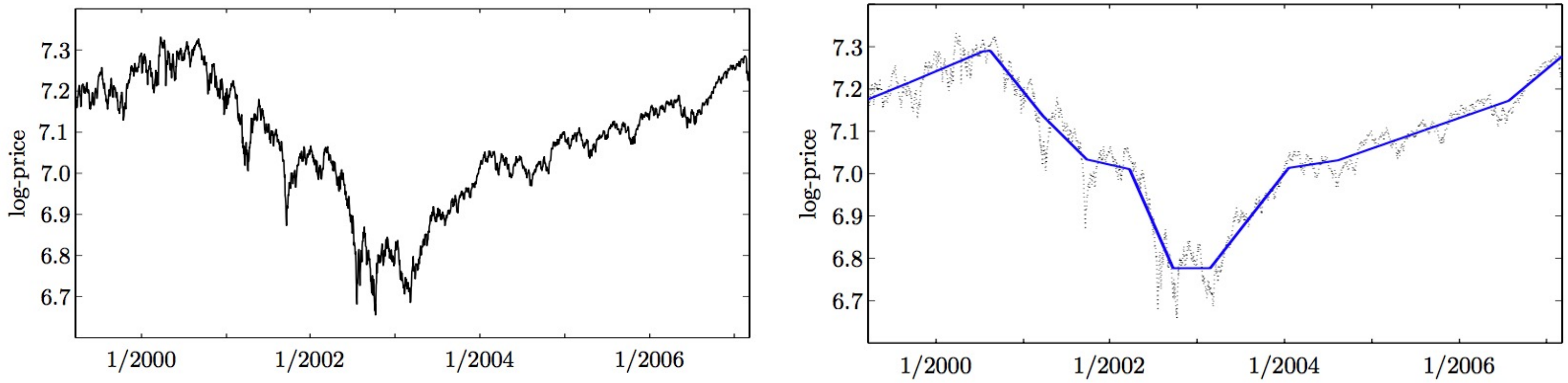
$$O_{\mathbb{P}}(n^{-(2k+2)/(2k+3)})$$

- Best achievable rate for linear smoothers

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

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:
  - Analysis: making senses of what happened.
  - **Forecasting: predict the future**

This talk: towards online trend filtering

1. Minimax rate for TV classes with an online estimator?
    - Stochastic environment + TV class
    - Stochastic environment + higher order TV class

2. Can we succeed in adversarial environments?
    - A reduction to strongly adaptive online learning
    - Universal dynamic regret and oracle inequalities
    - Adding covariates: Exponential concave losses and GLMs

"Online Nonparametric forecasting" in *stochastic* environments.

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
  - $y_t = \theta_t + \epsilon_t \quad , \epsilon_t \sim iid\, \mathrm{subgauss}(0, \sigma^2)$ is revealed

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

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

Arrows: Adaptive Restarting Rule for Online averaging using Wavelet Shrinkage

1. Keep predicting online averages
2. Apply Wavelet Shrinkage to the sequence so far
3. If $\frac{1}{\sqrt{k}} \sum_{l=0}^{\log_2(k)-1} 2^{l/2} \|\hat{\alpha}(t_h : t)[l]\|_1 > \frac{\sigma}{\sqrt{k}}$
   - then "restart"
   - Otherwise keep going!

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

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

- With nearly linear run-time of $O(n \log n)$
- Adapts to unknown Cn
- Adapts to the smaller Holder / Sobolev classes

Baby and W. (2019) *"Online Forecasting of Total Variance Bounded Sequences"* NeurIPS'19: https://arxiv.org/abs/1906.03364

How about higher order TV classes?

Adaptive Vovk-Azoury-Warmuth forecaster (AdaVAW)

1. Online least square (compete with the best polynomial fit)

$$\hat{y}_t = \langle \boldsymbol{x_t}, A_t^{-1} \sum_{s=t_h-k}^{t-1} y_s \boldsymbol{x_s} \rangle$$

2. Apply Wavelet Shrinkage to the sequence so far

Let $(y_1, y_2) = \texttt{pack}(\boldsymbol{y}_r)$

Let $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2) = (T(\boldsymbol{W}\mathbf{y}_1), T(\boldsymbol{W}\mathbf{y}_2))$

3. If $\|\hat{\boldsymbol{\alpha}}_1\|_2 + \|\hat{\boldsymbol{\alpha}}_2\|_2 > \sigma$

   - then "restart"
   - Otherwise keep going!

