UNIVERSAL PIECEWISE LINEAR REGRESSION OF INDIVIDUAL SEQUENCES: LOWER BOUND

Georg C. Zeitler, Andrew C. Singer

Department of ECE, University of Illinois, Urbana, IL
Email: {zeitler2, acsinger}@uiuc.edu

Suleyman S. Kozat

IBM, Yorktown Heights, NY
Email: kozat@us.ibm.com

ABSTRACT

We consider universal piecewise linear regression of real valued bounded sequences under the squared loss function. In this setting, we present a lower bound on the regret of a universal sequential piecewise linear regressor compared to the best piecewise linear regressor that has access to the entire sequence in advance. This lower bound is tight in that it achieves the corresponding upper bound, suggesting a minmax optimality of the sequential regressor, for every individual bounded sequence.

Index Terms— Regression, piecewise linear approximation, prediction methods, universal, minimax methods

1. INTRODUCTION

Consider the problem of piecewise linear p-th order regression of an arbitrary real-valued sequence. Both the outcome sequence x^n and the observation sequence y^n are assumed to be deterministic individual sequences which are bounded such that $|x[t]| < A_x$ and $|y[t]| < A_y$ for all t. At each time instant t, after forming an estimate $\hat{x}[t]$ based on observations $\underline{y}[t] = [y[t], y[t-1], \ldots, y[t-p+1]]^{\mathrm{T}}$, one observes the t-th sample $x[t]$ of the sequence x^n. The accumulated loss of the regressor $\hat{x}[t]$ with respect to the sequence x^t up to time t is given by $l(x^t, \hat{x}^t) = \sum_{k=1}^{t}(x[k] - \hat{x}[k])^2$. This regressor is strongly sequential in the sense that at time t, it has only access to the observations $y[1], y[2], \ldots, y[t]$ up to time t and the past values of the outcome sequence, i.e., $x[1], x[2], \ldots, x[t-1]$. The goal of the sequential regressor is to perform almost as well as the best batch regressor knowing the entire sequence x^n in advance.

Sequential regression and prediction algorithms as well as upper and lower bounds on the regret of those algorithms are, e.g., described in the machine learning literature [1, 2, 3], the signal processing literature [4, 5, 6] and the information theory literature [7].

2. PIECEWISE LINEAR REGRESSION AND UPPER BOUND ON THE REGRET

Restriction to linear regression algorithms considerably limits the modeling power of the regressor. To generalize the class of regressors, we focus on piecewise linear regression, where we parse the past observation space $[-A_y, A_y]^p$ spanned by the observations into J fixed known regions $\mathcal{R}_j$ such that $\bigcup_{j=1}^{J} \mathcal{R}_j = [-A_y, A_y]^p$. Then, using a piecewise linear regression $\tilde{x}[t]$ from a sequential algorithm, we try to minimize

$$\sup_{x^n, y^n} \left\{ \sum_{t=1}^{n} (x[t] - \tilde{x}[t])^2 - \min_{\underline{w} \in \mathbb{R}^{Jp}} \sum_{t=1}^{n} (x[t] - \underline{w}_{s[t]}^{\mathrm{T}} \underline{y}[t])^2 \right\}, \tag{1}$$

where the state indicator variable $s[t] = j$ if $\underline{y}[t] = [y[t], y[t-1], \ldots, y[t-p+1]]^{\mathrm{T}} \in \mathcal{R}_j$. The vector $\underline{w} = [\underline{w}_1^{\mathrm{T}}, \underline{w}_2^{\mathrm{T}}, \ldots, \underline{w}_J^{\mathrm{T}}]^{\mathrm{T}}$ collects the J linear regression vectors $\underline{w}_j \in \mathcal{R}^p$.

