Supplementary Material for:
An Optimal Reduction of TV-Denoising to Adaptive Online Learning

Dheeraj Baby
dheeraj@ucsb.edu

Xuandong Zhao
xuandongzhao@ucsb.edu

Yu-Xiang Wang
yuxiangw@cs.ucsb.edu

Dept. of Computer Science, UC Santa Barbara

A More on Related Work

For any forecasting strategy whose output $\hat{y}_t$ at time t depends only on past observations, we have $\mathbb{E}[(\hat{y}_t - y_t)^2] - \mathbb{E}[(f(x_{i_t}) - y_t)^2] = \mathbb{E}[(\hat{y}_t - f(x_{i_t}))^2]$. Hence any algorithm that minimizes the dynamic regret against the sequence $f(x_{i_1}), \ldots, f(x_{i_n})$ with $\ell_t(x) = (x - y_t)^2$ being the loss at time t, can be potentially applied to solve our problem. However as noted in (Baby and Wang, 2019) a wide array of techniques such as (Zinkevich, 2003; Hall and Willett, 2013; Besbes et al., 2015; Chen et al., 2018b; Jadbabaie et al., 2015; Yang et al., 2016; Zhang et al., 2018a,b; Chen et al., 2018a; Yuan and Lamperski, 2019) are unable to achieve the optimal rate. However, we note that many of these algorithms support general convex/strongly-convex losses. The existence of a strategy with $\tilde{O}(n^{1/3}C_n^{2/3})$ rate for R_n, even in the more general (in comparison to offline problem) online setting considered in Fig. 1 is implied by the results of (Rakhlin and Sridharan, 2014) on online non-parametric regression with Besov spaces via a non-constructive argument. (Kotłowski et al., 2016) studies the problem of forecasting isotonic sequences. However, the techniques are not extensible to forecasting the much richer family of TV bounded sequences.

We acknowledge that univariate TV-denoising is a simple and classical problem setting, and there had been a number of studies on TV-denoising in multiple dimensions and on graphs, and to higher order TV functional, while establishing the optimal rates in those settings (Tibshirani, 2014; Wang et al., 2016; Hutter and Rigollet, 2016; Sadhanala et al., 2016, 2017; Li et al., 2018). The problem of adaptivity in C_n is generally open for those settings, except for highly special cases where the optimal tuning parameter happens to be independent to C_n (see e.g., (Hutter and Rigollet, 2016)). Generalization of the techniques developed in this paper to these settings are possible but beyond the scope of this paper. That said, as (Padilla et al., 2017) establishes, an adaptive univariate fused lasso is already able to handle signal processing tasks on graphs with great generality by simply taking the depth-first-search order as a chain.

Using a specialist aggregation scheme to incur low adaptive regret was explored in (Adamskiy et al., 2016). However, the experts they use are same as that of (Hazan and Seshadhri, 2007). Due to this, their techniques are not directly applicable in our setting where the exogenous variables are queried in an arbitrary manner.

There are image denoising algorithms based on deep neural networks such as (Zhang et al., 2017). However, this body of work is complementary to our focus on establishing the connection between denoising and strongly adaptive online learning.

B Proofs of Technical Results

For the sake of clarity, we present a sequence of lemmas and sketch how to chain them to reach the main result in Section B.1. This is followed by proof of all lemmas in Section B.2 and finally the proof of Theorem 5 in Section B.3.

B.1 Proof strategy for Theorem 5

We first show that ALIGATOR suffers logarithmic regret against any expert in the pool $\mathcal{E}$ during its awake period. Then we exhibit a particular partition of the underlying TV bounded function such that number of chunks in the partition is $O(n^{1/3}C_n^{2/3})$. Following this, we cover each chunk with atmost $\log n$ experts and show that each expert in the cover suffers a $\tilde{O}(1)$ estimation error. The Theorem then follows by summing the estimation error across all chunks.

Some notations. In the analysis thereafter, we will use the following notations. Let $\tilde{\sigma} = \sigma\sqrt{2\log(4n/\delta)}$, $R_\sigma = 16(B+\tilde{\sigma})^2$ and $\mathcal{T}(I) = \{t \in [n] : i_t \in I\}$ for any $I \in \mathcal{I}|_{[n]}$, where $\mathcal{I}|_{[n]}$ is defined according to the terminology in Section 2.1. Let $\theta_t := f(x_{i_t})$.

First, we show that ALIGATOR is competitive against any expert in the pool $\mathcal{E}$.

Lemma 1. *For any interval $I \in \mathcal{I}|_{[n]}$ such that $\mathcal{T}(I)$ is non-empty, the predictions made by* ALIGATOR *$\hat{y}_t$ satisfy*

$$\sum_{t \in \mathcal{T}(I)} (\hat{y}_t - \theta_t)^2 \leq \frac{e-1}{3-e} \sum_{t \in \mathcal{T}(I)} (\mathcal{A}_I(t) - \theta_t)^2 + \frac{\log(n \log n) R_\sigma + 2R_\sigma^2 \log(2n \log n/\delta)}{3-e},$$

with probability atleast $1-\delta$.

Corollary 2. *Let $\mathbb{S} = \{P_1, \ldots, P_M\}$ be an arbitrary ordered set of consecutive intervals in $[n]$. For each $i \in [n]$ let $\mathcal{U}_i$ be the set containing elements of the GC that covers the interval P_i according to Proposition 1. Denote $\lambda := \frac{\log(n \log n) R_\sigma + 2R_\sigma^2 \log(2n \log n/\delta)}{3-e}$. Then* ALIGATOR *forecasts $\hat{y}_t$ satisfy*

$$\sum_{t=1}^{n} (\hat{y}_t - \theta_t)^2 \leq \min_{\mathbb{S}} \sum_{i=1}^{M} \sum_{I \in \mathcal{U}_i} \mathbf{1}\{|\mathcal{T}(I)| > 0\} \left(\frac{e-1}{3-e} \sum_{t \in \mathcal{T}(I)} (\mathcal{A}_I(t) - \theta_t)^2 + \lambda \right),$$

with probability atleast $1-\delta$.

The minimum across all partitions in the Corollary above hints to the novel ability of ALIGATOR to incur potentially very low estimation errors.

Next, we proceed to exhibit a partition of the set of exogenous variables queried by the adversary that will eventually lead to the minimax rate of $\tilde{O}(n^{1/3} C_n^{2/3})$. The existence of such partitions is a non-trivial matter.