$$\mathrm{TV}^k(C_n) := \{\boldsymbol{\theta}_{1:n} \in \mathbb{R}^n : n^k \|D^{k+1}\boldsymbol{\theta}_{1:n}\|_1 \leq C_n\}$$

$$\|\boldsymbol{\theta}_{1:n}\|_\infty \leq B$$

- AdaVAW achieves the minimax rate: $\tilde{O}\left(n^{\frac{1}{2k+3}} (C_n)^{\frac{2}{2k+3}}\right)$
- Adapts to unknown Cn
- Adaptive fast rates: Number of knots J. O(J) error.
- Adapts to the smaller Holder / Sobolev classes

Baby and W. (2020) *"Adaptive Online Estimation of Piecewise Polynomial Trends"* NeurIPS'20: https://arxiv.org/abs/2010.00073

Key idea behind these algorithms and Interesting analogy to *online learning*

- n*MSE ⇐==⇒ Dynamic Regret
- Total variation ⇐==⇒ Path length
- Haar Wavelets ⇐==⇒ Geometric cover
- Online averaging ⇐==⇒ Online Gradient Descent

- Key ideas in the algorithm: adaptively determine the length of the history to use!

Are there alternative approaches from online learning? Can we generalize our approach to handle a broader family of problems?

- Yes! We can obtain optimal TV denoising / fused lasso using “Strongly Adaptive Online Learning”.

- And we can get rid of the stochastic assumptions all together!

Baby, Zhao and W. (2021) *“An Optimal Reduction of TV-Denoising to Adaptive Online Learning” AISTATS’21*: https://arxiv.org/abs/2101.09438

Baby and W. *“Optimal Dynamic Regret in Exp-Concave Online Learning”* **COLT’21 Best Student Paper** https://arxiv.org/abs/2104.11824

Dynamic regret minimization in online learning

- For each $t \in [n] := \{1, \ldots, n\}$, learner predicts $\boldsymbol{x}_t \in \mathcal{D} \subset \mathbb{R}^d$.
- Adversary reveals a loss function $f_t : \mathbb{R}^d \to \mathbb{R}$

Goal: Learner aims to control its dynamic regret against any sequence of comparators $\boldsymbol{w}_1, \ldots \boldsymbol{w}_n$ where $\boldsymbol{w}_t \in \mathcal{W} \subseteq \mathcal{D}$ for all t.

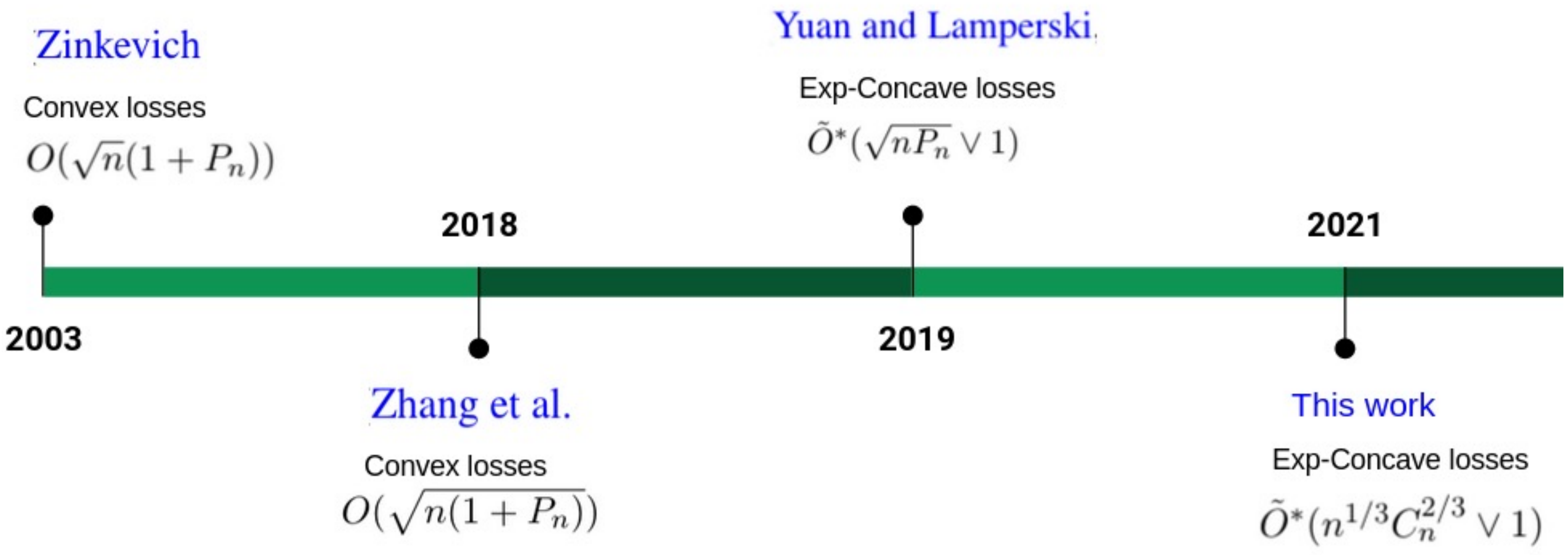
$$R_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) := \sum_{t=1}^{n} f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{w}_t),$$

