Online Forecasting of Total-Variation-bounded Sequences

UCSB

Dheeraj Baby, Yu-Xiang Wang

Introduction and Objective

◇ Nonparametric Online Forecasting model

1. Fix action time intervals $1, 2, \ldots, n$
2. The player declare a forecasting strategy $\mathcal{A}_i : \mathbb{R}^{i-1} \to \mathbb{R}$ for $i = 1, \ldots, n$.
3. An adversary chooses a sequence $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_n] \in \mathbb{R}^n$.
4. For every time point $i = 1, \ldots, n$:
   (a) We play $x_i = \mathcal{A}_i(y_1, \ldots, y_{i-1})$.
   (b) We receive a feedback $y_i = \theta_i + Z_i$, where Z_i is a zero-mean, independent subgaussian noise.
5. At the end, the player suffers a cumulative error $\sum_{i=1}^{n} (x_i - \theta_i)^2$.

◇ Assumptions

1. Knowledge of σ^2 of sub-gaussian noise.
2. Ground truth sequence $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^T \in TV(C_n)$, where $\mathrm{TV}(C_n) := \{\boldsymbol{\theta} \in \mathbb{R}^n \mid \|D\boldsymbol{\theta}\|_1 \le C_n\}$ and D is the discrete difference operator. C_n is not required to be known apriori.
3. $|\theta_1| \le U$

◇ Questions of interest

1. Information-theoretic limit
   - What is the optimal Total Squared Error (TSE) for any method?
2. Designing of locally adaptive policies:
   - $TV(C_n)$ features sequences that can exhibit spatially localized trends.
   - How to design a minimax policy that is locally adaptive to these non-uniform trends?

Performance of Existing Policies

Theorem 1 (A lowerbound on TSE) *Assume $\min\{U, C_n\} > 2\pi\sigma$ and $n > 3$, there is a universal constant c such that*

$$\inf_{x_{1:n}} \sup_{\theta_{1:n} \in \mathrm{TV}(C_n)} \mathbb{E}\left[\sum_{t=1}^{n} \left(x_t(y_{1:t-1}) - \theta_t\right)^2\right] \ge c(U^2 + C_n^2 + \sigma^2 \log n + n^{1/3} C_n^{2/3} \sigma^{4/3}).$$

◇ TSE for existing policies grows as $O(\sqrt{n})$

- Restarting OGD [1,2]
- Adaptive Optimistic Mirror Descent [3]
- Moving Averages

◇ Can we achieve the lowerbound?

Our Policy

ARROWS: inputs - observed y values, $\delta \in (0, 1]$, σ^2, a hyper-parameter $\beta > 6$

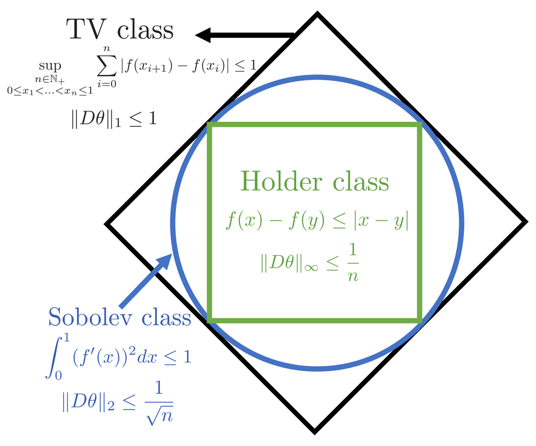
1. Initialize $t_h = 1$, $newBin = 1$, $y_0 = 0$
2. For $t = 1$ to n:
   (a) if $newBin == 1$, predict $x_t^{t_h} = y_{t-1}$, else predict $x_t^{t_h} = \bar{y}_{t_h:t-1}$
   (b) set $newBin = 0$, observe y_t and suffer loss $(x_t^{t_h} - \theta_t)^2$
   (c) Let $\hat{y} = pad_0(y_{t_h}, \ldots, y_t)$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

Theorem 2 (TSE of ARROWS) *Let the feedback be $y_t = \theta_t + Z_t$, $t = 1, \ldots, n$ and Z_t be independent, σ-subgaussian random variables. If $\beta = 24 + \frac{8 \log(8/\delta)}{\log(n)}$, then with probability at least $1 - \delta$, ARROWS achieves a dynamic regret of $\tilde{O}(n^{1/3} \|D\theta(1:n)\|_1^{2/3} \sigma^{4/3} + |\theta_1|^2 + \|D\theta(1:n)\|_2^2 + \sigma^2)$ where $\tilde{O}$ hides a logarithmic factor in n and $1/\delta$.*

