Spatially Adaptive Online Prediction of Piecewise Regular Functions

Sabyasachi Chatterjee   SC1706@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
117 Illini Hall
Champaign, IL 61820

Subhajit Goswami   GOSWAMI@MATH.TIFR.RES.IN
School of Mathematics
Tata Institute of Fundamental Research
1, Homi Bhabha Road
Colaba, Mumbai 400005, India

Abstract

We consider the problem of estimating piecewise regular functions in an online setting, i.e., the data arrive sequentially and at any round our task is to predict the value of the true function at the next revealed point using the available data from past predictions. We propose a suitably modified version of a recently developed online learning algorithm called the sleeping experts aggregation algorithm. We show that this estimator satisfies oracle risk bounds simultaneously for all local regions of the domain. As concrete instantiations of the expert aggregation algorithm proposed here, we study an online mean aggregation and an online linear regression aggregation algorithm where experts correspond to the set of dyadic subrectangles of the domain. The resulting algorithms are near linear time computable in the sample size. We specifically focus on the performance of these online algorithms in the context of estimating piecewise polynomial and bounded variation function classes in the fixed design setup. The simultaneous oracle risk bounds we obtain for these estimators in this context provide new and improved (in certain aspects) guarantees even in the batch setting and are not available for the state of the art batch learning estimators.

Keywords: Online Prediction, Spatial/Local Adaptivity, Adaptive Regret, Piecewise Polynomial Fitting, Bounded Variation Function Estimation, Oracle Risk Bounds.

1. Introduction

In this paper we revisit the classical problem of estimating piecewise regular functions from noisy evaluations. The theory discussed here is potentially useful for a general notion of piecewise regularity although we will specifically give attention to the problem of estimating piecewise constant or piecewise polynomial functions of a given degree $m \geq 0$ and bounded variation functions which are known to be well approximable by such piecewise constant/polynomial functions. A classical aim is to design adaptive estimators which adapt optimally to the (unknown) number of pieces of the underlying piecewise constant/polynomial function. For example, if the true signal is (exactly or close to) a piecewise constant function with unknown number of pieces k, then it is desirable that the estimator attains a near parametric $\tilde{O}(k/N)$ rate of convergence (N is the sample size, $\tilde{O}(\cdot)$ indicates "upto some *fixed* power of $\log N$"), for all possible values of k. This desired notion is often shown by demonstrating that an estimator satisfies the so called oracle risk bound trading off

a squared error approximation term and a complexity term. Several nonparametric regression estimators, such as Wavelet Shrinkage Donoho and Johnstone (1994), Donoho and Johnstone (1998), Trend Filtering Mammen and van de Geer (1997), Tibshirani et al. (2014), Tibshirani (2020), Dyadic CART Donoho (1997), Optimal Regression Tree Chatterjee and Goswami (2021a), are known to attain such an oracle risk bound in the context of estimating piecewise constant/polynomial functions. In this context, two natural questions arise which we address in this paper.

Q1: Consider the online version of the problem of estimating piecewise constant or piecewise polynomial functions, i.e., the data arrive sequentially and at any round our task is to predict the value of the true function at the next revealed point using the available data from past predictions. *Does there exist an estimator which attains an oracle risk bound (similar to what is known in the batch learning setting) in the online setting*? This seems a natural and perhaps an important question given applications of online learning to forecasting trends.

Q2: The oracle risk bounds available for batch learning estimators in the literature imply a notion of adaptivity of the estimator. This adaptivity can be thought of as a global notion of adaptivity as the risk bound is for the entire mean sum of squared error of the estimator. If it can be shown that an estimator satisfies an oracle risk bound locally, simultaneously over several subregions of the domain, then this will imply a local/spatial notion of adaptivity. We explain this more in Section 1.2. This gives rise to our second question. *Does there exist an estimator which attains an oracle risk bound simultaneously over several subregions of the domain in the online setting?*. Even in the batch learning setting, it is not known whether state of the art estimators such as Wavelet Shrinkage, Trend Filtering, Dyadic CART, Optimal Regression Tree satisfy such a simultaneous oracle risk bound.

The main purpose of this paper is to recognize, prove and point out that by using a suitably modified version of an online aggregation algorithm developed in the online learning community, it is possible to answer both the above questions in the affirmative.

1.1. Problem Setting

Throguhout this paper, we will work with regression or signal denoising in the fixed lattice design setup where the underlying domain is a d dimensional grid $L_{d,n} = [n]^d =: \{1,\ldots,n\}^d$. Here n can be thought of as the sample size per dimension and the total sample size can be thought of as $N = n^d$. All of what we do here is meaningful in the regime where d is moderately low and fixed but n is large. The specific dimensions of interest are $d = 1, 2$ or 3 which are relevant for sequence, image or video denoising or forecasting respectively.

We will focus on the problem of noisy online signal prediction. Let K denote the d-dimensional grid $L_{d,n}$ and we abbreviate $|K| = N$. Suppose $\theta^* \in \mathbb{R}^K$ is the true underlying signal and

$$y = \theta^* + \sigma\epsilon \tag{1.1}$$

where $\sigma > 0$ is unknown and ϵ consists of independent, mean zero sub-Gaussian entries with unit dispersion factor.

Consider the following online prediction protocol. At round $t \in [1 : N]$,

- An adversary reveals a index $\rho(t) \in K$ such that $\rho(t) \notin \{\rho(1),\ldots,\rho(t-1)\}$.

- Learner predicts $\hat{\theta}_{\rho(t)}$.

- Adversary reveals $y_{\rho(t)} = \theta^*_{\rho(t)} + \epsilon_{\rho(t)}$, a noisy version of $\theta^*_{\rho(t)}$.

Note that ρ turns out to be a possibly adversarially chosen permutation of the entries of K (see the paragraph preceding Theorem 7). At the end of N rounds, the predictions of the learner are measured with the usual expected mean squared loss criterion given by

$$\mathrm{MSE}(\hat{\theta}, \theta^*) := \frac{1}{N} \mathbb{E}\|\hat{\theta} - \theta^*\|^2.$$

Remark 1 *Clearly, this online setting is a harder problem than batch learning where we get to observe the whole array y at once and we need to estimate θ^* by denoising y. Therefore, any online learning algorithm can also be used in the batch learning setup as well.*

1.2. A Definition of Spatial Adaptivity

In nonparametric function estimation, the notion of spatial/local adaptivity for an estimator is a highly desirable property. Intuitively, an estimator is spatially/locally adaptive if it adapts to a notion of complexity of the underlying true function locally on every part of the spatial domain on which the true function is defined. It is accepted that wavelet shrinkage based estimators, trend filtering estimators or optimal decision trees are spatially adaptive in some sense or the other. However, the meaning of spatial adaptivity varies quite a bit in the literature. It seems that there is no universally agreed upon definition of spatial adaptivity. In this section, we put forward one way to give a precise definition of spatial adaptivity which is inspired from the literature on *strongly adaptive online algorithms* (see, e.g., Daniely et al. (2015), Adamskiy et al. (2012), Hazan and Seshadhri (2007)) developed in the online learning community. One of the goals of this paper is to convince the reader that with a fairly simple analysis of the proposed online learning algorithm it is possible to establish this notion of spatial adaptivity (precisely defined below) in a very general setting.

Several batch learning estimators $\hat{\theta}$ satisfy the so called oracle risk bounds of the following form:

$$\mathrm{MSE}(\hat{\theta}, \theta^*) \leq \frac{1}{N} \inf_{\theta \in \mathbb{R}^N} \left(\|\theta - \theta^*\|^2 + \sigma^2 k_{\mathrm{comp}}(\theta) p(\log n) \right)$$

where k_{comp} denotes a complexity function defined on the vectors in $\mathbb{R}^m$ and $p(\log n)$ is some power of $\log n$. This is a risk bound which implies that the estimator $\hat{\theta}$ adapts to the complexity of the true signal $k_{\mathrm{comp}}(\theta)$.

For example, optimal decision trees such as the Dyadic CART and the ORT estimator satisfy an oracle risk bound (see Section 7 in Donoho (1997) and Theorem 2.1 in Chatterjee and Goswami (2021a)), so does the Trend Filtering estimator (Remark 3.1 in Guntuboyina et al. (2020)) and more classically, the wavelet shrinkage based estimators (see Section 1.5 in Donoho and Johnstone (1994)). Here, the complexity function k_{comp} is directly proportional to the number of constant/polynomial pieces for univariate functions. For multivariate functions, k_{comp} is still proportional to the number of constant/polynomial pieces; measured with respect to an appropriate class of rectangular partitions of the domain, see Chatterjee and Goswami (2021a).

However, from our point of view, this type of oracle risk bound, while being highly desirable and guaranteeing adaptivity against the complexity function k_{comp}, is still a global adaptivity bound as the bound is for the mean squared error of the whole signal θ^*. A good notion of local/spatial adaptivity should reveal the adaptivity of the estimator to the local complexity of the underlying signal. This naturally motivates us to make the following definition of spatial adaptivity.

We say that an estimator $\hat{\theta}$ is spatially adaptive with respect to the complexity parameter k_{comp} and with respect to a class $\mathcal{S}$ of subregions or subsets of the domain $L_{d,n}$ if the following risk bound holds simultaneously for every $S \in \mathcal{S}$,

$$\text{MSE}(\hat{\theta}_S, \theta_S^*) \leq \frac{1}{|S|} \inf_{\theta \in \mathbb{R}^S} \left(\|\theta - \theta_S^*\|^2 + \sigma^2 k_{\text{comp}}(\theta)\, p(\log n) \right).$$

The above definition of spatial adaptivity makes sense because if the above holds simultaneously for every $S \in \mathcal{S}$, then the estimator $\hat{\theta}$ estimates at θ^* locally on S with a rate of convergence that depends on the local complexity $k_{\text{comp}}(\theta_S^*)$. We will prove that our proposed online learning estimator is spatially adaptive in the sense described above with respect to a large class of subregions S.

1.3. Summary of Our Results

1. We formulate a slightly modified version of the so called sleeping experts aggregation algorithm for a general class of experts and a general class of comparator signals. We then state and prove a general simultaneous oracle risk bound for the proposed online prediction algorithm; see Theorem 7. This is the main result of this paper and is potentially applicable to several canonical estimation/prediction settings.

2. We specifically study an online mean aggregation algorithm as a special instance of our general algorithm and show that it satisfies our notion of spatial adaptivity (see Theorem 12) with respect to the complexity parameter that counts the size of the minimal rectangular partition of the domain $L_{d,n}$ on which the true signal θ^* is piecewise constant. Even in the easier offline setting, natural competitor estimators like Dyadic CART and ORT are not known to satisfy our notion of spatial adaptivity. Equipped with the spatially adaptive guarantee we proceed to demonstrate that this online mean aggregation algorithm also attains spatially adaptive minimax rate optimal bounds (see Theorem 13) for the bounded variation function class in general dimensions. This is achieved by combining Theorem 12 with known approximation theoretic results. Such a spatially adaptive guarantee as in Theorem 13 is not known to hold for the TV Denosing estimator, the canonical estimator used for estimating bounded variation functions.

3. We then study an online linear regression aggregation algorithm based on the Vovk-Azoury-Warmouth (VAW) forecaster (see Vovk (1998), Azoury and Warmuth (2001)) as another instantiation of our general algorithm. We show that this algorithm satisfies our notion of spatial adaptivity (see Theorem 15) with respect to the complexity parameter which counts the size of the minimal rectangular partition of the domain $L_{d,n}$ on which the true signal θ^* is piecewise polynomial of any given fixed degree ≥ 1. Like in the case with piecewise constant signals discussed above, natural competitor estimators such as Trend Filtering or higher order

versions of Dyadic CART are not known to satisfy our notion of spatial adaptivity even in the easier offline setting. We then demonstrate that this online linear regression aggregation algorithm also attains spatially adaptive minimax rate optimal bounds (see Theorem 18) for univariate higher order bounded variation functions. This is again achieved by combining Theorem 15 with known approximation theoretic results. Such a spatially adaptive guarantee as in Theorem 18 is not known to hold for the state of the art Trend Filtering estimator.

1.4. Closely Related Works

In a series of recent papers Baby and Wang (2019), Baby and Wang (2020), Baby et al. (2021b), Baby and Wang (2021), the authors there have studied online estimation of univariate bounded variation and piecewise polynomial signals. In particular, the paper Baby et al. (2021b) brought forward sleeping experts aggregation algorithms Chernov and Vovk (2009), Daniely et al. (2015) in the context of predicting univariate bounded variation functions. These works have been a source of inspiration for this current paper.

On the other hand, in another recent paper Chatterjee and Goswami (2021a), the authors studied the offline estimation of piecewise polynomial and bounded variation functions with a particular attention on obtaining adaptive oracle risk bounds. The estimators considered in that paper are optimal decision trees such as Dyadic CART (Donoho (1997)) and related variants. After coming across the paper Baby et al. (2021b) we realized that by using sleeping experts aggregation algorithms, one can obtain oracle risk bounds in the online setting which would then be applicable to *online* estimation of piecewise polynomial and bounded variation functions in general dimensions as considered in Chatterjee and Goswami (2021a).

The main point of difference of this work with the papers Baby and Wang (2019), Baby and Wang (2020), Baby et al. (2021b) is that here we formulate a general oracle risk bound that works simultaneously over a collection of subsets of the underlying domain (see Theorem 7). We then show that this result can be used in conjunction with some approximation theoretic results (proved in Chatterjee and Goswami (2021a)) to obtain spatially adaptive near optimal oracle risk bounds for piecewise constant/polynomial and bounded variation functions in general dimensions. To the best of our understanding, the papers Baby and Wang (2019), Baby and Wang (2020), Baby et al. (2021b) have not addressed function classes beyond the univariate case. Even in the univariate case, neither have they given a unified treatment for bounded variation functions of all orders nor have they addressed oracle risk bounds for piecewise constant/polynomial functions. But *most importantly*, it appears that our work is the first, in the online setting, to formulate a simultaneous oracle risk bound as in Theorem 7 and realize that one can deduce from this near optimal risk bounds for several function classes of recent interest. We hope that Theorem 7 will find applications for several other function classes (see Section 6.2 below).

