Probability Bounds for Active Learning in the Regression Problem

A.-K. Fermin and C. Ludeña

Abstract In this contribution we consider the problem of active learning in the regression setting. That is, choosing an optimal sampling scheme for the regression problem simultaneously with that of model selection. We consider a batch type approach and an on-line approach adapting algorithms developed for the classification problem. Our main tools are concentration-type inequalities which allow us to bound the supreme of the deviations of the sampling scheme corrected by an appropriate weight function.

1 Introduction

Consider the following regression model

$$y_i = x_0(t_i) + \varepsilon_i,\ i = 1, \ldots, n \tag{1}$$

where the observation noise ε_i are i.i.d. realizations of a random variable ε.

The problem we consider in this chapter is that of estimating the real-valued function x_0 based on $t_1, \ldots, t_n$ and a subsample of size $N < n$ of the observations $y_1, \ldots, y_n$ measured at a well-chosen subsample of $t_1, \ldots, t_n$. This is relevant when, for example, obtaining the values of y_i for each sample point t_i is expensive or time consuming.

A.-K. Fermin (✉)
Université Paris Nanterre, laboratoire Modal'X, Nanterre, France
e-mail: aferminrodriguez@parisnanterre.fr

C. Ludeña
Universidad Jorge Tadeo Lozano, Dpto. de Ciencias Básicas y Modelado, Carrera, Bogotá, Colombia
e-mail: carennec.ludenac@utadeo.edu.co

P. Bertail et al. (eds.), *Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 250, https://doi.org/10.1007/978-3-319-96941-1_14

In this work we propose a statistical regularization approach for selecting a good subsample of the data in this regression setting by introducing a weighted sampling scheme (importance weighting) and an appropriate penalty function over the sampling choices.

We begin by establishing basic results for a fixed model, and then the problem of model selection and choosing a good sampling set simultaneously. This is what is known as active learning. We will develop two approaches. The first, a batch approach (see, for example, [7]), assumes the sampling set is chosen all at once, based on the minimization of a certain penalized loss function for the weighted sampling scheme. The second, an iterative approach [1], considers a two-step iterative method choosing alternatively the best new point to be sampled and the best model given the set of points.

The weighted sampling scheme requires each data point t_i to be sampled with a certain probability $p(t_i)$ which is assumed to be inferiorly bounded by a certain constant p_{min}. This constant plays an important role because it controls the expected sample size $\mathbb{E}(N) = \sum_{i=1}^{n} p(t_i) > np_{min}$. However, it also is inversely proportional to the obtained error terms in the batch procedure (see Theorems 2.1 and 2.2), so choosing p_{min} too small will lead to poor bounds. Thus essentially, the batch procedure aims at selecting the best subset of data points (points with high probability) for the user chosen error bound. In the iterative procedure this problem is addressed by considering a sequence of sampling probabilities $\{p_j\}$ where at each step j $p_j(t_i)$ is chosen to be as big as the greatest fluctuation for this data point over the hypothesis model for this step.

Following the active learning literature for the regression problem based on ordinary least squares (OLS) and weighted least squares learning (WLS) (see, for example [5–7] and the references therein) in this chapter we deal mainly with a linear regression setting and a quadratic loss function. This will be done by fixing a spanning family $\{\phi_j\}_{j=1}^{m}$ and considering the best L^2 approximation x_m of x_0 over this family. However, our approach is based on empirical error minimization techniques and can be readily extended to consider other models whenever bounds in probability are available for the error term.

Our results are based on concentration-type inequalities. Although variance minimization techniques for choosing appropriate subsamples are a well-known tool, giving adequate bounds in probability allowing for optimal non-asymptotic rates has been much less studied in the regression setting.

This is also true for the iterative procedure, where our results generalize previous ones obtained only in the classification setting for finite model spaces.

This chapter is organized as follows. In Sect. 2 we formulate the basic problem and study the batch approach for simultaneous sample and model selection. In Sect. 3 we study the iterative approach to sample selection and we discuss effective sample size reduction. All the proofs are available in the extended arXiv version [3].

2 Preliminaries

2.1 Basic Assumptions

We assume that the observations noise ε_i in (1) are i.i.d. realizations of a random variable ε satisfying the moment condition

MC Assume the r.v. ε satisfies $\mathbb{E}\varepsilon = 0$, $\mathbb{E}(|\varepsilon|^r/\sigma^r) \leq r!/2$ for all $r > 2$ and $\mathbb{E}(\varepsilon^2) = \sigma^2$.

It is important to stress that the observations depend on a fixed design $t_1, \ldots, t_n$. For this, we need some notation concerning this design. For any vectors u, v r, we define the normalized norm and the normalized scalar product by

$$\|u\|_{n,r}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i (u_i)^2, \quad \text{and} \quad < u, v >_{n,r} = \frac{1}{n}\sum_{i=1}^{n} r_i u_i v_i.$$

We drop the letter r from the notation when $r = 1$. With a slight abuse of notation, we will use the same notation when u, v, or r are functions by identifying each function (e.g. u) with the vector of values evaluates as t_i (e.g. $(u(t_1), \ldots, u(t_n))$. We also require the empirical max-norm $\|u\|_\infty = \max_i |u_i|$.

