ESAIM: PS 28 (2024) 227–257
https://doi.org/10.1051/ps/2024006

ESAIM: Probability and Statistics
www.esaim-ps.org

THE VANISHING LEARNING RATE ASYMPTOTIC FOR LINEAR L^2-BOOSTING

Clément Dombry[1] and Youssef Esstafa[2,*]

Abstract. We investigate the asymptotic behaviour of gradient boosting algorithms when the learning rate converges to zero and the number of iterations is rescaled accordingly. We mostly consider L^2-boosting for regression with linear base learner as studied in P. Bühlmann and B. Yu, *J. Am. Statist. Assoc.* **98** (2003) 324–339 and analyze also a stochastic version of the model where subsampling is used at each step (J.H. Friedman, *Computat. Statist. Data Anal.* **38** (2002) 367–378). We prove a deterministic limit in the vanishing learning rate asymptotic and characterize the limit as the unique solution of a linear differential equation in an infinite dimensional function space. Besides, the training and test error of the limiting procedure are thoroughly analyzed. We finally illustrate and discuss our result on a simple numerical experiment where the linear L^2-boosting operator is interpreted as a smoothed projection and time is related to its number of degrees of freedom.

Mathematics Subject Classification. 62G08, 60J20.

Received February 27, 2023. Accepted March 25, 2024.

1. Introduction

In the past decades, boosting has become a major and powerful prediction method in machine learning. The success of the classification algorithm AdaBoost by [1] demonstrated the possibility to combine many weak learners in a sequential way in order to produce better predictions, with widespread applications in gene expression [2] or music genre identification [3], to name only a few. [4] were able to see a wider statistical framework that lead to the gradient boosting [5], where a weak learner (*e.g.*, regression trees) is used to optimize a loss function in a sequential procedure akin to gradient descent. Choosing the loss function according to the statistical problem at hand results in a versatile and efficient tool that can handle classification, regression, quantile regression or survival analysis... The popularity of gradient boosting is also due to its efficient implementation in the R package `gbm` by [6] or in the recent and successful extreme gradient boosting algorithm XGBoost by [7]. We refer to the monograph by [8] for an overview on the numerous developments on boosting.

Strong theoretical results have justified the good performance of boosting. Consistency of boosting algorithm, *i.e.* their ability to achieve the optimal Bayes error rate for large samples, is considered in [9], [10], [11] or [12]. The present paper is strongly influenced by [13] that proposes an analysis of regression boosting algorithms built on linear base learners thanks to explicit formulas for the boosted predictor and its error rate.

Keywords and phrases: Boosting, non parametric regression, statistical learning, stochastic algorithm, Markov chain, convergence of stochastic process.

[1] Université de Franche-Comté, CNRS, LmB, F-25000 Besançon, France.
[2] Le Mans Université, Laboratoire Manceau de Mathématiques, Avenue Olivier Messiaen, 72085 Le Mans Cedex 09, France.

* Corresponding author: Youssef.Esstafa@univ-lemans.fr

In this paper, we focus on gradient boosting for regression with square loss and we briefly describe the corresponding algorithm. Consider a regression model with deterministic design, that is

$$Y_i = f(x_i) + \varepsilon_i, \quad 1 \le i \le n, \tag{1.1}$$

where the response Y is real valued, the explanatory variable X takes the values $X = x_1, \dots, x_n \in [0,1]^p$ assumed deterministic, the regression function $f : [0,1]^p \to \mathbb{R}$ is measurable and the errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed centered random variables. We assume the error variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$ is finite. Note that the assumption of i.i.d. errors is used only in Section 2.3 to compute the expected training and test errors; other than that, our results hold in the more general heteroscedastic setting where the errors are independent and such that $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma_i^2$, $1 \le i \le n$. Nevertheless we stick to the case of i.i.d. errors for the sake of simplicity.

Based on observations $(Y_i, x_i)_{1 \le i \le n}$ of the regression model (1.1), we aim at estimating the regression function f. Given a weak learner

$$\hat{f}(x) = \hat{f}(x; (Y_i, x_i)_{1 \le i \le n}),$$

the boosting algorithm with learning rate $\lambda \in (0,1)$ produces a sequence of models $\hat{F}_m^\lambda(x)$, $0 \le m \le \mathbf{M}$, by recursively fitting the weak learner to the current residuals and updating the model with a shrunken version of the fitted model. More formally, we define

$$\begin{cases} \hat{F}_0^\lambda(x) & = \bar{Y}_n, \\ \hat{F}_{m+1}^\lambda(x) & = \hat{F}_m^\lambda(x) + \lambda \hat{f}(x; (R_{m,i}^\lambda, x_i)_{1 \le i \le n}), \quad 0 \le m \le \mathbf{M} - 1, \end{cases} \tag{1.2}$$

where $\bar{Y}_n$ denotes the empirical mean of $(Y_i)_{1 \le i \le n}$ and $(R_{m,i}^\lambda)_{1 \le i \le n}$ the residuals

$$R_{m,i}^\lambda = Y_i - \hat{F}_m^\lambda(x_i), \quad 1 \le i \le n.$$

In practice, the shrinkage parameter λ and the number of iterations m are the main parameters and must be chosen suitably to achieve good performance. Common practice is to fix λ to a small value, typically $\lambda = 0.01$ or 0.001, and then to select m by cross-validation and the early stopping procedure. Citing [6], with slight modifications to match our notations:

> *"The issues that most new users of gbm struggle with are the choice of tree numbers m and shrinkage λ. It is important to know that smaller values of λ (almost) always give improved predictive performance. That is, setting $\lambda = 0.001$ will almost certainly result in a model with better out-of-sample predictive performance than setting $\lambda = 0.01$. However, there are computational costs, both storage and CPU time, associated with setting shrinkage to be low. The model with $\lambda = 0.001$ will likely require ten times as many iterations as the model with $\lambda = 0.01$, increasing storage and computation time by a factor of 10."*

This citation clearly emphasizes the role of small learning rates in boosting. The purpose of the present paper is to prove the existence of a vanishing learning rate limit ($\lambda \to 0$) for the boosting algorithm when the number of iterations is rescaled accordingly. More precisely, in the case when the base learner is linear, we prove the existence of the limit

$$\hat{F}_t(x) = \lim_{\lambda \downarrow 0} \hat{F}_{[t/\lambda]}^\lambda(x), \quad t \ge 0. \tag{1.3}$$

We furthermore characterize the limit as the solution of a linear differential equation in infinite dimensional space and also analyze the corresponding training and test errors. The case of stochastic gradient boosting [26], where

subsampling is introduced at each iteration, is also analyzed: we prove the existence of a deterministic vanishing learning rate limit that corresponds to a modified deterministic base learner defined in a natural way. The analysis of this stochastic framework requires involved tools of Markov chain theory and the characterization of their convergence through generators [14, 15]. It is interesting to see our results as an analogous of the gradient flow approximation of stochastic gradient descent that has recently been discussed in machine learning research and the vast literature on deep learning neural network training (see for instance [16–18] or [19]). In this perspective, boosting can be interpreted as a functional gradient descent and the vanishing learning rate limit as its gradient flow approximation. To our best knowledge, this is the first result in this direction for boosting.

A limitation of our work is the strong assumption of linearity of the base learner: the ubiquitous regression tree does not satisfy this assumption and further work is needed to deal with this important case. Our results are of probabilistic nature: we focus on the existence and properties of the limit (1.3) for fixed sample size $n \geq 1$, while statistical issues such as consistency as $n \to \infty$ is left aside for further research.

The paper is structured as follows. In Section 2, we prove the existence of the vanishing learning rate limit (1.3) for the boosting procedure with linear base learner (Prop. 2.6), we characterize the limit as the solution of a linear differential equation in a function space (Thm. 2.9) and we analyze the training and test errors (Props. 2.14 and 2.15). The stochastic gradient boosting where subsampling is introduced at each step is considered in Section 3. We prove that the vanishing learning rate limit still exists and that the convergence holds in quadratic mean (Cor. 3.6) and also in the sense of functional weak convergence in Skorokhod space (Thm. 3.7). A simple numerical experiment is presented in Section 4 in order to illustrate our theoretical findings, leading us to the interpretation of linear L^2-boosting as a smoothed projection where time is related to the degrees of freedom of the linear boosting operator. A conclusion summarizing our work and its interest from a practical point of view is given in Section 5. All the technical proofs are gathered in Section 6.

2. L^2-boosting with linear base learner

2.1. Framework

We consider the framework of boosting for regression with L^2-loss and linear base learner provided by [13] (see also [20] for interesting complements). This framework allows for explicit computations relying on linear algebra. The Banach space of bounded functions on $[0,1]^p$ is denoted by $\mathbb{B} = \mathbb{B}([0,1]^p, \mathbb{R})$ and is endowed with the uniform norm. Our main hypothesis is the following linearity assumption of the base learner $\hat{f}$.

Assumption 2.1. We assume that the base learner of the boosting algorithm (1.2) satisfies

$$\hat{f}(x; (Y_i, x_i)_{1 \leq i \leq n}) = \sum_{j=1}^{n} Y_j g_j(x), \quad x \in [0,1]^p, \tag{2.1}$$

where $g_1, \ldots, g_n \in \mathbb{B}$ may depend on $(x_i)_{1 \leq i \leq n}$.

It follows from Assumption 2.1 that g_j is the output of the base learner for input $(Y_i)_{1 \leq i \leq n} = (\delta_{ij})_{1 \leq i \leq n}$, where the Kronecker symbol δ_{ij} is equal to 1 if $i = j$ and 0 otherwise.