1.5. Further Related Literature

Question $Q1$ posed in Section 1 above has been addressed in the online learning literature by Gerchinovitz (2011) in the following sense. Their regret bound (1) with "basis" functions ϕ_j's that correspond to indicator functions of dyadic rectangles implies a regret bound similar to that of Theorem 4.2. However, the obtained regret bound is not locally adaptive, and the underlying algorithm

appears to be computationally intensive relying on monte carlo approximations. More generally, our setup here falls under the umbrella of online (fixed design) nonparametric regression. The general results of Rakhlin and Sridharan (2014), Gaillard and Gerchinovitz (2015) , Cesa-Bianchi et al. (2017) give general non-constructive ways to attain minimax rate optimal regrets for general function classes. The algorithms we propose here are explicit and efficiently computable. The recent work of Kuzborskij and Cesa-Bianchi (2020) is related to this current article in the sense that they also study local adaptivity of online prediction. However, their notion of local adaptivity is different from what is the focus of the current article.

2. Aggregation of Experts Algorithm

In this section, we describe our main prediction algorithm. Our algorithm is a slight modification of the so called *Strongly Adaptive* online algorithms discussed in Hazan and Seshadhri (2007), Adamskiy et al. (2012), Daniely et al. (2015). We then provide a regret bound in terms of the squared error loss for this algorithm against the squared error loss incurred by any expert; see Proposition 4.

In this section K could be any general finite domain like $L_{d,n}$. An *expert* will stand for a set $S \subset K$ equipped with an online rule defined on S where, by an *online rule* $r^{(S)}$ corresponding to S, we mean a collection of (measurable) maps $r_{U,s}^{(S)} : \mathbb{R}^U \to \mathbb{R}$ indexed by $U \subset S$ and $s \in S \setminus U$. Operationally, the expert corresponding to a subset $S \in \mathcal{S}$ containing $\rho(t)$ predicts at the revealed point $\rho(t)$ the number given by

$$\hat{y}_{\rho(t)}^{(S)} = \hat{y}_{\rho[1:t],\rho(t)}^{(S)} = r_{\rho[1:(t-1)]\cap S,\,\rho(t)}^{(S)} \left(y_{\rho[1:(t-1)]\cap S} \right). \tag{2.1}$$

As is clear from the above expression, the prediction at the point $\rho(t) \in S$ actually depends on $\rho[1:t]$, i.e., all the revealed points upto time t. However, in order to keep our notations light, we will omit the reference to $\rho[1:t]$ in the sequel and denote the prediction at $s \in S$ from the expert S as $\hat{y}_s^{(S)}$ while reinvoking this dependence whenever they are relevant. The display (2.1) defines a vector $\hat{y}^{(S)} \in \mathbb{R}^S$ containing the predictions of the expert corresponding to the subset S. A *family* of experts corresponds to a sub-collection $\mathcal{S}$ of subsets of K. For any choice of online rules for every $S \in \mathcal{S}$, we refer to them collectively as an online rule $\mathbf{r}$ associated to $\mathcal{S}$.

As per the protocol described in Section 1, at the beginning of any round $t \in [N]$ where $|K| = N$, the data index $\rho(t)$ is revealed. At this point, the experts corresponding to subsets $S \in \mathcal{S}$ either not containing $\rho(t)$ or not having any data index revealed previously, become inactive. All other experts provide a prediction of their own.

We are now ready to describe our aggregation algorithm $\mathcal{A}$ for a set of experts $\mathcal{S}$. The input to this algorithm is the data vector $y \in \mathbb{R}^K$ which is revealed sequentially in the order given by the permutation ρ. We denote the output of this algorithm here by $\hat{y} \in \mathbb{R}^K$. At each round $t \in [N]$, the algorithm outputs a prediction $\hat{y}_{\rho(t)}$. Below and in the rest of the article, we use $T_a(x)$ to denote the truncation map $T_a(x) = \min\{\max\{x, -a\}, a\}$.

Aggregation algorithm. Parameters - subset of experts $\mathcal{S}$, online rule $\mathbf{r} = \{r^{(S)} : S \in \mathcal{S}\}$ and truncation parameter $\lambda > 1$

Initialize $w_{S,1} = \frac{1}{|\mathcal{S}|}$ for all $S \in \mathcal{S}$.

For $t = 1, \ldots, N$:

1. Adversary reveals $\rho(t)$.

2. Choose a set of *active experts* A_t as

$$\{S \in \mathcal{S} : \rho(t) \in S \text{ and } \rho(t') \in S \text{ for some } t' < t\}$$

   if $t > 1$ and $\{S \in \mathcal{S} : \rho(t) \in S\}$ if $t = 1$. If $A_t = \emptyset$, predict $\hat{y}_{\rho(t)} = 0$ and skip to round $t + 1$.

3. Predict $\hat{y}_{\rho(t)} = \sum_{S \in A_t} \hat{w}_{S,t} T_\lambda(\hat{y}_{\rho(t)}^{(S)})$ where $\hat{w}_{S,t} := \frac{w_{S,t}}{\sum_{S \in A_t} w_{S,t}}$.

4. Update $w_{S,t}$'s so that $w_{S,t+1} = w_{S,t}$ for $S \notin A_t$ and

$$w_{S,t+1} = \frac{w_{S,t}\, \mathrm{e}^{-\alpha\, \ell_{S,t}}}{\sum_{S \in A_t} \hat{w}_{S,t}\, \mathrm{e}^{-\alpha\, \ell_{S,t}}} = \frac{w_{S,t}\, \mathrm{e}^{-\alpha\, \ell_{S,t}}}{\sum_{S \in A_t} w_{S,t}\, \mathrm{e}^{-\alpha\, \ell_{S,t}}} \sum_{S \in A_t} w_{S,t}$$

   otherwise, where $\ell_{S,t} := \left(T_\lambda(y_{\rho(t)}) - T_\lambda(\hat{y}_{\rho(t)}^{(S)})\right)^2$ and $\alpha := \frac{1}{8\lambda^2}$.

Remark 2 *This algorithm is similar to the sleeping experts aggregation algorithm discussed in Daniely et al. (2015) except that we apply this algorithm after truncating the data by the map T_λ. Since, we are interested in (sub-Gaussian) unbounded errors, our data vector y need not be bounded which necessitates this modification. See Remark 6 below.*

Remark 3 *It is perhaps more natural to define the set of experts A_t as $\{S \in \mathcal{S} : \rho(t) \in S\}$ in step 2 above. Although this definition yields the same theoretical bounds given in this article, we found in our simulation experiments that "taking the opinion of too many inexperienced experts" actually brings down the performance in practice.*

In the sequel we will refer to our aggregation algorithm as $\mathcal{A}(\mathbf{r}, \mathcal{S}, \lambda)$. The following proposition guarantees that the performance of the above algorithm is not much worse as compared to the performance of any expert $S \in \mathcal{S}$; for any possible input data y.

Proposition 4 (error comparison against individual experts for arbitrary data) *For any ordering ρ of K and $S \in \mathcal{S}$, we have*

$$\sum_{s \in S} \left(y_s - \hat{y}_s\right)^2 \leq \sum_{s \in S} \left(y_s - \hat{y}_s^{(S)}\right)^2 + 8\lambda^2 \log \mathrm{e}|\mathcal{S}| + 2\|y_S - \Pi_\lambda y_S\|^2 + 4\lambda^2 \sum_{s \in S} \mathbf{1}\left(|y_s| > \lambda\right), \quad (2.2)$$

where $\Pi_\lambda z$, for any vector $z \in \mathbb{R}^A$ with $|A| < \infty$, denotes the ℓ^2-projection of z onto the ℓ^∞-ball of radius λ, i.e., $(\Pi_\lambda z)_a = T_\lambda(z_a)$ for all $a \in A$.

Remark 5 *A remarkable aspect of Proposition 4 is that it holds for any input data $y \in \mathbb{R}^{L_{d,n}}$. In particular, no probabilistic assumption is necessary. In the terminology of online learning, this is said to be a prediction bound for individual sequences. Usually, such an individual sequence prediction bound is stated for bounded data; see, e.g., Hazan and Seshadri (2007), Daniely et al. (2015). On the other hand, Proposition 4 holds for any data because we have introduced a truncation parameter in our aggregation algorithm.*

Remark 6 *The effect of truncation is clearly reflected in the last two terms of (2.2). The issue of unbounded data points in the **noisy** setup was dealt earlier in the literature — see, e.g., (Baby et al., 2021a, Theorem 5) — by choosing a value of λ so that **all** the datapoints lie within the interval $[-\lambda, \lambda]$ with some prescribed (high) probability $1 - \delta$. The comparison bounds analogous to (2.2) (without the last two terms) then hold on this high probability event. However, the issue of how to choose λ such that **all** the datapoints lie within the interval $[-\lambda, \lambda]$ is not trivial to resolve unless one knows something about the data generating mechanism. There is a simple way to get around this problem in the offline version by setting $\lambda = \max_{j \in [N]} |y_j|$ (see (Baby et al., 2021a, Remark 8)) which is obviously **not** possible in the online setting. The problem of tuning the parameter α sequentially was also addressed by Gaillard and Gerchinovitz (2015), in his setup of online linear regression, by taking λ as the maximum of $|y_1|, \ldots, |y_{t-1}|$ at any round t, with a regret analysis. It is not clear to us whether this time varying choice of λ can be made to amalgamate with our proof techniques. Our version of the algorithm and the accompanying Proposition 4 provide an explicit bound on the error due to truncation for arbitrary $y \in \mathbb{R}^K$ in the online problem. To the best of our knowledge, such a bound was not available in the literature in the current setup. An operational implication of Proposition 4 is that even if a few data points exceed λ in absolute value by not too great a margin, we still get effective risk bounds.*

Proof The proof is similar to the proof of regret bounds for exponentially weighted average forecasters with exp-concave loss functions (see, e.g., Hazan and Seshadri (2007), Cesa-Bianchi and Lugosi (2006)). However, we need to take some extra care in order to deal with our particular activation rule (see step 2) and obtain the error terms as in (2.2).

Let us begin with the observation that the function $e^{-\eta(x-z)^2}$, where $\eta > 0$, is concave in x for all $x, z \in [-1/\sqrt{8\eta}, 1/\sqrt{8\eta}]$. Clearly, this condition is satisfied for $\eta = \alpha$ and all $x, z \in [-\lambda, \lambda]$. Therefore, since $\hat{y}_{\rho(t)}$ is an average of $T_\lambda(\hat{y}_{\rho(t)}^{(S)})$'s (which by definition lie in $[-\lambda, \lambda]$) with respective weights $\hat{w}_{S,t}$ (see step 3 in the algorithm), we can write using the Jensen's inequality,

$$\exp\left(-\alpha \left(\hat{y}_{\rho(t)} - T_\lambda(y_{\rho(t)})\right)^2\right) \geq \sum_{S \in A_t} \hat{w}_{S,t} \, e^{-\alpha \ell_{S,t}}$$

where we recall that $\ell_{S,t} := \left(T_\lambda(y_{\rho(t)}) - T_\lambda(\hat{y}_{\rho(t)}^{(S)})\right)^2$.

Fix a subset $S \in \mathcal{S}$ such that $S \in A_t$. Taking logarithm on both sides and using the particular definition of updates in step 4 of $\mathcal{A}$, we get

$$\left(\hat{y}_{\rho(t)} - T_\lambda(y_{\rho(t)})\right)^2 - \ell_{S,t} \leq \alpha^{-1} \log\left(\frac{e^{-\alpha \ell_{S,t}}}{\sum_{S' \in \mathcal{S}} \hat{w}_{S',t} e^{-\alpha \ell_{S',t}}}\right) = \alpha^{-1} \log\left(\frac{w_{S,t+1}}{w_{S,t}}\right).$$

However, since $w_{S,t+1} = w_{S,t}$ and hence the logarithm is 0 whenever $S \notin A_t$ (see step 4 in the algorithm), we can add the previous bound over all t such that $S \in A_t$ to deduce:

$$\sum_{t:S\in A_t} \left(\hat{y}_{\rho(t)} - T_\lambda(y_{\rho(t)}) \right)^2 \leq \sum_{t:S\in A_t} \ell_{S,t} + \alpha^{-1} \sum_{t\in[N]} \log\left(\frac{w_{S,t+1}}{w_{S,t}} \right) \leq \sum_{t:S\in A_t} \ell_{S,t} + \alpha^{-1}\log|\mathcal{S}|$$

where in the final step we used the fact that $w_{S,1} = \frac{1}{|\mathcal{S}|}$ and $w_{S,N+1} \leq 1$. Now let $\rho(t) \in S$. It follows from our activation rule in step 2 that unless t is the *first* round such that $\rho(t) \in S$ (call it t_S), one has $S \in A_t$. Therefore, the previous display gives us,

$$\sum_{t:S\ni\rho(t)} \left(\hat{y}_{\rho(t)} - T_\lambda(y_{\rho(t)}) \right)^2 = \sum_{t:S\in A_t} \left(\hat{y}_{\rho(t)} - T_\lambda(y_{\rho(t)}) \right)^2 + \left(\hat{y}_{\rho(t_S)} - T_\lambda(y_{\rho(t_S)}) \right)^2$$

$$\leq \sum_{t:S\ni\rho(t)} \ell_{S,t} + 4\lambda^2 + \alpha^{-1}\log|\mathcal{S}| = \sum_{t:S\ni\rho(t)} \ell_{S,t} + 8\lambda^2\log \mathrm{e}|\mathcal{S}| \tag{2.3}$$

where in the second step we used the fact that both $\hat{y}_{\rho(t_S)}$ and $T_\lambda(y_{\rho(t_S)})$ lie in $[-\lambda, \lambda]$. Now we have to get rid of the truncation function $T_\lambda(\cdot)$ from both sides of this expression. To this end we will use the following elementary inequality,

$$(x - z)^2 \leq (T_\lambda(x) - z)^2 + 2(T_\lambda(x) - x)^2 + 4\lambda^2 \mathbb{1}(|x| > \lambda), \quad \text{when } |z| \leq \lambda \tag{2.4}$$

and also,

$$(T_\lambda(x) - T_\lambda(z))^2 \leq (x - z)^2. \tag{2.5}$$

We apply (2.4) to $x = y_{\rho(t)}$ and $z = \hat{y}_{\rho(t)} \in [-\lambda, \lambda]$ for each t satisfying $\rho(t) \in S$ and subsequently use (2.3) to deduce

$$\sum_{s\in S} \left(y_s - \hat{y}_s \right)^2 \leq \sum_{s\in S} (\hat{y}_s - T_\lambda(y_s))^2 + 8\lambda^2\log \mathrm{e}|\mathcal{S}| + 2\sum_{s\in S}(y_s - T_\lambda(y_s))^2 + 4\lambda^2 \sum_{s\in S} \mathbb{1}\,(|y_s| > \lambda)$$

$$\leq \sum_{s\in S} \left(T_\lambda(y_s) - T_\lambda(\hat{y}_s^{(S)}) \right)^2 + 8\lambda^2\log \mathrm{e}|\mathcal{S}| + 2\|y_S - \Pi_\lambda y_S\|^2 + 4\lambda^2 \sum_{s\in S} \mathbb{1}\,(|y_s| > \lambda).$$

We can now conclude the proof from the above display by plugging $x = y_s$ and $z = \hat{y}_s^{(S)}$ into (2.5) for each $s \in S$ ∎

3. A General Simultaneous Oracle Risk Bound

In this section we will state a general simultaneous oracle risk bound for online prediction of noisy signals. As in the last section, in this section also K refers to any arbitrary finite domain. Recall from our setting laid out in the introduction that we observe a data vector $y \in \mathbb{R}^K$ in some order where we can write

$$y = \theta^* + \sigma\epsilon$$

where $\sigma > 0$ is unknown and ϵ_t's are independent, mean zero sub-Gaussian variables with unit dispersion factor. For specificity, we assume in the rest of the paper that

$$\max(\mathbb{P}[\epsilon_t \geq x], \mathbb{P}[\epsilon_t \leq -x]) \leq 2\mathrm{e}^{-x^2/2}, \quad \text{for all } x \geq 0 \text{ and } t \in K.$$

Let us emphasize that the constant 2 is arbitrary and changing the constant would only impact the absolute constants in our main result, i.e., Theorem 7 below . Our focus here is on estimating signals θ^* that are piecewise regular on certain sets as we now explain. Let $\mathcal{S}$ be a family of subsets of K (cf. the family of experts $\mathcal{S}$ in our aggregation algorithm) and for each $S \in \mathcal{S}$, let $F_S \subset \mathbb{R}^S$ denote a class of functions defined on S.