Dynamic regrets are parametrized by variation incurred by the comparator sequence

$$P_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) = \sum_{t=1}^{n} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2$$

$$C_n(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n) = \sum_{t=1}^{n} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_1$$

Brief history of dynamic regret problem



A Primer of Strongly Adaptive Online Learner

- Algorithms whose static regret in any local time window is controlled.
- Consider any interval $[i_s, i_t] := \{i_s, i_s + 1, \ldots, i_t\} \subseteq [n]$. An SA algorithm achieves logarithmic static regret on $[i_s, i_t]$ when the losses are exp-concave.
- Achieved by hedging over a pool of base learners of n ONS instances where instance t starts working from time t.
- Examples of such methods include FLH from Hazan and Seshadhri (2007) and IFLH from Zhang et al. (2018b).

Optimal dynamic regret for exp-concave losses

Theorem 1 (exp-concave losses)

Let

$$R_n^+(C_n) := \sup_{\substack{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n \in \mathcal{D}^- \\ \sum_{t=2}^n \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_1 \le C_n}} \sum_{t=1}^n f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{w}_t),$$

By running FLH with learning rate α and base learners as ONS with decision set $\mathcal{D}$ and parameter $\zeta = \min\left\{\frac{1}{4G^\dagger(2B\sqrt{d}+2G/\beta)}, \alpha\right\}$, we attain $R_n^+(C_n) = \tilde{O}\left(d^{3.5}(n^{1/3}C_n^{2/3} \vee 1)\right)$ if $C_n > 1/n$ and $O(d^{1.5}\log n)$ otherwise. Here $a \vee b := \max\{a, b\}$ and $\tilde{O}(\cdot)$ hides dependence on the constants $B, G, G^\dagger, \alpha$ and factors of $\log n$.

Exp-concave losses: why do they matter?

Definition: A twice differentiable function f is α-exp-concave *if and only if*

$$\nabla^2 f(\mathbf{x}) \succcurlyeq \alpha \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top$$

Online linear regression: $$f(x) = (y_i - \phi_i^T x)^2$$

Portfolio optimization: $$f(\mathbf{x}) = -\log(\mathbf{r}_t^\top \mathbf{x}).$$

Now we can optimally compete with **any arbitrary changing sequences** of linear predictors / portfolio choices!

Back to TV denoising, but in an adversarial environment

- At time $t \in [n]$ learner predicts $x_t \in \mathcal{D} := [-B, B]$.
- Adversary reveals a label $y_t \in [-B, B]$.
- Learner suffers loss $(y_t - x_t)^2$.

Define a non-parametric sequence class as:

$$\mathcal{TV}^B(C_n) := \left\{ w_{1:n} \middle| TV(w_{1:n}) := \sum_{t=2}^{n} |w_t - w_{t-1}| \leq C_n, \ |w_t| \leq B \ \forall t \in [n] \right\}.$$

Learner aims to control:

$$R_n(C_n) := \sum_{t=1}^{n} (y_t - x_t)^2 - \inf_{w_1, \dots, w_n \in \mathcal{TV}^B(C_n)} \sum_{t=1}^{n} (y_t - w_t)^2$$

Dynamic regret of SA learner

Theorem 2 (squared error losses)

Let x_t be the prediction at time t of FLH with learning rate $\zeta = 1/(8B^2)$ and base learners as FTL. Then for any comparator $(w_1, \ldots, w_n) \in \mathcal{TV}^B(C_n)$

$$\sum_{t=1}^{n} (y_t - x_t)^2 - (y_t - w_t)^2 = \tilde{O}\left(n^{1/3} C_n^{2/3} B^{4/3} \vee B^2\right),$$

where the labels obey $|y_t| \leq B$, $\tilde{O}(\cdot)$ hides dependence on logarithmic factors of horizon n and $a \vee b := \max\{a, b\}$.

