Proofs for non-even design points

Anonymous Author(s)
Affiliation
Address
email

1 General principles

First, we cite a result that tells us that draws are roughly evenly spaced when they come from a distribution that is (mutually) continuous with respect to Lebesgue measure. We do not have control over the probability of the good event in this lemma.

Lemma 1.1 (Lemma 5 of Wang et al. [2014])**.** *Suppose p is a pdf with support in $[0, 1]$ and such that $p(x) \geq p_0 > p$. Let $x_1, ...x_n$ be a sorted list of iid draws from p. Then, with probability at least $1 - 2p_0 n^{-10}$, the maximum gap between two draws satisfies*

$$\max_{i>1} |x_i - x_{i-1}| \leq \frac{c \log n}{p_0 n}$$

where c is a universal constant.

We can then generalize the analysis for the bias of linear regression on a subset of our data. The following is a cleaner version of the computations in Appendix E.1. of the original submission.

Lemma 1.2 (Key lemma: "Lemma K")**.** *Suppose $x_1, ...x_n$ are sorted covariates such that $\max_{j=2,...n}(x_j - x_{j-1}) \leq O(\log n/n)$. Let $\theta_j := f(x_j)$, so that our data is $\{(x_i, \theta_i)\}_{i=1}^n$. Further, consider some subset $\{(x_i, \theta_i)\}_{i=r}^N$. Let $\hat{l}(z) = \hat{a} + \hat{b}x$ be the linear least squares fit trained on this subset. Then the error of $\hat{l}$ is given by*

$$\sum_{i=r}^{N} (\hat{l}(x_i) - \theta_i)^2 \leq O((N-r)^3 TV_1(\theta[r:N])^2/n^2)$$

Proof. WLOG suppose $r = 1$.

Define $\bar{a}$ to be equal to θ_1 and $\bar{b}$ to be $\frac{1}{N}\sum_{j=1}^{N} s_j$, where for $j > 1$ we let $s_j = \frac{\theta_j - \theta_{j-1}}{x_j - x_{j-1}}$ be the slope from the datapoint $j-1$ to j. We then have

$$\sum_{i=1}^{N} (\hat{a} + \hat{b}x_i - \theta_i)^2$$

$$\leq \sum_{i=1}^{N} (\bar{a} + \bar{b}x_i - \theta_1 - \sum_{k=2}^{i} s_k(x_k - x_{k-1})^2$$

$$= \sum_{i=1}^{N} (\bar{b}\sum_{k=2}^{i}(x_k - x_{k-1}) - \sum_{k=2}^{i}(x_k - x_{k-1})s_k)^2$$

$$= \sum_{i=1}^{N} (\sum_{k=2}^{i}(\bar{b} - s_k)(x_k - x_{k-1}))^2$$

$$\leq \sum_{i=1}^{N} \sum_{k=2}^{i} (\bar{b} - s_k)^2 \sum_{k=2}^{i} (x_k - x_{k-1})^2$$

$$\leq \sum_{i=1}^{N} NTV_1(\theta[1:R])^2 \times O(\frac{NR^2 \log^2 n}{n^2})$$

$$\leq \tilde{O}(N^3 TV_1(\theta[1:R])^2/n^2)$$

□

2 Extension of ADDLE

We now explain how these Lemmas allow us to extend ADDLE to handle uneven data points.

2.1 Theorem statement

First, we state a strict generalization of Theorem 6.1, where we assume that the spacing of the points can be somewhat uneven.

Theorem 2.1. *Let f be such that $TV_1(f) = C < \infty$. Consider sorted design points $0 \leq x_1, ...x_n \leq 1$ such that $\max_{j=2,...n} |x_j - x_{j-1}| \leq O(\frac{\log n}{n})$, and responses $\{y_t\}$ coming from the regression model. Let $\{\hat{y}_t\}_{t=1}^n$ be the the predictions generated by ADDLE, now with (clipped) VAW forcasters as experts. With probability $1 - \delta$, the total squared error satisfies:*

$$\sum_{t=1}^{n} (\hat{y}_t - f(x_t))^2 = \tilde{O}(n^{1/5} C^{2/5})$$

where $\tilde{O}$ hides constants (including σ) and polylog factors of n and δ.

By Lemma 1.1, we have the corollary that ADDLE is (near-)optimal for covariates coming from a continuous distribution.

Corollary 2.2. *Suppose the same setting as Theorem 2.1, except that $x_1, ..., x_n$ are sorted draws from a pdf p with support in $[0, 1]$ and such that $p(x) \geq p_0 > 0$. Then, with probability at least $1 - p_0 n^{-10} - \delta$, the error satisfies*

$$\sum_{t=1}^{n} (\hat{y}_t - f(x_t))^2 = \tilde{O}(n^{1/5} C^{2/5})$$

where $\tilde{O}$ hides constants (including σ) and polylog factors of n, p_0 and δ.