Lemma 3. *Let $\mathbb{S} = \{x_{k_1} < \ldots, < x_{k_m}\} \subseteq \mathcal{X}$ be the exogenous variables queried by the adversary over n rounds where each $k_i \in [n]$. Denote $\theta^{(i)} := f(x_{k_i})$ and $p(i) := \#\{t : x_{i_t} = x_{k_i}\}$ for each $i \in [m]$. Denote $[x_i, x_j] := \{x_{k_i}, x_{k_{i+1}}, \ldots, x_{k_j}\}$. For any $[x_i, x_j] \subseteq \mathbb{S}$, define $V(x_i, x_j) = \sum_{k=i}^{j-1} |\theta^{(i)} - \theta^{(i+1)}|$. There exists a partitioning $\mathcal{P} = \{[x_1, x_{r_1}], [x_{r_1+1}, x_{r_2}], \ldots, [x_{r_{M-1}+1}, x_m]\}$ of $\mathbb{S}$ that satisfies*

1. *For any $[x_i, x_j] \in \mathcal{P} \setminus \{[x_{r_{M-1}+1}, x_m]\}$, $V(x_i, x_j) \leq \frac{B}{\sqrt{\sum_{k=i}^{j} p(k)}}$.*
2. *$V(x_{r_{M-1}+1}, x_{m-1}) \leq \frac{B}{\sqrt{\sum_{k=r_{M-1}+1}^{m-1} p(k)}}$.*
3. *Number of partitions $M \leq \max\{3n^{1/3} C_n^{2/3} B^{-2/3}, 1\}$.*

The next lemma controls the estimation error incurred by an expert during its awake period.

Lemma 4. *Let $\{\underline{x}, < \ldots, < \bar{x}\}$ be the exogenous variables queried by the adversary over n rounds in an arbitrary interval $I \in \mathcal{I}|_{[n]}$. Then with probability atleast $1-\delta$*

$$\sum_{t \in \mathcal{T}(I)} (\theta_t - \mathcal{A}_I(t))^2 \leq 2V(\underline{x}, \bar{x})^2 |\mathcal{T}(I)| + 2\sigma^2 \log(2n^3 \log n/\delta) \log(|\mathcal{T}(I)|),$$

where $V(\cdot, \cdot)$ is defined as in Lemma 3.

To prove Theorem 5, our strategy is to apply Corollary 2 to the partition in Lemma 3. By the construction of the GC, each chunk in the partition can be covered using atmost $\log n$ intervals. Now consider the estimation error incurred by an expert corresponding to one such interval. Due to statements 1 and 2 in Lemma 3 the $V(\underline{x}, \bar{x})^2 |\mathcal{T}(I)|$ term of error bound in Lemma 4 can be shown to $O(1)$. When summed across all intervals that cover a chunk, the total estimation error within a chunk becomes $\tilde{O}(1)$. Now appealing to statement 3 of Lemma 3, we get a total error of $\tilde{O}(n^{1/3} C_n^{2/3})$ when the error is summed across all chunks in the partition.

B.2 Omitted Lemmas and Proofs

Lemma 5. *Let $\mathcal{V}$ be the event that for all $t \in [n]$, $|\epsilon_t| \leq \sigma\sqrt{2\log(4n/\delta)}$. Then $\mathbb{P}(\mathcal{V}) \geq 1 - \delta/2$.*

Proof. By gaussian tail inequality, we have for a fixed t $P(|\epsilon_t| > \sigma\sqrt{2\log(4n/\delta)}) \leq \delta/2n$. By taking a union bound we get $P(|\epsilon_t| \geq \sigma\sqrt{2\log(4n/\delta)}) \leq \delta/2$ for all $t \in [n]$. ☐

Some notations. In the analysis thereafter, we will use the following filtration.

$$\mathcal{F}_j = \sigma((i_1, y_{i_1}), \ldots, (i_{j-1}, y_{i_{j-1}})).$$

Let's denote $\mathbb{E}_j[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_j]$ and $\mathrm{Var}_j[\cdot] := \mathrm{Var}[\cdot|\mathcal{F}_j]$. Let $\theta_j = f(x_{i_j})$ and $\tilde{\sigma} = \sigma\sqrt{2\log(4n/\delta)}$. Let $R_\sigma = 16(B+\tilde{\sigma})^2$ and $\mathcal{T}(I) = \{t \in [n] : i_t \in I\}$

Lemma 6. *(Freedman type inequality, (Beygelzimer et al., 2011)) For any real valued martingale difference sequence $\{Z_t\}_{t=1}^T$ with $|Z_t| \leq R$ it holds that,*

$$\sum_{t=1}^T Z_t \leq \eta(e-2)\sum_{t=1}^T \mathrm{Var}_t[Z_t] + \frac{R\log(1/\delta)}{\eta},$$

with probability atleast $1-\delta$ for all $\eta \in [0, 1/R]$.

Lemma 7. *For any $j \in [n]$, we have*

1. $\mathbb{E}_j[(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2|\mathcal{V}] = \mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}]$.
2. $\mathrm{Var}_j[(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2|\mathcal{V}] \leq R_\sigma \mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}]$.

Proof. We have,

$$\begin{aligned}\mathbb{E}_j[(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2|\mathcal{V}] &=_{(a)} \mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}] - 2\mathbb{E}_j[\epsilon_j|\mathcal{V}]\mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)|\mathcal{V}],\\ &= \mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}],\end{aligned}$$

where line (a) is due to the independence of ϵ_j with the past. Since $(\mathcal{A}_I(j) + \theta_j - 2y_j)^2 \leq 16(B+\tilde{\sigma})^2$ under the event $\mathcal{V}$, it holds that

$$\begin{aligned}\mathrm{Var}_j[(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2|\mathcal{V}] &\leq \mathbb{E}_j[(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2|\mathcal{V}]^2,\\ &\leq 16(B+\tilde{\sigma})^2\mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}].\end{aligned}$$

☐

Lemma 8. *For any interval $I \in \mathcal{I}$, it holds with probability atleast $1-\delta$ that*

1. $\sum_{j\in\mathcal{T}(I)}(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2 \leq \sum_{j\in\mathcal{T}(I)}(e-1)(\mathcal{A}_I(j) - \theta_j)^2 + R_\sigma^2\log(2n\log n/\delta)$,
2. $\sum_{j\in\mathcal{T}(I)}(y_j - \hat{y}_j)^2 - (y_j - \theta_j)^2 \geq \sum_{j\in\mathcal{T}(I)}(3-e)(\hat{y}_j - \theta_j)^2 - R_\sigma^2\log(2n\log n/\delta)$.