2.2 Discretization Scheme

To start with we will consider the approximation of function x_0 over a finite-dimensional subspace S_m. This subspace will be assumed to be linearly spanned by the set $\{\phi_j\}_{j \in \mathscr{I}_m} \subset \{\phi_j\}_{j \geq 1}$, with $\mathscr{I}_m$ a certain index set. Moreover, we shall, in general, be interested only in the vector $(x_0(t_i))_{i=1}^n$ which we shall typically denote just by x_0 stretching notation slightly.

We will assume the following properties hold:

AB There exists an increasing sequence c_m such that $\|\phi_j\|_\infty \leq c_m$ for $j \leq m$.

AQ There exist a certain density q and a positive constant Q such that $q(t_i) \leq Q, i = 1, \ldots, n$ and

$$\int \phi_l(t)\phi_k(t)q(t)\,dt = \delta_{k,l},$$

where δ is the Kronecker delta.

We will also require the following discrete approximation assumption. Let $G_m = [\phi_j(t_i)]_{i,j}$ be the associated empirical $n \times m$ Gram matrix. We assume that $G_m^t D_q G_m$ is invertible and moreover that $\frac{1}{n} G_m^t D_q G_m \to I_m$, where D_q is the diagonal matrix with entries $q(t_i)$, for $i = 1, \ldots, n$ and I_m the identity matrix of size m. More precisely, we will assume

AS There exist positive constants α and c, such that

$$\| I_m - \frac{1}{n} G_m^t D_q G_m \| \leq c n^{-1-\alpha}.$$

Given [AQ], assumption [AS] is a numerical approximation condition which is satisfied under certain regularity assumptions over q and $\{\phi_j\}$. To illustrate this condition we include the following example.

Example 2.1 Haar Wavelets: let $\phi(t) = \mathbf{1}_{[0,1]}(t)$, $\psi(t) = \phi(2t) - \phi(2t-1)$ (see, for example, [4]), with $q(t) = \mathbf{1}_{[0,1]}(t)$. Define

$$\phi_{j,k}(t) = 2^{j/2}\phi(2^j t - k)\,,\ t \in [0,1]\,,\ j \geq 0 \text{ and } k \in \mathbb{Z}\,;$$
$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)\,,\ t \in [0,1]\,,\ j \geq 0 \text{ and } k \in \mathbb{Z}\,.$$

For all $m \geq 0$, S_m denotes the linear space spanned by the functions $(\phi_{m,k}, k \in \mathbb{Z})$. In this case $c_m \leq 2^{m/2}$ and condition [AS] is satisfied for the discrete sample $t_i = i/2^m$, $i = 0, \cdots, 2^{m-1}$.

We will denote by $\hat{x}_m \in S_m$ the function that minimizes the weighted norm $\|x - y\|_{n,q}^2$ over S_m evaluated at points $t_1, \ldots, t_n$. This is,

$$\hat{x}_m = \arg\min_{x \in S_m} \frac{1}{n} \sum_{i=1}^{n} q(t_i)(y_i - x(t_i))^2 = R_m y,$$

with $R_m = G_m (G_m^t D_q G_m)^{-1} G_m^t D_q$ the orthogonal projector over S_m in the q-empirical norm $\| \cdot \|_{n,q}$.

Let $x_m := R_m x_0$ be the projection of x_0 over S_m in the q-empirical norm $\| \cdot \|_{n,q}$, evaluated at points $t_1, \ldots, t_n$. Our goal is to choose a good subsample of the data collection such that the estimator of the unobservable vector x_0 in the finite-dimensional subspace S_m, based on this subsample, attains near optimal error bounds. For this we must introduce the notion of subsampling scheme and importance weighted approaches (see [1, 7]), which we discuss below.

2.3 Sampling Scheme and Importance Weighting

In order to sample the data set we will introduce a sampling probability $p(t)$ and a sequence of Bernoulli$(p(t_i))$ random variables w_i, $i = 1, \ldots, n$ independent of ε_i

with $p(t_i) > p_{\min}$. Let $D_{w,q,p}$ be the diagonal matrix with entries $q(t_i)w_i/p(t_i)$. So that $\mathbb{E}\left(D_{w,q,p}\right) = D_q$. Sometimes it will be more convenient to rewrite $w_i = \mathbf{1}_{u_i < p(t_i)}$ for $\{u_i\}_i$ an i.i.d. sample of uniform random variables, independent of $\{\varepsilon_i\}_i$ in order to stress the dependence on p of the random variables w_i.

The next step is to construct an estimator for $x_m = R_m x_0$, based on the observation vector y and the sampling scheme p. For this, we consider a modified version of the estimator $\hat{x}_m$.