In [6], a sequential piecewise linear regressor $\tilde{x}[t]$ is presented whose regret with respect to the best piecewise linear batch regressor of order p satisfies

$$\frac{1}{n}\sum_{t=1}^{n}(x[t] - \tilde{x}[t])^2 \leq \frac{1}{n} \min_{\underline{w}} \left\{ \sum_{t=1}^{n} (x[t] - \underline{w}_{s[t]}^{\mathrm{T}} \underline{y}[t])^2 + \delta \|\underline{w}\|_2^2 \right\} + \frac{pJA_x^2}{n} \ln\left(\frac{n}{J}\right) + O\left(\frac{1}{n}\right), \tag{2}$$

for any $x^n \in [-A_x, A_x]^n$, $y^n \in [-A_y, A_y]^n$ and $\delta > 0$. Defining J time vectors of length n_j, $t_j^{n_j} = \{t : s[t] = j\}$, and sequences $x_j^{n_j} = \{x[t_j[k]]\}_{k=1}^{n_j}$ and $\underline{y}_j^{n_j} = \{\underline{y}[t_j[k]]\}_{k=1}^{n_j}$, the regressor $\tilde{x}[t]$ achieving this bound is given by [6]

$$\tilde{x}[t] = \underline{\tilde{w}}_{s[t]}^{\mathrm{T}}[t-1]\underline{y}[t], \tag{3}$$

with

$$\underline{\tilde{w}}_j[t-1] = \left(\sum_{k=1}^{t} \underline{y}_j[k]\underline{y}_j^{\mathrm{T}}[k] + \delta_j I_p \right)^{-1} \sum_{k=1}^{t-1} \underline{y}_j[k] x_j[k], \tag{4}$$

where $\delta_j > 0$ is a positive constant and I_p the $p \times p$ identity matrix. Equation (2) states that there exists a sequential regressor given in Eq. (3) that can predict $x[t]$ as well as the best piecewise linear regressor with J regions whose p-th order regression vectors $\underline{w}_j$, $j = 1, 2, \ldots, J$, could have been selected based on observing the entire sequence x^n.

3. LOWER BOUND ON THE REGRET FOR SCALAR REGRESSION

In the following, we first focus on piecewise scalar regressors, i.e., $p = 1$, and derive a lower bound for

$$\inf_{q\in\mathcal{Q}} \sup_{x^n,y^n} \left\{ \sum_{t=1}^{n} (x[t] - \tilde{x}_q[t])^2 - \inf_{\underline{w}} \sum_{t=1}^{n} (x[t] - w_{s[t]} y[t])^2 \right\}, \tag{5}$$

where $\mathcal{Q}$ is the class of all sequential regressors. This lower bound coincides with the upper bound given in Eq. (2), suggesting that the regressor described in Eq. (3) is optimal in a sense that no sequential regressor can do much better, i.e., in a min-max sense. This is stated in the following theorem.

Theorem 1: *Let x^n and y^n be individual bounded sequences with $|x[t]| < A_x$ and $|y[t]| < A_y$, and let $\tilde{x}_q[t]$ form the output of a sequential regression algorithm. Then*

$$\inf_{q\in\mathcal{Q}} \sup_{x^n,y^n} \frac{1}{n} \left\{ \sum_{t=1}^{n} (x[t] - \tilde{x}_q[t])^2 - \inf_{\underline{w}} \sum_{t=1}^{n} (x[t] - w_{s[t]} y[t])^2 \right\} \geq \frac{JA_x^2(1-\epsilon)}{n} \ln\left(\frac{n}{J}\right) - \frac{G}{n} - O\left(\frac{1}{n^2}\right), \tag{6}$$

where $\mathcal{Q}$ is the class of all sequential regressors, for all $\epsilon > 0$ and a positive constant $G > 0$.

Hence, for every sequential regressor there exists a pair of sequences x^n and y^n such that the normalized accumulated regression error is at least $O(n^{-1}\ln(n))$ worse than that of the best batch regressor. This means that the regressor of Eq. (3) cannot be improved upon, in a min-max sense. The proof of the theorem is based on results in [3] for linear regression.