Under Assumption 2.1, the boosting algorithm with input $(Y_i, x_i)_{1 \leq i \leq n}$ and learning rate $\lambda \in (0,1)$ outputs a sequence of bounded functions $(\hat{F}_m^\lambda)_{m \geq 1}$. The sequence remains in the finite dimensional linear space spanned in $\mathbb{B}$ by the functions $g_1, \ldots, g_n$ and the constant functions (due to the initialization equal to the constant function $\bar{Y}_n$). A straightforward recursion based on Equation (1.2) yields

$$\hat{F}_m^\lambda(x) = \bar{Y}_n + \sum_{i=1}^{n} w_{m,i}^\lambda g_i(x) \tag{2.2}$$

where the weights $w_m^\lambda = (w_{m,i}^\lambda)_{1\le i\le n}$ satisfy

$$\begin{cases} w_{0,i}^\lambda & = 0 \\ w_{m+1,i}^\lambda & = w_{m,i}^\lambda + \lambda(Y_i - \bar{Y}_n) - \lambda \sum_{j=1}^n w_{m,j}^\lambda g_j(x_i) \end{cases}.$$

This linear recursion system can be rewritten in vector form as

$$\begin{cases} w_0^\lambda & = 0 \\ w_{m+1}^\lambda & = (I - \lambda S) w_m^\lambda + \lambda \tilde{Y} \end{cases}, \tag{2.3}$$

with $S = (g_j(x_i))_{1\le i,j\le n}$, $\tilde{Y} = (Y_i - \bar{Y}_n)_{1\le i\le n}$ the centered observations and I the $n \times n$ identity matrix. This linear recursion is easily solved, yielding the following proposition.

Proposition 2.2. *Under Assumption 2.1, the boosting algorithm output $\hat{F}_m^\lambda$ is given by Equation (2.2) with weights*

$$w_m^\lambda = \lambda \sum_{j=0}^{m-1} (I - \lambda S)^j \tilde{Y}, \quad m \ge 0. \tag{2.4}$$

If the matrix S is invertible, then

$$w_m^\lambda = S^{-1} \left[I - (I - \lambda S)^m \right] \tilde{Y}, \quad m \ge 0.$$

Note that this result is similar to Proposition 1 in Bühlmann and Yu [13], but they consider only the values on the observed sample $(x_i)_{1\le i\le n}$ while we provide the extrapolation to $x \in [0,1]^p$ more explicitly. Also we consider a different initialization to the empirical mean instead of initialization to zero, which seems more relevant in practice.

Example 2.3. A simple example satisfying Assumption 2.1 is the Nadaraya-Watson estimator (see [21, 22])

$$\hat{f}(x) = \frac{\sum_{i=1}^n K_h(x - x_i) Y_i}{\sum_{i=1}^n K_h(x - x_i)}, \quad x \in [0,1]^p,$$

where $h > 0$ is the bandwidth, $K : \mathbb{R}^p \to (0, +\infty)$ is the kernel, *i.e.* a density function, and $K_h(z) = h^{-d} K(z/h)$ the rescaled kernel.

Example 2.4. A more involved example of base learner, discussed in [13] Section 3.2, is the smoothing spline in dimension $p = 1$. For $r \ge 1$ and $\nu > 0$, the smoothing spline $\hat{f}$ is the unique minimizer over $\mathcal{W}_2^{(r)}$ of the penalized criterion

$$\sum_{i=1}^n (Y_i - \hat{f}(x_i))^2 + \nu \int_0^1 (\hat{f}^{(r)}(x))^2 \mathrm{d}x,$$

where $\mathcal{W}_2^{(r)}$ denotes the Sobolev space of functions that are continuously differentiable of order $r-1$ with square integrable weak derivative of order r. Assuming $0 < x_1 < \cdots < x_n < 1$, the solution is known to be piecewise polynomial function of degree $r+1$ with constant derivative of order $r+1$ on $n+1$ intervals $(0, x_1), \ldots, (x_n, 1)$. It is used in [13] that the matrix S is symmetric positive definite with positive eigenvalues $1 = \mu_1 = \ldots = \mu_r > \ldots > \mu_n > 0$, see [23].

Example 2.5. Closely related with regression trees discussed in the introduction is the class of base learners based on a partition $(A_k)_{1\le k\le K}$ of the prediction space $[0,1]^p$, see Györfi et al. [24, Chapter 4] for instance. As an example, a partition of size $K = 2^p$ can be constructed from the (x_i) by splitting $[0,1]^p$ along each dimension $j = 1,\dots,p$ at a threshold given by the the median value of the $(x_i^j)_{1\le i\le n}$. Given any partition $(A_k)_{1\le k\le K}$ such that each region A_k contains at least one observation x_i, the base learner is defined as

$$\hat{f}(x) = \sum_{k=1}^{K} \bar{Y}(A_k)\mathbb{1}_{A_k}(x)$$

with

$$\bar{Y}(A_k) = \frac{\sum_{i=1}^n \mathbb{1}_{A_k}(x_i)Y_i}{\sum_{i=1}^n \mathbb{1}_{A_k}(x_i)}.$$

In other words, the learner is piecewise constant on each region A_k with value $\bar{Y}(A_k)$ equal to the mean of the Y_i's corresponding to observations with $x_i \in A_k$. Clearly, Assumption 2.1 holds with

$$g_j(x) = \sum_{k=1}^{K} \frac{\mathbb{1}_{A_k}(x_j)\mathbb{1}_{A_k}(x)}{n_k}$$

where $n_k = \sum_{i=1}^n \mathbb{1}_{A_k}(x_i)$ denotes the number of observations in A_k.