Consider a uniform random sample $u_1, \ldots, u_n$ and let $w_i = w_i(p) = \mathbf{1}_{u_i < p(t_i)}$ for a given p. For the given realization of $u_1, \ldots, u_n$, $D_{w,q,p}$ will be strictly positive for those $w_i = 1$. Moreover, as follows from the singular value decomposition, the matrix $(G_m^t D_{w,q,p} G_m)$ is invertible as long as at least one $w_i \neq 0$. Set $R_{m,p} = G_m(G_m^t D_{w,q,p} G_m)^{-1} G_m^t D_{w,q,p}$. Then $R_{m,p}$ is the orthogonal projector over S_m in the wq/p-empirical norm $\|\cdot\|_{n,wq/p}$ and it is well defined if at least one $w_i \neq 0$. If all $w_i = 0$, the projection is defined to be 0.

As the approximation of x_m, we then consider (for a fixed m, p and $(u_1, \ldots, u_n)$) the random quantity

$$\hat{x}_{m,p} = \arg\min_{x \in S_m} \|x - y\|^2_{n,\frac{qw}{p}} = \arg\min_{x \in S_m} \frac{1}{n} \sum_{i=1}^{n} \frac{w_i}{p(t_i)} q(t_i)(y_i - x(t_i))^2.$$

Note that

$$\hat{x}_{m,p} = R_{m,p} y, \tag{2}$$

This estimator depends on y_i only if $w_i = 1$. However, as stated above, this depends on $p(t_i)$ for the given probability p.

2.4 Choosing a Good Sampling Scheme

To begin with, given n, we will assume that S_m is fixed with dimension $|\mathscr{I}_m| = d_m$ and $d_m = o(n)$. Remark that the bias $\|x_0 - x_m\|^2_{n,q}$ is independent of p so for our purposes it is only necessary to study the approximation error $\|x_m - \hat{x}_{m,p}\|^2_{n,q}$ which does depend on how p is chosen.

Let $\mathscr{P} := \{p_k, k \geq 1\}$ be a numerable collection of $[0, 1]$ valued functions over $\{t_1, \ldots, t_n\}$. Set $p_{k,min} = \min_i p_k(t_i)$. We will assume that $\min_k p_{k,min} > p_{\min}$. The way the candidate probabilities are ordered is not a major issue, although in practice it is sometimes convenient to incorporate prior knowledge (certain sample points are known to be needed in the sample, for example) letting favourite candidates appear first in the order. To get the idea of what a sampling scheme may be, consider the following toy example:

Example 2.2 Let $\Pi = \{0.1, 0.4, 0.6, 0.9\}$ and set $\mathscr{P} = \{p, p(t_i) = \pi_j \in \Pi, i = 1, \ldots, n\}$ which is a set of $|\Pi|^n$ functions. In this example, any given p will tend to favour the appearance of points t_i with $p(t_i) = 0.9$ and disfavour the appearance of those t_i with $p(t_i) = 0.1$.

A good sampling scheme p, based on the data, should be the minimizer over $\mathscr{P}$ of the non-observable quantity $\|x_m - \hat{x}_{m,p}\|_{n,q}^2$. In order to find a reasonable observable equivalent we start by writing,

$$\begin{aligned}[\hat{x}_{m,p} - x_m] &= R_{m,p}[x_0 - x_m] + R_{m,p}\varepsilon \\ &= \mathbb{E}\left(R_{m,p}\right)[x_0 - x_m] + (R_{m,p} - \mathbb{E}\left(R_{m,p}\right))[x_0 - x_m] + R_{m,p}\varepsilon. \end{aligned} \tag{3}$$

Consider first the deterministic term $\mathbb{E}\left(R_{m,p}\right)[x_0 - x_m]$ in (3). We have the next lemma which is proved in the extended arXiv version.

Lemma 2.1 *Under condition [AS] if $m = o(n)$, then*

$$\|\mathbb{E}\left(R_{m,p}\right)[x_0 - x_m]\|_{n,q} = O\left(\frac{n^{-1-\alpha}\|x_0 - x_m\|_{n,q}}{p_{min}}\right).$$

From Lemma 2.1, we can derive that the deterministic term is small with respect to the other terms. Thus, it is sufficient for a good sampling scheme to take into account the second and third terms in (3). We propose to use an upper bound with high probability of those two last terms as in a penalized estimation scheme and to base our choice on this bound.

Define

$$\widetilde{B}_1(m, p_k, \delta) = \|x_0 - x_m\|_{n,q}^2 (\widetilde{\beta}_{m,k}(1 + \widetilde{\beta}_{m,k}^{1/2}))^2 \tag{4}$$

with

$$\widetilde{\beta}_{m,k} = \frac{c_m(\sqrt{17} + 1)}{2}\sqrt{\frac{d_m Q}{n p_{k,\min}}}\sqrt{2\log(2^{7/4} d_m k(k+1)/\delta)}. \tag{5}$$

The second square root appearing in the definition of $\widetilde{\beta}_{m,k}$ is included in order to give uniform bounds over the numerable collection $\mathscr{P}$.