3.1. Proof of the Theorem

Defining $x_{\underline{w}}[t] = w_{s[t]} y[t]$, we have for any distribution on x^n and y^n that

$$\inf_{q\in\mathcal{Q}} \sup_{x^n,y^n} \left\{ l(x^n, \tilde{x}_q^n) - \inf_{\underline{w}} l(x^n, x_{\underline{w}}^n) \right\} \geq L(n), \tag{7}$$

where

$$L(n) \triangleq \inf_{q\in\mathcal{Q}} \mathrm{E}\left[l(x^n, \tilde{x}_q^n) \right] - \mathrm{E}\left[\inf_{\underline{w}} l(x^n, x_{\underline{w}}^n) \right]. \tag{8}$$

Hence, it is enough to lower bound $L(n)$ to find a lower bound on Eq. (5).We now consider the following distribution on x^n and y^n. The sequence y^n is constructed such that $s[t] = 1$ for the first n_1 points, $s[t] = 2$ for the next n_2 points up to the last n_J points, where $s[t] = J$. The constraint on n_j is that $\sum_{j=1}^{J} n_j = n$. Hence, we have J independent least squares problems in J regions, and we solve the computation of the lower bound for each region independently. In each region j, pick a random variable θ_j from a beta distribution with parameters (C_j, C_j), given by

$$p(\theta_j) = \frac{\Gamma(2\mathrm{C_j})}{\Gamma(C_j)\Gamma(C_j)} \theta_j^{C_j - 1} (1-\theta_j)^{C_j - 1}. \tag{9}$$

At time instant m, $m \in \{1, 2, \ldots, n_j\}$, let the observation $y_j[m] = y_j$ for all m, where y_j is such that $y_j[m] \in \mathcal{R}_j$. The signal $x_j[m]$ is generated such that $x_j[m] = A_x$ with probability $(1-\theta_j)$ and $x_j[m] = -A_x$ with probability θ_j, independently of the previous trials. Note that the outcome sequence is independent of the observation sequence, by construction.

3.2. Loss of Sequential Regressor

We note that the accumulated expected loss of the best sequential regressor is lower bounded by the loss of the best sequential estimator which under the squared error loss is given by the MMSE estimator

$$\hat{x}_{j,q}[m] = \mathrm{E}\left[x_j[m] | x_j^{m-1} \right] = \mathrm{E}\left[(1-2\theta_j) A_x | x_j^{m-1} \right], \tag{10}$$

Applying Bayes' rule to $p(\theta_j | x_j^{m-1})$ and using the properties of the beta distribution, we obtain that

$$\mathrm{E}\left[\theta_j | x_j^{m-1} \right] = \frac{m - 1 - N_a + C_j}{m - 1 + 2C_j}, \tag{11}$$

with N_a denoting the number of occurrences of A_x in the sequence x_j^{m-1}. Then, the best sequential estimator can be written as

$$\hat{x}_{j,q}[m] = \frac{\sum_{k=1}^{m-1} x_j[k]}{m - 1 + 2C_j}. \tag{12}$$

The expected loss of this estimator in region j can then be computed as

$$\sum_{m=1}^{n_j} \mathrm{E}\left[\left(x_j[m] - \frac{1}{m - 1 + 2C_j} \sum_{k=1}^{m-1} x_j[k] \right)^2 \right]. \tag{13}$$

Expanding the square, we find that

$$\mathrm{E}[x_j^2[m]] = \mathrm{E}[\mathrm{E}[x_j^2[m] | \theta_j]] = A_x^2. \tag{14}$$

Since given θ_j, $x_j[m]$ and $x_j[k]$, $1 \leq k \leq m-1$, are independent, we obtain that

$$\begin{aligned} \mathrm{E}\left[x_j[m] \sum_{k=1}^{m-1} x_j[k] \right] &= \mathrm{E}\left[\mathrm{E}\left[x_j[m] | \theta_j \right] \mathrm{E}\left[\sum_{k=1}^{m-1} x_j[k] \middle| \theta_j \right] \right] \\ &= A_x^2 (m-1) \mathrm{E}\left[(1-2\theta_j)^2 \right] \\ &= \frac{A_x^2 (m-1)}{2C_j + 1}. \end{aligned} \tag{15}$$