Interestingly, the matrix S has a simple structure in this case. Without loss of generality, the points x_i can be reordered so that $x_1,\dots,x_{n_1}$ are in A_1, $x_{n_1+1},\dots,x_{n_1+n_2}$ are in A_2 and so forth. Then the matrix S is block diagonals with K blocks, the k-th block having size $n_k \times n_k$ with entries all equal to $1/n_k$. This matrix satisfies $S^j = S$ for all $j \ge 1$ so that Equation (2.4) can be rewritten as

$$w_m^\lambda = \frac{1-(1-\lambda)^m}{\lambda} S\tilde{Y}, \quad m \ge 0.$$

2.2. The vanishing learning rate asymptotic

We next consider the existence of a limit in the vanishing learning rate asymptotic $\lambda \to 0$. The explicit simple formulas from Proposition 2.2 allows for a simple analysis. The results in this section mostly relies on the simple fact that the linear dynamical system (2.3) is the Euler discretization of the linear ODE

$$\begin{cases} w(t) &= 0 \\ w'(t) &= -Sw(t) + \tilde{Y} \end{cases}. \tag{2.5}$$

For this reason the weights $(w^\lambda_{[t/\lambda]})_{t\ge 0}$ converge to the ODE solution $(w(t))_{t\ge 0}$ as the discretization step λ converges to 0. We recall that the exponential of a square matrix M is defined by

$$\exp(M) = \sum_{k\ge 0} \frac{1}{k!} M^k.$$

Proposition 2.6. *Under Assumption 2.1 and for any $0 < \zeta < 1$, we have*

$$\lambda^{-\zeta}\left(\hat{F}^\lambda_{[t/\lambda]}(x) - \hat{F}_t(x)\right) \xrightarrow[\lambda\to 0]{} 0, \quad t \ge 0,\ x \in [0,1]^p, \tag{2.6}$$

uniformly on compact sets $[0,T]\times[0,1]^p$, $T>0$, *where the limit satisfies*

$$\hat{F}_t(x)=\bar{Y}_n+\sum_{i=1}^{n}w_{t,i}g_i(x) \tag{2.7}$$

with weights $w_t=(w_{t,i})_{1\le i\le n}$ *given by*

$$w_t=-\sum_{j\ge 1}\frac{(-t)^j}{j!}S^{j-1}\tilde{Y},\quad t\ge 0. \tag{2.8}$$

If the matrix S *is invertible, then*

$$w_t=S^{-1}\left(I-\mathrm{e}^{-tS}\right)\tilde{Y},\quad t\ge 0. \tag{2.9}$$

Proposition 2.6 states the uniform convergence $\hat{F}^\lambda\to\hat{F}$ as $\lambda\to 0$ on compact sets in space and time $[0,1]^p\times[0,T]$ but it provides no information about the long time behaviour $T\to+\infty$. The interplay between the two limits $\lambda\to 0$ and $T\to+\infty$ is another interesting question left for further research.

The formulas are even more explicit in the case when S is a symmetric matrix because it can then be diagonalized in an orthonormal basis of eigenvectors.

Corollary 2.7. *Suppose Assumption 2.1 is satisfied and* $S=(g_j(x_i))_{1\le i,j\le n}$ *is a symmetric matrix. Denote by* $(\mu_j)_{1\le j\le n}$ *the eigenvalues of* S *and by* $(u_j)_{1\le j\le n}$ *the corresponding eigenvectors. Then the vanishing learning rate asymptotic yields the weights*

$$w_t=\sum_{j=1}^{n}\frac{1-\mathrm{e}^{-\mu_j t}}{\mu_j}u_j u_j^\top\tilde{Y}$$

and the limit

$$\hat{F}_t(x)=\bar{Y}_n+\sum_{1\le i,j\le n}\frac{1-\mathrm{e}^{-\mu_j t}}{\mu_j}\left(v_i^\top u_j u_j^\top\tilde{Y}\right)g_i(x), \tag{2.10}$$

with $(v_i)_{1\le i\le n}$ *the canonical basis in* $\mathbb{R}^n$. *When* $\mu=0$, *we use extension by continuity, that is the convention* $(1-\mathrm{e}^{-\mu t})/\mu=t$.

Example 2.8. For a partition based learner as defined in Example 2.5, the matrix S is a projection of rank K so that the eigenvalues are 1 with multiplicity K and 0 with multiplicity $n-K$. The eigenvectors associated to 1 are $u_k=(1_{A_k}(x_i)/\sqrt{n_k})_{1\le i\le n}$, $k=1,\dots,K$. Then we have $\sum_{k=1}^{K}u_k u_k^\top=S$ and Corollary 2.7 yields the simple formulas $w_t=(1-\mathrm{e}^{-t})S\tilde{Y}$ and $\hat{F}_t(x)=\bar{Y}_n+(1-\mathrm{e}^{-t})S\tilde{Y}$.

