Preprints of the
15th IFAC Symposium on System Identification
Saint-Malo, France, July 6-8, 2009

L_∞ Layers and the Probability of False Prediction ⋆

Simone Garatti * **Marco C. Campi** **

* *Dipartimento di Elettronica ed Informazione, Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italy. E-mail:* sgaratti@elet.polimi.it, *web-site:* http://www.elet.polimi.it/upload/sgaratti/.
** *Dipartimento di Elettronica per l'Automazione, Università di Brescia, via Branze 38, 25123 Brescia, Italy. E-mail:* marco.campi@ing.unibs.it, *web-site:* http://bsing.ing.unibs.it/∼campi/.

Abstract: In this paper, we consider the L_∞ criterion of best fit for identifying linear regression models from data. This criterion has much appeal as it protects against the worst since it selects the regression model that minimizes the largest deviation from observations. Moreover, based on the obtained optimal loss, one can build prediction intervals, as opposed to single prediction values, which are guaranteed to contain newly generated output values with high probability. Our main result is the exact quantification of the probability of making a false prediction valid in conditions of independent and stationary observations, a result that provides an exhaustive characterization of the confidence of prediction based on L_∞ regression models. It turns out that this false prediction probability is independent of the mechanism through which observations are generated. This result bears implications on the possibility of obtaining general and non-conservative evaluations for the probability of false prediction with no a-priori knowledge of the data generation mechanism properties.

Keywords: System identification; Prediction; Probability of false prediction; L_∞ regression.

1. INTRODUCTION

We consider linear regression models of a variable $y \in \mathbb{R}$ on a p-dimensional *explanatory* variable $x \in \mathbb{R}^p$. Precisely, given q *regressor* functions $f_j : \mathbb{R}^p \to \mathbb{R}$, $j = 1,\dots,q$, the linear regression model is given by

$$y(x) = \sum_{j=1}^{q} \beta_j f_j(x) = f(x)^T \beta, \tag{1}$$

where $f(x) := [f_1(x) \ \cdots \ f_q(x)]^T$ is the vector of regression functions and $\beta = [\beta_1 \ \cdots \ \beta_q]^T$ is the vector of tunable coefficients. As a simple example, (1) encompasses affine models in x, that is $y(x) = \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_p x^{(p)} + \beta_{p+1}$, where superscript (i) indicates the i-th component of vector x.

Given a batch of N independent observations (x_i, y_i), $i = 1,\dots,N$, the model is tuned according to the L_∞ criterion of best fit. This amounts to select the coefficients β_j so as to minimize the maximum deviation of the observed y_i's from $y(x_i)$, namely

$$\min_{\beta=[\beta_1,\ \cdots\ \beta_q]^T} \max_{i=1,\dots,N} \left| y_i - f(x_i)^T \beta \right|. \tag{2}$$

The optimal solution of (2) is indicated with $\beta^* = [\beta_1^* \ \cdots \ \beta_q^*]^T$ while the optimal cost value is h^*.

L_∞ regression has a long history which dates back to some works of Euler in the late 17th century, as described in Harter [1975]. Since the 1950s there has been a renewed interest spurred by the development of linear programming techniques to compute the L_∞ regression solution, see e.g. Karst [1958], Kelley [1958], Wagner [1959], Appa and Smith [1973], Barrodale and Phillips [1975], Armstrong and Kung [1979], Planitz and Gates [1991], Ruzinsky and Olsen [1989], Zhang [1993], Narula and Wellington [2007]; see also the monographs Birkes and Dodge [1993], Arthanari and Dodge [1993], Cheney [1999].

⋆ Paper supported by the MIUR national project "Identification and adaptive control of industrial systems".

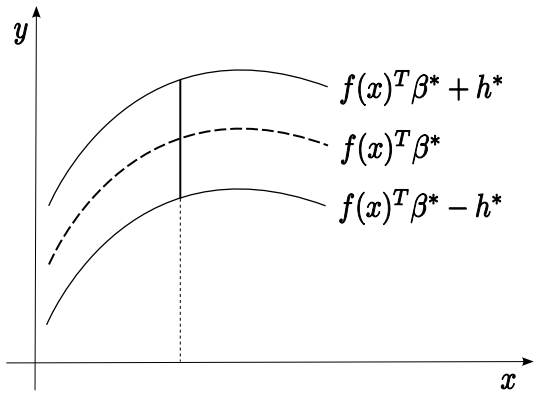

Figure 1. L_∞ layers for the construction of prediction intervals.