Finally, the second square term can be rewritten as

$$\mathrm{E}\left[\left(\sum_{k=1}^{m-1} x_j[k]\right)^2\right] = \mathrm{E}\Big[\mathrm{E}\left[(A_x(m-1) - 2A_x N_{m-1})^2 | \theta_j\right]\Big], \tag{16}$$

where N_{m-1} denotes the number of occurrences of $-A_x$ in the sequence x_j^{m-1}. Given θ_j, the variable N_{m-1} is a binomially distributed random variable with size $(m-1)$ and parameter θ_j. Then we can evaluate $\mathrm{E}[(\sum_{k=1}^{m-1} x_j[k])^2]$ as

$$\begin{aligned}\mathrm{E}\left[\left(\sum_{k=1}^{m-1} x_j[k]\right)^2\right] \\ = A_x^2\left(2(m-1)\frac{C_j}{2C_j+1} + (m-1)^2\frac{1}{2C_j+1}\right).\end{aligned} \tag{17}$$

Combining Eqs. (14) to (17) yields for the expected loss of the best sequential estimator in region j

$$\begin{aligned}&\sum_{m=1}^{n_j} \mathrm{E}\left[(x_j[m] - \hat{x}_{j,q}[m])^2\right] \\ &= \sum_{m=1}^{n_j}\Bigg\{A_x^2 - \frac{2A_x^2(m-1)}{(m-1+2C_j)(2C_j+1)} + \frac{A_x^2}{(m-1+2C_j)^2} \\ &\qquad \left(2(m-1)\frac{C_j}{2C_j+1} + (m-1)^2\frac{1}{2C_j+1}\right)\Bigg\}.\end{aligned} \tag{18}$$

3.3. Loss of Batch Regressor

We now proceed and compute the expected loss of the best batch regressor in region j which is given by

$$w_j^* = \left(\sum_{k=1}^{n_j} y_j[k] y_j^{\mathrm{T}}[k]\right)^{-1} \sum_{k=1}^{n_j} y_j[k] x_j[k]. \tag{19}$$

Exploiting the structure of y^n, the expression $w_j^* y_j[m]$ can be simplified to $w_j^* y_j[m] = 1/n_j \sum_{k=1}^{n_j} x_j[k]$. Now, the loss of the batch regressor in region j can be written as

$$\sum_{m=1}^{n_j} \mathrm{E}\left[\left(x_j[m] - \frac{1}{n_j}\sum_{k=1}^{n_j} x_j[k]\right)^2\right]. \tag{20}$$

As before, $\mathrm{E}[x_j^2[m]]$ is given by A_x^2. The expectation of the cross term of Eq. (20) can computed as

$$\begin{aligned}&\mathrm{E}\left[x_j[m]\sum_{k=1}^{n_j} x_j[k]\right] = \\ &= \mathrm{E}\left[\mathrm{E}\left[x_j^2[m]|\theta_j\right] + \mathrm{E}\left[x_j[m]|\theta_j\right]\mathrm{E}\left[\sum_{\substack{k=1\\k\neq m}}^{n_j} x_j[k]\middle|\theta_j\right]\right] \\ &= \mathrm{E}\left[A_x^2 + A_x^2(n_j-1)(1-2\theta_j)^2\right] = A_x^2 + A_x^2\frac{n_j-1}{2C_j+1}.\end{aligned} \tag{21, 22}$$

It remains to compute $\mathrm{E}\left[\left(\sum_{k=1}^{n_j} x_j[k]\right)^2\right]$, which can be rewritten using the variable N_{n_j} denoting the number of times that $x_j[m] = -A_x$ in the sequence $x_j^{n_j}$, as

$$\mathrm{E}\left[\left(\sum_{k=1}^{n_j} x_j[k]\right)^2\right] = \mathrm{E}\left[\mathrm{E}\left[\left(n_j A_x - 2N_{n_j}A_x\right)^2\middle|\theta_j\right]\right]. \tag{23}$$