Interestingly, the limit function $(\hat{F}_t)_{t\ge 0}$ appearing in the vanishing learning rate asymptotic can be characterized as the solution of a linear differential equation. The intuition is quite clear from the following heuristic: the boosting dynamic

$$\hat{F}^\lambda_{m+1}=\hat{F}^\lambda_m+\lambda\sum_{i=1}^{n}(Y_i-\hat{F}^\lambda_m(x_i))g_i$$

implies, for $t = \lambda m$,

$$\lambda^{-1}\left(\hat{F}^{\lambda}_{[(t+\lambda)/\lambda]} - \hat{F}^{\lambda}_{[t/\lambda]}\right) = \sum_{i=1}^{n}(Y_i - \hat{F}^{\lambda}_{[t/\lambda]}(x_i))g_i.$$

Letting $\lambda \to 0$, the convergence $\hat{F}^{\lambda}_{[t/\lambda]} \to \hat{F}_t$ suggests

$$\hat{F}'_t = \sum_{i=1}^{n}(Y_i - \hat{F}_t(x_i))g_i.$$

We make this rigorous in the following theorem. The main argument is that the weights (2.3) correspond to the Euler discretization of the linear ODE (2.5) and therefore converge as $\lambda \to 0$ to the ODE solution.

For $t \geq 0$, we consider $\hat{F}_t$ as an element of the Banach space of bounded functions $\mathbb{B} = \mathbb{B}([0,1]^p, \mathbb{R})$. We prove that $(\hat{F}_t)_{t\geq 0}$ is the unique solution of a linear differential equation. More precisely, it is easily seen that the linear operator $\mathcal{L} : \mathbb{B} \to \mathbb{B}$ defined by

$$\mathcal{L}(Z) = \sum_{i=1}^{n} Z(x_i)g_i, \quad Z \in \mathbb{B},$$

is bounded and we consider the differential equation in the Banach space $\mathbb{B}$

$$Z'(t) = -\mathcal{L}(Z(t)) + G, \quad t \geq 0, \tag{2.11}$$

with $G = \sum_{i=1}^{n} Y_i g_i$.

Theorem 2.9. *i) For all $Z_0 \in \mathbb{B}$, the differential equation (2.11) has a unique solution satisfying $Z(0) = Z_0$. Furthermore, if there exists $\mathcal{Y} \in \mathbb{B}$ such that $\mathcal{L}(\mathcal{Y}) = G$, this solution is explicitly given by*

$$Z(t) = (\mathrm{e}^{-t\mathcal{L}})Z_0 + (\mathrm{Id} - \mathrm{e}^{-t\mathcal{L}})\mathcal{Y}, \quad t \geq 0. \tag{2.12}$$

ii) The function $(\hat{F}_t)_{t\geq 0}$ is the solution of (2.11) with initial condition $\bar{Y}_n$. Assuming there exists $\mathcal{Y} \in \mathbb{B}$ such that $\mathcal{L}(\mathcal{Y}) = G$, we thus have

$$\hat{F}_t = (\mathrm{e}^{-t\mathcal{L}})\bar{Y}_n + (\mathrm{Id} - \mathrm{e}^{-t\mathcal{L}})\mathcal{Y}, \quad t \geq 0.$$

Remark 2.10. The condition $\mathcal{L}(\mathcal{Y}) = G$ is satisfied as soon as $\mathcal{Y}(x_i) = Y_i$, $1 \leq i \leq n$. In particular, it holds if the x_i's are pairwise distinct. It is used mostly for convenience and elegance of notations. Indeed we have

$$(\mathrm{Id} - \mathrm{e}^{-t\mathcal{L}})(\mathcal{Y}) = -\sum_{k\geq 1}\frac{(-t)^k}{k!}\mathcal{L}^k(\mathcal{Y}) = \sum_{k\geq 1}\frac{(-1)^{k-1}t^k}{k!}\mathcal{L}^{k-1}(G)$$

and, if the existence of $\mathcal{Y}$ is not granted, one can replace in formula (2.12) the term involving $\mathcal{Y}$ by the series in the right hand side of the previous equation and check that this provides a solution of (2.11) in the general case.

Finally, we discuss the notion of stability of the boosting procedure. It requires that the output of the boosting algorithm does not explode for large time values.

Definition 2.11. The boosting algorithm is called *stable* if, for all possible input $(Y_i)_{1\le i\le n}$, the output $(\hat{F}_t)_{t\ge 0}$ remains uniformly bounded in $\mathbb{B}$ as $t \to \infty$.

It is here convenient to assume the following:

Assumption 2.12. In Equation (2.1), the functions $(g_i)_{1\le i\le n}$ are linearly independent and such that $\sum_{i=1}^n g_i(x) = 1$ for all $x \in [0,1]^p$.

The linear independence is sensible if the points $(x_i)_{1\le i\le n}$ are pairwise distinct. The constant sum implies that for constant input $Y_i = 1$, $1 \le i \le n$, the output $\hat{f}(x) \equiv 1$ is also constant. Both are mild assumptions satisfied by most learners in practice.