We now define $\mathcal{P}$ to be the set of all partitions of K all of whose constituent sets are elements of $\mathcal{S}$. For any partition $P \in \mathcal{P}$, define the class of signals

$$\Theta_P = \Theta_P\big((F_S : S \in \mathcal{S})\big) = \{\theta \in \mathbb{R}^K : \theta_S \in F_S \ \forall S \in P\}. \tag{3.1}$$

In this section, when we mention a piecewise regular function, we mean a member of the set Θ_P for a partition $P \in \mathcal{P}$ with not too many constituent sets.

For example, in this paper we will be specifically analyzing the case when $K = L_{d,n} = [n]^d$ is the d dimensional lattice or grid, $\mathcal{S}$ is the set of all (dyadic) rectangles of $L_{d,n}$ and F_S is the set of polynomial functions of a given degree $m \geq 0$ on the rectangular domain S. Then, $\mathcal{P}$ becomes the set of all (dyadic) rectangular partitions of $L_{d,n}$ and Θ_P becomes the set of piecewise polynomial functions on the partition P.

Coming back to the general setting, to describe our main result, we need to define an additional quantity which is a property of the set of online rules $\{r^{(S)} : S \in \mathcal{S}\}$. For any partition $P \in \mathcal{P}$ and any $\theta \in \Theta_P$, let $\mathcal{R}(y, \theta, P) = \mathcal{R}(\mathbf{r}, y, \theta, P) > 0$ be defined as,

$$\mathcal{R}(y, \theta, P) = \sup_{\rho} \sum_{S \in P} \Big(\sum_{t:\rho(t)\in S} (y_{\rho(t)} - \hat{y}^{(S)}_{\rho(t)})^2 - \|y_S - \theta_S\|^2 \Big) \tag{3.2}$$

(recall the definition of $\hat{y}^{(S)}_{\rho(t)}$ from (2.1)). Clearly $\mathcal{R}(y, \theta, P)$ is a random variable and we denote by

$$\overline{\mathcal{R}}(\theta, P) = \mathbb{E}\, \mathcal{R}(y, \theta, P), \tag{3.3}$$

its expected value.

Here is how we can interpret $\mathcal{R}(y, \theta, P)$. Given a partition $P \in \mathcal{P}$ consider the following prediction rule r_P. For concreteness, let the partition $P = (S_1, \ldots, S_k)$. At round t, there will be only one index $i \in [k]$ such that $\rho(t)$ is in S_i. Then the prediction rule r_P predicts a value $\hat{y}^{(S_i)}_{\rho(t)}$. In other words, r_P uses the prediction of the expert corresponding to the subset S_i in this round. Also, consider the prediction rule r_θ which at round t predicts by $\theta_{\rho(t)}$ for any fixed vector $\theta \in \Theta_P$. If we have the extra knowledge that the true signal θ^* indeed lies in Θ_P then it may be natural to use the above online learning rule r_P if the experts are good at predicting signals (locally on the domain S) which lie in F_S. *We can now interpret $\mathcal{R}(y, \theta, P)$ as the excess squared loss or regret (when the array revealed sequentially is y) of the online rule r_P as compared to the online rule r_θ.*

In the sequel we use $\|\theta\|_\infty$ to denote the ℓ_∞-norm of the vector θ. We also extend the definitions of $\mathcal{P}$ and Θ_P (see around (3.4)) to include partitions of *subsets* of K comprising only sets from $\mathcal{S}$. In particular, for any $\mathsf{T} \subset K$, define $\mathcal{P}_\mathsf{T}$ to be the set of all partitions P of T all of whose constituent sets are elements of $\mathcal{S}$. For any partition $P \in \mathcal{P}_\mathsf{T}$, define the class of signals

$$\Theta_P = \Theta_P\big((F_S : S \in \mathcal{S}, S \subset \mathsf{T})\big) = \{\theta \in \mathbb{R}^\mathsf{T} : \theta_S \in F_S \ \forall S \in P\}. \tag{3.4}$$

Notice that the dependence Θ_P on the subset T is *implicit* in the choice of $P \in \mathcal{P}_\mathsf{T}$.

Let us now say a few words about the choice of ordering ρ which we can generally think of as a stochastic process taking values in K. We call ρ as *non-anticipating* if, conditionally on $(\rho[1 : t], y_{\rho[1:t-1]})$, $\epsilon_{\rho(t)}$ is distributed as ϵ_s on the event $\{\rho(t) = s\}$ for any $t \in [N]$. Such orderings include deterministic orderings and orderings that are independent of the data. But more generally, any ordering where $\rho(t)$ is allowed to depend on the data only through $y_{\rho[1:t-1]}$ is non-anticipating. In particular, an adversary can choose to reveal the next index after observing all of the past data and our actions.

We are now ready to state our general result. In order to emphasize our primary interest in predicting the vector θ^*, below we use $\hat{\theta}$ *intechangeably* with the output $\hat{y}$ of the algorithm $\mathcal{A}(\mathbf{r}, \mathcal{S}, \lambda)$ applied to the data y for some suitable ordering ρ.

Theorem 7 (General Simultaneous Oracle risk bound) *Let K be a finite set. Fix a set of experts $\mathcal{S}$ equipped with online learning rule $\mathbf{r}$. For each $S \in \mathcal{S}$, fix $F_S \subset \mathbb{R}^S$ to be a class of functions defined on S. Suppose y is generated from the model* (1.1) *and is input to the algorithm $\mathcal{A}(\mathbf{r}, \mathcal{S}, \lambda)$. Let T be any subset of K. There exist absolute constants $c \in (0,1)$ and $C > 1$ such that for any non-anticipating ordering ρ of K,*

$$\mathbb{E}\|\hat{\theta}_{\mathsf{T}} - \theta^*_{\mathsf{T}}\|^2 \leq \inf_{\substack{P \in \mathcal{P}_{\mathsf{T}}, \\ \theta \in \Theta_P}} \left(\|\theta^*_{\mathsf{T}} - \theta\|^2 + C\lambda^2 |P| \log e|\mathcal{S}| + \overline{\mathcal{R}}(\theta, P)\right) + C\lambda^2 |\{s \in \mathsf{T} : |\theta_s| > \lambda\}|$$

$$+ C\|\theta^*_{\mathsf{T}} - \Pi_\lambda \theta^*_{\mathsf{T}}\|^2 + C(\sigma^2 + \lambda^2) \sum_{s \in \mathsf{T}} e^{-c \min\left(\frac{|\theta^*_s - \lambda|^2}{\sigma^2}, \frac{|\theta^*_s + \lambda|^2}{\sigma^2}\right)}.$$

(3.5)

In particular, there exists an absolute constant $C > 1$ such that for $\lambda \geq C(\sigma\sqrt{\log|\mathsf{T}|} \vee \|\theta^*\|_\infty)$, one has

$$\mathrm{MSE}(\hat{\theta}_{\mathsf{T}}, \theta^*_{\mathsf{T}}) \leq \inf_{P \in \mathcal{P}_{\mathsf{T}}, \theta \in \Theta_P} \frac{1}{|\mathsf{T}|}\left(\|\theta^*_{\mathsf{T}} - \theta\|^2 + C\lambda^2 |P| \log e|\mathcal{S}| + \overline{\mathcal{R}}(\theta, P)\right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2}.$$

(3.6)

Remark 8 *Our truncation threshold $C(\sigma\sqrt{\log|\mathsf{T}|} \vee \|\theta^*\|_\infty)$ is comparable to the threshold given in (Baby et al., 2021a, Theorem 5) for Gaussian errors.*

Remark 9 *Notice that $\mathcal{P}_{\mathsf{T}}$ can be empty if T does not admit of any partition with elements only in $\mathcal{S}$. In this case, the bounds* (3.5) *and* (3.6) *continue to hold by adopting the standard convention that infimum over an empty set is ∞.*

We now explain various features and aspects of the above theorem.

- The reader can read the bound in (3.6) as

$$\mathrm{MSE}(\hat{\theta}_{\mathsf{T}}, \theta^*_{\mathsf{T}}) \leq \inf_{\substack{P \in \mathcal{P}_{\mathsf{T}}, \\ \theta \in \Theta_P}} \frac{1}{|\mathsf{T}|}\left(\underbrace{\|\theta^*_{\mathsf{T}} - \theta\|^2}_{T_1} + \underbrace{C\lambda^2 |P| \log e|\mathcal{S}|}_{T_2} + \underbrace{\overline{\mathcal{R}}(\theta, P)}_{T_3}\right) + \textit{lower order term.}$$

Indeed, the only importance of the factor $\frac{1}{|\mathsf{T}|^2}$ in the last term of (3.6) is that it is $o(\frac{1}{|\mathsf{T}|})$, i.e., of lower order than the principal term. Indeed, by suitably increasing the constant C, we can get any given power of $|\mathsf{T}|$ in the denominator.

- To understand and interpret the above bound, it helps to first consider $\mathsf{T} = K$ and then fix a partition $P \in \mathcal{P}_K = \mathcal{P}$ and a piecewise regular signal $\theta \in \Theta_P$. We also keep in mind two prediction rules, the first one is the online rule r_P and the second one is r_θ (both described before the statement of Theorem 7). The bound inside the infimum is a sum of three terms, T_1, T_2 and T_3 as in the last display.

  1. The first term T_1 is simply the squared distance between θ and θ^*. This term is obviously small or big depending on whether θ is close or far from θ^*.

  2. The second term captures the complexity of the partition P where the complexity is simply the cardinality or the number of constituent sets/experts $|P|$ multiplied by log cardinality of the total number of experts $\log \mathrm{e}|\mathcal{S}|$. The reader can think of this term as the ideal risk bound achievable and anything better is not possible when the true signal θ^* is piecewise regular on P. This term is small or big depending on whether $|P|$ is small or big.

  3. The third term T_3 is $\overline{\mathcal{R}}(\theta, P)$ which can be interpreted as the expected excess squared loss or regret of the online rule r_P as compared to the prediction rule r_θ. This term is small or big depending on how good or bad is the online rule r_P compared to the prediction rule r_θ.

- Our bound is an infimum over the sum of three terms T_1, T_2, T_3 for any $P \in \mathcal{P}$ and $\theta \in \Theta_P$ which is why we can think of this bound as an oracle risk bound in the following sense. Consider the case when $\mathsf{T} = K$ and θ^* lies in Θ_{P^*} for some P^* which is of course unknown. In this case, an oracle who knows the true partition P^* might naturally trust experts locally and use the online prediction rule r_{P^*}. In this case, our bound reduces (by setting $P = P^*, \theta = \theta^*$) to the MSE incurred by this oracle prediction rule plus the ideal risk $|P^*| \log \mathrm{e}|\mathcal{S}|$ term which is unavoidable. Because of the term T_1, this argument holds even if θ^* does not exactly lie in Θ_{P^*} but is very close to it. To summarize, our MSE bound ensures that we nearly perform as well as an oracle prediction rule which knows the *true* partition corresponding to the target signal θ^*.

- The term T_3 in the MSE bound in (3.6) behooves us to find experts with good online prediction rules. If each expert $S \in \mathcal{S}$ indeed is equipped with a good prediction rule, then under the assumption that θ^* is exactly (or is close to) piecewise regular on a partition $P^* \in \mathcal{P}$, the term $T_3 = \overline{\mathcal{R}}(\theta^*, P^*)$ will be small and our bound will thus be better. This is what we do in our example applications, where we use provably good online rules such as running mean or the online linear regression forecaster of Vovk Vovk (1998). Infact, in each of the examples that we discuss subsequently in this paper, we bound the term $\overline{\mathcal{R}}(\theta, P)$ in two stages. First, we write

$$\overline{\mathcal{R}}(\theta, P) = \mathbb{E}\,\mathcal{R}(y, \theta, P) \le |P|\,\mathbb{E} \sup_{\rho, S \in P} \left(\sum_{t : \rho(t) \in S} (y_{\rho(t)} - \hat{y}_{\rho(t)}^{(S)})^2 - \|y_S - \theta_S\|^2 \right).$$

Then in the second stage we obtain a bound on the expectation in the right side above by a log factor. Thus there is no real harm if the reader thinks of $\overline{\mathcal{R}}(\theta, P)$ as also being of the order $|P|$ up to log factors which is the ideal and unavoidable risk as mentioned before.

- A remarkable feature of Theorem 7 is that the MSE bound in (3.6) holds *simultaneously* for all sets $\mathsf{T} \subset K$. Therefore, the interpretation that our prediction rule performs nearly as well as an oracle prediction rule holds locally for every subset or subregion T of the domain K. This fact makes our algorithm provably *spatially adaptive* to the class of all subsets of K with respect to the complexity parameter proportional to $|P|$ in the sense described in Section 1.2. The implications of this will be further discussed when we analyze online prediction of specific function classes in the next two sections.