One reason of appeal of the L_∞ criterion is that it protects against the worst since it selects the regression model that minimizes the largest deviation from observations. The layer of vertical height $2h^*$ centered around the optimal solution $y = f(x)^T\beta^*$ (hereafter called the L_∞ *layer*) is the most thin layer containing all observations and it can be used to predict future output data, as graphically represented in Figure 1. To be precise, given a further observation of the explanatory variable x, the regressor model produces the prediction $f(x)^T\beta^*$ for the unknown value of variable y, while the L_∞ layer determines a confidence interval $[f(x)^T\beta^* - h^*, f(x)^T\beta^* + h^*]$ around the

predicted value which should contain y with high (guaranteed) probability.

For the sake of completeness, it must be said that, for this prediction to be really useful, the vertical dispersion of the observations around the curve of best fit must be moderate, since, as observed as early as in 1809 by Gauss, the optimal L_∞ regressor only depends on few observations and is quite sensitive to even sporadic noise (that is to the fact that other variables, besides those listed in x, can in some observations significantly contribute to form the value of the y variable). Likewise, the presence of significant noise even in few observations drastically impoverish the quality of the layer by making it wide and therefore little accurate. Despite these drawbacks, the minimax method can be an interesting alternative to least-squares in applications with moderate noise.

1.1 Results of this paper

The goal of this paper is to study the reliability of the returned prediction intervals by quantifying the probability of false prediction of the L_∞ layer, that is the probability that the next data point falls outside the layer. Our main result is that the false prediction probability is independent of the probability that generated the data, a result bearing important implications on the applicability of L_∞ layers.

In more formal terms, assume that observations (x_i, y_i) are independently generated according to an invariant probability **Pr** in $\mathbb{R}^p \times \mathbb{R}$ with density $p(x,y)$ and consider the L_∞ layer $\{(x,y) \text{ such that } \left|y - f(x)^T\beta^*\right| \le h^*\}$. The probability of false prediction is then the probability of the L_∞ layer complement:

$$\begin{aligned}\eta &:= \text{probability of false prediction}\\ &= \mathbf{Pr}\{(x,y) \text{ such that } \left|y - f(x)^T\beta^*\right| > h^*\},\end{aligned}$$

and this η is a random variable since the layer is, as it depends on the observations (x_i, y_i). Therefore, η is characterized by its probability distribution, and one would like this distribution to concentrate around zero in order for the L_∞ layer to exhibit good prediction properties. We here establishes the fact that the probability distribution of η is a *Beta* distribution, and that it is independent of **Pr** and of the choice of the regressor functions f_j.

Theory-wise, this result

1. establishes the fact that η is a *pivotal variable*, in the sense that its distribution remains unchanged under variation of the problem elements;
2. supports in a quantitatively exact manner the intuitive idea that η is small with high probability.

From a practical point of view, the distribution of η provides exact information on the false prediction properties of the L_∞ layer. Two aspects deserve to be further highlighted in this connection:

1. since the η distribution is pivotal, the fact that the user has no knowledge of **Pr**, that is no knowledge of the data generation mechanism, does not introduce any conservatism in the evaluation of the dependability of the L_∞ layer. The general applicability of the result is important, quoting Hogg [1974]: *"... we know in practice the most models will seldom fit exactly the real situations. Thus, for the sake of application, it seems ridiculous to try to get the last ounce of mathematical efficiency out of some assumed situation. A more realistic approach would be to seek statistical procedures good for a broad class of possible underlying models, but which are not necessarily best for any of them."*. The nice additional feature of the theory here developed is that generality is obtained at no cost, due to that the distribution is pivotal;
2. since the selection of the regressor functions f_j does not impact the false prediction probability, when choosing the f_j functions one should only keep in mind the objective of obtaining a L_∞ layer of small width, and any a-priori knowledge should be addressed to achieve this result.

More information on the practical use of the result will be provided in due course after the result is established.

