A More on Related Work

For any forecasting strategy whose output $\hat{y}_t$ at time $t$ depends only on past observations, we have $\mathbb{E}[(\hat{y}_t - y_t)^2] = \mathbb{E}[(\hat{y}_t - f(x_i))^2]$. Hence any algorithm that minimizes the dynamic regret against the sequence $f(x_i), \ldots, f(x_n)$ 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; Jadabaie 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 $\theta_{i_t} := f(x_{i_t})$.

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

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

with probability at least $1 - \delta$.

**Corollary 2.** Let $\mathcal{S} = \{P_1, \ldots, P_M\}$ 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 $\lambda : = \log(n \log n)R_\sigma + 2R_\sigma^2 \log(2n \log n/\delta)$, Then ALIGATOR forecasts $\hat{y}_t$ satisfy

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

with probability at least $1 - \delta$.

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 $\tilde{\sigma}$. The existence of such partitions is a non-trivial matter.

**Lemma 3.** Let $\mathcal{S} = \{x_{k_1}, \ldots, x_{k_m}\} \subseteq \mathcal{X}$ 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 $[x_i, x_j] := \{x_{k_1}, x_{k_{i+1}}, \ldots, x_{k_j}\}$. For any $[x_i, x_j] \subseteq \mathcal{S}$, define $V(x_i, x_j) = \sum_{k=1}^{j-i} |\theta^{(i)} - \theta^{(i+1)}|$. 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 $\mathcal{S}$ 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. $V(x_{r_{M-1}+1}, x_m) \leq \frac{B}{\sqrt{\sum_{k=r_{M-1}+1}^{m} p(k)}}$.

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 $\{x, < \ldots, < \bar{x}\}$ be the exogenous variables queried by the adversary over $n$ rounds in an arbitrary interval $I \in \mathcal{I}_{|n|}$. Then with probability at least $1 - \delta$

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

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 almost $\log n$ 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 $V(x, \bar{x})^2|\mathcal{T}(I)|$ 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 $O(1)$. Now appealing to statement 3 of Lemma 3, we get a total error of $O(n^{1/3}C_n^{2/3})$ 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 $P(\mathcal{V}) \geq 1 - \delta/2$.

*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 $t \in [n]$. \hfill \qed

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

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

Let’s denote $E_j[\cdot] := E[\cdot | F_j]$ and $\text{Var}_j[\cdot] := \text{Var}[\cdot | F_j]$. 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 $T(I) = \{t \in [n] : i_t \in I\}$

**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,

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

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

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

1. $E_j[(y_j - A_t(j))^2 - (y_j - \theta_j)^2 | \mathcal{V}] = E_j[(A_t(j) - \theta_j)^2 | \mathcal{V}]$.
2. $\text{Var}_j[(y_j - A_t(j))^2 - (y_j - \theta_j)^2 | \mathcal{V}] \leq R_\sigma E_j[(A_t(j) - \theta_j)^2 | \mathcal{V}]$.

*Proof.* We have,

$$E_j[(y_j - A_t(j))^2 - (y_j - \theta_j)^2 | \mathcal{V}] = \text{(a)} \ E_j[(A_t(j) - \theta_j)^2 | \mathcal{V}] - 2E_j[\epsilon_j | \mathcal{V}]E_j[(A_t(j) - \theta_j)| \mathcal{V}],$$

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

$$\text{Var}_j[(y_j - A_t(j))^2 - (y_j - \theta_j)^2 | \mathcal{V}] \leq E_j[(y_j - A_t(j))^2] - (y_j - \theta_j)^2 | \mathcal{V}]^2, \leq 16(B + \tilde{\sigma})^2 E_j[(A_t(j) - \theta_j)^2 | \mathcal{V}].$$

\hfill \qed

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

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

*Proof.* Define $Z_j = (y_j - A_t(j))^2 - (y_j - \theta_j)^2 - (A_t(j) - \theta_j)^2$.

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 $\{Z_j\}_{j \in T(I)}$ is a martingale difference sequence and $|Z_j| \leq 16(B + \tilde{\sigma})^2 = R_\sigma$. Note that once we condition on the filtration $F_j$, there is no randomness remaining in the terms $(A_t(j) - \theta_j)^2$ and $(\hat{y}_j - \theta_j)^2$. Hence $E_j[(A_t(j) - \theta_j)^2 | \mathcal{V}] = (A_t(j) - \theta_j)^2$ and $E_j[(\hat{y}_j - \theta_j)^2 | \mathcal{V}] = (\hat{y}_j - \theta_j)^2$. Using Lemma 6 and taking $\eta = 1/R_\sigma$ we get,