Proof [Proof of Theorem 7] For convenience of referencing, we will use the notation $\hat{y}$ instead of $\hat{\theta}$. Since $y = \theta^* + \sigma\epsilon$, we can write for any $S \in \mathcal{S}$,

$$\|y_S - \hat{y}_S\|^2 = \|\hat{y}_S - \theta_S^*\|^2 + 2\sigma\langle\epsilon_S, \hat{y}_S\rangle + \sigma^2\|\epsilon_S\|^2.$$

However, since ρ is non-anticipating and $\hat{y}_{\rho(t)}$ is measurable relative to $(\rho[1:t], y_{\rho[1:(t-1)]})$ and ϵ_s's have mean zero, it follows from the previous display that

$$\mathbb{E}\|y_S - \hat{y}_S\|^2 = \mathbb{E}\|\hat{y}_S - \theta_S^*\|^2 + \sigma^2\mathbb{E}\|\epsilon_S\|^2. \tag{3.7}$$

On the other hand, adding up the upper bound on $\|y_S - \hat{y}_S\|^2$ given by Proposition 4 over all $S \in P$ for some $P \in \mathcal{P}$ we get

$$\|y_\mathsf{T} - \hat{y}_\mathsf{T}\|^2 \leq \sum_{S \in P}\sum_{t:\rho(t)\in S}\left(y_{\rho(t)} - \hat{y}_{\rho(t)}^{(S)}\right)^2 + 2\|y_\mathsf{T} - \Pi_\lambda y_\mathsf{T}\|^2 + 8\lambda^2|P|\log\mathrm{e}|\mathcal{S}|$$
$$+ 4\lambda^2\sum_{s\in\mathsf{T}}\mathbb{1}\{|y_s| > \lambda\}.$$

Now taking expectations on both sides and using the definition of $\overline{\mathcal{R}}(\theta, P)$ from (3.3), we can write

$$\mathbb{E}\|y_\mathsf{T} - \hat{y}_\mathsf{T}\|^2 \leq \mathbb{E}\|y_\mathsf{T} - \theta\|^2 + \overline{\mathcal{R}}(\theta, P) + 8\lambda^2|P|\log\mathrm{e}|\mathcal{S}| + 2\mathbb{E}\|y_\mathsf{T} - \Pi_\lambda y_\mathsf{T}\|^2$$
$$+ 4\lambda^2\sum_{s\in\mathsf{T}}\mathbb{P}(|y_s| > \lambda). \tag{3.8}$$

Since ϵ_s's have mean zero, we get by expanding $\|y_\mathsf{T} - \theta\|^2$,

$$\mathbb{E}\|y_\mathsf{T} - \theta\|^2 \leq \|\theta_\mathsf{T}^* - \theta\|^2 + \sigma^2\mathbb{E}\|\epsilon_\mathsf{T}\|^2.$$

Plugging this bound into the right hand side of (3.8), we obtain

$$\mathbb{E}\|y_\mathsf{T} - \hat{y}_\mathsf{T}\|^2 \leq \|\theta_\mathsf{T}^* - \theta\|^2 + \sigma^2\mathbb{E}\|\epsilon_\mathsf{T}\|^2 + \overline{\mathcal{R}}(\theta, P) + 8\lambda^2|P|\log\mathrm{e}|\mathcal{S}| + 2\mathbb{E}\|y_\mathsf{T} - \Pi_\lambda y_\mathsf{T}\|^2$$
$$+ 4\lambda^2\sum_{s\in\mathsf{T}}\mathbb{P}(|y_s| > \lambda).$$

Together with (3.7), this gives us

$$\mathbb{E}\|y_\mathsf{T} - \theta_\mathsf{T}^*\|^2 \leq \|\theta_\mathsf{T}^* - \theta\|^2 + \overline{\mathcal{R}}(\theta, P) + 8\lambda^2|P|\log\mathrm{e}|\mathcal{S}| + 2\mathbb{E}\|y_\mathsf{T} - \Pi_\lambda y_\mathsf{T}\|^2$$
$$+ 4\lambda^2\sum_{s\in\mathsf{T}}\mathbb{P}(|y_s| > \lambda).$$

Minimizing the right hand side in the above display over all $P \in \mathcal{P}$ and $\theta \in \Theta_P$, we get

$$\mathbb{E}\|y_{\mathsf{T}} - \theta_{\mathsf{T}}^*\|^2 \leq \inf_{P \in \mathcal{P}, \, \theta \in \Theta_P} \left(\|\theta_{\mathsf{T}}^* - \theta\|^2 + \overline{\mathcal{R}}(\theta, P) + 8\lambda^2 \, |P| \log |\mathcal{S}|\right) + 2\mathbb{E}\|y_{\mathsf{T}} - \Pi_\lambda y_{\mathsf{T}}\|^2$$
$$+ 4\lambda^2 \sum_{s \in \mathsf{T}} \mathbb{P}(|y_s| > \lambda). \tag{3.9}$$

It only remains to verify the bounds on the error terms due to truncation. Since $|\hat{y}_s| \leq \lambda$ by the design of our algorithm, we have

$$(\hat{y}_s - \theta_s^*)^2 \leq 2(\theta_s^* - T_\lambda(\theta_s^*))^2 + 8\lambda^2 \tag{3.10}$$

for all $s \in \mathsf{T}$. We will apply this naive bound whenever $|\theta_s^*| > \lambda$. So let us assume that s is such that $|\theta_s^*| \leq \lambda$. Let us start by writing

$$\mathbb{E}(y_s - T_\lambda(y_s))^2 = I_+ + I_-,$$

where, with $x_+ := \max(x, 0)$ and $x_- := -\min(x, 0)$,

$$I_+ := \mathbb{E}(y_s - \lambda)_+^2 = 2\sigma^2 \int_{x > \frac{\lambda}{\sigma}} \left(x - \frac{\lambda}{\sigma}\right) \mathbb{P}\left[\epsilon_s > x - \frac{\theta_s^*}{\sigma}\right] dx, \text{ and}$$

$$I_- := \mathbb{E}(y_s + \lambda)_-^2 = -2\sigma^2 \int_{x < -\frac{\lambda}{\sigma}} \left(x + \frac{\lambda}{\sigma}\right) \mathbb{P}\left[\epsilon_s < x - \frac{\theta_s^*}{\sigma}\right] dx.$$

In writing these expressions we used the standard fact that

$$\mathbb{E}(X - a)_+^2 = 2 \int_{x > a} (x - a)\mathbb{P}[X > x] \, dx.$$

We deal with I_+ first. Since $\theta_s^* \leq \lambda$ and ϵ_s has sub-Gaussian decay around 0 with unit dispersion factor, we can bound I_+ as follows:

$$I_+ = 2\sigma^2 \int_{x > \frac{\lambda - \theta_s^*}{\sigma}} \left(x - \frac{\lambda - \theta_s^*}{\sigma}\right) \mathbb{P}\left[\epsilon_s > x\right] dx \leq C\sigma^2 \, e^{-\frac{|\theta_s^* - \lambda|^2}{2\sigma^2}} \tag{3.11}$$

where C is an absolute constant. Similarly we can deduce

$$I_- \leq C\sigma^2 \, e^{-\frac{|\theta_s^* - \lambda|^2}{2\sigma^2}}. \tag{3.12}$$

On the other hand, for any $|\theta_s^*| \leq \lambda$, we have

$$\mathbb{P}[|y_s| > \lambda] \leq \mathbb{P}\left[\epsilon_s > \frac{\lambda - \theta_s^*}{\sigma}\right] + \mathbb{P}\left[\epsilon_s < \frac{-\lambda - \theta_s^*}{\sigma}\right] \leq 2e^{-\frac{|\theta_s^* - \lambda|^2}{2\sigma^2} \wedge \frac{|\theta_s^* + \lambda|^2}{\sigma^2}}. \tag{3.13}$$

Finally we plug the estimates (3.11), (3.12) and (3.13) into (3.9) when $|\theta_s^*| \leq \lambda$, and the estimate (3.10) and the trivial upper bound on probabilities when $|\theta_s^*| > \lambda$ to get (3.5). (3.6) then follows immediately from (3.5) by choosing C large enough. ∎

In the next few sections we will introduce and discuss online prediction rules for several classes of functions. In each case we will apply Theorem 7 to derive the corresponding risk bounds. The proofs of all the results are given in the Appendix section in the supplementary material.

4. Online Mean Aggregation over Dyadic Rectangles (OMADRE)

In this section, we will specifically study a particular instantiation of our general algorithm (laid out in Section 2) which we tentatively call as the Online Mean Aggregation over Dyadic Rectangles estimator/predictor (OMADRE). Here, $K = L_{d,n}$ and the set of experts corresponds to the set of all *dyadic rectangles* of $L_{d,n}$. Some precise definitions are given below.

An *axis aligned rectangle* or simply a rectangle R is a subset of $L_{d,n}$ which is a product of intervals, i.e., $R = \prod_{i=1}^{d}[a_i, b_i]$ for some $1 \le a_i \le b_i \le n$; $i \in [d]$. A sub-interval of $[1, n]$ is called *dyadic* if it is of the form $((a-1)2^s, a2^s]$ for some integers $0 \le s \le k$ and $1 \le a \le 2^{k-s}$ where we assume $n = 2^k$ for simplicity of exposition. *We call a rectangle dyadic if it is a product of dyadic intervals.*

Now we take our experts to be the dyadic sub-rectangles of $L_{d,n}$, i.e., in the terminology of Section 3, $\mathcal{S}$ is the set of dyadic sub-rectangles of $L_{d,n}$. We set $F_S = \text{span}(\{1\})$ —— the space of all constant functions on S — for all $S \in \mathcal{S}$. We also let $\mathcal{P}_{dp}$ be the set of all *dyadic rectangular partitions* of $L_{d,n}$ where a (dyadic) rectangular partition P is a partition of $L_{d,n}$ comprising only (respectively, dyadic) rectangles. Since there are at most $2n$ dyadic sub-intervals of $[n]$, we note that

$$|\mathcal{S}| = (2n)^d = 2^d N. \tag{4.1}$$

Under this setting, for any partition P of $L_{d,n}$ the set Θ_P refers to the set of all arrays $\theta \in \mathbb{R}^{L_{d,n}}$ such that θ is constant on each constituent set of P.

Finally we come to the choice of our online rule $\mathbf{r}$. It is very natural to consider the online averaging rule $\mathbf{r}$ defined as:

$$r_{U,s}^{(S)}((y_u : u \in U)) = \overline{y}_U \text{ for all } U \subset S \in \mathcal{S} \text{ and } s \in S \setminus U. \tag{4.2}$$

By convention, we set $\overline{y}_\emptyset = 0$.

Lemma 10 (Computational complexity of OMADRE**)** *There exists an absolute constant $C > 0$ such that the computational complexity, i.e., the number of elementary operations involved in the computation of OMADRE is bounded by $CN(\log_2 2n)^d$.*

Remark 11 *The above computational complexity is near linear in the sample size N but exponential in the dimension d. Therefore, the estimators we are considering here are very efficiently computable in low dimensions which are the main cases of interest here.*

The OMADRE, being an instance of our general algorithm will satisfy our simultaneous oracle risk bound in Theorem 7. This oracle risk bound can then be used to derive risk bounds for several function classes of interest. We now discuss two function classes of interest for which the OMADRE performs near optimally.

4.1. Result for Rectangular Piecewise Constant Functions in General Dimensions

Suppose θ^* is piecewise constant on some unknown rectangular partition P^* of the domain $K = L_{d,n}$. For concreteness, let the partition $P^* = (R_1, \ldots, R_k)$. An *oracle predictor* $\hat{\theta}_{(\text{oracle})}$ — which *knows* the minimal rectangular partition $(R_1, \ldots, R_k)$ of θ^* *exactly* —can simply use the online averaging prediction rule given in (4.2) separately within each of the rectangles$(R_1, \ldots, R_k)$. By a

15

basic result about online mean prediction, (see Lemma 21), it can be shown that the MSE of this oracle predictor is bounded by $O(\frac{k\|\theta^*\|_\infty \log n}{N})$. In words, the MSE of this oracle predictor scales (up to a log factor which is necessary) like the *number of constant pieces of θ^* divided by the sample size N which is precisely the parametric rate of convergence.*

A natural question is whether there exists an online prediction rule which a) adaptively achieves a MSE bound similar to the oracle prediction rule b) is computationally efficient. In the batch set up, this question is classical (especially in the univariate setting when $d = 1$) and has recently been studied thoroughly in general dimensions in Chatterjee and Goswami (2021a). It has been shown there that the Dyadic CART estimator achieves this near (up to log factors) oracle performance when $d \leq 2$ and a more computationally intensive version called the Optimal Regression Tree estimator (ORT) can achieve this near oracle performance in all dimensions under some assumptions on the true underlying partition. However, we are not aware of this question being explicitly answered in the online setting. We now state a theorem saying that the OMADRE essentially attains this objective. Below we denote the set of all partitions of $L_{d,n}$ into rectangles by $\mathcal{P}_{\text{all}}$. Note that the set $\mathcal{P}_{\text{dp}}$ is strictly contained in the set $\mathcal{P}_{\text{all}}$.

Theorem 12 (Oracle Inequality for Arbitrary Rectangular Partitions) *Let T be any subset of K and $\hat{\theta}^{OM}$ denote the OMADRE predictor. There exists an absolute constant C such that for any $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta^*\|_\infty)$, one has for any non-anticipating ordering ρ of $L_{d,n}$,*

$$\mathbb{E}\frac{1}{|\mathsf{T}|}\|\hat{\theta}_{\mathsf{T}}^{OM} - \theta_{\mathsf{T}}^*\|^2 \leq \inf_{\substack{P \in \mathcal{P}_{\text{all},\mathsf{T}} \\ \theta \in \Theta_P \subset \mathbb{R}^{\mathsf{T}}}} \left(\frac{1}{|\mathsf{T}|}\|\theta_{\mathsf{T}}^* - \theta\|^2 + C\lambda^2 \frac{|P|}{|\mathsf{T}|}(\log en)^d \log 2^d N\right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2}. \quad (4.3)$$

We now discuss some noteworthy aspects of the above theorem.

1. It is worth emphasizing that the above oracle inequality holds over all subsets T of K simultaneously. Therefore, the OMADRE is a *spatially adaptive* estimator in the sense of Section 1.2. Such a guarantee is not available for any existing estimator, even in the batch learning setup. For example, in the batch learning setup, all available oracle risk bounds for estimators such as Dyadic CART and related variants are known only for the full sum of squared errors over the entire domain.

2. To the best of our knowledge, the above guarantee is the first of its kind explicitly stated in the online learning setup. Therefore, the above theorem shows it is possible to attain a near (up to log factors) oracle performance by a near linear time computable estimator in the online learning set up as well; thereby answering our first main question laid out in Section 1.