2. MAIN RESULT

Before stating the main theorem, we establish some preliminary results of independent interest that are also instrumental to the theorem derivation.

2.1 Existence and uniqueness of β^*

The existence of β^* immediately follows by the observation that

$$\max_{i=1,\dots,N} \left|y_i - f(x_i)^T\beta\right|$$

is non-negative and piecewise-linear.

Uniqueness is more involved and requires the following natural condition.

Condition 1. The functions $f_j(x)$ are linearly independent on every set $A \subseteq \mathbb{R}^p$ of nonzero Lebesgue measure, i.e. for any $\overline{\beta} \in \mathbb{R}^q$, $\overline{\beta} \neq 0$, it holds that $f(x)^T\overline{\beta} \neq 0$ on every set $A \subseteq \mathbb{R}^p$ of nonzero Lebesgue measure. ∗

Condition 1 is verified for standard choices of regressor functions such as polynomial and trigonometric sums. It corresponds to requiring that none of the regressor functions is superfluous for the description of the relationship between x and y over a set A.

Under Condition 1, for any $\overline{\beta} \neq 0$, relation $f(x)^T\overline{\beta} = 0$ holds on a zero Lebesgue measure set only. Since (x,y) admits density, x also does, that is the marginal probability $\mathbf{Pr}_x$ of x is absolutely continuous with respect to the Lebesgue measure, so that the fact that $f(x)^T\overline{\beta} = 0$ holds only on a zero Lebesgue measure set implies that

$$\mathbf{Pr}_x\left\{f(x)^T\overline{\beta} = 0\right\} = 0, \quad \forall\overline{\beta} \in \mathbb{R}^q,\ \overline{\beta} \neq 0. \tag{3}$$

Uniqueness of the solution of (2) follows from (3), as established in the following proposition.

Proposition 1. Problem (2) with $N \ge q$ admits with probability 1 a unique solution if and only if (3) holds. ∗

Proof. Suppose that (3) holds. Then, the probability that the vector $f(x) = [f_1(x) \ \cdots \ f_q(x)]^T$ belongs to a given subspace of $\mathbb{R}^q$ of dimension less than q is zero. This in turn implies that, with probability one, for every choice of q different indexes $i_1, i_2, \dots, i_q$ from $1, \dots, N$, the vectors $f(x_{i_1})$, $f(x_{i_2})$, ..., $f(x_{i_q})$ are linearly independent. But this is the well-known Haar's condition (see Cheney [1999]) for the uniqueness of the solution,

so that the "if" part of the proposition is established.
Suppose instead that (3) does not hold, that is

$$\mathbf{Pr}_x\Big\{f(x)^T\overline{\beta}=0\Big\}>0$$

for some given $\overline{\beta}$. Then, there is a non-zero probability that all $x_1, x_2, \ldots, x_N$ are extracted where $f(x)^T\overline{\beta}=0$, in which case $f(x_1)^T, f(x_2)^T, \ldots, f(x_N)^T$ belong to a subspace of $\mathbb{R}^q$ of dimension less than q. If so, given a solution β^* of (2), we show that $\beta^*+\alpha\overline{\beta}$ is also a solution for every $\alpha\in\mathbb{R}$, that is the solution of (2) is not unique and the proof is complete.
To show that $\beta^*+\alpha\overline{\beta}$ is also a solution, simply note that

$$f(x_i)^T(\beta^*+\alpha\overline{\beta})=f(x_i)^T\beta^*+\alpha f(x_i)^T\overline{\beta}=f(x_i)^T\beta^*,$$

which means that the regressors obtained with $\beta^*+\alpha\overline{\beta}$ is the same as the regressor obtained with β^* and being the latter a solution, also the former must be. *

2.2 Probability distribution of η

We are now ready to state the main result of this paper.

Theorem 1. Let $N\geq q+1$ and assume that Condition 1 holds. Then, the probability distribution of η is

$$F_\eta(z):=\mathbf{Pr}^N\{\eta\leq z\}=\sum_{i=q+1}^{N}\binom{N}{i}z^i(1-z)^{N-i}.$$

Note that $F_\eta(z)$ does not depend on the probability $\mathbf{Pr}$ according to which data are generated, nor does it depend on the regression functions f_j used. *

In the theorem, $\mathbf{Pr}^N=\mathbf{Pr}\times\cdots\times\mathbf{Pr}$ refers to the product probability for the multi-extraction $(x_1,y_1),\ldots,(x_N,y_N)$. The proof is given in the next Section 3; we here proceed to a discussion on the significance the theorem.