The stability can be characterized in terms of the Jordan normal form of the matrix S, see for instance [25]. We recall that the Jordan normal form of S is a block diagonal matrix where each block, called Jordan block, is an upper triangular matrix of size $s \times s$ with a complex eigenvalue μ on the main diagonal and ones on the superdiagonal. The matrix can be diagonalized if and only if all its Jordan blocks have size 1.

Proposition 2.13. *Suppose Assumptions 2.1 and 2.12 are satisfied. Then the boosting procedure algorithm is stable if and only if all the blocks of the Jordan normal form of S satisfy:*

- *the eigenvalue has a positive real part;*
- *the eigenvalue has a null real part and the block has size* 1.

In particular, if S is symmetric, the boosting procedure is stable if and only if all the eigenvalues of S are non-negative.

2.3. Training and test error

We next consider the performance of the boosting regression algorithm in terms of L^2-loss, also known as mean squared error. We focus mostly on the vanishing learning rate asymptotic, although version of the results below could be derived for positive learning rate λ.

The training error is assessed on the training set used to fit the boosting predictor and compares the observations Y_i to their predicted values $\hat{F}_t(x_i)$, *i.e.*

$$\mathrm{err}_{train}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{F}_t(x_i))^2. \tag{2.13}$$

The generalization capacity of the algorithm is assessed on new observations that are not used during the fitting procedure. For test observations $(Y_i', X_i')_{1\le i\le n'}$, independent of the training sample, the test error is defined by

$$\mathrm{err}_{test}(t) = \frac{1}{n'}\sum_{i=1}^{n'}(Y_i' - \hat{F}_t(X_i'))^2. \tag{2.14}$$

We also consider a simpler version of the test error where extrapolation in the feature space is not evaluated and we take $n' = n$ and $X_i' = x_i$. Then, the test error writes

$$\mathrm{err}_{test}(t) = \frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{F}_t(x_i))^2, \tag{2.15}$$

and allows for simpler formulas with nice interpretation.

We first consider the behavior of the training error as defined in Equation (2.13). Note that

$$\mathrm{err}_{train}(t) = \frac{1}{n}\|R_t\|^2$$

where R_t is the vector of residuals at time t defined by

$$R_t = (Y_i - \hat{F}_t(x_i))_{1\le i\le n}, \quad t \ge 0,$$

and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. Furthermore, Proposition 2.6 implies $R_t = \mathrm{e}^{-tS}\tilde{Y}$, $t \ge 0$, so that

$$\mathrm{err}_{train}(t) = \frac{1}{n}\|\mathrm{e}^{-tS}\tilde{Y}\|^2, \quad t \ge 0.$$

The following proposition is related to Proposition 3 and Theorem 1 in [13].

Proposition 2.14. *Suppose Assumptions 2.1 and 2.12 are satisfied.*

i) We have $\lim_{t\to\infty} \mathrm{err}_{train}(t) = 0$ *for all possible input* $(Y_i)_{1\le i\le n}$ *if and only if all the eigenvalues of* S *have a positive real part.*

ii) The training error satisfies

$$\begin{aligned} \mathbb{E}[\mathrm{err}_{train}(t)] &= \mathrm{bias}^2(t) + \mathrm{var}_{train}(t), \\ \mathrm{bias}^2(t) &= \frac{1}{n}\|\mathrm{e}^{-tS}\tilde{f}\|^2, \\ \mathrm{var}_{train}(t) &= \frac{\sigma^2}{n}\mathrm{Trace}\left(\mathrm{e}^{-tS}J\mathrm{e}^{-tS^\top}\right), \end{aligned} \tag{2.16}$$

with $J = I - \frac{1}{n}1_n 1_n^T$, $\tilde{f} = f - \bar{f}1_n$, $f = (f(x_i))_{1\le i\le n}$ *and* $\bar{f} = \frac{1}{n}\sum_{i=1}^n f(x_i)$.

iii) If S *is symmetric with positive eigenvalues* $(\mu_i)_{1\le i\le n}$ *and corresponding eigenvectors* $(u_i)_{1\le i\le n}$,

$$\mathbb{E}[\mathrm{err}_{train}(t)] = \frac{1}{n}\sum_{i=1}^n (u_i^\top \tilde{f})^2 \mathrm{e}^{-2t\mu_i} + \frac{\sigma^2}{n}\sum_{i=1}^n \|Ju_i\|^2 \mathrm{e}^{-2t\mu_i}.$$

The expected training error is strictly decreasing and converges to 0 *exponentially fast as* $t \to \infty$.

The convergence of the training error to zero implies that the boosting procedure is stable as considered in Proposition 2.13 but the converse is not true since some eigenvalues may have a real part equal to zero. When S is symmetric definite positive, the expected training error converges exponentially fast to 0 (this was already proved in [13] Thm. 1 for $\lambda > 0$) but this exponential rate of convergence has to be taken with care since S may have very small eigenvalues, see the numerical illustration in Section 4.

The fact that the residuals converge to zero suggests that the boosting procedure eventually overfits the training observations and loses generalization power. A simple analysis of this overfit is provided by the test error with fixed covariates $X_i' = x_i$, as defined by Equation (2.15). For the sake of simplicity, we emphasize the case when S is symmetric.