3. We also reiterate that the infimum in the R.H.S in (4.3) is over the space of *all* rectangular partitions $\mathcal{P}_{\text{all}}$. This means that if the true signal θ^* is piecewise constant on an *arbitrary* rectangular partition with k rectangles, the OMADRE attains the desired $\tilde{O}(k/N)$ rate. Even in the batch learning set up, it is not known how to attain this rate in full generality. For example, it has been shown in Chatterjee and Goswami (2021a) that the Dyadic CART (or ORT) estimator enjoys a similar bound where the infimum is over the space of all recursive dyadic rectangular partitions (respectively decision trees) of K which is a *stirct* subset of $\mathcal{P}_{\text{all}}$. Thus, the bound presented here is stronger in this sensethan both these bounds known for Dyadic CART/ORT. More details about comparisons with Dyadic CART and ORT is given in Section 6.1.

4.2. Result for Functions with Bounded Total Variation in General Dimensions

Consider the function class whose total variation (defined below) is bounded by some number. This is a classical function class of interest in offline nonparametric regression since it contains functions which demonstrate spatially heterogenous smoothness; see Section 6.2 in Tibshirani (2015) and references therein. In the offline setting, the most natural estimator for this class of functions is what is called the Total Variation Denoising (TVD) estimator. The two dimensional version of this estimator is also very popularly used for image denoising; see Rudin et al. (1992). It is known that a well tuned TVD estimator is minimax rate optimal for this class in all dimensions; see Hütter and Rigollet (2016) and Sadhanala et al. (2016).

In the online setting, to the best of our knowledge, the paper Baby and Wang (2019) gave the first online algorithm attaining the minimax optimal rate. This algorithm is based on wavelet shrinkage. Recently, the paper Baby et al. (2021b) studied a version of the OMADRE in the context of online estimation of univariate bounded variation functions. In this section we state a result showing that with our definition of the OMADRE, it is possible to predict/forecast bounded variation functions online in general dimensions at nearly the same rate as is known for the batch set up.

We can think of $K = L_{d,n}$ as the d dimensional regular lattice graph. Then, thinking of $\theta \in \mathbb{R}^{L_{d,n}}$ as a function on $L_{d,n}$ we define

$$\mathrm{TV}(\theta) = \sum_{(u,v) \in E_{d,n}} |\theta_u - \theta_v| \tag{4.4}$$

where $E_{d,n}$ is the edge set of the graph $L_{d,n}$. The above definition can be motivated via the analogy with the continuum case. If we think $\theta[i_1, \ldots, i_n] = f(\frac{i_1}{n}, \ldots, \frac{i_d}{n})$ for a differentiable function $f : [0,1]^d \to \mathbb{R}$, then the above definition divided by n^{d-1} is precisely the Reimann approximation for $\int_{[0,1]^d} \|\nabla f\|_1$. In the sequel we denote,

$$\mathcal{BV}_{d,n}(V^*) := \{\theta \in \mathbb{R}^n : \mathrm{TV}(\theta) \leq V^*\}.$$

We are now ready to state:

Theorem 13 (Prediction error for $\mathcal{BV}_{d,n}(V^*)$ with online averages) *Fix any $\mathsf{T} \subset K$ that is a dyadic square and denote $V_{\mathsf{T}}^* = TV(\theta_{\mathsf{T}}^*)$. If $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta^*\|_\infty)$ as in Theorem 7, we have for some absolute constant $C > 1$ and any non-anticipating ordering ρ of $L_{d,n}$,*

$$\mathbb{E}\frac{1}{|\mathsf{T}|}\|\hat{\theta}_{\mathsf{T}}^{OM} - \theta_{\mathsf{T}}^*\|^2 \leq \frac{C}{|\mathsf{T}|}\big(\lambda^2(\log 2^d N)^2 + \lambda V_{\mathsf{T}}^*(\log 2^d N)^{3/2}\big) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2}. \tag{4.5}$$

when $d > 1$. On the other hand, for $d = 1$ we have

$$\mathbb{E}\frac{1}{|\mathsf{T}|}\|\hat{\theta}_{\mathsf{T}}^{OM} - \theta_{\mathsf{T}}^*\|^2 \leq C\,\lambda^{4/3}(\log 2^d N)^{4/3}\left(\frac{V_{\mathsf{T}}^*}{|\mathsf{T}|}\right)^{2/3} + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2}. \tag{4.6}$$

Here are some noteworthy aspects of the above theorem.

1. The above theorem ensure that the OMADRE matches the known minimax rate of estimating bounded variation functions in any dimension. To the best of our knowledge, this result is new in the in the online setting for the multivariate (i.e., $d \geq 2$) case.

2. Note that our MSE bounds hold simultaneously for all dyadic square regions. Thus, the OMADRE adapts to the unknown variation of the signal V_T^*, for any local dyadic square region T. In this sense, the OMADRE is spatially adaptive. Even in the batch setting, this type of simultaneous guarantee over a class of subsets of $L_{d,n}$ is not available for the canonical batch TVD estimator.

3. We require T to be a dyadic square because of a particular step in our proof where we approximate a bounded variation array with an array that is piecewise constant over a recursive dyadic partition of $L_{d,n}$ with pieces that have bounded aspect ratio. See Proposition 23 in the appendix.

5. Online Linear Regression Aggregation over Dyadic Rectangles (OLRADRE)

In this section, we consider another instantiation of our general prediction algorithm which is based on the Vovk, Azoury and Warmuth online linear regression forecaster, e.g see Vovk (1998). Similar to Section 4, we take our set of experts $\mathcal{S}$ to be the set of all dyadic sub-rectangles of $L_{d,n}$. However, the main difference is that we now take F_S to be the subspace spanned by a finite set $\mathcal{F}$ of basis functions on $\mathbb{R}^d$ restricted to S. In the next two subsections, we will focus specifically on the case when $\mathcal{F}$ is the set of all monomials in d variables with a maximum degree (see (5.2) below).

Next we need to choose an online rule $\mathbf{r}$ to which end the VAW linear regression forecaster leads to an online rule defined as:

$$r_{U,s}^{(S)}((y_u : u \in U)) = \hat{\beta}_s \cdot x_s \text{ with } \hat{\beta}_s := \Big(I + \sum_{u \in U \cup \{s\}} x_u x_u^T\Big)^{-1} \Big(\sum_{u \in U} y_u x_u\Big) \qquad (5.1)$$

for all $U \subset S \in \mathcal{S}$ and $s \in S \setminus U$ where x_u is the vector $(f(u) : f \in \mathcal{F}) \in \mathbb{R}^{\mathcal{F}}$ and $\cdot$ denotes the canonical inner product in $\mathbb{R}^{\mathcal{F}}$. By convention, we interpret an empty summation as 0.

The next lemma gives the computational complexity of the OLRADRE which is the same as that of the OMADRE except that it scales cubically with the cardinality of the basis function class $\mathcal{F}$.

Lemma 14 (Computational complexity of OLRADRE) *There exists an absolute constant $C > 0$ such that the computational complexity of OLRADRE is bounded by $C|\mathcal{F}|^3 N(\log_2 2n)^d$.*

We reiterate here that the set of basis functions can be taken to be anything (e.g relu functions, wavelet basis etc.) and yet a simultaneous oracle risk bound such as Theorem 7 will hold for the OLRADRE. We now move on to focus specifically on piecewise polynomial and univariate higher order bounded variation functions where OLRADRE performs near optimally.

5.1. Result for Rectangular Piecewise Polynomial Functions in General Dimensions

The setup for this subsection is essentially similar to that in Section 4.1 except that θ^* can now be piecewise *polynomial* of degree at most m on the (unknown) partition P^*. More precisely, we let

$$\mathcal{F} := \Big\{ \Big(\frac{u}{n}\Big)^{\mathbf{m}} := \prod_{i \in [d]} \Big(\frac{u_i}{n}\Big)^{m_i} : |\mathbf{m}| \leq m \Big\} \qquad (5.2)$$

18

where $\mathbf{m} = (m_1, \ldots, m_d)$ is the *multidegree* of the monomial $\left(\frac{u}{n}\right)^{\mathbf{m}}$ and $|\mathbf{m}| = \sum_{i \in [d]} m_i$ is the corresponding degree.

As before, an *oracle predictor* $\hat{\theta}_{(\text{oracle})}$ — which *knows* the minimal rectangular partition $(R_1, \ldots, R_k)$ of θ^* *exactly* —can simply use the online VAW online linear regression rule given in (5.1) separately within each of the rectangles $(R_1, \ldots, R_k)$. By a basic result about VAW online linear regression, (see Proposition 25), it can be shown that the MSE of this oracle predictor is bounded by $O_d(\frac{k\|\theta^*\|_\infty \log n}{N})$. In words, the MSE of this oracle predictor again scales (up to a log factor which is necessary) like the *number of constant pieces of θ^* divided by the sample size N which is precisely the parametric rate of convergence*. We will now state a result saying that OLRADRE, which is computationally efficient, can attain this oracle rate of convergence, up to certain additional multiplicative log factors.

Since any $\theta \in \Theta_P$, where $P \in \mathcal{P}_{\text{all},\mathsf{T}}$ for some $\mathsf{T} \subset K$ (cf. the statement of Theorem 12), is piecewise polynomial on P, we can associate to any such θ the number

$$s_{m,\infty}(\theta) = \max_{S \in P, |\mathbf{m}| \leq m} n^{|\mathbf{m}|} \beta_{\mathbf{m},S} \text{ where } \theta_S \equiv \sum_{|\mathbf{m}| \leq m} \beta_{\mathbf{m},S}\, u^{\mathbf{m}} = \sum_{|\mathbf{m}| \leq m} n^{|\mathbf{m}|} \beta_{\mathbf{m},S} \left(\frac{u}{n}\right)^{\mathbf{m}}. \quad (5.3)$$

The reader should think of θ_s as $g(\frac{s}{n})$ where g is some piecewise polynomial function defined on the unit cube $[0,1]^d$ and hence of $s_{m,\infty}(\theta)$ as its maximum coefficient which is a bounded number, i.e., it does not grow with n. Let us keep in mind that the OLRADRE depends on the underlying degree m which we keep implicit in our discussions below. We can now state the analogue of Theorem 12 in this case.

Theorem 15 (Oracle Inequality for Arbitrary Rectangular Partitions) *Let T be any subset of K and $\hat{\theta}^{OL}$ denote the OLRADRE predictor. Then there exist an absolute constant C and a number $C_{m,d} > 1$ depending only on m and d such that for $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta\|_\infty)$, one has for any non-anticipating ordering ρ of $L_{d,n}$,*

$$\mathbb{E}\, \frac{1}{|\mathsf{T}|}\|\hat{\theta}_{\mathsf{T}}^{OL} - \theta_{\mathsf{T}}^*\|^2 \leq \inf_{\substack{P \in \mathcal{P}_{\text{all},\mathsf{T}} \\ \theta \in \Theta_P \subset \mathbb{R}^{\mathsf{T}}}} \left(\frac{1}{|\mathsf{T}|}\|\theta_{\mathsf{T}}^* - \theta\|^2 + C_{m,d}\lambda_{m,*}^2 \frac{|P|}{|\mathsf{T}|}(\log en)^d \log 2^d N\right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2}$$

$$(5.4)$$

where $\lambda_{m,*} = \lambda + s_{m,\infty}(\theta)$.

We now make some remarks about this theorem.

Remark 16 *We are not aware of such a simultaneous oracle risk bound explicitly stated before in the literature for piecewise polynomial signals in general dimensions in the online learning setting.*

Remark 17 *Even in the batch learning setting, the above oracle inequality is a stronger result than available results for higher order Dyadic CART or ORT (Chatterjee and Goswami (2021a)) in the sense that the infimum is taken over the space of all rectangular partitions $\mathcal{P}_{\text{all}}$ instead of a more restricted class of partitions.*

5.2. Result for Univariate Functions of Bounded Variation of Higher Orders

One can consider the univariate function class of all m times (weakly) differentiable functions, whose m th derivative is of bounded variation. This is also a canonical function class in offline nonparametric regression. A seminal result of Donoho and Johnstone (1998) shows that a wavelet threshholding estimator attains the minimax rate in this problem. Locally adaptive regression splines, proposed by Mammen and van de Geer (1997), is also known to achieve the minimax rate in this problem. Recently, Trend Filtering, proposed by Kim et al. (2009), has proved to be a popular nonparametric regression method. Trend Filtering is very closely related to locally adaptive regression splines and is also minimax rate optimal over the space of higher order bounded variation functions; see Tibshirani et al. (2014) and references therein. Moreover, it is known that Trend Filtering adapts to functions which are piecewise polynomials with regularity at the knots. If the number of pieces is not too large and the length of the pieces is not too small, a well tuned Trend Filtering estimator can attain near parametric risk as shown in Guntuboyina et al. (2020). In the online learning setting, this function class has been studied recently by Baby and Wang (2020) using online wavelet shrinkage methods. We now state a spatially adaptive oracle risk bound attained by the OLRADRE for this function class.

Let $K = L_{1,n} = [[1, n]]$ and for any vector $\theta \in \mathbb{R}^n$, let us define its m-th order (discrete) derivative for any integer $r \geq 0$ in a recursive manner as follows. We start with $D^{(0)}(\theta) = \theta$ and $D^{(1)}(\theta) = (\theta_2 - \theta_1, \ldots, \theta_n - \theta_{n-1})$. Having defined $D^{(m-1)}(\theta)$ for some $m \geq 2$, we set $D^{(m)}(\theta) = D^{(1)}(D^{(m-1)}(\theta))$. Note that $D^{(m)}(\theta) \in \mathbb{R}^{n-m}$. For sake of convenience, we denote the operator $D^{(1)}$ by D. For any positive integer $m \geq 1$, let us also define the m-th order variation of a vector θ as follows:

$$V^{(m)}(\theta) = n^{m-1}|D^{(m)}(\theta)|_1 \tag{5.5}$$

where $|.|_1$ denotes the usual ℓ^1-norm of a vector. Notice that $V^1(\theta)$ is the total variation of a vector defined in (4.4). Like our definition of total variation, our definition in (5.5) is also motivated by the analogy with the continuum. If we think of θ as an evaluation of an m times differentiable function $f : [0, 1] \to \mathbb{R}$ on the grid $(1/n, 2/n \ldots, n/n)$, then the Reimann approximation to the integral $\int_{[0,1]} f^{(m)}(t)dt$ is precisely equal to $V^{(m)}(\theta)$. Here $f^{(m)}$ denotes the m-th order derivative of f. Thus, the reader should assume that $V^{(m)}(\theta)$ is of constant order for a generic θ. Analogous to the class $\mathcal{BV}_{d,n}(V^*)$, let us define for any integer $m \geq 1$,

$$\mathcal{BV}_n^{(m)}(V^*) = \{\theta \in \mathbb{R}^n : V^{(m)}(\theta) \leq V^*\}.$$

In the spirit of our treatment of the class $\mathcal{BV}_{d,n}(V^*)$ in Section 4.2, we take

$$\mathcal{F} = \{1, x, \ldots, x^{m-1}\}.$$

We now state the main result of this subsection.