In the following, the expression $\mathrm{tr}(A)$ stands for the trace of the matrix A. Set $T_{m,p_k} = \mathrm{tr}((R_{m,p_k} D_q^{1/2})^t R_{m,p_k} D_q^{1/2})$ and define

$$\widetilde{B}_2(m, p_k, \delta) = \sigma^2 r(1 + \theta_k)\frac{T_{m,p_k} + Q}{n} + \sigma^2 Q \frac{log^2(2/\delta)}{dn}, \tag{6}$$

with $r > 1$ and $d = d(r) < 1$ a positive constant that depends on r. The sequence $\theta_k \geq 0$ is such that $\sum_k e^{-\sqrt{dr\theta_k(d_m+1)}} < 1$ holds.

It is thus reasonable to consider the best p as the minimizer

$$\hat{p} = \underset{p_k \in \mathscr{P}}{\operatorname{argmin}} \widetilde{B}(m, p_k, \delta, \gamma, n), \tag{7}$$

where, for a given $0 < \gamma < 1$,

$$\widetilde{B}(m, p_k, \delta, \gamma, n) = \{(1+\gamma)\widetilde{B}_1(m, p_k, \delta) + (1+1/\gamma)\widetilde{B}_2(m, p_k, \delta)\}.$$

The different roles of $\widetilde{B}_1$ and $\widetilde{B}_2$ appear in the following lemmas:

Lemma 2.2 *Assume that the conditions [AB], [AS], and [AQ] are satisfied and that there is a constant $p_{min} > 0$ such that for all $i = 1, \ldots, n$, $p(t_i) > p_{k,\min} > p_{\min}$. Assume $\widetilde{B}_1$ to be selected according to (4). Then for all $\delta > 0$ we have*

$$P\left[\sup_{\mathscr{P}}\{\|(R_{m,p} - \mathbb{E}\left(R_{m,p}\right))[x_0 - x_m]\|_{n,q}^2 - \widetilde{B}_1(m, p, \delta)\} > 0\right] \leq \delta/2$$

Lemma 2.3 *Assume the observation noise in Eq. (1) is an i.i.d. collection of random variables satisfying the moment condition [MC]. Assume that the condition [AQ] is satisfied and assume that there is a constant $p_{min} > 0$ such that $p(t_i) > p_{min}$ for all $i = 1, \ldots, n$. Assume $\widetilde{B}_2$ to be selected according to (6) with $r > 1$, $d = d(r)$ and $\theta_k \geq 0$, such that the following Kraft inequality $\sum_k e^{-\sqrt{dr\theta_k(m+1)}} < 1$ holds. Then,*

$$P(\sup_{\mathscr{P}}\{\|R_{m,p}\varepsilon\|_{n,q}^2 - \widetilde{B}_2(m, p, \delta)\} > 0) < \delta/2.$$

Those two lemmas together with Lemma 2.1 assure that the proposed estimation procedure, based on the minimization of $\widetilde{B}$, is consistent establishing non-asymptotic rates in probability.

We may now state the main result of this section, namely, non-asymptotic consistency rates in probability of the proposed estimation procedure. The proof follows from Lemmas 2.2 and 2.3 and is given in the extended arXiv version along with the proof of the lemmas.

Theorem 2.1 *Assume that the conditions [AB], [AS], and [AQ] are satisfied. Assume $\hat{p}$ to be selected according to (7). Then the following inequality holds with probability greater than $1 - \delta$*

$$\|x_m - \hat{x}_{m,\hat{p}}\|_{n,q}^2 \leq \inf_{p \in \mathscr{P}} 6\left(\|\mathbb{E}\left(R_{m,p}\right)(x_m - x_0)\|_{n,q}^2 + \widetilde{B}(m, p, \delta, \gamma, n)\right).$$

Remark 2.1 In the minimization scheme given above it is not necessary to know the term $\|x_0 - x_m\|_{n,q}^2$ in $\widetilde{B}_1$ as this term is constant with regard to the sampling scheme p. Including this term in the definition of $\widetilde{B}_1$, however, is important because it leads

to optimal bounds in the sense that it balances p_{min} with the mean variation, over the sample points, of the best possible solution x_m over the hypothesis model set S_m. This idea shall be pursued in depth in Sect. 3.

Moreover, minimizing $\widetilde{B}_1$ essentially just requires selecting k such that $p_{k,\min}$ is largest and doesn't intervene at all if $p_{k,\min} = p_{min}$ for all k. Minimization based on $p_k(t_i)$ for all sample points is given by the trace T_{m,p_k} which depends on the initial random sample u independent of $\{(t_i, y_i), i = 1, \ldots, n\}$. A reasonable strategy in practice, although we do not have theoretical results for it, is to consider several realizations of u and select sample points which appear more often in the selected sampling scheme $\hat{p}$.