In words, Theorem 1 says that η is a random variable with a *Beta* distribution with parameters $q+1$ and $N-q$, independently of the probability with which (x_i,y_i) are extracted and of the functional form of the regressors f_j. The property that

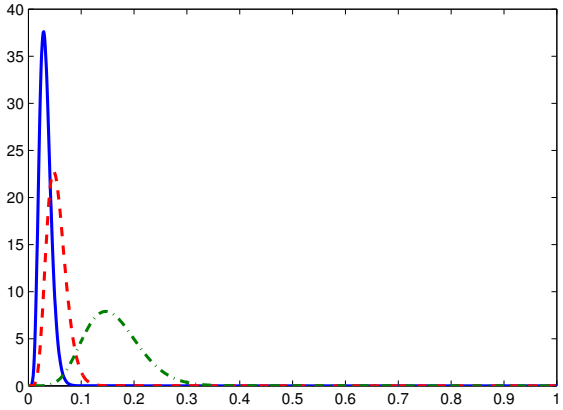

Figure 2. $f_\eta(z)$ for $q=7$ and $N=50$ (dash-dotted line), $N=150$ (dashed line), $N=250$ (continuous line).

the distribution of η does not depend on the distribution of the observations can be phrased by saying that "η is a pivotal random variable". The probability density of η is

$$f_\eta(z):=\frac{\mathrm{d}}{\mathrm{d}z}F_\eta(z)=(N-q)\binom{N}{q}z^q(1-z)^{N-q-1},$$

and it is graphically visualized for different values of N in Figure 2.

Using $f_\eta(z)$, or $F_\eta(z)$, permits one to exactly quantify the probability of false prediction of the L_∞ layer for any finite N and without any knowledge of the data generation mechanism. For an easier, though approximated, evaluation of the probability of false prediction, the Chernoff bound for the Beta tail, see Chernoff [1952], can be used:

$$\begin{aligned}\mathbf{Pr}^N\{\eta\leq z\}&=\sum_{i=q+1}^{N}\binom{N}{i}z^i(1-z)^{N-i}\\&=1-\sum_{i=0}^{q}\binom{N}{i}z^i(1-z)^{N-i}\\&\geq 1-e^{-2N(z-\frac{q}{N})^2},\end{aligned}$$

where the last inequality is in fact the Chernoff bound. An inspection of the last formula reveals that, for any fixed z, $\mathbf{Pr}^N\{\eta\leq z\}$ tends to 1 exponentially fast as N increases.

We end this section with an example that helps gain insight in the presented results.

2.3 An example

Let $y\in\mathbb{R}$ and $x\in\mathbb{R}$. $N=250$ points $(x_1,y_1),\ldots,(x_{250},y_{250})$ have been collected, by independent extractions in $\mathbb{R}^2$, according to an unknown probability density. The data are shown in Figure 3.

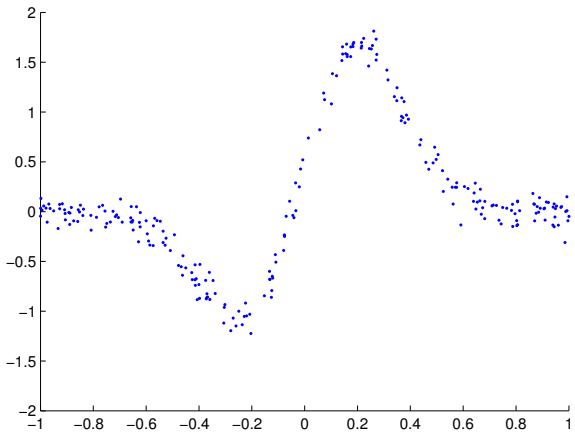

Figure 3. Observations (x_i,y_i).