◇ Runtime of ARROWS is $O(n \log n)$

Minimax rates for TSE

Class		Minimax rate for Forecasting	Minimax rate for Smoothing[4]	Minimax rate for Linear Forecasting
TV	$\|D\theta\|_1 \le C_n$	$n^{1/3} C_n^{2/3} \sigma^{4/3} + C_n^2 + \sigma^2$	$n^{1/3} C_n^{2/3} \sigma^{4/3} + \sigma^2$	$n^{1/2} C_n \sigma + C_n^2 + \sigma^2$
Sobolev	$\|D\theta\|_2 \le C_n'$	$n^{2/3} [C_n']^{2/3} \sigma^{4/3} + [C_n']^2 + \sigma^2$	$n^{2/3} [C_n']^{2/3} \sigma^{4/3} + \sigma^2$	$n^{2/3} [C_n']^{2/3} \sigma^{4/3} + [C_n']^2 + \sigma^2$
Holder	$\|D\theta\|_\infty \le L_n$	$n L_n^{2/3} \sigma^{4/3} + n L_n^2 + \sigma^2$	$n L_n^{2/3} \sigma^{4/3} + \sigma^2$	$n L_n^{2/3} \sigma^{4/3} + n L_n^2 + \sigma^2$
Minimax Algorithm		ARROWS	Wavelet Smoothing[4] Trend Filtering[5]	Restarting OGD[1,2] Moving Averages



Canonical Scaling[a]		Forecasting	Smoothing	Linear Forecasting
TV	$C_n \asymp 1$	$n^{1/3}$	$n^{1/3}$	$n^{1/2}$
Sobolev	$C_n' \asymp 1/\sqrt{n}$	$n^{1/3}$	$n^{1/3}$	$n^{1/3}$
Holder	$L_n \asymp 1/n$	$n^{1/3}$	$n^{1/3}$	$n^{1/3}$

[a] The "canonical scaling" are obtained by discretizing functions in canonical function classes. Under the canonical scaling, Holder class $\subset$ Sobolev class $\subset$ TV class, as shown in the figure on the left.

1. *For compactness we hide the dependence of U and $\log n$ from all forecasting rates.*
2. ARROWS is adaptively minimax over the described classes.
3. Linear forecasters are fundamentally limited in predicting TV bounded sequences
   (a) Policies such as Restarting OGD/MA are unable to come up with a single window size that performs optimally through out the duration.