Remark 2.2 Albeit the appearance of weight terms which depend on k both in the definition of $\widetilde{B}_1$ and $\widetilde{B}_2$, actually the ordering of $\mathscr{P}$ does not play a major role. The weights are given in order to assure convergence over the numerable collection $\mathscr{P}$. Thus in the definition of $\widetilde{\beta}_{m,k}$ any sequence of weights θ'_k (instead of $[k(k+1)]^{-1}$) assuring that the series $\sum_k \theta'_k < \infty$ is valid. Of course, in practice $\mathscr{P}$ is finite. Hence for $M = |\mathscr{P}|$ a more reasonable bound is just to consider uniform weights $\theta'_k = 1/M$ instead.

Remark 2.3 Setting $H_{m,p_k} := (G_m^t D_{w,q,p_k} G_m)^{-1} G_m^t D_{w,q,p_k}$ we may write $T_{m,p_k} = \mathrm{tr}(G_m^t D_q G_m H_{m,p_k} H_{m,p_k}^t)$ in the definition of B_2. Thus our convergence rates are as in Lemma 1, [5]. Our approach, however, provides non-asymptotic bounds in probability as opposed to asymptotic bounds for the quadratic estimation error.

Remark 2.4 As mentioned at the beginning of this section, the expected "best" sample size given u is $\hat{N} = \sum_i \hat{p}(t_i)$, where u is the initial random sample independent of $\{(t_i, y_i), i = 1, \ldots, n\}$. Of course, a uniform inferior bound for this expected sample size is $\mathbb{E}\left(\hat{N}\right) > np_{min}$, so that the expected size is inversely proportional to the user chosen estimation error. In practice, considering several realizations of the initial random sample provides an empirical estimator of the non-conditional "best" expected sample size.

2.5 Model Selection and Active Learning

Given a model and n observations $(t_1, y_1), \ldots, (t_n, y_n)$ we know how to estimate the best sampling scheme $\hat{p}$ and to obtain the estimator $\hat{x}_{m,\hat{p}}$. The problem is that the model m might not be a good one. Instead of just looking at *fixed* m we would like to consider simultaneous model selection as in [7]. For this we shall pursue a more global approach based on loss functions.

We start by introducing some notation. Set $l(u, v) = (u - v)^2$ the squared loss and let $L_n(x, y, p) = \frac{1}{n} \sum_{i=1}^n q(t_i) \frac{w_i}{p(t_i)} l(x(t_i), y_i)$ be the empirical loss function for the quadratic difference with the given sampling distribution. Set $L(x) := \mathbb{E}\left(L_n(x, y, p)\right)$ with the expectation taken over all the random variables involved.

Let $L_n(x, p) := \mathbb{E}_\varepsilon (L_n(x, y, p))$ where $\mathbb{E}_\varepsilon ()$ stands for the conditional expectation given the initial random sample u, that is the expectation with respect to the random noise ε. It is not hard to see that

$$L(x) = \frac{1}{n} \sum_{i=1}^{n} q(t_i) \mathbb{E} (l(x(t_i), y_i)) ,$$

and

$$L_n(x, p) = \frac{1}{n} \sum_{i=1}^{n} q(t_i) \frac{w_i}{p(t_i)} \mathbb{E} (l(x(t_i), y_i)) .$$

Recall that $\hat{x}_{m,p} = R_{m,p} y$ is the minimizer of $L_n(x, y, p)$ over each S_m for given p and that $x_m = R_m x_0$ is the minimizer of $L(x)$ over S_m. Our problem is then to find the best approximation of the target x_0 over the function space $S_0 := \bigcup_{m \in \mathscr{J}} S_m$. In the notation of Sect. 2.2 we assume for each m that S_m is a bounded subset of the linearly spanned space of the collection $\{\phi_j\}_{j \in I_m}$ with $|\mathscr{I}_m| = d_m$.

Unlike the fixed m setting, model selection requires controlling not only the variance term $\|x_m - \hat{x}_{m,p}\|_{n,q}$ but also the unobservable bias term $\|x_0 - x_m\|^2_{n,q}$ for each possible model S_m. If all samples were available this would be possible just by looking at $L_n(x, y, p)$ for all S_m and p, but in the active learning setting labels are expensive.

Set $e_m := \|x_0 - x_m\|_\infty$. In what follows we will assume that there exists a positive constant C such that $\sup_m e_m \le C$. Remark this implies $\sup_m \|x_0 - x_m\|_{n,q} \le QC$, with Q defined in [AQ].

As above $p_k \in \mathscr{P}$ stands for the set of candidate sampling probabilities and $p_{k,\min} = \min_i (p_k(t_i))$.