2.2 Proof steps

Change to linear VAW experts and analyze their error

We now consider as our expert a linear Vovk-Azoury-Warmuth (VAW) forcaster (Cesa-Bianchi and Lugosi [2006]) starting at time r and terminating at time s. This is a very minor change from the original linear regression experts, and does not affect computational or statistical efficiency.

The VAW expert is fed data $D_{s,r} := \{(x_j, \theta_j)\}_{j=r}^{s}$ in an online fashion, and produces estimates $\hat{w}_r^r, ...\hat{w}_s^r$. Let $\hat{l}(x_i) = \hat{u}^T z_i$ be the linear least squares estimate trained on $D_{s,r}$, where $z_i = [1, x_i]^T$. By Theorem 11.8 of Cesa-Bianchi and Lugosi [2006], we have:

$$\sum_{j=r}^{s} (\theta_j - \hat{w}_j^r)^2 \leq \sum_{j=r}^{s} (\theta_j - \hat{u}^T z_i)^2 + \frac{1}{2} \|\hat{u}\|_2^2$$

By "Lemma K" (Lemma 1.2) we have that the first term is bounded by $|r - s|^3 TV_1(\theta[r:s])^2/n^2$. By Corollary 40 in Baby and Wang [2020], the second term is $O(1)$.

Thus

$$\sum_{j=r}^{s} (\theta_j - \hat{w}_j^r)^2 \leq \sum_{j=r}^{s} (\theta_j - \hat{l}(x_j))^2 + O(1) \tag{1}$$

Oracle partition Construction proceeds in the same manner as before, where TV_1 of a bin is computed with respect to realized covariate spacing.

Wrapping up proof for ADDLE The aggregation algorithm is agnostic to the spacing of the covariates. Notice that we can reduce the error of ADDLE on an interval $[r, s]$ to the error of the (unclipped) expert that starts at r using the same argument as before (i.e., the proofs in Appendix D of the submission stand up to Equation (15)). In line 1155 of the submission, we instead use Equation 1 to get the following bound on the (unclipped) expert's error.

$$\sum_{j=r}^{s} (\hat{w}_j^r - \theta_j)^2 \leq \sum_{j=r}^{s} (\hat{l}_{r:s}(x_j) - \theta_j)^2 + O(1)$$

We then have

$$\sum_{j=r}^{s} (\hat{l}_{r:s}(x_j)^2 - \theta_j)^2 \leq \sigma^2 \sum_{j=r}^{s} x_j^T (X_{r:s} X_{r:s}^T)^{-1} x_j + \sum_{j=r}^{s} (l_{r:s}(x_j) - \theta_j)^2 \leq 2\sigma^2 + TV_1(\theta[r:s])^2 |r-s|^3 / n^2$$

where $l_{r:s} = \mathbb{E}[\hat{l}_{r:s}]$

We then conclude the proof of ADDLE in exactly the same way as in the submission. That is, we use the oracle partition to produce a set of intervals of size $O(n^{1/5} C^{2/5})$ together with experts who achieve constant error on each interval.

3 Extension of AKORN

All of the spline results of Appendix C go through without technical changes. Now that ADDLE has been generalized to the uneven covariate setting, Lemma C.1. also goes through by an application of Lemma "K" to the bias of the linear fits $\hat{a}_t$. The result is the theorem/corollary pair, which are analogous to the previous results for ADDLE:

Theorem 3.1. *Let f be such that $TV_1(f) = C < \infty$. Consider sorted design points $0 \leq x_1, ...x_n \leq 1$ such that $\max_{j=2,...n} |x_j - x_{j-1}| \leq O(\frac{\log n}{n})$, and responses $\{y_t\}$ coming from the regression model. Let $\hat{f}$ be the function returned by AKORN. Then, with probability $1 - \delta$, the average square error satisfies:*

$$\frac{1}{n} \sum_{t=1}^{n} (\hat{f}(x_t) - f(x_t))^2 = \tilde{O}(n^{-4/5} C^{2/5})$$

where $\tilde{O}$ hides constants (including σ) and polylog factors of n and δ.

Corollary 3.2. *Suppose now that $x_1, ..., x_n$ are sorted draws from a pdf p with support in $[0, 1]$ and such that $p(x) \geq p_0 > 0$. Then, with probability at least $1 - p_0 n^{-10} - \delta$, the error of $\hat{f}$ satisfies*

$$\sum_{t=1}^{n} (\hat{f}(x_t) - f(x_t))^2 = \tilde{O}(n^{1/5} C^{2/5})$$

where $\tilde{O}$ hides constants (including σ) and polylog factors of n, p_0 and δ.

References

Dheeraj Baby and Yu-Xiang Wang. Adaptive online estimation of piecewise polynomial trends, 2020. URL https://arxiv.org/abs/2010.00073.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006. ISBN 978-0-521-84108-5. doi: 10.1017/CBO9780511546921.

Yu-Xiang Wang, Alex Smola, and Ryan Tibshirani. The falling factorial basis and its statistical applications. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.