A polynomial regressor $y(x)=\beta_1+\beta_2x+\ldots+\beta_7x^6$ is tuned according to the L_∞ criterion

$$\min_{\beta=[\beta_1\cdots\beta_7]^T}\max_{i=1,\ldots,250}\left|y_i-\beta_1-\beta_2x_i-\ldots-\beta_7x_i^6\right|,$$

and the corresponding L_∞ layer is obtained as shown in Figure 4.

How reliable is the claim that a next, still unseen, point will fall in the layer with probability at least 90%? This question is the same as asking for the probability that $\eta\leq 0.1$, and the answer can be found in Theorem 1: this probability is equal to $1-\sum_{i=8}^{250}\binom{250}{i}0.1^i(1-0.1)^{250-i}\approx 1-10^{-5}$. In other words, it is extremely likely that the obtained L_∞ layer contains at least 90% of the probability mass with which data are generated.

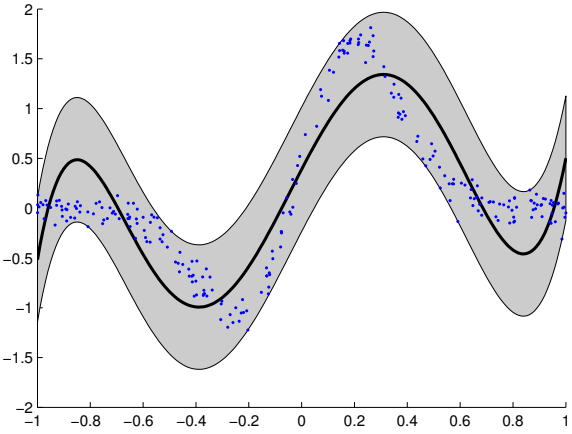

Figure 4. Polynomial regressor and corresponding L_∞ layer.

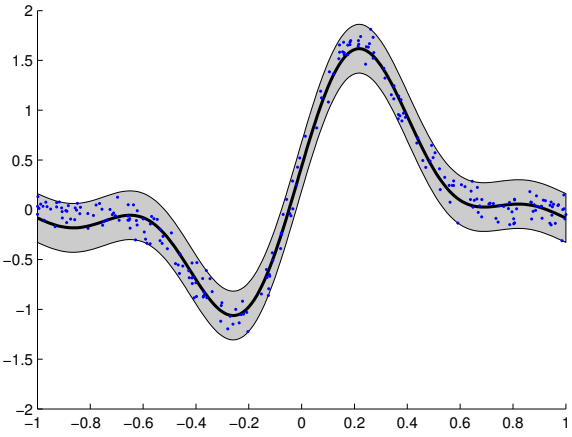

Figure 5. Trigonometric regressor and corresponding L_∞ layer.

Upon an inspection of Figure 4, it is apparent that the constructed layer is not tight around the data points, that is the data seem to have some pattern that is not captured by the model. Considering instead a trigonometric regressor, we can set out to solve the minimization problem

$$\min_{\beta=[\beta_1\ \cdots\ \beta_7]^T} \max_{i=1,\dots,250} \big|y_i-\beta_1^*-\beta_2^*\sin(\pi x_i)-\beta_3^*\cos(\pi x_i)- \cdots-\beta_6^*\sin(3\pi x_i)-\beta_7^*\cos(3\pi x_i)\big|.$$

The obtained layer is in Figure 5, and it appears to tightly fit the observations. As already noted, Theorem 1 holds irrespective of the chosen regression functions, so that we can still claim that the constructed layer contains at least 90% of the probability mass with confidence $1-10^{-5}$.

3. PROOF OF THE MAIN THEOREM

We need the following preliminary definition and lemma.

Definition 1. (point of support). We say that a data point

$$(x_\ell,y_\ell),\quad \ell\in\{1,2,\dots,N\},$$

is of *support* if

$$\min_\beta \max_{i=1,\dots,\ell-1,\ell+1,\dots,N}\left|y_i-f(x_i)^T\beta\right| < \min_\beta \max_{i=1,\dots,N}\left|y_i-f(x_i)^T\beta\right|.$$

In other words, a point is a point of support if its removal improves the solution. ∗