Experimental results

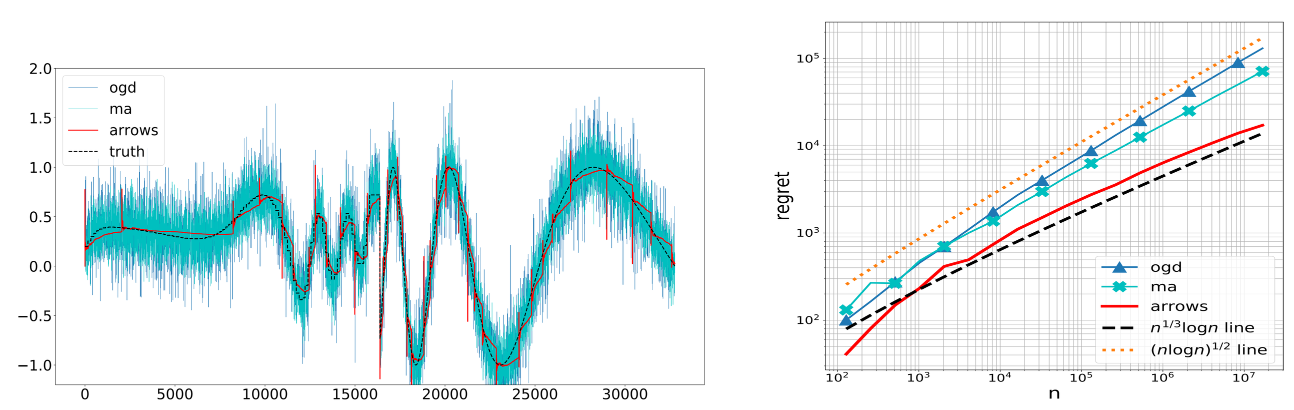
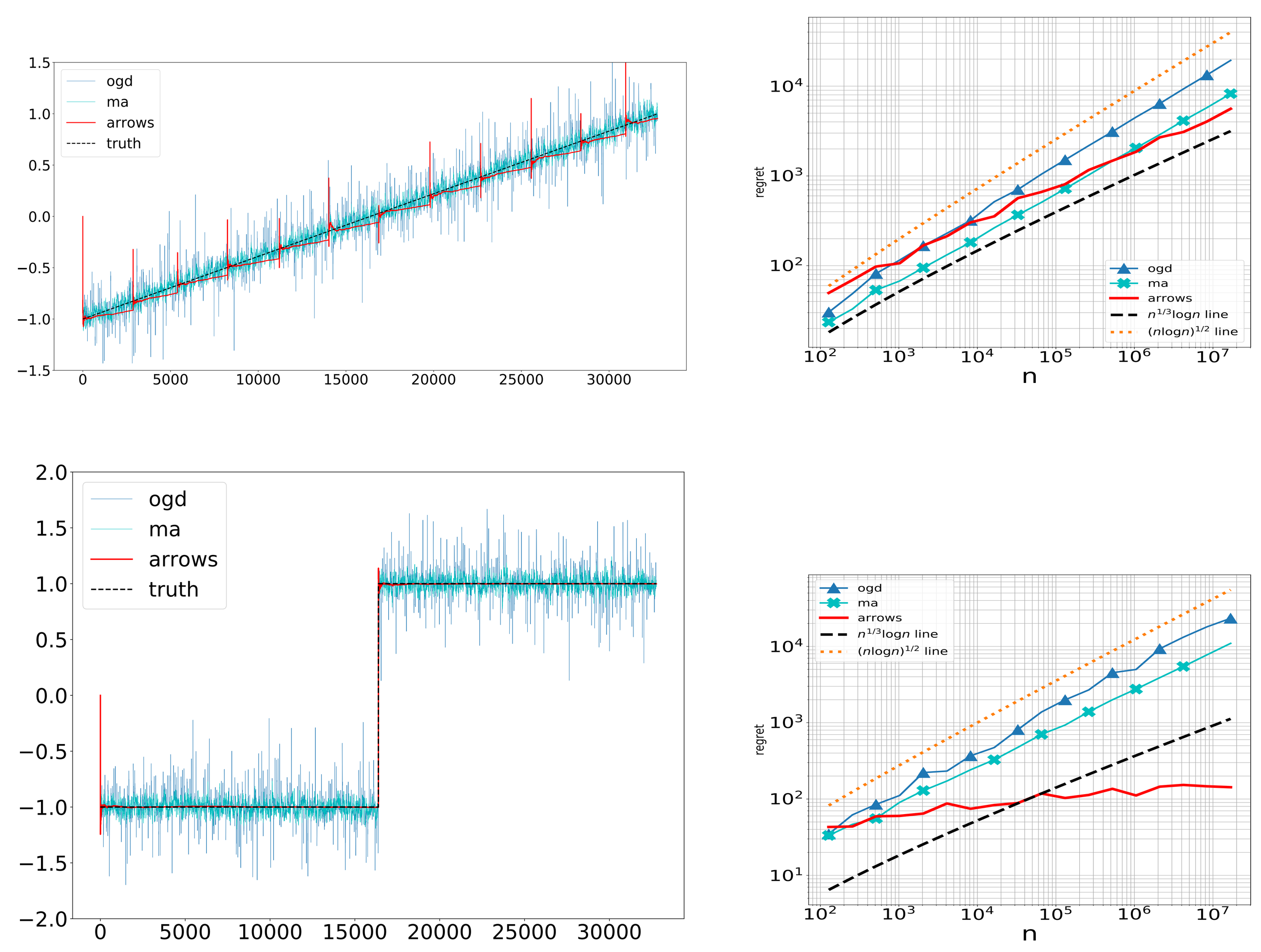


Figure description: The figure shows the results on a function with heterogeneous smoothness with the hyperparameters selected according to their theoretical optimal choice for the TV class. The top panel illustrates that ARROWS is locally adaptive to heterogeneous smoothness of the ground truth. Red peaks in the figure signifies restarts. During the initial and final duration, the signal varies smoothly and ARROWS chooses a larger window size for online averaging. In the middle, signal varies rather abruptly. Consequently ARROWS chooses a smaller window size. On the other hand, the linear smoothers OGD and MA use a constant width and cannot adapt to the different regions of the space. This differences are also reflected in the quantitative evaluation on the bottom panel, which clearly shows that OGD and MA has a suboptimal $\tilde{O}(\sqrt{n})$ TSE while ARROWS attains the $\tilde{O}(n^{1/3})$ minimax TSE!



References

[1] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. In *Operations Research*, 2015.

[2] Xi Chen, Yining Wang, and Yu-Xiang Wang. Non-stationary Stochastic Optimization under Lp, q-Variation Measures In *Operations Research*, 2018

[3] Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398-406, 2015

[4] David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. In *Annals of statistics*, 1998.

[5] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. In *SIAM Review*, 2009.