Proof. Define $Z_j = (y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2 - (\mathcal{A}_I(j) - \theta_j)^2$.

Condition on the event $\mathcal{V}$ that $|\epsilon_t| \leq \sigma\sqrt{2\log(4n/\delta)}.\forall t \in [n]$ which happens with probability atleast $1-\delta/2$ by Lemma 5. By Lemma 7, we have $\{Z_j\}_{j\in\mathcal{T}(I)}$ is a martingale difference sequence and $|Z_j| \leq 16(B+\tilde{\sigma})^2 = R_\sigma$. Note that once we condition on the filtration $\mathcal{F}_j$, there is no randomness remaining in the terms $(\mathcal{A}_I(j) - \theta_j)^2$ and $(\hat{y}_j - \theta_j)^2$. Hence $\mathbb{E}_j[(\mathcal{A}_I(j) - \theta_j)^2|\mathcal{V}] = (\mathcal{A}_I(j) - \theta_j)^2$ and $\mathbb{E}_j[(\hat{y}_j - \theta_j)^2|\mathcal{V}] = (\hat{y}_j - \theta_j)^2$. Using Lemma 6 and taking $\eta = 1/R_\sigma$ we get,

$$\sum_{j\in\mathcal{T}(I)}(y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2 \leq \sum_{j\in\mathcal{T}(I)}(e-1)(\mathcal{A}_I(j) - \theta_j)^2 + R_\sigma^2\log(4n\log n/\delta),$$

with probability atleast $1 - \delta/(4n \log n)$ for a fixed expert A_I. Taking a union bound across all $O(n \log n)$ experts in $\mathcal{E}$ leads to,

$$\mathbb{P}\left(\sum_{j \in \mathcal{T}(I)} (y_j - \mathcal{A}_I(j))^2 - (y_j - \theta_j)^2 \geq \sum_{j \in \mathcal{T}(I)} (e-1)(\mathcal{A}_I(j) - \theta_j)^2 + R_\sigma^2 \log(2n \log n/\delta) | \mathcal{V} \right) \leq \delta/4,$$

for any expert $\mathcal{A}_I$.

By similar arguments on the martingale difference sequence $(\hat{y}_j - \theta_j)^2 - (y_j - \hat{y}_j)^2 - (y_j + \theta_j)^2$, it can be shown that

$$\mathbb{P}\left(\sum_{j \in \mathcal{T}(I)} (y_j - \hat{y}_j)^2 - (y_j - \theta_j)^2 \leq \sum_{j \in \mathcal{T}(I)} (3-e)(\hat{y}_j - \theta_j)^2 - R_\sigma^2 \log(2n \log n/\delta) | \mathcal{V} \right) \leq \delta/4,$$

for any interval $I \in \mathcal{I}|_{[n]}$. Taking union bound across the previous two bad events and multiplying the probability of noise boundedness event $\mathcal{V}$ leads to the lemma. $\square$

Lemma 1. *For any interval $I \in \mathcal{I}|_{[n]}$ such that $\mathcal{T}(I)$ is non-empty, the predictions made by* ALIGATOR *$\hat{y}_t$ satisfy*

$$\sum_{t \in \mathcal{T}(I)} (\hat{y}_t - \theta_t)^2 \leq \frac{e-1}{3-e} \sum_{t \in \mathcal{T}(I)} (\mathcal{A}_I(t) - \theta_t)^2 + \frac{\log(n \log n) R_\sigma + 2R_\sigma^2 \log(2n \log n/\delta)}{3-e},$$

with probability atleast $1 - \delta$.

Proof. Condition on the event $\mathcal{V}$. Then the losses $f_t(x) = (y_t - x)^2$ are $\frac{1}{4(B+\sigma\sqrt{\log(2n/\delta)})^2} := \eta$ exp-concave (Haussler et al., 1998; Cesa-Bianchi and Lugosi, 2006). Since we pass $\eta \cdot f_t(x)$ as losses to SAA in ALIGATOR, Lemma 2 gives

$$\sum_{t \in \mathcal{T}(I)} -\log\left(\sum_{J \in A_t} w_{t,J} e^{-\eta f_t(\mathcal{A}_J(t))} \right) - \eta f_t(\mathcal{A}_I(t)) \leq \log(n \log n). \tag{1}$$

By η exp-concavity of $f_t(x)$, we have

$$\begin{aligned} -\log\left(\sum_{J \in A_t} w_{t,J} e^{-\eta f_t(\mathcal{A}_J(t)))} \right) &\geq \eta f_t\left(\sum_{J \in A_t} w_{t,J} \mathcal{A}_I(t) \right), \\ &= \eta f_t(\hat{y}_t). \end{aligned} \tag{2}$$

Combining (1) and (2) gives,

$$\begin{aligned} \sum_{t \in \mathcal{T}(I)} f_t(\hat{y}_t) - f_t(\mathcal{A}_I(t)) &\leq \frac{\log(n \log n)}{\eta}, \\ &\leq \log(n \log n) R_\sigma. \end{aligned}$$

So,

$$\sum_{t \in \mathcal{T}(I)} (y_t - \hat{y}_t)^2 - (y_t - \theta_t)^2 \leq \sum_{t \in \mathcal{T}(I)} (y_t - \mathcal{A}_I(t))^2 - (y_t - \theta_t)^2 + \log(n \log n) R_\sigma,$$

Now invoking Lemma (8) followed by a trivial rearrangement completes the proof.

$\square$

Lemma 3. *Let $\mathbb{S} = \{x_{k_1} < \ldots, < x_{k_m}\} \subseteq \mathcal{X}$ be the exogenous variables queried by the adversary over n rounds where each $k_i \in [n]$. Denote $\theta^{(i)} := f(x_{k_i})$ and $p(i) := \#\{t : x_{i_t} = x_{k_i}\}$ for each $i \in [m]$. Denote $[x_i, x_j] := \{x_{k_i}, x_{k_{i+1}}, \ldots, x_{k_j}\}$. For any $[x_i, x_j] \subseteq \mathbb{S}$, define $V(x_i, x_j) = \sum_{k=i}^{j-1} |\theta^{(i)} - \theta^{(i+1)}|$. There exists a partitioning $\mathcal{P} = \{[x_1, x_{r_1}], [x_{r_1+1}, x_{r_2}], \ldots, [x_{r_{M-1}+1}, x_m]\}$ of $\mathbb{S}$ that satisfies*

1. *For any* $[x_i, x_j] \in \mathcal{P} \setminus \{[x_{r_{M-1}+1}, x_m]\}$, $V(x_i, x_j) \leq \frac{B}{\sqrt{\sum_{k=i}^{j} p(k)}}$.