Theorem 18 (Prediction error for $\mathcal{BV}_N^{(m)}(V^*)$, $m > 1$) *Fix any interval* $\mathsf{T} \subset K$ *and denote* $V_\mathsf{T}^* = V^{(m)}(\theta_\mathsf{T}^*)$. *Also let* $\|\theta^*\|_{m-1,\infty} := \max_{0 \leq j < m} N^j \|D^j(\theta^*)\|_\infty$. *Then there exist an absolute constant C and a number $C_m > 1$ depending only on m such that for $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta^*\|_\infty)$, we have for any non-anticipating ordering ρ of $L_{1,n}$,*

$$\mathbb{E}\frac{1}{|\mathsf{T}|}\|\hat{\theta}_\mathsf{T}^{QL} - \theta_\mathsf{T}\|^2 \leq C_m \, \lambda_{m,*}^{\frac{4m}{2m+1}} \log eN \left(\frac{(V_\mathsf{T}^*)^{1/m}}{|\mathsf{T}|}\right)^{\frac{2m}{2m+1}} + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|^2} \tag{5.6}$$

where $\lambda_{m,*} := \lambda + \|\theta^*\|_{m-1,\infty}$ (cf. the statement of Theorem 15).

We now make some remarks about the above theorem.

Remark 19 *The above spatially adaptive risk bound for bounded variation functions of a general order is new even in the easier batch learning setting. State of the art batch learning estimators like Trend Filtering or Dyadic CART are not known to attain such a spatially adaptive risk bound.*

6. Discussion

In this section we discuss some natural related matters.

6.1. Detailed Comparison with Dyadic CART

The Dyadic CART is a natural offline analogue of the OMADRE described in Section 4. Similarly, higher order versions of Dyadic CART and Trend Filtering are natural offline analogues of the univariate piecewise polynomial OLRADRE described in Section 5. Therefore, it makes sense to compare our oracle risk bound (notwithstanding simultaneity and the fact that OMADRE/OLRADRE are online algorithms) in Theorems 12, 15, 18 with the available offline oracle risk bound for Dyadic CART, see Theorem 2.1 in Chatterjee and Goswami (2021a). This result is an oracle risk bound where the infimum is over all recursive dyadic partitions (see a precise definition in Section 1.2.1 of Chatterjee and Goswami (2021a)) of $L_{d,n}$. On the other hand, our oracle risk bounds are essentially an infimum over all dyadic partitions $\mathcal{P}_{dp}$. In dimensions $d = 1, 2$ these two classes of partitions coincide (see Lemma 8.2 in Chatterjee and Goswami (2021a) in the arxiv version) but for $d > 2$, the class of partitions $\mathcal{P}_{dp}$ strictly contain the class of recursive dyadic partitions (see Remark 8.3 in Chatterjee and Goswami (2021a) in the arxiv version). Therefore, the oracle risk bounds in Theorems 12, 15 are stronger in this sense.

The above fact also allows us to convert the infimum over all dyadic partitions $\mathcal{P}_{dp}$ to the space of all rectangular partitions $\mathcal{P}_{\text{all}}$ since any partition in $\mathcal{P}_{\text{all}}$ can be refined into a partition in $\mathcal{P}_{dp}$ with the number of rectangles inflated by a $(\log n)^d$ factor. In dimensions $d \geq 3$, such an offline oracle risk bound (where the infimum is over $\mathcal{P}_{\text{all}}$) is not known for Dyadic CART. As far as we are aware, the state of the art result here is shown in Chatterjee and Goswami (2021a) where the authors show that a significantly more computationally intensive version of Dyadic CART, called the ORT estimator is able to adaptively estimate signals which are piecewise constant on fat partitions. In contrast, Theorems 12, 15 hold for all dimensions d, the infimum in the oracle risk bound is over the set of all rectangular partitions $\mathcal{P}_{\text{all}}$ and no fatness is needed.

It should also be mentioned here that compared to batch learning bounds for Dyadic CART, our bounds have an extra log factor and some signal dependent factors which typically scale like $O(1)$. Note that the computational complexity of our algorithm is also worse by a factor $(\log n)^d$, compare Lemma 10 to Lemma 1.1 in Chatterjee and Goswami (2021a). However, it should be kept in mind that we are in the online setup which is a more difficult problem setting than the batch learning setting.

6.2. Some Other Function Classes

Our simultaneous oracle risk bounds are potentially applicable to other function classes as well not considered in this paper. We now mention some of these function classes.

A similar batch learning oracle risk bound with an infimum over the set of all recursive dyadic partitions was used by Donoho (1997) to demonstrate minimax rate optimality of Dyadic CART for some anisotropically smooth bivariate function classes. Using our result, it should be possible to attain a simultaneous version of minimax rate optimal bounds for these types of function classes.

Consider the class of bounded monotone signals on $L_{d,n}$ defined as

$$\mathcal{M}_{d,n} = \{\theta \in [0,1]^{L_{n,d}} : \theta[i_1, \ldots, i_d] \leq \theta[j_1, \ldots, ,j_d] \text{ whenever } i_1 \leq j_1, \ldots, i_d \leq j_d\}.$$

Estimating signals within this class falls under the purview of Isotonic Regression. Isotonic Regression has been a topic of recent interest in the online learning community; see Kotłowski et al. (2016), Kotlowski et al. (2017). It can be checked that the total variation for any d dimensional isotonic signal with range $O(1)$ grows like $O(n^{d-1})$ which is of the same order as a canonical bounded variation function. Therefore, the bound in Theorem 12 would give spatially adaptive minimax rate optimal bounds for Isotonic Regression as well. In the offline setup, a lot of recent papers have investigated Isotonic regression with the aim of establishing minimax rate optimal rates as well as near optimal adaptivity to rectangular piecewise constant signals; see Deng and Zhang (2020), Han et al. (2019). Theorem 12 establishes that such adaptivity to rectangular piecewise constant signals as well as maintaining rate optimality over isotonic functions is also possible in the online setting by using the OMADRE proposed here.

Let us now consider univariate convex regression. In the offline setting, it is known that the least squares estimator LSE is minimax rate optimal, attaining the $\tilde{O}(n^{-4/5})$ rate, over convex functions with bounded entries, see e.g. Guntuboyina and Sen (2013), Chatterjee et al. (2016). It is also known that the LSE attains the $\tilde{O}(k/n)$ rate if the true signal is piecewise linear in addition to being convex. Theorem 15 and Theorem 18 imply both these facts also hold for the OLRADRE (since a convex function automatically has finite second order bounded variation) where we fit linear functions (polynomial of degree 1) on intervals. To the best of our knowledge, such explicit guarantees for online univariate convex regression were not available in the literature before this work.

6.3. Computation Risk Tradeoff

The main reason for us considering dyadic rectangles (instead of all rectangles) as experts is to save computation. In particular, if one uses the set of all rectangles as experts, the computational complexity of the resulting algorithm would be $O_d(N^3)$. One can think of this estimator as the online analogue of the ORT estimator defined in Chatterjee and Goswami (2021a). For this estimator, the risk bounds would be better. For example, the $(\log n)^d$ term multiplying $|P|$ in the bound in Theorems 12, 15 would now no longer be present. In particular, the exponent of $\log n$ would be 2 for all dimensions d which is only one log factor more than a known minimax lower bound for the space of all rectangular piecewise constant functions; see Lemma 3.1 in Chatterjee and Goswami (2021a).

One can also easily interpolate and take the set of experts somewhere between the set of dyadic rectangles and the set of all rectangles, say by considering all rectangles with side lengths a multiple of some chosen integer l. Thus one can choose the set of experts by trading off computational time and the desired statistical prediction performance.

6.4. Open Problems

In our opinion, our work here raises some interesting open questions which we leave for future research.

1. It appears that if a function class is well approximable by rectangular piecewise constant/polynomial functions then the type of oracle risk bounds proved here may be used to derive some nontrivial prediction bounds. However, for many function classes, this kind of approximability may not hold. For example, we can consider the class of Hardy Krause Bounded Variation Functions (see Fang et al. (2021)) or its higher order versions (see Ki et al. (2021)) where the existing covering argument produces nets (to estimate metric entropy) which are not necessarily rectangular piecewise constant/linear respectively. These function classes are also known not to suffer from the curse of dimensionality in the sense that the metric entropy does not grow exponentially in $\frac{1}{\epsilon}$ with the dimension d. More generally, it would be very interesting to come up with computationally efficient and statistically rate optimal online prediction algorithms for such function classes.

2. The analysis presented here relies a lot on the light-tailed nature of the noise. It can be checked that Theorem 7 can also be proved when the noise is mean 0 sub-exponential, we would only get an appropriate extra log factor. However, the proof would break down for heavy-tailed noise. This seems to be an open area and not much attention has been given to the noisy online prediction problem with heavy tailed noise. Most of the existing results in the online learning community assume bounded but arbitrary data. The heavy tailed setting we have in mind is that the data y is not arbitrary but of the form signal plus noise, except that the noise can be heavy tailed. It would be very interesting to obtain an analogue of Theorem 7 in this setting. Clearly, the algorithm has to change as well in the sense that instead of aggregating means one should aggregate medians of various rectangles in some appropriate way.

Acknowledgments

References

Dmitry Adamskiy, Wouter M Koolen, Alexey Chernov, and Vladimir Vovk. A closer look at adaptive regret. In *International Conference on Algorithmic Learning Theory*, pages 290–304. Springer, 2012.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. *Advances in Neural Information Processing Systems*, 32, 2019.

Dheeraj Baby and Yu-Xiang Wang. Adaptive online estimation of piecewise polynomial trends. *Advances in Neural Information Processing Systems*, 33:20462–20472, 2020.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *Conference on Learning Theory*, pages 359–409. PMLR, 2021.

Dheeraj Baby, Xuandong Zhao, and Yu-Xiang Wang. An optimal reduction of tv-denoising to adaptive online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2899–2907. PMLR, 2021a.

Dheeraj Baby, Xuandong Zhao, and Yu-Xiang Wang. An optimal reduction of tv-denoising to adaptive online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2899–2907. PMLR, 2021b.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Conference on Learning Theory*, pages 465–481. PMLR, 2017.

Sabyasachi Chatterjee and Subhajit Goswami. Adaptive estimation of multivariate piecewise polynomials and bounded variation functions by optimal decision trees. *Ann. Statist.*, 49(5):2531–2551, 2021a. ISSN 0090-5364. doi: 10.1214/20-aos2045. URL https://doi.org/10.1214/20-aos2045.

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2D total variation denoising. *IEEE Trans. Inform. Theory*, 67(6, part 2):4060–4091, 2021b. ISSN 0018-9448. doi: 10.1109/TIT.2021.3059657. URL https://doi.org/10.1109/TIT.2021.3059657.

Sabyasachi Chatterjee et al. An improved global risk bound in concave regression. *Electronic Journal of Statistics*, 10(1):1608–1629, 2016.

Alexey Chernov and Vladimir Vovk. Prediction with expert evaluators' advice. In *International Conference on Algorithmic Learning Theory*, pages 8–22. Springer, 2009.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411. PMLR, 2015.

Hang Deng and Cun-Hui Zhang. Isotonic regression in multi-dimensional spaces and graphs. *The Annals of Statistics*, 48(6):3672–3698, 2020.

David L Donoho. Cart and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870–1911, 1997.

David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

Billy Fang, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy–krause variation. *The Annals of Statistics*, 49(2):769–792, 2021.

Pierre Gaillard and Sébastien Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory*, pages 764–796. PMLR, 2015.

Sébastien Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 377–396. JMLR Workshop and Conference Proceedings, 2011.

Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields*, 2013. *To appear*, available at http://arxiv.org/abs/1305.1648.

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, and Bodhisattva Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics*, 48 (1):205–229, 2020.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, Richard J Samworth, et al. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

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

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 2016.

Dohyeong Ki, Billy Fang, and Adityanand Guntuboyina. Mars via lasso. *arXiv preprint arXiv:2111.11694*, 2021.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. l_1 trend filtering. *SIAM Rev.*, 51(2):339–360, 2009. ISSN 0036-1445. doi: 10.1137/070690274. URL http://dx.doi.org.ezproxy.cul.columbia.edu/10.1137/070690274.

Wojciech Kotłowski, Wouter M Koolen, and Alan Malek. Online isotonic regression. In *Conference on Learning Theory*, pages 1165–1189. PMLR, 2016.

Wojciech Kotlowski, Wouter M Koolen, and Alan Malek. Random permutation online isotonic regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Ilja Kuzborskij and Nicolò Cesa-Bianchi. Locally-adaptive nonparametric online learning. *Advances in Neural Information Processing Systems*, 33:1679–1689, 2020.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Alexander Rakhlin and Karthik Sridharan. Statistical learning theory and sequential prediction. *Lecture Notes in University of Pennsyvania*, 2012.

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

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J Tibshirani. Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Advances in Neural Information Processing Systems*, pages 3513–3521, 2016.

Ryan Tibshirani. Nonparametric regression (and classification). 2015.

Ryan J Tibshirani. Divided differences, falling factorials, and discrete splines: Another look at trend filtering and related problems. *arXiv preprint arXiv:2003.03886*, 2020.

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

Volodya Vovk. Competitive on-line linear regression. *Advances in Neural Information Processing Systems*, pages 364–370, 1998.

Appendix A. Simulations

A.1. 1D Plots

We provide plots of the OMADRE for a visual inspection of its performance. There are three plots for scenarios corresponding to different true signals θ^*, where for any $i \in [n]$, we have $\theta_i^* = f(i/n)$ for some function $f : [0,1] \to \mathbb{R}$, specified below and the errors are generated from $N(0,1)$. The sample size is taken to be $n = 2^{16}$ for these plots, given in Figure 1. The truncation parameter λ has been taken to be $2 \max\{\|\theta^*\|_\infty, \sigma (2 \log n)^{1/2}\}$ for all our $1D$ simulations. It may be possible to get better predictions by choosing a smaller value of λ but we have not done any systematic search for these simulations as this particular choice seemed to work well.

The ordering of the revealed indices is taken to be the forward ordering $1, 2, 3, \ldots$ and the backward ordering $n, n-1, n-2, \ldots$. The predictions corresponding to the two orderings are then averaged in the plots.