The points of support fully characterize the solution of the optimization problem, in the sense that the solution (β^*,h^*) of the original problem

$$\min_\beta \max_{i=1,\dots,N} |y_i-f(x_i)^T\beta| \tag{4}$$

is the same as the solution of problem

$$\min_\beta \max_{i=i_1,\dots,i_d} |y_i-f(x_i)^T\beta|, \tag{5}$$

where $(x_{i_1},y_{i_1}),\dots,(x_{i_d},y_{i_d})$ are the points of support. To see this, one can note that the removal of a point which is not of support from the initial set of points $(x_1,y_1),\dots,(x_N,y_N)$ does not change the solution; this is a tautological fact that descends from the very definition of point of support. Moreover, one can easily see that the points of support for the problem with the remaining $N-1$ data points are the same as the points of support for the original problem with N points. Proceeding iteratively and eliminating each time a point which is not of support, the conclusion is eventually drawn that the solution of the problem with only the points of support is the same as the solution of the original problem with N observations, i.e. (4) and (5) have the same solution.

To proceed, we need the following lemma.

Lemma 1. The number of points of support is almost surely equal to $q+1$. ∗

Proof. We first show that the number of points of support can be less than $q+1$ with probability zero only.
Consider the observations such that the number of points of support d is less than $q+1$. As we have just seen, the points of support determine the solution of the original problem, so that

$$h^*=\min_\beta \max_{i=i_1,\dots,i_d}\left|y_i-f(x_i)^T\beta\right|, \tag{6}$$

where $(x_{i_1},y_{i_1}),\dots,(x_{i_d},y_{i_d})$ are the points of support. But then, since β has at least as many components as there are points of support, equation (6) implies that $h^*=0$ whenever $f(x_{i_1}),\dots,f(x_{i_d})$ are linearly independent, a situation that occurs with probability one (see the proof of Proposition 1). On the other hand, h^* is also given by

$$h^*=\max_{i=1,\dots,N}|y_i-f(x_i)^T\beta^*|,$$

so that $h^*=0$ implies $y_i=f(x_i)^T\beta^*$ for all $i=1,\dots,N$. This means that all points $(f_1(x_i),\dots,f_q(x_i),y_i)$, $i\neq i_1,\dots,i_d$, belong to the proper subspace of $\mathbb{R}^{q+1}$ generated by the d points $(f_1(x_i),\dots,f_q(x_i),y_i)$, $i=i_1,\dots,i_d$, and this situation happens with probability zero since (x,y) have density and Condition 1 holds.
Hence, the number of points of support is less than $q+1$ with probability zero only.
Suppose now that the number of points of support is instead greater than $q+1$ and consider the following $N+1$ regions in $\mathbb{R}^{q+1}$:

$$F_i=\left\{(\beta,h)\in\mathbb{R}^{q+1}:\ |y_i-f(x_i)^T\beta|\le h\right\},\quad i=1,\dots,N.$$

and

$$F_{N+1}=\left\{(\beta,h)\in\mathbb{R}^{q+1}:\ h<h^*\right\}.$$

For any choice $\{i_1,i_2,\dots,i_{q+2}\}$ of $q+2$ indexes from the set $\{1,2,\dots,N+1\}$, we have that

$$\bigcap_{i=i_1,\dots,i_{q+2}} F_i\neq\emptyset. \tag{7}$$

Indeed, if $\{i_1, i_2, \ldots, i_{q+2}\} \in \{1,2,\ldots,N\}$, then (β^*, h^*) is a point in $\bigcap_{i=i_1,\ldots,i_{q+2}} F_i$ and hence (7) holds. Suppose instead that one of the indexes $i_1, \ldots, i_{q+2}$ is $N+1$, say $i_{q+2} = N+1$. Then, we certainly have

$$\min_{\beta} \max_{i=i_1,\ldots,i_{q+1}} \left| y_i - f(x_i)^T \beta \right| < h^*, \tag{8}$$

since at least one point of support is missing in the list of $q+1$ points with respect to which max is taken (recall that we have supposed that the number of points of support is greater than $q+1$). This means that $\bigcap_{i=i_1,\ldots,i_{q+1}} F_i$ contains a point $(\bar{\beta}, \bar{h})$ with $\bar{h} < h^*$. Thus, this point is also in F_{N+1} and (7) remains proven in this case too.
Since (7) holds and since all sets F_i, $i = 1,\ldots,N+1$ are convex, resorting to Helly's theorem (see Rockafellar [1970]) now yields

$$\bigcap_{i=1,\ldots,N+1} F_i \neq \emptyset.$$

Thus, we can find a point (β^{**}, h^{**}) which is simultaneously in all F_i, $i = 1,\ldots,N$, so that it satisfies $|y_i - f(x_i)^T \beta^{**}| \leq h^{**}$, $i = 1,\ldots,N+1$, and that is also in F_{N+1}, so that $h^{**} < h^*$. But then this (β^{**}, h^{**}) would outperform (β^*, h^*), the optimal solution, and this is a contradiction. This concludes the proof of the lemma. *

We are now ready to prove Theorem 1.