2. $V(x_{r_{M-1}+1}, x_{m-1}) \leq \frac{B}{\sqrt{\sum_{k=r_{M-1}+1}^{m-1} p(k)}}$.

3. *Number of partitions* $M \leq \max\{3n^{1/3}C_n^{2/3}B^{-2/3}, 1\}$.

Proof. We provide below a constructive proof. Consider the following scheme of partitioning $\mathbb{S}$.

1. Set $\text{pings} = p(1), \text{TV} = 0, M = 1$.

2. Start a partition from x_1.

3. For i = 2 to m
   (a) If $\text{TV} + |\theta^{(i)} - \theta^{(i-1)}| > \frac{B}{\sqrt{\text{pings}+p(i)}}$:
      i. $\text{pings} = p(i), \text{TV} = 0$ // start a new bin (partition) from position x_i.
      ii. $M = M + 1$ // increase the bin counter
   (b) Else:
      i. $\text{pings} = \text{pings} + p(i), \text{TV} = \text{TV} + |\theta^{(i)} - \theta^{(i-1)}|$

Statements 1 and 2 of the Lemma trivially follows from the strategy. Next, we provide an upper bound on number of bins M spawned by the above scheme. Let $[x_1, x_{r_1}], [x_{r_1+1}, x_{r_2}], \ldots, [x_{r_{M-1}, x_{r_M}}]$ be the partition of $\mathbb{S}$ discovered by the above scheme.

Define the quantity $\text{TV}_1 := \sum_{i=1}^{r_1} |\theta^{(i)} - \theta^{(i+1)}|$ associated with bin 1. Similarly define $\text{TV}_2, \ldots, \text{TV}_{M-1}$ for other bins.

Define $N(1) = \sum_{i=1}^{r_1+1} p(i)$. Similarly define $N(2), \ldots N(M-1)$. It is immediate that $\sum_{i=1}^{M-1} N(i) \leq 2n$.

We have,

$$
\begin{aligned}
C_n &\geq \sum_{i=1}^{M-1} \text{TV}_i, \\
&\geq_{(1)} \sum_{i=1}^{M-1} \frac{B}{\sqrt{N(i)}}, \\
&\geq_{(2)} \frac{(M-1)^{3/2} \cdot B}{\sqrt{2n}},
\end{aligned}
$$

where (1) follows from step 3(a) of the partitioning scheme and (2) is due to convexity of $1/\sqrt{x}, x > 0$ and applying Jensen's inequality. Rearranging and noting that $M - 1 \geq M/2$, when $M > 1$, we obtain

$$
M \leq 3n^{1/3}C_n^{2/3}B^{-2/3}.
$$

Note that when $C_n = 0$, M will remain 1 as a result of the partitioning scheme.

□

Lemma 4. *Let* $\{\underline{x}, < \ldots, < \bar{x}\}$ *be the exogenous variables queried by the adversary over* n *rounds in an arbitrary interval* $I \in \mathcal{I}|_{[n]}$. *Then with probability atleast* $1 - \delta$

$$
\sum_{t \in \mathcal{T}(I)} (\theta_t - \mathcal{A}_I(t))^2 \leq 2V(\underline{x}, \bar{x})^2 |\mathcal{T}(I)| + 2\sigma^2 \log(2n^3 \log n/\delta) \log(|\mathcal{T}(I)|),
$$

where $V(\cdot, \cdot)$ *is defined as in Lemma 3.*

Proof. Let $q(t) = \sum_{s=1}^{t-1} \mathbf{1}\{i_s \in I\}$. Assume $q(t) > 0$. Fix a particular expert $\mathcal{A}_I$ and a time t. Since $y_t \sim N(\theta_t, \sigma^2)$ by gaussian tail inequality we have,

$$\mathbb{P}\left(\left|\frac{\sum_{s=1}^{t-1}(y_s - \theta_s)\mathbf{1}\{i_s \in I\}}{\sum_{s=1}^{t-1}\mathbf{1}\{i_s \in I\}}\right| \geq \frac{\sigma}{\sqrt{q(t)}}\sqrt{\log\left(\frac{2n^3\log n}{\delta}\right)}\right) \leq \frac{\delta}{(n^3\log n)}.$$

Applying a union bound across all time points and all experts implies that for any expert $\mathcal{A}_I$ and $t \in \mathcal{T}(I)$ with q(t) > 0,

$$\left|\mathcal{A}_I(t) - \frac{\sum_{s=1}^{t-1}\theta_s\mathbf{1}\{i_s \in I\}}{q(t)}\right| \leq \frac{\sigma}{\sqrt{q(t)}}\sqrt{\log\left(\frac{2n^3\log n}{\delta}\right)}$$

with probability atleast $1-\delta$.

Now adding and subtracting θ_t inside the $|\cdot|$ on LHS and using $|a-b| \geq |a| - |b|$ yields,

$$|\mathcal{A}_I(t) - \theta_t| \leq \left|\theta_t - \frac{\sum_{s=1}^{t-1}\theta_s\mathbf{1}\{i_s \in I\}}{q(t)}\right| + \frac{\sigma}{\sqrt{q(t)}}\sqrt{\log\left(\frac{2n^3\log n}{\delta}\right)}.$$

Hence,

$$\begin{aligned}\sum_{t\in\mathcal{T}(I)}(\theta_t - \mathcal{A}_I(t))^2 &\leq_{(a)} \sum_{t\in\mathcal{T}(I)} 2\left(\theta_t - \frac{\sum_{s=1}^{t-1}\theta_s\mathbf{1}\{i_s \in I\}}{q(t)}\right)^2 + 2\frac{\sigma^2}{q(t)}\log\left(\frac{2n^3\log n}{\delta}\right)\\ &\leq \sum_{t\in\mathcal{T}(I)} 2\left(\theta_t - \frac{\sum_{s=1}^{t-1}\theta_s\mathbf{1}\{i_s \in I\}}{q(t)}\right)^2 + 2\sigma^2\log(|\mathcal{T}(I)|)\log\left(\frac{2n^3\log n}{\delta}\right), \quad (3)\end{aligned}$$

with probability atleast $1-\delta$. In (a) we used the relation $(a+b)^2 \leq 2a^2 + 2b^2$.