$$\sum_{j \in T(I)} (y_j - A_t(j))^2 - (y_j - \theta_j)^2 \leq \sum_{j \in T(I)} (e - 1)(A_t(j) - \theta_j)^2 + R_\sigma^2 \log(4n \log n/\delta),$$
with probability at least $1 - \delta/(4n \log n)$ for a fixed expert $A_I$. Taking a union bound across all $O(n \log n)$ experts in $E$ leads to,
\[
\mathbb{P}\left( \sum_{j \in \mathcal{T}(I)} (y_j - A_I(j))^2 - (y_j - \theta_j)^2 \geq \sum_{j \in \mathcal{T}(I)} (e - 1)(A_I(j) - \theta_j)^2 + R^2 R_{\sigma}^2 \log(2n \log n/\delta) \right) \leq \delta/4,
\]
for any expert $A_I$.

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

**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
\[
\sum_{t \in \mathcal{T}(I)} (\hat{y}_t - \theta_t)^2 \leq \frac{e - 1}{3 - e} \sum_{t \in \mathcal{T}(I)} (A_I(t) - \theta_t)^2 + \frac{\log(n \log n) R_{\sigma} + 2 R^2 R_{\sigma}^2 \log(2n \log n/\delta)}{3 - e},
\]
with probability at least $1 - \delta$.

**Proof.** Condition on the event $V$. Then the losses $f_t(x) = (y_t - x)^2$ are $\frac{1}{4(B + \sigma \sqrt{\log(2n/\delta)})^2}$-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
\[
\sum_{t \in \mathcal{T}(I)} -\log \left( \sum_{j \in A_I} w_{t,j} e^{-\eta f_t(A_I(j))} \right) - \eta f_t(A_I(t)) \leq \log(n \log n).
\]

(1)

By $\eta$ exp-concavity of $f_t(x)$, we have
\[
-\log \left( \sum_{j \in A_I} w_{t,j} e^{-\eta f_t(A_I(j))} \right) \geq \eta f_t \left( \sum_{j \in A_I} w_{t,j} A_I(t) \right),
\]
\[
= \eta f_t(\hat{y}_t).
\]

(2)

Combining (1) and (2) gives,
\[
\sum_{t \in \mathcal{T}(I)} f_t(\hat{y}_t) - f_t(A_I(t)) \leq \frac{\log(n \log n)}{\eta},
\]
\[
\leq \log(n \log n) R_{\sigma}.
\]

So,
\[
\sum_{t \in \mathcal{T}(I)} (y_t - \hat{y}_t)^2 - (y_t - \theta_t)^2 \leq \sum_{t \in \mathcal{T}(I)} (y_t - A_I(t))^2 - (y_t - \theta_t)^2 + \log(n \log n) R_{\sigma},
\]

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

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

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

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 $S$.

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

2. Start a partition from $x_1$.

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

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 $[x_1, x_{r_1}], [x_{r_1+1}, x_{r_2}], \ldots, [x_{rM-1}, x_r]$ be the partition of $\mathbb{S}$ discovered by the above scheme.

Define the quantity $TV_1 := \sum_{i=1}^{r_1} |\theta(i) - \theta(i+1)|$ associated with bin 1. Similarly define $TV_2, \ldots, TV_{M-1}$ 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,

$$C_n \geq \sum_{i=1}^{M-1} TV_i,$$

$$\geq (1) \sum_{i=1}^{M-1} \frac{B}{\sqrt{N(i)}},$$

$$\geq (2) \frac{(M-1)^{3/2} \cdot B}{\sqrt{2n}},$$

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 $\{x, < \ldots, < \bar{x}\}$ be the exogenous variables queried by the adversary over $n$ rounds in an arbitrary interval $I \in \mathcal{I}[n]$. Then with probability at least $1 - \delta$

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

where $V(\cdot, \cdot)$ is defined as in Lemma 3.
Proof. Let \( q(t) = \sum_{s=1}^{t-1} 1\{i_s \in I\} \). Assume \( q(t) > 0 \). Fix a particular expert \( A_t \) and a time \( t \). Since \( y_t \sim N(\theta_t, \sigma^2) \) by gaussian tail inequality we have,
\[
\mathbb{P} \left( \frac{\sum_{s=1}^{t-1} (y_s - \theta_s) 1\{i_s \in I\}}{\sum_{s=1}^{t-1} 1\{i_s \in I\}} \geq \frac{\sigma}{\sqrt{q(t)}} \sqrt{\log \left( \frac{2n^3 \log n}{\delta} \right)} \right) \leq \frac{\delta}{(n^3 \log n)}.
\]
Applying a union bound across all time points and all experts implies that for any expert \( A_t \) and \( t \in T(I) \) with \( q(t) > 0 \),
\[
|A_t(t) - \frac{\sum_{s=1}^{t-1} \theta_s 1\{i_s \in I\}}{q(t)}| \leq \frac{\sigma}{\sqrt{q(t)}} \sqrt{\log \left( \frac{2n^3 \log n}{\delta} \right)}
\]
with probability at least \( 1 - \delta \).