Let us start by computing the quantity:

$$\mu_k = \int_0^1 (1-z)^k F_\eta(\mathrm{d}z).$$

Recalling that $\eta := \mathbf{Pr}\left\{(x,y) : \left|y - f(x)^T \beta^*\right| > h^*\right\}$ and that the extractions are independent, μ_k is the probability that k further extracted observations fall inside the L_∞ layer determined by β^* and h^*. In other words, assuming that $N+k$ observations $(x_1,y_1),\ldots,(x_N,y_N),(x_{N+1},y_{N+1}),\ldots,(x_{N+k},y_{N+k})$ are extracted and letting β^* and h^* be the optimal solution for the first N observations , μ_k is given by

$$\begin{aligned}
&\mu_k \\
&= \mathbf{Pr}^{N+k}\left\{\left|y_i - f(x_i)^T \beta^*\right| \leq h^*, \text{ for all } i = N+1,\ldots,N+k\right\} \\
&= \mathbf{E}\left[\mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*\right| \leq h^*, \text{ for all } i=N+1,\ldots,N+k\right\}}\right],
\end{aligned}$$

where $\mathbf{E}$ denotes the expected value jointly over the N observations determining (β^*, h^*) and over the additional k observations, and $\mathbf{1}_A$ is the indicator function of set A.
Now, let $S = \{i_1,\ldots,i_k\}$ be a generic subset of k indexes from $\{1,2,\ldots,N+k\}$ and let $\mathscr{S}$ be the family of all possible choices of S ($\mathscr{S}$ contains $\binom{N+k}{k}$ elements). Moreover let $\overline{S} = \{1,2,\ldots,N+k\} - S$.
If we indicate by $\beta^*_{\overline{S}}$ and $h^*_{\overline{S}}$ the optimal solution and the optimal value of the problem

$$\min_{\beta} \max_{i \in \overline{S}} \left| y_i - f(x_i)^T \beta \right|,$$

then, owing to the independence of observations, we have that

$$\begin{aligned}
&\mathbf{E}\left[\mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*\right| \leq h^*, \text{ for all } i=N+1,\ldots,N+k\right\}}\right] \\
&= \mathbf{E}\left[\mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*_{\overline{S}}\right| \leq h^*_{\overline{S}}, \text{ for all } i \in S\right\}}\right], \ \forall S \in \mathscr{S}.
\end{aligned}$$

Whence,

$$\begin{aligned}
\mu_k &= \mathbf{E}\left[\mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*\right| \leq h^*, \text{ for all } i=N+1,\ldots,N+k\right\}}\right] \\
&= \frac{1}{\binom{N+k}{k}} \sum_{S \in \mathscr{S}} \mathbf{E}\left[\mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*_{\overline{S}}\right| \leq h^*_{\overline{S}}, \text{ for all } i \in S\right\}}\right] \\
&= \frac{1}{\binom{N+k}{k}} \mathbf{E}\left[\sum_{S \in \mathscr{S}} \mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*_{\overline{S}}\right| \leq h^*_{\overline{S}}, \text{ for all } i \in S\right\}}\right]. \qquad (9)
\end{aligned}$$