Further we have,

$$\sum_{t\in\mathcal{T}(I)} 2\left(\theta_t - \frac{\sum_{s=1}^{t-1}\theta_s\mathbf{1}\{i_s \in I\}}{q(t)}\right)^2 \leq 2V(\underline{x}, \bar{x})^2|\mathcal{T}(I)|. \quad (4)$$

Combining (3) and (4) completes the proof.

□

B.3 Proof of the main result: Theorem 5

Proof. Throughout the proof we carry forward all notations used in Lemmas 3 and 4.

We will apply Corollary 2 to the partition in Lemma 3. Take a specific partition $[x_i, x_j] \in \mathcal{P}$ with $j \neq m$. Consider a set of indices $F = \{k_i, k_i+1, \ldots, k_j\}$ of consecutive natural numbers between k_i and k_j. By Proposition 1 F can be covered using elements in $\mathcal{I}|_{[n]}$. Let this cover be $\mathcal{U}$. For any $I \in \mathcal{U}$, we have

$$\begin{aligned}\sum_{t\in\mathcal{T}(I)}(\theta_t - \mathcal{A}_I(t))^2 &\leq_{(a)} 2V(\underline{x}, \bar{x})^2|\mathcal{T}(I)| + 2\sigma^2\log(2n^3\log n/\delta)\log(|\mathcal{T}(I)|)\\ &\leq 2V(\underline{x}, \bar{x})^2|\mathcal{T}(F)| + 2\sigma^2\log(2n^3\log n/\delta)\log(|\mathcal{T}(I)|)\\ &\leq_{(b)} 2B^2 + 2\sigma^2\log(2n^3\log n/\delta)\log(n),\end{aligned}$$

, with probability atleast $1-\delta$. Step (a) is due to Lemma 4 and (b) is due to statement 1 of Lemma 3.

Using Lemma 1 and a union bound on the bad events in Lemmas 1 and 4 yields,

$$\sum_{t\in\mathcal{T}(I)}(\hat{y}_t - \theta_t)^2 \leq \frac{e-1}{3-e}\left(2B^2 + 2\sigma^2\log(2n^3\log n/\delta)\log(n)\right) + \lambda,$$

with probability atleast $1-2\delta$ and λ is as defined in Corollary 2. Due to the property of exponentially decaying lengths as stipulated by Proposition 1, there are only atmost $2\log|F| \leq 2\log n$ intervals in $\mathcal{U}$. So,

$$\sum_{t \in \mathcal{T}(F)} (\hat{y}_t - \theta_t)^2 \leq 2\log n \left(\frac{e-1}{3-e} \left(2B^2 + 2\sigma^2 \log(2n^3 \log n/\delta) \log(n)\right) + \lambda \right).$$

Similar bound can be obtained for the last bin $[x_{r_{M-1}+1}, x_m]$ in $\mathcal{P}$. There are two cases to consider. In case 1, we consider the scenario when $V(x_{r_{M-1}+1}, x_m)$ obeys relation 1 of Lemma 3. Then the analysis is identical to the one presented above. In case 2, we consider the scenario when $V(x_{r_{M-1}+1}, x_{m-1})$ obeys relation 2 of Lemma 3 while $V(x_{r_{M-1}+1}, x_m)$ doesn't. Then the error incurred within the interior $[x_{r_{M-1}+1}, x_{m-1}]$ can be bounded as before. To bound the error at last point, we only need to bound the error of expert that performs mean estimation of iid gaussians. It is well known that the cumulative squared error for this problem is atmost $\sigma^2 \log(n/\delta)$ with probability atleast $1-\delta$.

By Lemma 3, $|\mathcal{P}| = \max\{3n^{1/3}C_n^{2/3}B^{-2/3}, 1\}$. Hence the total error summed across all partitions in $\mathcal{P}$ becomes,

$$\begin{aligned}
\sum_{t=1}^{n} (\hat{y}_t - \theta_t)^2 &\leq 2\log n \left(\frac{e-1}{3-e} \left(4n^{1/3}C_n^{2/3}B^{4/3} + 4\sigma^2 \log(2n^3 \log n/\delta) \log(n) n^{1/3} C_n^{2/3} B^{-2/3}\right)\right) \\
&\quad + 4\log(n) \frac{e-1}{3-e} \lambda n^{1/3} C_n^{2/3} B^{-2/3} \\
&\quad + 2\log(n) \left(\frac{e-1}{3-e} \left(2B^2 + 2\sigma^2 \log(2n^3 \log n/\delta) \log(n)\right) + \lambda \right) + \sigma^2 \log(n/\delta), \\
&= \tilde{O}(n^{1/3} C_n^{2/3}),
\end{aligned} \tag{5}$$

with probability atleast $1-2\delta$. A change of variables from $2\delta \to \delta$ completes the proof. As a closing note, we remark that the aggressive dependence of B in (5) on cases when B is too small can be dampened by using a threshold of $\frac{1}{\sqrt{\text{pings}_{+p(i)}}}$ in the partition scheme presented in proof of Lemma 3. $\square$

C Excluded details in Experimental section

Waveforms. The waveforms shown in Fig. 1 and 2 are borrowed from (Donoho and Johnstone, 1994). Note that both functions exhibit spatially inhomogeneous smoothness behaviour.