Now adding and subtracting \( \theta_t \) inside the \(| \cdot |\) on LHS and using \(|a - b| \geq |a| - |b|\) yields,
\[
|A_t(t) - \theta_t| \leq \left| \theta_t - \frac{\sum_{s=1}^{t-1} \theta_s 1\{i_s \in I\}}{q(t)} \right| + \frac{\sigma}{\sqrt{q(t)}} \sqrt{\log \left( \frac{2n^3 \log n}{\delta} \right)}.
\]
Hence,
\[
\sum_{t \in T(I)} (\theta_t - A_t(t))^2 \leq_{(a)} \sum_{t \in T(I)} 2 \left( \theta_t - \frac{\sum_{s=1}^{t-1} \theta_s 1\{i_s \in I\}}{q(t)} \right)^2 + 2 \frac{\sigma^2}{q(t)} \log \left( \frac{2n^3 \log n}{\delta} \right)
\]
\[
\leq \sum_{t \in T(I)} 2 \left( \theta_t - \frac{\sum_{s=1}^{t-1} \theta_s 1\{i_s \in I\}}{q(t)} \right)^2 + 2\sigma^2 \log(|T(I)|) \log \left( \frac{2n^3 \log n}{\delta} \right),
\]
with probability at least \( 1 - \delta \). In (a) we used the relation \((a + b)^2 \leq 2a^2 + 2b^2\).

Further we have,
\[
\sum_{t \in T(I)} 2 \left( \theta_t - \frac{\sum_{s=1}^{t-1} \theta_s 1\{i_s \in I\}}{q(t)} \right)^2 \leq 2V(x, \bar{x}|T(I)|).
\]
(4)

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

\(\square\)

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 \([x_i, x_j] \in P\) 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 \( T[n] \). Let this cover be \( U \). For any \( I \in U \), we have
\[
\sum_{t \in T(I)} (\theta_t - A_t(t))^2 \leq_{(a)} 2V(x, \bar{x}|T(I)|) + 2\sigma^2 \log(2n^3 \log n/\delta) \log(|T(I)|)
\]
\[
\leq 2V(x, \bar{x}|T(F)|) + 2\sigma^2 \log(2n^3 \log n/\delta) \log(|T(I)|)
\]
\[
\leq_{(b)} 2B^2 + 2\sigma^2 \log(2n^3 \log n/\delta) \log(n),
\]
, with probability at least \( 1 - \delta \). 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,
\[
\sum_{t \in T(I)} (\hat{\theta}_t - \theta_t)^2 \leq \frac{e - 1}{3 - e} (2B^2 + 2\sigma^2 \log(2n^3 \log n/\delta) \log(n)) + \lambda,
\]
with probability at least $1 - 2\delta$ and $\lambda$ is as defined in Corollary 2. Due to the property of exponentially decaying lengths as stipulated by Proposition 1, there are only at most $2\log|F| \leq 2\log n$ intervals in $\mathcal{U}$. So,

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

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 at most $\sigma^2 \log(n/\delta)$ with probability at least $1 - 2\delta$.

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,

$$\sum_{t=1}^{n} (\hat{y}_t - \theta_t)^2 \leq 2\log n \left( \frac{e - 1}{3 - e} \left( 4n^{1/3}C_n^{2/3}B^{1/3} + 4\sigma^2 \log(2n^3 \log n/\delta) \log(n) n^{1/3}C_n^{2/3}B^{-2/3} \right) + 4 \log(n) \frac{e - 1}{3 - e} \lambda n^{1/3}C_n^{2/3}B^{-2/3} \right) + 2 \log(n) \left( \frac{e - 1}{3 - e} \left( 2B^2 + 2\sigma^2 \log(2n^3 \log n/\delta) \log(n) \right) + \lambda \right) + \sigma^2 \log(n/\delta),$$

with probability at least $1 - 2\delta$. A change of variables from $2\delta \rightarrow \delta$ 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 $\frac{1}{\sqrt{\text{pings}^{1/2}\sigma}}$ in the partition scheme presented in proof of Lemma 3.

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.

![Doppler function](image1.png)

**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. ground truth. $\text{ALIGATOR}$ incurs lower bias than wavelets. The bias incurred by dof fused lasso is roughly comparable to $\text{ALIGATOR}$ 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 $\lambda$ 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 $\text{Loss}/(\text{para} \ast (\sigma^2 + \sigma'^2/m))$. As Fig. 7 shows, $\text{para} = 2$ 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.

![Daily COVID cases in New Mexico](image)

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

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