A new type of oracle inequality

- Theorem 2 implies the following oracle inequality

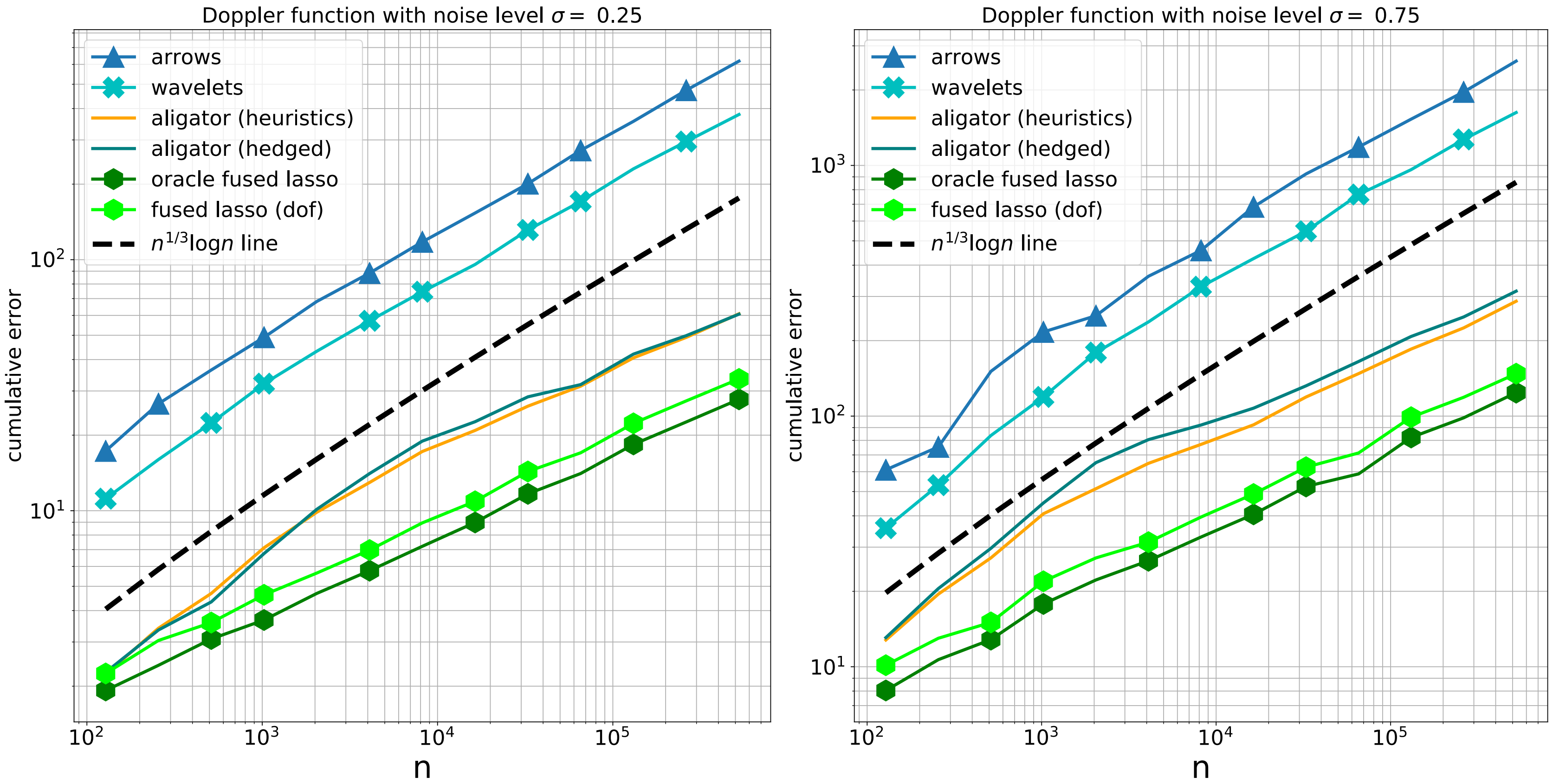
$$\sum_{t=1}^{n}(y_t - x_t)^2 \leq \min_{w_1,\ldots,w_n} \sum_{t=1}^{n}(y_t - w_t)^2 + \tilde{O}\left(n^{1/3}\mathrm{TV}(w_{1:n})^{2/3}B^{4/3} \vee B^2\right).$$