Given θ_j, the distribution of N_{n_j} is binomial with size n_j and parameter θ_j, and we get that

$$\mathrm{E}\left[\left(\sum_{k=1}^{n_j} x_j[k]\right)^2\right] = A_x^2\left(2n_j\frac{C_j}{2C_j+1} + n_j^2\frac{1}{2C_j+1}\right). \tag{24}$$

Now, the loss of the batch regressor in region j can be written as

$$\sum_{m=1}^{n_j}\left\{A_x^2 - \frac{2A_x^2}{n_j}\left(1 + \frac{n_j-1}{2C_j+1}\right) + \frac{A_x^2}{n_j^2}\left(\frac{2n_jC_j + n_j^2}{2C_j+1}\right)\right\}. \tag{25}$$

We can now combine the results obtained in Eqs. (18) and (25) for the expected loss of the best sequential and batch regressor in each region to express $L(n)$, after some algebra, as

$$\begin{aligned}L(n) \geq A_x^2\sum_{j=1}^{J}\sum_{m=1}^{n_j}\Bigg\{&\frac{2C_j}{(2C_j+1)(m-1+2C_j)} \\ &+ \frac{2C_j}{n_j(2C_j+1)}\Bigg\}.\end{aligned} \tag{26}$$

The sum over m is lower bounded by its integral, and yields that

$$\begin{aligned}L(n) &\geq A_x^2\sum_{j=1}^{J}\frac{2C_j}{(2C_j+1)}\int_{m=0}^{n_j}\frac{1}{m-1+2C_j}\mathrm{d}m \\ &\geq A_x^2\frac{2C}{2C+1}\sum_{j=1}^{J}\ln(n_j) - G,\end{aligned}$$

by choosing $C_j = C \geq 1/2$ for all j and a constant $G = JA_x^2\frac{2C}{2C+1}\ln(2C-1)$. This lower bound is valid for all integer values n_j satisfying $\sum_{j=1}^{J} n_j = n$. We now let $n_j = \lfloor(n/J)\rfloor$ for $j = 1, 2, \ldots, J-1$, and $n_J = n-(J-1)\lfloor(n/J)\rfloor$. Since $(n/J)-1 \leq \lfloor(n/J)\rfloor \leq (n/J)$, application of Taylor's theorem to $\ln((n/J)-1)$ about (n/J) yields for the lower bound

$$L(n) \geq JA_x^2\frac{2C}{2C+1}\left(\ln\left(\frac{n}{J}\right) - \frac{1}{n-J}\right) - G \tag{27}$$

$$\geq JA_x^2\frac{2C}{2C+1}\ln\left(\frac{n}{J}\right) - G - \mathrm{O}\left(\frac{1}{n}\right). \tag{28}$$

Hence, for every $0 < \epsilon' \leq 1/2$, we can pick C large enough such that

$$L(n) \geq JA_x^2(1-\epsilon')\ln\left(\frac{n}{J}\right) - G - \mathrm{O}\left(\frac{1}{n}\right) \quad (29)$$

$$\geq JA_x^2(1-\epsilon)\ln\left(\frac{n}{J}\right) - G - \mathrm{O}\left(\frac{1}{n}\right), \quad (30)$$

for every $\epsilon > 0$. This concludes the proof of Theorem 1.

4. LOWER BOUND ON THE REGRET FOR P-TH ORDER REGRESSION

In this section, we extend the previous results to piecewise linear regression of order p. Now, we assume that the observation space $[-A_y, A_y]^p$ is partitioned in J disjoint known regions that are concentric around the origin. Then the following theorem holds.