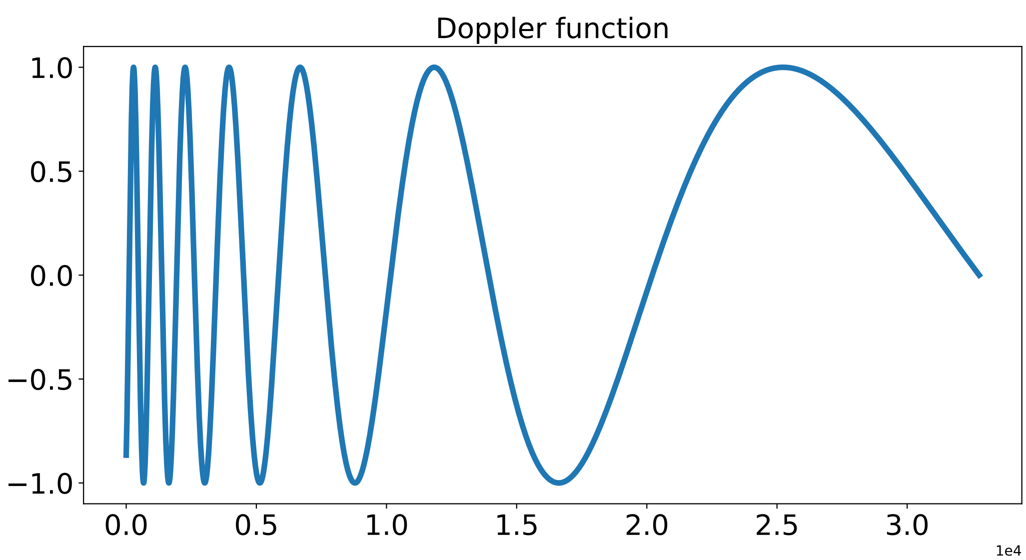

Figure 1: *Doppler function, TV = 27*

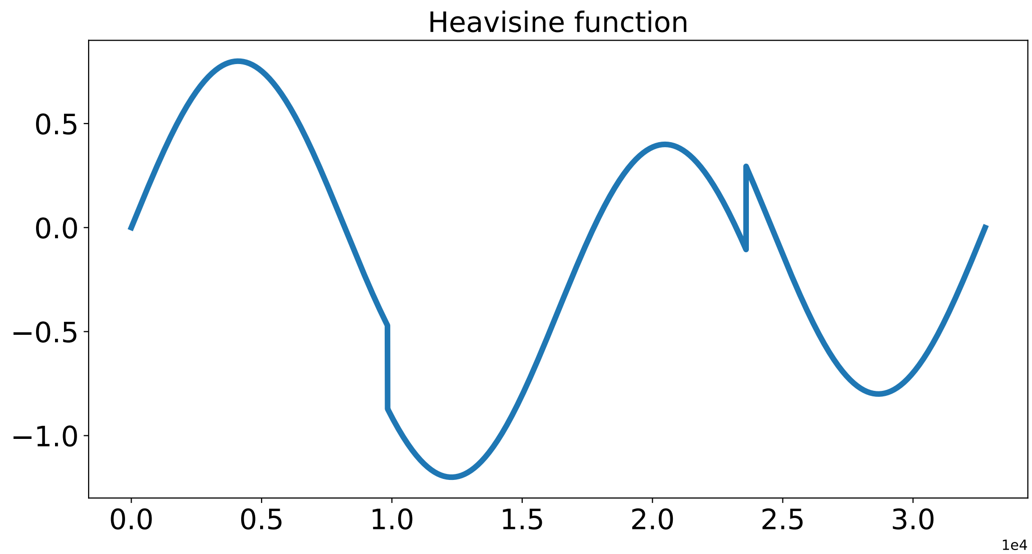

Figure 2: *Heavisine function, TV = 7.2*

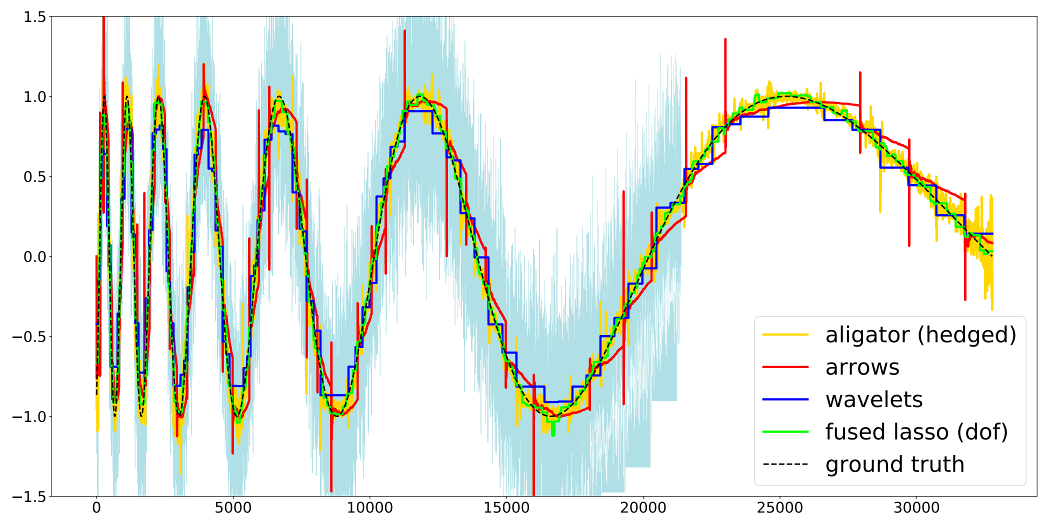

Figure 3: *Fitted signals for Doppler function with noise level* $\sigma = 0.35$

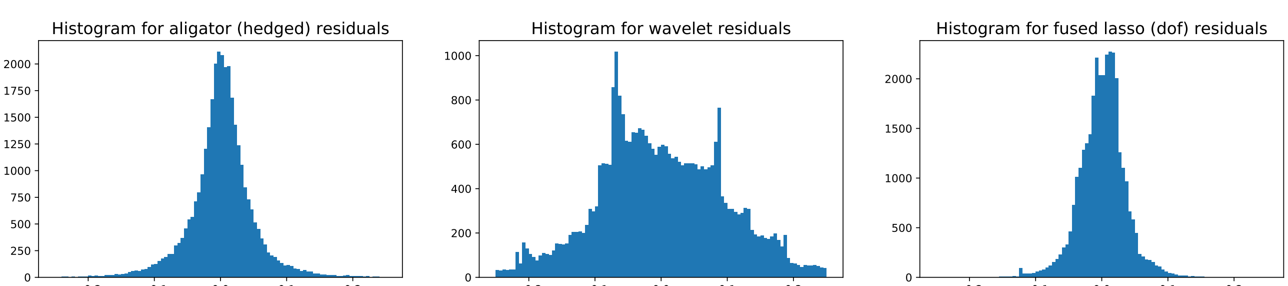

Figure 4: *Histogram of residuals for various algorithms when run on Doppler function with noise level* $\sigma = 0.35$*. Note that they are residuals w.r.t to ground truth.* ALIGATOR *incurs lower bias than wavelets. The bias incurred by dof fused lasso is roughly comparable to* ALIGATOR *while former is more compute intensive.*