1. Scenario 1 [Piecewise Constant Signal]: We consider the piecewise constant function

$$f(x) = 2(1(x \in [1/5, 2/5])) + 1(x \in [2/5, 3/5]) + 21(x \in [3/5, 4/5]),$$

   and consider the the 1D OMADRE. The corresponding plot is shown in the second diagram of Figure 1.

2. Scenario 2 [Piecewise Linear Signal]: We consider the piecewise linear function

$$f(x) = 6x(1(x \in [0, 1/3])) + (-12x + 6)1(x \in [1/3, 2/3]) + (x - 8/3)(1(x \in [2/3, 1])),$$

   and consider the 1D OMADRE. The corresponding plot is shown in the second diagram of Figure 1.

3. Scenario 3 [Piecewise Quadratic Signal]: We consider the piecewise quadratic function

$$f(x) = \begin{cases} 18x^2 & \text{if } x \in [0, 1/3] \\ -36(x - 1/2 - 1/\sqrt{12})(x - 1/2 + \sqrt{12}) & \text{if } x \in [1/3, 2/3] \\ 18(x-1)^2 & \text{if } x \in [2/3, 1] \end{cases}.$$

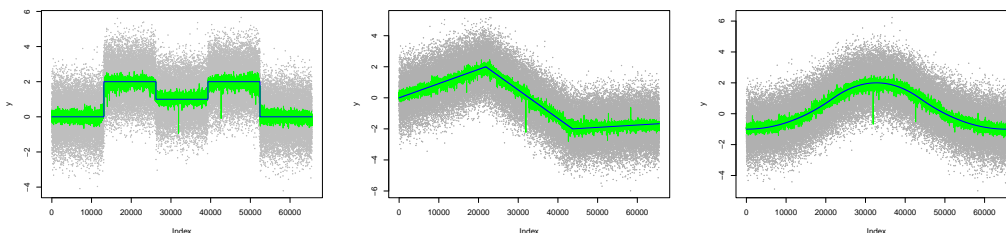and consider the 1D OMADRE estimator. The corresponding plot is shown in the third diagram of Figure 1.

Figure 1: The blue curve is the true signal, the grey points are data points and the green curve constitutes the OMADRE predictions. The plotted predictions are averaged over two predictions when the data are revealed in the forward and backward order.

A.2. 1D Comparisons

We conduct a simulation study to compare the performance of the OMADRE and the OLRADRE of order $1, 2$ which aggregates linear function predictions and quadratic function predictions. We consider the ground truth signal as the smooth sinusoidal function

$$f(x) = \sin 2\pi x + \cos 5\pi x.$$

We considered various signal to noise ratios by setting the noise standard deviation σ to be $0.5, 1$ or 2. We also considered sample sizes $n = 2^{10}, 2^{12}, 2^{14}$. In each case, we estimated the MSE by 50 Monte Carlo replications. Here also, the predictions corresponding to the forward and backward orderings are averaged. We report the MSE's in Tables 1, 2 and 3 respectively.

Table 1: MSEs of OMADRE estimator in different scenarios

n	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$
2^{10}	0.045	0.076	0.191
2^{12}	0.027	0.048	0.127
2^{14}	0.014	0.031	0.087

Table 2: MSEs of OLRADRE (linear) in different scenarios

n	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$
2^{10}	0.088	0.099	0.143
2^{12}	0.049	0.057	0.087
2^{14}	0.025	0.030	0.050

Table 3: MSEs of OLRADRE (quadratic) in different scenarios

n	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$
2^{10}	0.079	0.091	0.136
2^{12}	0.040	0.048	0.079
2^{14}	0.020	0.025	0.044

It is reasonable to expect that the OLRADRE aggregating quadratic function predictions would perform no worse than the OLRADRE aggregating linear function predictions which in turn would perform no worse than the OMADRE estimator. From the tables 1, 2 and 3 we see that when the noise variance is low, the opposite happens and the OMADRE gives a better performance. It is only when the noise variance becomes high, the OLRADRE aggregating quadratic functions starts to perform the best. We see a similar phenomenon for other ground truth functions as well. We are not sure what causes this but we believe that in the low noise regime, the weights of the local experts are high (for the OMADRE estimator) and for smooth functions these predictions would be very accurate. Since the OLRADRE has a shrinkage effect (note the presence of I in the gram matrix), there is bias for the predictions of the local experts which is why the local experts in this case predict slightly worse than for the OMADRE estimator. In the case when the signal to noise ratio is low, the algorithms are forced to use experts corresponding to wider intervals for which case the bias of the OLRADRE predictions become negligible.

A.3. 2D Plots

We conduct a simulation study to observe the performance of the proposed OMADRE estimator in three different scenarios each corresponding to a different true signal θ^*. In every case, the errors are generated from a centered normal distribution with standard deviation 0.25, the dimension $d = 2$ and we take the number of pixels in each dimension to be $n = 64, 128, 256$. We estimate the MSE by 50 Monte Carlo replications and they are reported in Table 4. The truncation parameter λ has been taken to be $2\max\{\|\theta^*\|_\infty, \sigma\left(2\log(n^2)\right)^{1/2}\}$. In each of the cases, a uniformly random ordering of the vertices of $L_{2,n}$ has been taken to construct the OMADRE estimator. Overall, we see that our OMADRE estimator performs pretty well.

1. Scenario 1 [Rectangular Signal]: The true signal θ^* is such that for every $(i_1, i_2) \in L_{2,n}$, we have
$$\theta^*_{(i_1,i_2)} = \begin{cases} 1 & \text{if } n/3 \leq i_1, i_2 \leq 2n/3 \\ 0 & \text{otherwise} \end{cases}.$$
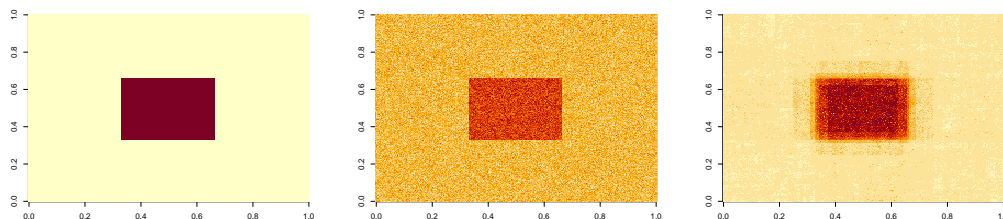The corresponding plots are shown in Figure 2 when $n = 256$.

Figure 2: The first diagram refers to the true signal, the second one to the noisy signal and the third one to the estimated signal by the OMADRE estimator.

2. Scenario 2 [Circular Signal]: The true signal θ^* is such that for every $(i_1, i_2) \in L_{2,n}$, we have

$$\theta^*_{(i_1,i_2)} = \begin{cases} 1 & \text{if } \sqrt{(i_1 - n/2)^2 + (i_2 - n/2)^2} \leq n/4 \\ 0 & \text{otherwise} \end{cases}.$$

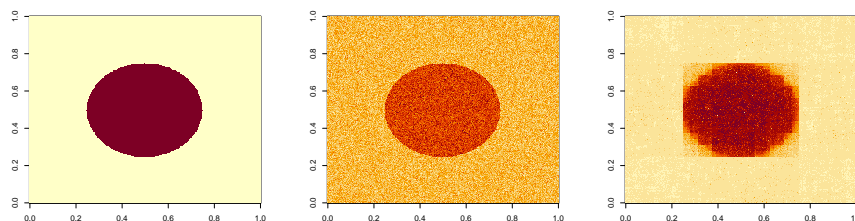The corresponding plots are shown in Figure 3 when $n = 256$.

Figure 3: The first diagram refers to the true signal, the second one to the noisy signal and the third one to the estimated signal by the OMADRE estimator.

3. Scenario 3 [Sinusoidal Smooth Signal]: The true signal θ^* is such that for every $(i_1, i_2) \in L_{2,n}$, we have $\theta^*_{(i_1,i_2)} = f(i_1/n, i_2/n)$, where

$$f(x, y) = \sin(\pi x) \sin(\pi y).$$

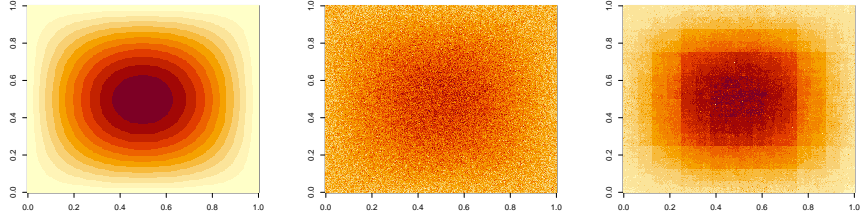The corresponding plots are shown in Figure 4 when $n = 256$.

Figure 4: The first diagram refers to the true signal, the second one to the noisy signal and the third one to the estimated signal by the OMADRE estimator.

Table 4: MSEs of CV Dyadic CART estimator in different scenarios

$n \times n$	Scenario 1	Scenario 2	Scenario 3
64×64	0.035	0.037	0.014
128×128	0.022	0.022	0.008
256×256	0.012	0.013	0.005

Appendix B. Appendix

B.1. Proofs of Lemma 10 and Lemma 14

We only prove Lemma 14 since it contains the proof of Lemma 10. In the remainder of this subsection the constant C always stands for an absolute constant whose precise value may change from one occurrence to the next. For every $s \in L_{d,n}$, we let $\mathcal{S}(s)$ denote the subcollection of all dyadic rectangles $S \subset L_{d,n}$ containing s.

At the outset of every round $t = 1, \ldots, N$, we maintain several objects for every $S \in \mathcal{S}$. These include the weight $w_{S,t}$, the $L \times L$ matrix $X_{S,t} := I + \sum_{s \in \rho[1:(t-1)] \cap S} x_s x_s^T$ where $L = |\mathcal{F}|$ and the vector $z_{S,t} = \sum_{s \in \rho[1:(t-1)] \cap S} y_s x_s \in \mathbb{R}^L$. We also store the indicator $I_{S,t} \in \{0, 1\}$ whether S has had any datapoint upto round $t-1$ which is required to determine the set of active experts A_t (recall step 2 of $\mathcal{A}$). In the beginning, $w_{S,1} = \frac{1}{|\mathcal{S}|}$ (recall the initialization step of $\mathcal{A}$), $X_{S,1} = I$, $z_{S,1} = 0$ and $I_{S,1} = 0$ for all $S \in \mathcal{S}$. We first analyze the number of elementary operations necessary for computing the estimate $\hat{y}_{\rho(t)}$ and updating the matrices $(X_{S,t}; S \in \mathcal{S}(\rho(t)))$ as well as the indicators $I_{S,t}$ after the adversary reveals $\rho(t)$.

To this end observe that, we can visit all the rectangles in $A_t \subset \mathcal{S}(\rho(t))$ by performing binary search on each coordinate of $\rho(t) \in L_{d,n}$ in the lexicographic order and checking for the value of $I_{S,t}$. This implies, firstly, that $|\mathcal{S}(\rho(t))| \leq (\log_2 2n)^d$ and secondly, that the number of operations required to update the indicators $I_{S,t}$'s is bounded by $(\log_2 2n)^d$. Now let us recall from (5.1) that,

$$\hat{y}_{\rho(t)}^{(S)} = X_{S+1,t}^{-1} z_{S,t} \cdot x_{\rho(t)}, \quad \text{where } X_{S,t+1} = X_{S,t} + x_{\rho(t)} x_{\rho(t)}^T.$$

Computing $X_{S+1,t}$ and its inverse, and the subsequent multiplication with $z_{S,t}$ require at most CSL^3 and CL^2 many basic operations respectively. Evaluating the inner product with x_s afterwards take at most CL many basic steps. Thus, we incur CL^3 as the total cost for computing $w_{S,t} T_\lambda(\hat{y}_{\rho(t)}^{(S)})$ and

updating $X_{S,t}$ for each $S \in \mathcal{S}(\rho(t))$. Calculating $\hat{y}_{\rho(t)}$ from the numbers $w_{S,t} T_\lambda(\hat{y}_{\rho(t)}^{(S)})$'s (see step 3 of $\mathcal{A}$), where $S \in \mathcal{S}(\rho(t))$, requires $C|\mathcal{S}(\rho(t))|$ many additional steps. Therefore, the combined cost for computing $\hat{y}_{\rho(t)}$ and updating $X_{S,t}$'s for all $S \in \mathcal{S}(\rho(t))$ is bounded by $CL^3|\mathcal{S}(\rho(t))| = CL^3(\log_2 2n)^d$.

After the adversary reveals $y_{\rho(t)}$, we need to update the weights $w_{S,t}$, the vectors $z_{S,t}$ and the indicators $I_{S,t}$ for all $S \in \mathcal{S}(t)$ (see step 4 of $\mathcal{A}$). For this we first need to compute the numbers $w_{S,t} e^{-\alpha \ell_{S,t}}$ for all $S \in A_t$ and this takes $C|\mathcal{S}(\rho(t))| = C(\log_2 2n)^d$ many basic operations. It takes an additional $C(\log_2 2n)^d$ many basic operations in order to compute the sums $\sum_{S \in A_t} w_{S,t}$ and $\sum_{S \in A_t} w_{S,t} e^{-\alpha \ell_{S,t}}$. Using these numbers, we can now update the weights as

$$w_{S,t+1} = \frac{w_{S,t} e^{-\alpha \ell_{S,t}}}{\sum_{S \in A_t} w_{S,t} e^{-\alpha \ell_{S,t}}} \sum_{S \in A_t} w_{S,t}$$

and this also involves $C(\log_2 n)^d$ many elementary operations. Updating the vector $z_{S,t}$ to $z_{S,t+1} = z_{S,t} + y_{\rho(t)} x_{\rho(t)}$ takes at most CL many basic steps for every S and hence $CL(\log_2 n)^d$ many steps in total.

Putting everything together, we get that the computational complexity of OLRADRE is bounded by $CL^3 N(\log_2 n)^d$.

B.2. Proof of Theorem 12

Recall the definition of $\overline{R}(\theta, P)$ for any partition P of $K = L_{d,n}$ and a $\theta \in \Theta_P$ given right after (3.2). It turns out that for the online averaging rule, one can give a clean bound on $\overline{R}(\theta, P)$ which is stated next as a proposition.