Proposition 2.15. *i) The test error with fixed covariates defined by Equation (2.15) satisfies*

$$\begin{aligned} \mathbb{E}[\mathrm{err}_{test}(t)] &= \mathrm{bias}^2(t) + \mathrm{var}_{test}(t), \\ \mathrm{bias}^2(t) &= \frac{1}{n}\|\mathrm{e}^{-tS}\tilde{f}\|^2, \\ \mathrm{var}_{test}(t) &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\mathrm{Trace}\left((I - \mathrm{e}^{-tS})J(I - \mathrm{e}^{-tS})^\top\right). \end{aligned}$$

ii) If S is symmetric with positive eigenvalues $(\mu_i)_{1\leq i\leq n}$ and associated eigenvectors $(u_i)_{1\leq i\leq n}$,

$$\text{bias}^2(t) = \frac{1}{n}\sum_{i=1}^{n}(u_i^\top \tilde{f})^2 \mathrm{e}^{-2t\mu_i},$$
$$\text{var}_{test}(t) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\sum_{i=1}^{n}\|Ju_i\|^2(1-\mathrm{e}^{-t\mu_i})^2,$$

so that the the following properties hold:
- *the squared bias is decreasing, convex and vanishes as $t\to\infty$;*
- *the variance is increasing and with limit $2\sigma^2$ as $t\to\infty$;*
- *the expected test error is decreasing in the neighborhood of zero, eventually increasing and with limit $2\sigma^2$ as $t\to\infty$.*

We retrieve with explicit theoretical formulas the known behavior of boosting in practice: the choice of $t\geq 0$ is crucial in the bias/variance trade-off. Small values of $t\geq 0$ lead to underfitting while overfitting appears for larger time values. In the early stage of the procedure, the bias decreases more rapidly that the variance increases, leading to a reduced test error. In practice, cross-validation and early stopping is used to estimate the test error and choose when to stop the boosting procedure, see [11].

Remark 2.16. When the boosting algorithm is initialized at $\hat{F}_0 = 0$ as in [13], the expected training and test error from Propositions 2.14 and 2.15 become

$$\mathbb{E}[\text{err}_{train}(t)] = \frac{1}{n}\sum_{i=1}^{n}(u_i^\top f)^2\mathrm{e}^{-2t\mu_i} + \frac{\sigma^2}{n}\sum_{i=1}^{n}\|u_i\|^2\mathrm{e}^{-2t\mu_i}$$

and

$$\mathbb{E}[\text{err}_{test}(t)] = \frac{1}{n}\sum_{i=1}^{n}(u_i^\top f)^2\mathrm{e}^{-2t\mu_i} + \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\sum_{i=1}^{n}\|u_i\|^2\mathrm{e}^{-2t\mu_i}.$$

These values are always larger than those with initialization $\hat{F}_0 = \bar{Y}_n$, whence we recommend initialization to the empirical mean.

When the test error includes extrapolation in the predictor space – *i.e.* the new test observations $(Y_i', X_i')_{1\leq i\leq n'}$ are i.i.d. and independent of the training observation as in Equation (2.14) – the formula we obtain for its expectation is more difficult to analyze.

Proposition 2.17. *Assume S is symmetric with positive eigenvalues. The test error defined by Equation (2.14) has expectation*

$$\begin{aligned}\mathbb{E}[\text{err}_{test}(t)] = \frac{n+1}{n}\sigma^2 &+ \mathbb{E}\left[\left(f(X') - \bar{f} - \tilde{f}^\top S^{-1}\left(I-\mathrm{e}^{-tS}\right)g(X')\right)^2\right] \\ &+ \sigma^2\mathbb{E}\left[g(X')^\top\left(I-\mathrm{e}^{-tS}\right)S^{-1}JS^{-1}\left(I-\mathrm{e}^{-tS}\right)g(X')\right]\end{aligned}$$

with $g(X') = (g_i(X'))_{1\leq i\leq n}$.

3. Stochastic gradient boosting

Following [26], it is common practice to use a stochastic version of the boosting algorithm where subsampling is introduced at each step of the procedure. The package `gbm` by [6] uses the subsampling rate equal to 50%

by default, meaning that each step involves only a subsample with half of the observations randomly chosen. This subsampling is known to have a regularization effect and we consider in this section the existence of the vanishing learning rate limit for such stochastic boosting algorithms.

3.1. Framework

We consider the following stochastic boosting algorithm that encompasses stochastic gradient boosting, see Example 3.2 below. We assume the weak learner $\hat{f}(x) = \hat{f}(x; (y_i, x_i)_{1\le i\le n}, \xi)$ depends on the observations $(y_i, x_i)_{1\le i\le n}$ and on an external source of randomness ξ with a finite set Ξ of possible values. We define the stochastic boosting algorithm by the recursion

$$\begin{cases} \hat{F}_0^\lambda(x) & = \bar{Y}_n, \\ \hat{F}_{m+1}^\lambda(x) & = \hat{F}_m^\lambda(x) + \lambda \hat{f}(x; (R_{m,i}^\lambda, x_i)_{1\le i\le n}, \xi_{m+1}), \quad m \ge 0, \end{cases} \tag{3.1}$$

where ξ_m, $m \ge 1$, are i.i.d. Ξ-valued random variables independent of $(Y_i, x_i)_{1\le i\le n}$ and $R_{m,i}^\lambda = Y_i - \hat{F}_m^\lambda(x_i)$, $1 \le i \le n$, are the residuals.