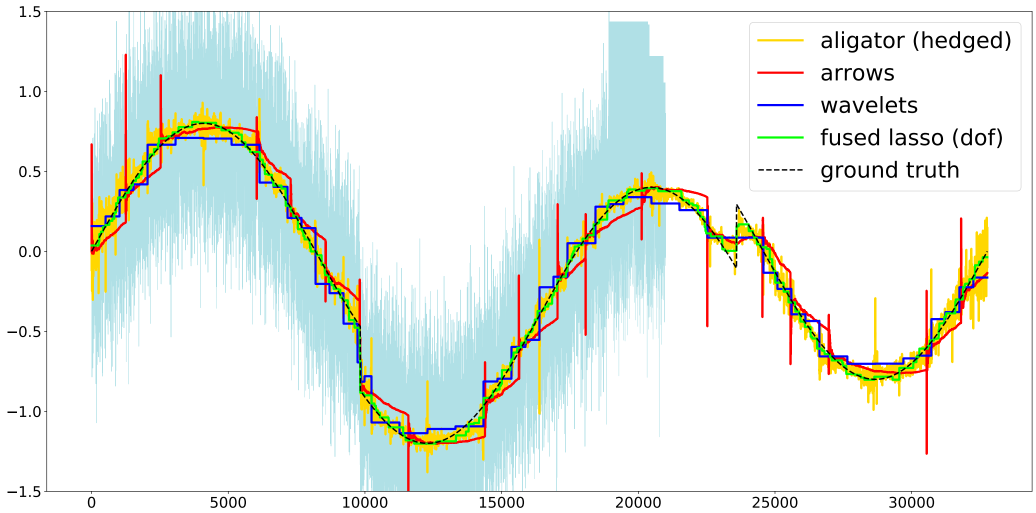

Figure 5: *Fitted signals for Heavisine function with noise level* $\sigma = 0.35$

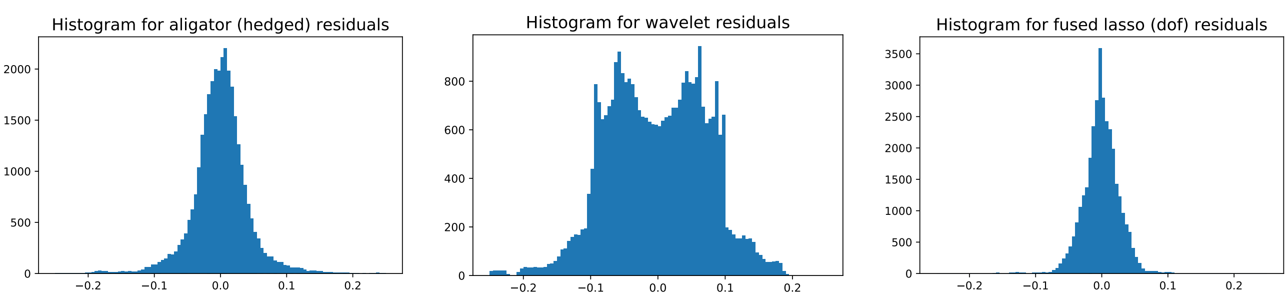

Figure 6: *Histogram of residuals for various algorithms when run on Heavisine function with noise level* $\sigma = 0.35$. *Note that they are residuals w.r.t to ground truth.* ALIGATOR *incurs lower bias than wavelets. The bias incurred by dof fused lasso is roughly comparable to* ALIGATOR *while former is more compute intensive.*

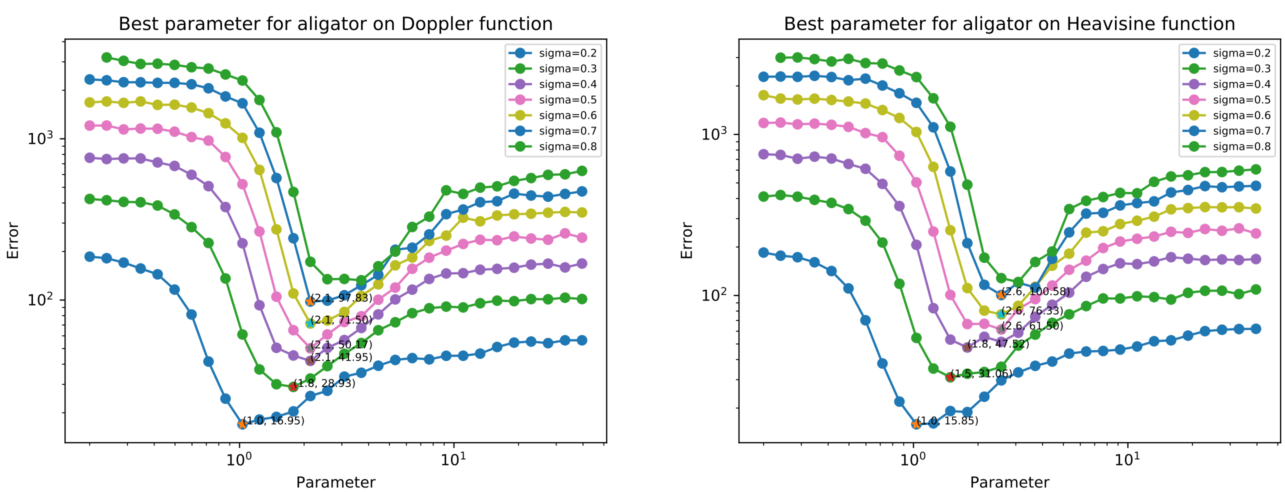

Figure 7: *Hyper-parameter search for learning rate in* ALIGATOR *(heuristics).*

Hyper-parameter search. Initially we used a grid search on an exponential grid to realize that the optimal λ across all experiments fall within the range [0.125, 8]. Then we used a fine-tuned grid

$[0.125, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 10, 12, 14, 16]$ to search for the final hyper parameter value. For ALIGATOR *(heuristics)*, we searched for different noise levels in order to find best learning rate. We set search method as Loss$/(para * (\sigma^2 + \sigma^2/m))$. As Fig. 7 shows, $para = 2$ is found to provide good results across all signals we consider.

Padding for wavelets. For "wavelet" estimator in Fig. 6, when data length is not a power of 2, we used the reflect padding mode in (Lee et al., 2019), though the results are similar for other padding schemes.

Experiments on Real Data. We follow the experimental setup described in Section 5. A qualitative comparison of the forecasts for the state of New Mexico, USA is illustrated in Fig. 8. The average RMSE of ALIGATOR and Holt ES for all states in USA is reported in Table 1.