- Fused Lasso denoiser attains the following oracle inequality: $\sum_{t=1}^{n}(u_t - \hat{x}_t)^2 \leq \min_{w_1,\ldots,w_n} \sum_{t=1}^{n}(u_t - w_t)^2 + \tilde{O}_P\left(\lambda \mathrm{TV}(w_{1:n})\right)$, (See (Guntuboyina et al., 2017; Ortelli and van de Geer, 2019))

- When $\lambda \asymp n^{1/3}/C_n^{1/3}$, it implies the optimal statistical estimation rate of $\tilde{O}(n^{1/3}C_n^{2/3})$

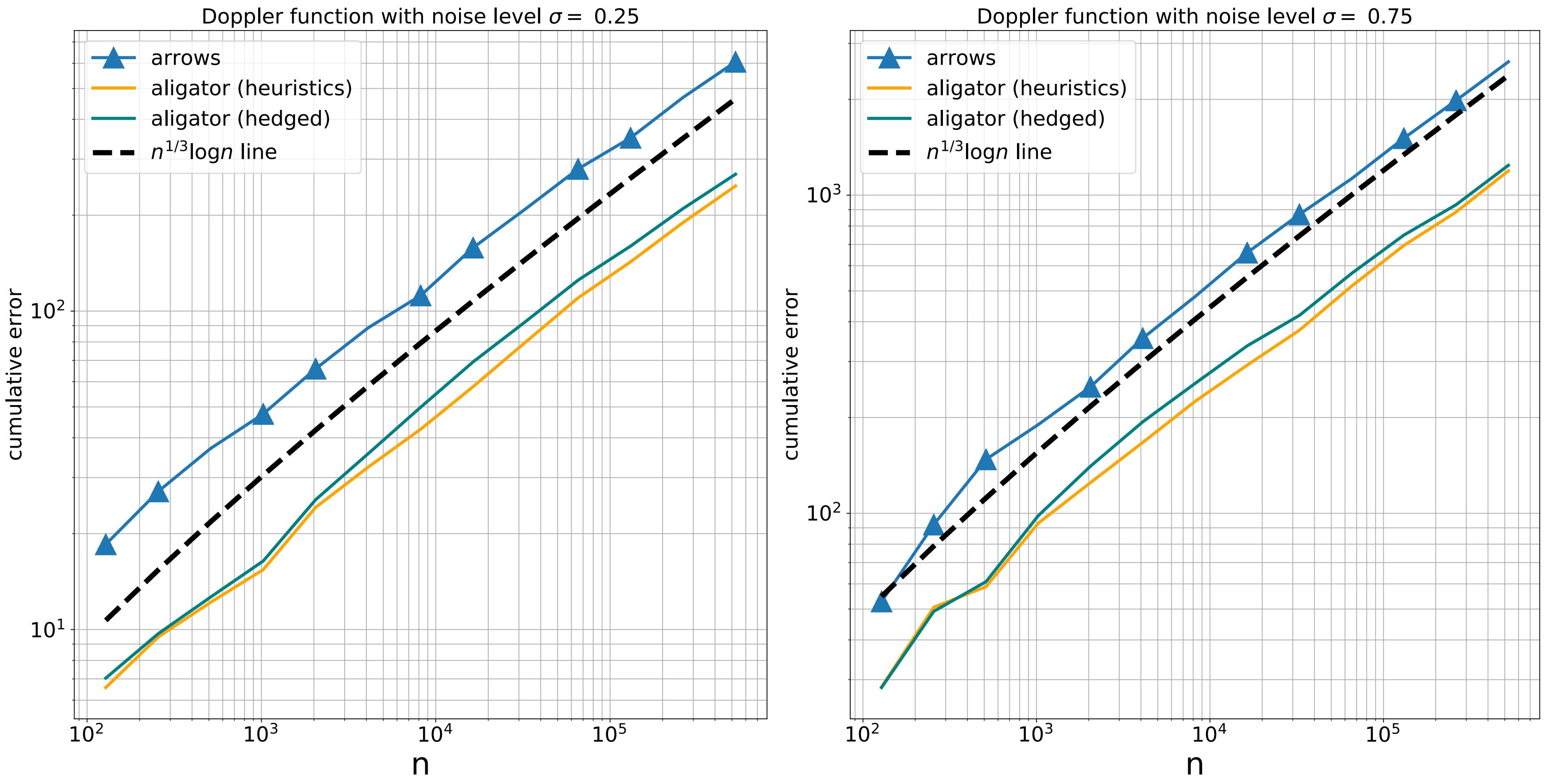
- Our results don't require any statistical assumptions on y_t, eliminate the need to choose hyperparameter λ and also imply the same estimation rate achievable by the optimal choice of λ for the iid setting.