Define

$$pen_0(m, p_k, \delta) = \frac{QC^2}{p_{k,\min}} \sqrt{\frac{1}{2n} \ln(\frac{6d_m(d_m + 1)}{\delta})}, \tag{8}$$

$$pen_1(m, p_k, \delta) = QC\beta^2_{m,k}(1 + \beta^{1/2}_{m,k})^2, \tag{9}$$

with

$$\beta_{m,k} = \frac{c_m(\sqrt{17} + 1)}{2} \sqrt{\frac{d_m Q}{n p_{k,\min}}} \sqrt{2 \log(\frac{3 * 2^{7/4} d^2_m (d_m + 1)k(k + 1)}{\delta})},$$

and finally setting $T_{p_k,m} = \mathrm{tr}((R_{m,p_k} D_q^{1/2})^t R_{m,p_k} D_q^{1/2})$, define

$$pen_2(m, p_k, \delta) = \sigma^2 \left\{ r(1 + \theta_{m,k}) \frac{T_{p_k,m} + Q}{n} + \frac{Q \ln^2(6/\delta)}{dn} \right\} \tag{10}$$

where $\theta_{m,k} \geq 0$ is a sequence such that $\sum_{m,k} e^{-\sqrt{dr\theta_{m,k}(d_m+1)}} < 1$ holds.

We remark that the change from δ to $\delta/(d_m(d_m+1))$ in pen_0 and pen_1 is required in order to account for the supremum over the collection of possible model spaces S_m.

Also, we remark that introducing simultaneous model and sample selection results in the inclusion of term $pen_0 \sim C^2/p_{k,\min}\sqrt{1/n}$ which includes an L_∞ type bound instead of an L_2 type norm which may yield non-optimal bounds. Dealing more efficiently with this term would require knowing the (unobservable) bias term $\|x_0 - x_m\|_{n,q}$. A reasonable strategy is selecting $p_{k,min} = p_{k,min}(m) \geq \|x_0 - x_m\|_{n,q}$ whenever this information is available.

In practice, $p_{k,min}$ can be estimated for each model m using a previously estimated empirical error over a subsample if this is possible. However this yields a conservative choice of the bound. One way to avoid this inconvenience is to consider iterative procedures, which update on the unobservable bias term. This course of action shall be pursued in Sect. 3.

With these definitions, for a given $0 < \gamma < 1$ set

$$pen(m, p, \delta, \gamma, n) = 2p_0(m, p, \delta) + (\frac{1}{p_{\min}} + \frac{1}{\gamma})pen_1(m, p, \delta)$$
$$+(\frac{1}{p_{\min}^2}(\frac{2}{\gamma} + 1) + \frac{1}{\gamma})pen_2(m, p, \delta) + 2((c+1)\frac{n^{-(1+\alpha)}QC}{p_{min}})^2.$$

and define

$$L_{n,1}(x, y, p) = L_n(x, y, p) + pen(m, p, \delta, \gamma, n).$$

The appropriate choice of an optimal sampling scheme simultaneously with that of model selection is a difficult problem. We would like to choose simultaneously m and p, based on the data in such a way that optimal rates are maintained. We propose for this a penalized version of $\hat{x}_{m,\hat{p}}$, defined as follows.

We start by choosing, for each m, the best sampling scheme

$$\hat{p}(m) = \arg\min_p pen(m, p, \delta, \gamma, n), \tag{11}$$

computable before observing the output values $\{y_i\}_{i=1}^n$, and then calculate the estimator $\hat{x}_{m,\hat{p}(m)} = R_{m,\hat{p}(m)}y$ which was defined in (2).

Finally, choose the best model as

$$\hat{m} = \arg\min_m L_{n,1}(y, \hat{x}_{m,\hat{p}(m)}, \hat{p}(m)). \tag{12}$$

The penalized estimator is then $\hat{x}_{\hat{m}} := \hat{x}_{\hat{m},\hat{p}(\hat{m})}$. It is important to remark that for each model m, $\hat{p}(m)$ is independent of y and hence of the random observation error structure. The following result assures the consistency of the proposed estimation

procedure, although the obtained rates are not optimal as observed at the beginning of this section.

Theorem 2.2 *With probability greater than* $1-\delta$*, we have*

$$\begin{aligned} L(\hat{x}_{\hat{m}}) &\leq \frac{1+\gamma}{1-4\gamma}[L(x_m) + \min_{m,k}(2p_0(m, p_k, \delta) + \frac{1}{p_{\min}} pen_1(m, p_k, \delta) \\ &\quad + \frac{1}{p_{\min}^2}(1+2/\gamma) pen_2(m, p_k, \delta))] \\ &\leq \frac{1+\gamma}{1-4\gamma} \min_m [L(x_m) + \min_k pen(m, p_k, \delta, \gamma, n)] \end{aligned}$$

Remark 2.5 In practice, a reasonable alternative to the proposed minimization procedure is estimating the overall error by cross-validation or leave one out techniques and then choose m minimizing the error for successive essays of probability $\hat{p}$. Recall that in the original procedure of Sect. 2.5, labels are not required to obtain $\hat{p}$ for a fixed model. Cross-validation or empirical error minimization techniques do, however, require a stock of "extra" labels, which might not be affordable in the active learning setting. Empirical error minimization is specially useful for applications where what is required is a subset of very informative sample points, as for example when deciding what points get extra labels (new laboratory runs, for example) given a first set of complete labels is available. Applications suggest that $\hat{p}$ obtained with this methodology (or a threshold version of $\hat{p}$ which eliminates points with sampling probability $\hat{p}_i \leq \eta$ a certain small constant) is very accurate in finding "good" or informative subsets, over which model selection may be performed.