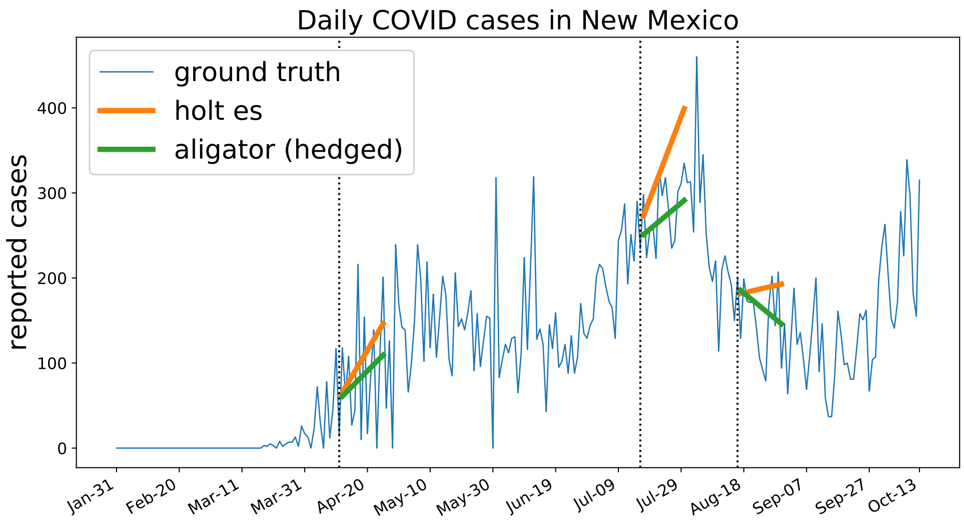

Figure 8: *A demo on forecasting COVID cases based on real world data. We display the two weeks forecasts of hedged* ALIGATOR *and Holt ES, starting from the time points identified by the dotted lines. Both the algorithms are trained on a 2 month data prior to each dotted line. We see that hedged* ALIGATOR *detects changes in trends more quickly than Holt ES. Further, hedged* ALIGATOR *attains a 12% reduction in the average RMSE from that of Holt ES (see Table 1).*

State	RMSE Aligator	RMSE Holt ES	% improvement
New Jersey	411.87	546.89	24.69
Ohio	216.24	280.24	22.84
Florida	1330.33	1671.23	20.4
Alabama	290.71	362.13	19.72
New York	876.35	1054.2	16.87
Rhode Island	85.11	98.23	13.35
Vermont	7.59	8.7	12.76
Kansas	142.17	162.16	12.33
New Mexico	57.88	65.99	12.29
Connecticut	206.79	235.6	12.23
California	1456.48	1650.25	11.74
Pennsylvania	258.21	290.6	11.14
Kentucky	145.61	163.59	10.99
New Hampshire	25.16	27.99	10.1
Minnesota	161.41	179.12	9.89
Michigan	315.86	350.24	9.82
Hawaii	30.24	33.18	8.86
Texas	1510.42	1650.73	8.5
South Dakota	56.83	61.8	8.04
Utah	118.97	128.96	7.74
Alaska	17.54	18.96	7.52
Washington	188.8	202.74	6.88
North Carolina	265.74	284.47	6.58
Nebraska	98.49	105.41	6.56
Montana	28.31	30.28	6.51
Missouri	224.51	239.9	6.42
Iowa	205.77	219.28	6.16
District of Columbia	33.58	35.74	6.04
Virginia	194.29	206.44	5.89
Nevada	159.88	168.92	5.35
Wyoming	16.43	17.25	4.73
Georgia	493.93	518.27	4.7
Oregon	55.48	58.21	4.68
Louisiana	562.89	590.49	4.67
Maryland	209.95	218.22	3.79
Illinois	475.49	492.09	3.37
West Virginia	37.34	38.63	3.33
Delaware	64.1	66.26	3.26
Tennessee	384.55	396.95	3.12
Arizona	481.91	493.73	2.39
South Carolina	271.87	277.42	2.0
Idaho	93.83	95.44	1.68
Colorado	142.58	144.53	1.35
Mississippi	206.67	209.11	1.16
Arkansas	164.83	164.88	0.03
Massachusetts	302.79	301.8	-0.32
Oklahoma	151.82	146.65	-3.41
Indiana	185.1	178.2	-3.73
North Dakota	42.14	40.49	-3.92
Wisconsin	219.04	203.37	-7.15
Maine	14.59	13.37	-8.36

Table 1: Average RMSE across all states in USA. The experimental setup and computation of error metrics are as described in Section 5. The % improvement tab is computed as follows. Let x_1 and x_2 be the RMSE of ALIGATOR and Holt ES respectively. Then % improvement $= (x_2 - x_1)/\max\{x_1, x_2\}$.

References

Dmitry Adamskiy, Wouter M. Koolen, Alexey Chernov, and Vladimir Vovk. A closer look at adaptive regret. *Journal of Machine Learning Research*, 2016.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5): 1227–1244, 2015.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Niangjun Chen, Gautam Goel, and Adam Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. In *Conference on Learning Theory (COLT-18)*, 2018a.

Xi Chen, Yining Wang, and Yu-Xiang Wang. Non-stationary stochastic optimization under lp, q-variation measures. 2018b.

David Donoho and Iain Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Eric Hall and Rebecca Willett. Dynamical models and tracking regret in online convex programming. In *International Conference on Machine Learning (ICML-13)*, pages 579–587, 2013.

D. Haussler, J. Kivinen, and M. K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inf. Theor.*, 1998.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Jan-Christian Hutter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory (COLT-16)*, 2016.

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

Wojciech Kotłowski, Wouter M. Koolen, and Alan Malek. Online isotonic regression. In *Annual Conference on Learning Theory (COLT-16)*, volume 49, pages 1165–1189. PMLR, 2016.

Gregory Lee, Ralf Gommers, Filip Waselewski, Kai Wohlfahrt, and Aaron O Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 2019. URL https://doi.org/10.21105/joss.01237.

Yuan Li, Benjamin Mark, Garvesh Raskutti, and Rebecca Willett. Graph-based regularization for regression problems with highly-correlated designs. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 740–742. IEEE, 2018.

Oscar Hernan Madrid Padilla, James Sharpnack, and James G Scott. The dfs fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research*, 18(1):6410–6445, 2017.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. *Advances in Neural Information Processing Systems (NIPS-16)*, 2016.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, James Sharpnack, and Ryan Tibshirani. Higher-order total variation classes on grids: Minimax theory and trend filtering methods. *Advances in Neural Information Processing Systems (NIPS-17)*, 2017.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning (ICML-16)*, pages 449–457, 2016.

Jianjun Yuan and Andrew Lamperski. Trading-off static and dynamic regret in online least-squares and beyond. 2019.

K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 2017.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS-18)*, pages 1323–1333, 2018a.

Lijun Zhang, Tianbao Yang, Zhi-Hua Zhou, et al. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning (ICML-18)*, pages 5877–5886, 2018b.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.