SA learner is reasonably practical even in an *offline setting,* matching optimally tuned fused lasso up to a constant.



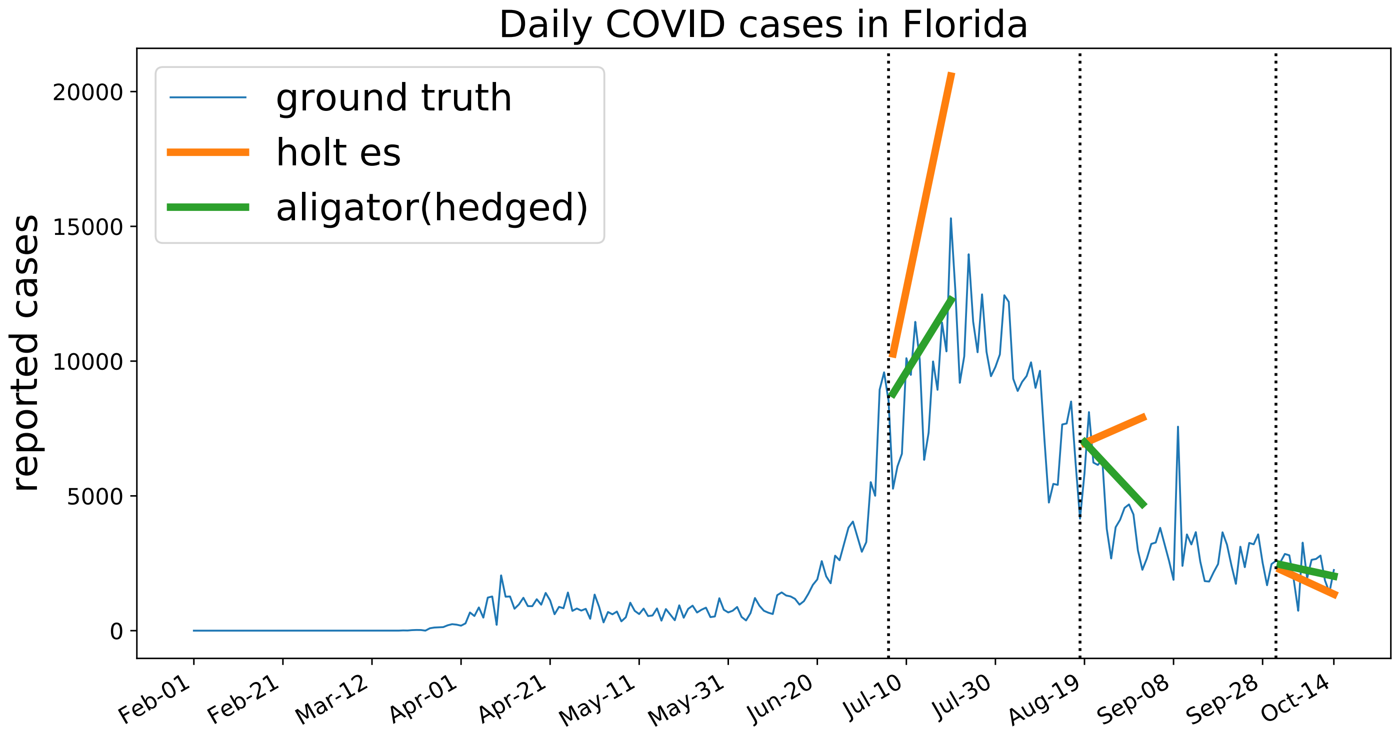
(a) Offline experiments

SA learner beats Arrows in the online setting



(b) Online experiments

Using SA learner to "online trend removal" for COVID hospitalization forecasts



Sketch of the proof: offline optimal sequence

Consider the offline convex optimization problem:

$$\min_{\tilde{u}_1, \ldots, \tilde{u}_n} \quad \frac{1}{2} \sum_{t=1}^{n} (y_t - \tilde{u}_t)^2$$
$$\text{s.t.} \quad \sum_{t=1}^{n-1} |\tilde{u}_{t+1} - \tilde{u}_t| \leq C_n$$

Let $u_1, \ldots, u_n$ be the optimal primal variables and let $\lambda \geq 0$ be the optimal dual variable corresponding to the TV constraint.