3 Iterative Procedure: Updating the Sampling Probabilities

A major drawback of the batch procedure is the appearance of $p_{\min}$ in the denominator of error bounds, since typically $p_{\min}$ must be small in order for the estimation procedure to be effective. Indeed, since the expected number of effective samples is given by $\mathbb{E}(N) := \mathbb{E}\left(\sum_i p(t_i)\right)$, small values of $p(t_i)$ are required in order to gain in sample efficiency.

Proofs in Sect. 2.5 depend heavily on bounding expressions such as

$$\frac{1}{n}\sum_{i=1}^{n} q(t_i)\frac{w_i}{p(t_i)}\varepsilon_i(x-x')(t_i)$$

or

$$\frac{1}{n}\sum_{i=1}^{n} q(t_i)\left(\frac{w_i}{p(t_i)}-1\right)(x-x')^2(t_i)$$

where x and x' belong to a given model family S_m. Thus, it seems like a reasonable alternative to consider iterative procedures for which at time j, $p_j(t_i) \sim \max_{x,x' \in S_j} |x(t_i) - x'(t_i)|$ with S_j the current hypothesis space. In what follows we develop this strategy, adapting the results of [1] from the classification to the regression problem. Although we continue to work in the setting of model selection over bounded subsets of linearly spanned spaces, results can be readily extended to other frameworks such as additive models or kernel models. Once again, we will require certain additional restrictions associated to the uniform approximation of x_0 over the target model space.

More precisely, we start with an initial model set $S(= S_{m_0})$ and set x^* to be the overall minimizer of the loss function $L(x)$ over S. Assume additionally

AU $\quad \sup_{x \in S} \max_{t \in \{t_1, \ldots, t_n\}} |x_0(t) - x(t)| \leq B$

Let $L_n(x) = L_n(x, y, p)$ and $L(x)$ be as in Sect. 2.5. For the iterative procedure introduce the notation

$$L_j(x) := \frac{1}{n_j} \sum_{i=1}^{nj} q(t_{j_i}) \frac{w_i}{p(t_{j_i})} (x(t_{j_i}) - y_{j_i})^2, \quad j = 0, \ldots, n$$

with $n_j = n_0 + j$ for $j = 0, \ldots, n - n_0$.

In the setting of Sect. 2 for each $0 \leq j \leq n$, S_j will be the linear space spanned by the collection $\{\phi_\ell\}_{\ell \in \mathscr{I}_j}$ with $|\mathscr{I}_j| = d_j$, $d_j = o(n)$.

In order to bound the fluctuations of the initial step in the iterative procedure we consider the quantities defined in Eqs. (4) and (6) for $r = \gamma = 2$. That is,

$$\Delta_0 = 2\sigma^2 Q \left\{ \frac{2(d_0 + 1)}{n_0} + \frac{\log^2(2/\delta)}{n_0} \right\}$$
$$+2(\widetilde{\beta}_{m_0}(1 + \widetilde{\beta}_{m_0}))^2 B^2.$$

with

$$\widetilde{\beta}_{m_0} = \frac{c_{m_0}(\sqrt{17} + 1)}{2} \sqrt{\frac{d_0 Q}{n_0 p_{\min}}} \sqrt{2 \log(2^{7/4} m_0/\delta)}.$$

As discussed in Sect. 2.4, Δ_0 requires some initial guess of $\|x_0 - x_{m_0}\|_{n,q}^2$. Since this is not available, we consider the upper bound B^2. Of course this will possibly slow down the initial convergence as Δ_0 might be too big, but will not affect the overall algorithm. Also remark we do not consider the weighting sequence θ_k of Eq. (6) because the sampling probability is assumed fixed.

Next set $B_j = \sup_{x,x' \in S_{j-1}} \max_{t \in \{t_1,\ldots,t_n\}} |x(t) - x'(t)|$ and define

$$\Delta_j = \sqrt{\sigma^2 Q[(\frac{2(d_j+1)}{n_j}) + \frac{\log^2(4n_j(n_j+1)/\delta)}{n_j}]}$$
$$+\sqrt{\log(4n_j(n_j+1)/\delta)\frac{16B_j^2(2B_j \wedge 1)^2 Q^2}{n_j}} + 4\sqrt{4\frac{(d_j+1)\log n}{n_j}}.$$

The iterative procedure is stated as follows:

1. For $j = 0$:
   - Choose (randomly) an initial sample of size n_0, $M_0 = \{t_{k_1}, \ldots, t_{k_{n_0}}\}$.
   - Let $\hat{x}_0$ be the chosen solution by minimization of $L_0(x)$ (or possibly a weighted version of this loss function).
   - Set $S_0 \subset \{x \in S : L_0(x) < L_0(\hat{x}_0) + \Delta_0\}$
2. At step j:
   - Select (randomly) a sample candidate point t_j, $t_j \notin M_{j-1}$.
     Set $M_j = M_{j-1} \cup \{t_j\}$
   - Set $p(t_j) = (\max_{x,x' \in S_{j-1}} |x(t_j) - x'(t_j)| \wedge 1)$ and generate $w_j \sim Ber(p(t_j))$.
     If $w_j = 0$, set $j = j + 1$ and go to (2) to choose a new sample candidate.
     If $w_j = 1$ sample y_j and continue.
   - Let $\hat{x}_j = \arg\min_{x \in S_{j-1}} L_j(x) + \Delta_{j-1}(x)$
   - Set $S_j \subset \{x \in S_{j-1} : L_j(x) < L_j(\hat{x}_j) + \Delta_j\}$
   - Set $j = j + 1$ and go to (2) to choose a new sample candidate.

Remark that, such as it is stated, the procedure can continue only up until time n (when there are no more points to sample). If the process is stopped at time $T < n$, the term $\log(n(n+1))$ can be replaced by $\log(T(T+1))$. We have the following result, which generalizes Theorem 2 in [1] to the regression case.

Theorem 3.1 *Let $x^* = arg\min_{x \in S} L(x)$. Set $\delta > 0$. Then, with probability at least $1 - \delta$ for any $j \leq n$*

- $|L(x) - L(x')| \leq 2\Delta_{j-1}$, *for all* $x, x' \in S_j$
- $L(\hat{x}_j) \leq [L(x^*) + 2\Delta_{j-1}]$

Remark 3.1 An important issue is related to the initial choice of m_0 and n_0. As the overall precision of the algorithm is determined by $L(x^*)$, it is important to select a sufficiently complex initial model collection. However, if $d_{m_0} >> n_0$, then Δ_0 can be big and $p_j \sim 1$ for the first samples, which leads to a more inefficient sampling scheme.

3.1 Effective Sample Size

For any sampling scheme the expected number of effective samples is, as already mentioned, $\mathbb{E}\left(\sum_i p(t_i)\right)$. Whenever the sampling policy is fixed, this sum is not random and effective reduction of the sample size will depend on how small sampling probabilities are. However, this will increase the error bounds as a consequence of the factor $1/p_{\min}$. The iterative procedure allows a closer control of both aspects and under suitable conditions will be of order $\sum_j \sqrt{L(x^*) + \Delta_j}$. Recall from the definition of the iterative procedure we have $p_j(t_i) \sim \max_{x,x' \in S_j} |x(t_i) - x'(t_i)|$, whence the expected number of effective samples is of the order of $\sum_j \max_{x,x' \in S_j} |x(t_i) - x'(t_i)|$. It is then necessary to control $\sup_{x,x' \in S_{j-1}} |x(t_i) - x'(t_i)|$ in terms of the (quadratic) empirical loss function L_j. For this we must introduce some notation and results relating the supremum and L_2 norms [2].

Let $S \subset L_2 \cap L_\infty$ be a linear subspace of dimension d, with basis $\Phi := \{\phi_j,\ j \in m_S\}$, $|m_S| = d$. Set $\overline{r} := \inf_\Lambda r_\Lambda$, where Λ stands for any orthonormal basis of S.

We have the following result

Lemma 3.1 *Let $\hat{x}_j$ be the sequence of iterative approximations to x^* and $p_j(t)$ be the sampling probabilities in each step of the iteration, $j = 1, \ldots, T$. Then, the effective number of samples, that is, the expectation of the required samples $N_e = \mathbb{E}\left(\sum_{j=1}^T p_j(t_j)\right)$ is bounded by*

$$N_e \leq 2\sqrt{2}\overline{r}(\sqrt{L(x*)} \sum_{j=1}^{T} \sqrt{d_j} + \sum_{j=1}^{T} \sqrt{d_j \Delta_j}).$$

References

1. Beygelzimer, A., Dasgupta, S., & Langford, J. (2009). Importance weighted active learning. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 49–56). New York: ACM.
2. Birgé, L., & Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli, 4*, 329–395.
3. Fermin, A. K., & Ludeña, C. (2018). Probability bounds for active learning in the regression problem. arXiv: 1212.4457
4. Härdle, W., Kerkyacharian, G., Picard, D., & Tsybakov, A. (1998). *Wavelets, approximation and statistical applications: Vol. 129. Lecture notes in statistics.* New York: Springer.
5. Sugiyama, M. (2006). Active learning in approximately linear regression based on conditional expectation generalization error with model selection. *Journal of Machine Learning Research, 7*, 141–166.
6. Sugiyama, M., Krauledat, M., & Müller, K. R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research, 8*, 985–1005.
7. Sugiyama, M., & Rubens, N. (2008). A batch ensemble approach to active learning with model selection. *Neural Networks, 21*(9), 1278–1286.