Theorem 2: *Let x^n and y^n be individual bounded sequences with $|x[t]| < A_x$ and $|y[t]| < A_y$, and let $\tilde{x}_q[t]$ form the output of a sequential regression algorithm. Assume further that the regions $\mathcal{R}_j$, $j = 1, 2, \ldots, J$, are concentric around the origin. Then*

$$\inf_{q\in\mathcal{Q}} \sup_{x^n, y^n} \frac{1}{n}\left\{\sum_{t=1}^{n}(x[t]-\tilde{x}_q[t])^2 - \inf_{\underline{w}}\sum_{t=1}^{n}(x[t]-\underline{w}_{s[t]}^T\underline{y}[t])^2\right\}$$

$$\geq \frac{JA_x^2p(1-\epsilon)}{n}\ln\left(\frac{n}{J}\right) - \frac{G}{n} - O\left(\frac{1}{n^2}\right), \quad (31)$$

where $\mathcal{Q}$ is the class of all sequential regressors, for all $\epsilon > 0$ and a positive constant $G > 0$.

The proof of Theorem 2 follows from the proof of Theorem 1 by interleaving p independent subsequences in each region to generate x^n, and by constructing the sequence y^n such that the observation vector $\underline{y}_j[m] \in \mathbb{R}^p$ has its only entry $y_j \neq 0$ such that $\underline{y}_j[m] \in \mathcal{R}_j$.

5. SIMULATION RESULTS

In this section we present simulation results validating that there exists a pair of sequences x^n and y^n for which the regret between the normalized accumulated regression error of the piecewise linear regressor given in Eq. (3) and the batch regressor is at least as large as the bound given in Theorem 1, as shown in Fig. 1. These results where obtained for $p = 1$ by generating the sequences x^n and y^n as described in the proof of Theorem 1, for $J = 11$ uniform regions parsing the observation space. The bounds on the magnitude of the sequences are $A_x = 4$ and $A_y = 6$, and $\epsilon = 0.2$.

6. CONCLUSIONS

Establishing a tight lower bound on the regret of the best sequential regressor with respect to the regressor with access to the entire sequence in advance, we have shown that the piecewise linear regressor presented in [6] is optimal in a min-max sense, in that the regret of any sequential predictor cannot be much better.

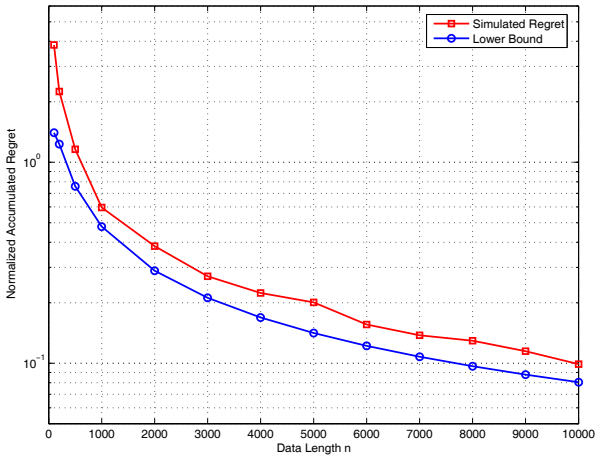

Fig. 1. Simulated regret and the corresponding lower bound

7. REFERENCES

[1] N. Cesa-Bianchi, P.M. Long, and M.K. Warmuth, "Worst-case quadratic loss bounds for prediction using linear functions and gradient descent," *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 604–619, 1996.

[2] V. Vovk, "Aggregating strategies," in *Proc. 3rd Annual Workshop Comp. Learning Theory*, 1990, pp. 371–383.

[3] V. Vovk, "Competitive online statistics," *International Statistical Review*, vol. 69, pp. 213–248, 2001.

[4] A.C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Information Theory*, vol. 47, no. 10, pp. 2685–2699, October 1999.

[5] A.C. Singer, S.S. Kozat, and M. Feder, "Universal linear least squares prediction: Upper and lower bounds," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2354–2362, August 2002.

[6] S.S. Kozat, A.C. Singer, and G.C. Zeitler, "Universal piecewise linear prediction via context trees," submitted to IEEE Transactions on Signal Processing.

[7] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1289–1292, July 1993.