Proposition 20 *Let $y_t = \theta_t^* + \sigma \epsilon_t$ for $t \in K = L_{d,n}$ where $\sigma > 0$ and ϵ_t's are independent, mean zero sub-Gaussian variables with unit dispersion factor. Then we have for any partition $P \in \mathcal{P}_\mathsf{T}$, where $\mathsf{T} \subset K$, and any $\theta \in \Theta_P$,*

$$\overline{R}(\theta, P) \leq C|P| \left(\|\theta_\mathsf{T}^*\|_\infty^2 + \sigma^2 \log eN \right) \log eN. \tag{B.1}$$

where $C > 1$ is some absolute constant.

Proof We have

$$\overline{R}(\theta, P) = \mathbb{E} \, \mathcal{R}(y, \theta, P) \leq |P| \, \mathbb{E} \sup_{\rho, S \in P} \left(\sum_{t:\rho(t) \in S} (y_{\rho(t)} - \hat{y}_{\rho(t)}^{(S)})^2 - \|y_S - \theta_S\|^2 \right).$$

Now, the following deterministic lemma is going to be of use to us.

Lemma 21 *Let $z_1, \ldots, z_T$ be an arbitrary sequence of numbers and $\hat{z}_t := \frac{1}{t-1} \sum_{s=1}^{t-1} z_s$ for $t = 2, \ldots, T$ where $\hat{z}_1 = 0$. Then, we have*

$$\|z - \hat{z}\|^2 - \|z - \bar{z}\|^2 \leq 4\|z\|_\infty^2 \log eT. \tag{B.2}$$

For a proof of the above lemma, see, e.g., Theorem 1.2 in Orabona (2019). Using the above deterministic lemma and the previous display, we can write for any partition $P \in \mathcal{P}_\mathsf{T}$ and $\theta \in \Theta_P$,

$$\overline{\mathcal{R}}(\theta, P) \leq 4|P| \, \mathbb{E} \sup_{\rho, S \in P} \Big(\sum_{t:\rho(t) \in S} (y_{\rho(t)} - \hat{y}_{\rho(t)}^{(S)})^2 - \|y_S - \overline{y_S}\|^2 \Big)$$

$$\leq 4|P| \, \mathbb{E}\|y\|_\infty^2 \, \log \mathrm{e}n \leq C|P| \, (\|\theta^*\|_\infty^2 + \sigma^2 \log \mathrm{e}N) \log \mathrm{e}N. \qquad \text{(B.3)}$$

where $\overline{y_S}$ denotes the mean of the entries of y_S and we deduce the last inequality from a standard upper bound on the tail of sub-Gaussian random variables. ■

The following corollary is a direct implication of Theorem 7 and Proposition 20 applied to the particular setting described at the beginning of Section 4.

Corollary 22 *Let T be any subset of K. Let $\hat{\theta}^{OM}$ denote the OMADRE predictor. There exists an absolute constant $C > 1$ such that for $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta^*\|_\infty)$, one has for any non-anticipating ordering ρ of K,*

$$\mathbb{E}\|\hat{\theta}_\mathsf{T}^{OM} - \theta_\mathsf{T}^*\|^2 \leq \inf_{\substack{P \in \mathcal{P}_{\mathrm{dp}, T} \\ \theta \in \Theta_P \subset \mathbb{R}^T}} \big(\|\theta_\mathsf{T}^* - \theta\|^2 + C\lambda^2 |P| \log 2^d N\big) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|}. \qquad \text{(B.4)}$$

We are now ready to prove Theorem 12.
Proof The proof directly follows from (3.6) and Proposition 20. ■

Proof [Proof of Theorem 12] Fix any partition $P \in \mathcal{P}_{\mathrm{all}, \mathsf{T}}$ and any $\theta \in \Theta_\mathsf{P}$. Consider a dyadic refinement of P which we denote by $\tilde{P}$. By definition, $\tilde{P} \in \mathcal{P}_{\mathrm{dp}, \mathsf{T}}$ and $\theta \in \Theta_{\tilde{P}}$. Therefore, we can use the bound in (B.4) given in Corollary 22. The proof is then finished by noting that $|\tilde{P}| \leq |P|(\log \mathrm{e}n)^d$. ■

B.3. Proof of Theorem 13

It has been shown in Chatterjee and Goswami (2021a) that the class of functions $\mathcal{BV}_{d,n}(V^*)$ is well-approximable by piecewise constant functions with dyadic rectangular level sets which makes it natural to study the OMADRE estimator for this function class.

The following result was proved in Chatterjee and Goswami (2021a) (see Proposition 8.5 in the arxiv version).

Proposition 23 *Let $\theta \in \mathbb{R}^{L_{d,n}}$ and $\delta > 0$. Then there exists a dyadic partition $P_{\theta, \delta} = (R_1, \ldots, R_k)$ in $\mathcal{P}_{\mathrm{rdp}}$ such that*
a) $k = |P_{\theta, \delta}| \leq 1 + \log_2 N \, \big(1 + \frac{\mathrm{TV}(\theta)}{\delta}\big)$, and for all $i \in [k]$,
c) $\mathrm{TV}(\theta_{R_i}) \leq \delta$, and
d) $\mathcal{A}(R_i) \leq 2$
where $\mathcal{A}(R)$ denotes the aspect ratio of a generic rectangle R.

Let $\Pi_{\Theta_{P_{\theta, \delta}}} := \Pi_{P_{\theta, \delta}}$ denote the orthogonal projector onto the subspace $\Theta_{P_{\theta, \delta}}$ of $\mathbb{R}^{L_{d,n}}$ comprising functions that are constant on each $R_i \in P_{\theta, \delta}$. It is clear that $\Pi_{P_{\theta, \delta}}\theta(a) = \overline{\theta}_{R_i}$ — the average value of θ over R_i — for all $a \in R_i$ and $i \in [k]$. We will use $\Pi_{P_{\theta^*, \delta}}\theta^*$ as θ in our application of (B.4) in this case. In order to estimate $\|\theta^* - \Pi_{\Theta_{P_{\theta^*, \delta}}}\theta^*\|$, we would need the following approximation theoretic result.

Proposition 24 *Let* $\theta \in \mathbb{R}^{\otimes_{i \in [d]}[n_i]}$ *and*

$$\overline{\theta} := \sum_{(j_1, j_2, \ldots, j_d) \in \otimes_{i \in [d]}[n_i]} \theta[j_1, j_2, \ldots, j_d] / \prod_{i \in [d]} n_i$$

be the average of the elements of θ. Then for every $d > 1$ we have,

$$\sum_{(j_1, j_2, \ldots, j_d) \in \otimes_{i \in [d]}[n_i]} |\theta[j_1, j_2, \ldots, j_d] - \overline{\theta}|^2 \leq \left(1 + \max_{i, j \in [d]} \frac{n_i}{n_j}\right)^2 \text{TV}(\theta)^2. \tag{B.5}$$

For $d = 1$, on the other hand, we have

$$\sum_{j \in [N]} |\theta[j] - \overline{\theta}|^2 \leq N \text{TV}(\theta)^2. \tag{B.6}$$

See Chatterjee and Goswami (2021a) (Proposition 8.7 in the arxiv version) for a proof of (B.5) and Chatterjee and Goswami (2021b) (Lemma 10.3 in the arxiv version) for (B.6). Propositions 23 and 24 together with the description of $\Pi_{P_{\theta,\delta}}$ as the operator that projects θ onto its average value on each rectangle R_i, imply that

$$\|\theta - \Pi_{P_{\theta,\delta}}\theta\|^2 \leq C|P_{\theta,\delta}|\delta^2 = C \log_2 2N \left(\delta^2 + \delta \text{TV}(\theta)\right) \tag{B.7}$$

for $d > 1$ whereas for $d = 1$,

$$\|\theta - \Pi_{P_{\theta,\delta}}\theta\|^2 \leq N\delta^2. \tag{B.8}$$

Proof [Proof of Theorem 13] We get by plugging the bounds from (B.7) and item (a) in Proposition 23 — both evaluated at $\theta = \Pi_{P_{\theta_\mathsf{T}^*,\delta}}\theta_\mathsf{T}^*$ — into Corollary 22

$$\mathbb{E}\|\hat{\theta}_\mathsf{T}^{QM} - \theta_\mathsf{T}^*\|^2 \leq \inf_{\delta > 0} C\left((\delta^2 + \delta V_\mathsf{T}^*) \log 2^d N + \lambda^2 (\log 2^d N)^2 \left(1 + \frac{V_\mathsf{T}^*}{\delta}\right)\right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|}.$$

Now putting $\delta = \lambda (\log 2^d N)^{1/2}$ in the above display, we obtain (4.5).

For $d = 1$, we follow the exact same steps except that we now use the bound (B.8) in lieu of (B.7) to deduce

$$\mathbb{E}\|\hat{\theta}_\mathsf{T}^{QM} - \theta_\mathsf{T}^*\|^2 \leq \inf_{\delta > 0} C\left(|\mathsf{T}|\delta^2 + \lambda^2 (\log 2^d N)^2 \left(1 + \frac{V_\mathsf{T}^*}{\delta}\right)\right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|}.$$

This immediately leads to (4.6) upon setting $\delta = (V_\mathsf{T}^*)^{1/3} \lambda^{2/3} (\log 2^d N)^{2/3} |\mathsf{T}|^{-1/3}$. ∎

B.4. Proof of Theorem 15

We take a similar approach as in the proof of Theorem 12. Let us begin with an upper bound on the regret of the estimator (see, e.g., (Rakhlin and Sridharan, 2012, pp. 38–40) for a proof).

Proposition 25 (Regret bound for Vovk-Azoury-Warmuth forecaster) *Let* $(z_1, x_1), \ldots, (z_T, x_T)$ $\in \mathbb{R} \times \mathbb{R}^d$ *and define for* $t = 1, \ldots, T$ *(cf. (5.1)),*

$$\hat{z}_t := \hat{\beta}_s \cdot x_s \text{ where } \hat{\beta}_s = \left(I + \sum_{s=1}^{t-1} x_s x_s^T \right)^{-1} \left(\sum_{s=1}^{t-1} y_s x_s \right).$$

Then, we have

$$\sum_{t \in [T]} (\hat{z}_t - z_t)^2 - \inf_{\beta \in \mathbb{R}^d} \left\{ \sum_{t \in [T]} (z_t - \beta \cdot x_t)^2 + \|\beta\|^2 \right\} \leq d \|z_\infty\|^2 \log(1 + T \max_{t \in [T]} \|x_t\|^2 / d). \quad \text{(B.9)}$$

Using similar arguments as in (B.3), but applying (B.9) instead of (B.2) for bounding the regret, we get for any $P \in \mathcal{P}_\mathsf{T}$ and $\theta \in \Theta_P$,

$$\overline{R}(\theta, P) \leq C_{m,d} |P| \left(s_{m,\infty}(\theta)^2 + \sigma^2 \log eN \right) \log eN$$

where $C_{m,d} > 1$ depends only on m and d. The remaining part of the proof is similar to that of Theorem 12.

B.5. Proof of Theorem 18

The proof requires, first of all, that the class $\mathcal{BV}_n^{(m)}(V^*)$ is well-approximable by piecewise polynomial functions with degree at most $m - 1$. To this end we present the following result which was proved in Chatterjee and Goswami (2021a) (see Proposition 8.9 in the arxiv version).

Proposition 26 *Fix a positive integer* $m > 1$ *and* $\theta \in \mathbb{R}^n$, *and let* $V^m(\theta) := V$. *For any* $\delta > 0$, *there exists a partition* $P_{\theta,m,\delta}$ *(of* $L_{1,n}$*) in* $\mathcal{P}_{\mathrm{dp}}$ *and* $\theta' \in \Theta_{P_{\theta,m,\delta}}(\{F_S : S \in \mathcal{S}\})$ *such that*
a) $|P_{\theta,m,\delta}| \leq C \delta^{-1/m}$ *for an absolute constant* C,
b) $\|\theta - \theta'\|_\infty \leq V\delta$, *and*
c) $\max_{S \in P_{\theta,m,\delta}, 0 \leq j < m} n^j \beta_{j,S} \leq C_m \max_{0 \leq j < m} n^j \|D^j(\theta)\|_\infty$ *where* $\theta' \equiv \beta_{0,S} + \ldots + \beta_{m-1,S} x^{m-1}$ *on* S *and* C_m *is a constant depending only on* m.

Proposition 26 immediately gives us

$$\|\theta - \Pi_{\Theta_{P_{\theta,m,\delta}}} \theta\|^2 \leq NV^2\delta^2. \quad \text{(B.10)}$$

(recall that $N = n$ since $d = 1$).
Proof [Proof of Theorem 18] Given any $\delta > 0$, let θ'_δ denote the vector given by Proposition 26 for $\theta = \theta^*_\mathsf{T}$. Then from Proposition 25 and the item c) in Proposition 26, we get, for some constant C_m depending only on m,

$$\overline{\mathcal{R}}(\theta'_\delta, P_{\theta,m,\delta}) \leq C_m |P_{\theta,m,\delta}| \left(\|\theta^*_\mathsf{T}\|_{m-1,\infty}^2 + \mathbb{E}\|y_\infty\|^2 \log eN \right)$$
$$\leq C_m |P_{\theta,m,\delta}| \left(\|\theta^*_\mathsf{T}\|_{m-1,\infty}^2 + \sigma^2 \log eN \right) \log eN \quad \text{(B.11)}$$

where the last step follows from a similar computation as in (B.3) (observe that $\|\theta^*_\mathsf{T}\|_{m-1,\infty} \geq \|\theta^*_\mathsf{T}\|_\infty$). Now, since $\log |\mathcal{S}| = \log 2N$ (see around (4.1)), we get by plugging the bounds from (B.10), item (a) in Proposition 26 and (B.11) into (3.6) that for any $\lambda \geq C(\sigma\sqrt{\log N} \vee \|\theta^*_\mathsf{T}\|_\infty)$,

$$\mathbb{E}\|\hat{\theta}^{QL}_\mathsf{T} - \theta^*_\mathsf{T}\|^2 \leq \inf_{\delta > 0} C \left(|\mathsf{T}| (V^*_\mathsf{T})^2 \delta^2 + C_m \delta^{-1/m} (\lambda + \|\theta^*_\mathsf{T}\|_{m-1,\infty})^2 \log eN \right) + \frac{\sigma^2 + \lambda^2}{|\mathsf{T}|}$$

where $C_m > 0$ depends only on m. Now putting

$$\delta = C_m^{\frac{m}{2m+1}}(\lambda + \|\theta_{\mathsf{T}}^*\|_{m-1,\infty})^{\frac{2m}{2m+1}}(V_{\mathsf{T}}^*)^{-\frac{2m}{2m+1}}(\log eN)^{\frac{m}{2m+1}}|\mathsf{T}|^{-\frac{m}{2m+1}}$$

in the above display, we obtain (5.6). $\blacksquare$