For a fixed multi-sample $(x_1,y_1),\ldots,(x_{N+k},y_{N+k})$, the quantity

$$\sum_{S \in \mathscr{S}} \mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*_{\overline{S}}\right| \leq h^*_{\overline{S}}, \text{ for all } i \in S\right\}}$$

counts the number of choices of S such that the L_∞ layer constructed on the observations with indexes in $\overline{S}$ contains all the remaining points in S. These S are those such that $(\beta^*_{\overline{S}}, h^*_{\overline{S}})$ is also the solution of the problem with all $N+k$ observations

$$\min_{\beta} \max_{i \in \{1,\ldots,N+k\}} \left| y_i - f(x_i)^T \beta \right|, \tag{10}$$

and this happens if and only if S does not contain any of the support points for the problem (10) (see Definition 1). Since in Lemma 1 we proved that the number of support points is almost surely equal to $q+1$ (note that the lemma holds irrespectively of the actual number of data points in the L_∞ regression problem), the number of those S is the same as the number of all possible choices of k indexes out of $N+k-q-1$, i.e. $\binom{N+k-q-1}{k}$, and we therefore have that

$$\sum_{S \in \mathscr{S}} \mathbf{1}_{\left\{\left|y_i - f(x_i)^T \beta^*_{\overline{S}}\right| \leq h^*_{\overline{S}}, \text{ for all } i \in S\right\}} = \binom{N+k-q-1}{k},$$

almost surely. By substituting this latter expression in (9) gives

$$\mu_k = \int_0^1 (1-z)^k F_\eta(\mathrm{d}z) = \frac{\binom{N+k-q-1}{k}}{\binom{N+k}{k}}, \quad k = 1,2,\ldots. \tag{11}$$

Expression

$$F_\eta(z) = \sum_{i=q+1}^{N} \binom{N}{i} z^i (1-z)^{N-i}$$

(which corresponds to $F_\eta(\mathrm{d}z) = (N-q)\binom{N}{q} z^q (1-z)^{N-q-1} \mathrm{d}z$) indeed satisfies (11), as it can be seen by an integration by parts. On the other hand, no other expressions $F_\eta(z)$ are admissible since determining an F_η satisfying (11) is a moment problem for a distribution with finite support (recall that η takes values in $[0,1]$) and its solution is unique (see e.g. Corollary 1, §12.9, Chapter II of Shiryaev [1996]).
Thus, it remains proven that $F_\eta(z) = \sum_{i=q+1}^{N} \binom{N}{i} z^i (1-z)^{N-i}$. □

REFERENCES

G. Appa and C. Smith. On L_1 and Chebyshev estimation. *Journal of Mathematical Programming*, 5:73–87, 1973.

R.D. Armstrong and D.S. Kung. Min-max estimates for a linear multiple regression problem. *Applied Statistics*, 28:93–100, 1979.

T.S. Arthanari and Y. Dodge. *Mathematical programming in statistics*. Wiley Classics Library. John Wiley and Sons, New York, NY, 1993.

I. Barrodale and C. Phillips. Solution of an over-determined system of linear equations in the Chebyshev norm. *ACM Transactions on Mathematical Software*, 1:264–270, 1975.

D. Birkes and Y. Dodge. *Alternative methods of regression*. John Wiley and Sons, New York, NY, 1993.

E.W. Cheney. *Introduction to Approximation Theory*. AMS, Providence, RI, 1999.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

H.L. Harter. The method of least squares and some alternatives – part iii. *International Statistical Reviews*, 43:1–44, 1975.

R.V. Hogg. Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association*, 69:909–923, 1974.

O.J. Karst. Linear curve fitting using least deviation. *Journal of the American Statistical Association*, 53:118–132, 1958.

J.E. Kelley. An application of linear programming to curve fitting. *Journal of industrial and applied mathematics*, 6: 15–22, 1958.

S.C. Narula and J.F. Wellington. Multiple criteria linear regression. *European Journal of Operational Research*, 181: 767–772, 2007.

M. Planitz and J. Gates. Strict discrete approximation in the L_1 and L_∞ norms. *Applied statistics*, 40:113–122, 1991.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

S. Ruzinsky and E. Olsen. L_1 and L_∞ minimization via a variant of Karmarkar's algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:245–253, 1989.

A.N. Shiryaev. *Probability*. Springer, New York, NY, USA, 1996.

H.M. Wagner. Linear programming techniques for regression analysis. *Journal of the American Statistical Association*, 54 (285):206–212, 1959.

Y. Zhang. Primal-dual interior point approach for computing L_1 solutions and L_∞ solutions of over-determined systems. *Journal of Optimization Theory and Applications*, 77:323–341, 1993.