The sequence $u_1, \ldots, u_n$ will be referred as the offline optimal.

Adaptive partitioning of the sequence into bins according to the offline optimal comparator

We construct a partitioning of $[n]$ into M bins as follows $\{[1_s, 1_t], \ldots, [i_s, i_t], \ldots, [M_s, M_t]\}$ satisfying:

- $C_i := \sum_{j=i_s}^{i_t - 1} |u_{j+1} - u_j| \leq B/\sqrt{n_i}$ where $n_i := i_t - i_s + 1$, $i \in [M]$.
- Number of bins obeys $M = O(n^{1/3} C_n^{2/3} B^{-2/3} \vee 1)$.

Regret decomposition into three terms

$$R_n(C_n) = \underbrace{\sum_{i=1}^{M} \sum_{j=i_s}^{i_t} (x_j - y_j)^2 - (y_j - \bar{y}_i)^2}_{T_{1,i}} +$$

$$\underbrace{\sum_{i=1}^{M} \sum_{j=i_s}^{i_t} (y_j - \bar{y}_i)^2 - (y_j - \bar{u}_i)^2}_{T_{2,i}} +$$

$$\underbrace{\sum_{i=1}^{M} \sum_{j=i_s}^{i_t} (y_j - \bar{u}_i)^2 - (y_j - u_j)^2}_{T_{3,i}}$$

By Strong Adaptivity $T_{1,i} = O(B^2 \log n)$.

By KKT conditions

$$\begin{aligned} T_{3,i} &\leq n_i C_i^2 + 3\lambda C_i \\ &\leq B^2 + 3\lambda C_i, \end{aligned}$$

Turns out that T2 can be very negative when we need it to be.

$$T_{2,i} = \sum_{j=i_s}^{i_t} (y_j - \bar{y}_i)^2 - (y_j - \bar{u}_i)^2$$ is always negative.

$T_{2,i} \leq -\frac{\lambda^2}{n_i}$ when $u_{i_s:i_t}$ is not isotonic.

Nice cancellation:

$$\begin{aligned} T_{1,i} + T_{2,i} + T_{3,i} &\leq -\frac{\lambda^2}{n_i} + 3\lambda C_i + \tilde{O}(B^2) \\ &= -\left(\frac{\lambda}{\sqrt{n_i}} - \frac{3C_i\sqrt{n_i}}{2}\right)^2 + \frac{9n_iC_i^2}{4} + \tilde{O}(B^2) \\ &= \tilde{O}(B^2), \end{aligned}$$

- Similarly $T_{1,i} + T_{2,i} + T_{3,i} = O(B^2)$ even when the sequence $u_{i_s:i_t}$ is isotonic.
- Summing across all $O(n^{1/3}C_n^{2/3}B^{-2/3} \vee 1)$ bins in the partition yields a regret of $\tilde{O}\left(n^{1/3}C_n^{2/3}B^{4/3} \vee B^2\right)$ of Theorem 2.

Conclusions

- Online locally adaptive nonparametric estimators that make sequential predictions while achieving the optimal rates for offline estimators.
- New techniques that show “strongly adaptive online learners” achieve an optimal dynamic regret for strongly convex and exponential concave losses.
- A lot of possibilities and open problems at the intersection of adaptive nonparametric regression and adaptive online learning.

Thank you for your attention!

References

1. Baby and W. *“Online Forecasting of Total Variance Bounded Sequences”* NeurIPS’19
2. Baby and W. *“Adaptive Online Estimation of Piecewise Polynomial Trends”* NeurIPS’20
3. Baby, Zhao and W. *“An Optimal Reduction of TV-Denoising to Adaptive Online Learning” AISTATS’21*
4. Baby and W. *“Optimal Dynamic Regret in Exp-Concave Online Learning”* COLT’21 Best Student Paper