Assumption 3.1. We assume that the base learner of the stochastic boosting algorithm (3.1) satisfies

$$\hat{f}(x; (Y_i, x_i)_{1\le i\le n}, \xi) = \sum_{j=1}^n Y_j g_j(x, \xi), \quad x \in [0,1]^p,$$

where $g_1, \dots, g_n \in \mathbb{B}$ may depend on $(x_i)_{1\le i\le n}$ and $\xi \in \Xi$.

We assume that Ξ is finite mostly for simplicity and also because it is enough to cover two particularly important cases.

Example 3.2. Starting from a base learner $\hat{f}$ satisfying Assumption 2.1 (with n replaced by $[sn]$) and applying stochastic subsampling [26], we obtain a stochastic setting that satisfies Assumption 3.1. Let the sample size $n \ge 1$ be fixed and consider subsambling with rate $s \in (0,1)$, *e.g.* $s = 50\%$. Define Ξ as the set of all subsets ξ of $\{1, \dots, n\}$ with fixed size $[sn]$. Note that Ξ is finite with cardinality $\binom{n}{[sn]}$. The learner $\hat{f}$ fitted on subsample $\xi \in \Xi$ is written

$$\hat{f}(x; (Y_i, x_i)_{1\le i\le n}, \xi) = \hat{f}(x; (Y_i, x_i)_{i\in\xi}).$$

We use here a mild abuse of notation: in the left hand side, $\hat{f}$ denotes the randomized learner, the sample size is n and subsampling is introduced by ξ; in the right hand side, $\hat{f}$ denotes the deterministic base learner and the sample size is $[sn]$. Stochastic boosting corresponds to Algorithm (3.1) with the sequence $(\xi_m)_{m\ge1}$ uniformly distributed on Ξ, which corresponds to uniform subsampling.

Example 3.3. Another important example covered by the stochastic boosting algorithm (3.1) is the design of additive models. The idea is to provide an approximation of the regression function $f(x)$, $x \in [0,1]^p$, by an additive model of the form $f_1(x^{(1)}) + \cdots + f_p(x^{(p)})$, where $x^{(j)}$ denotes the jth component of x and f_j the principal effect of $x^{(j)}$. Such an additive model does not include interactions between different components. Assume that a base learner $\hat{f}$ with one-dimensional covariate space $[0,1]$ is given and that $\hat{f}$ satisfies Assumption 2.1 with $p = 1$. For instance, $\hat{f}$ can be a smoothing spline as in Example 2.4, see [13] Section 4. We consider stochastic regression boosting where the base learner $\hat{f}$ is sequentially applied with a randomly chosen predictor. Formally, set

$$\hat{f}(x; (Y_i, x_i)_{1\le i\le n}, \xi) = \hat{f}(x; (Y_i, x_i^{(\xi)})_{1\le i\le n}), \quad \xi = 1, \dots, p.$$

It is easily checked that the learner in the left hand side satisfies Assumption 3.1 and that algorithm (3.1) with $(\xi_m)_{m\geq 1}$ uniformly distributed on $\Xi = \{1, \ldots, p\}$ outputs a sequence of additive models. This strategy is often used with a more involved procedure where, at each step, the p different possible predictors are considered and the best one is kept, see [13] Section 4. But this falls beyond Assumption 3.1 because choosing the optimal component is not a linear operation and the randomized choice proposed here is a sensible alternative satisfying Assumption 3.1.

Example 3.4. A simple class of stochastic learner consists in learner based on random partitions $\xi = (A_k)_{1\leq k\leq K}$ of $[0,1]^p$ into K subsets. As in Example 2.5, the learner writes

$$\hat{f}(x; (Y_i, x_i)_{1\leq i\leq n}, \xi) = \sum_{k=1}^{K} \bar{Y}(A_k) \mathbb{1}_{A_k}(x)$$

but now the partition is randomly chosen. Clearly, Assumption 3.1 holds with

$$g_j(x, \xi) = \sum_{k=1}^{K} \frac{\mathbb{1}_{A_k}(x_j)\mathbb{1}_{A_k}(x)}{n(A_k)}.$$

3.2. Convergence of finite dimensional distributions

For fixed input $(Y_i, x_i)_{1\leq i\leq n}$, the stochastic boosting algorithm (3.1) provides a sequence of stochastic processes $\hat{F}_m^\lambda$, $m \geq 1$, and we consider the vanishing learning rate limit (1.3) under Assumption 3.1. We first prove convergence of the finite dimensional distributions thanks to elementary moment computations formulated in the next proposition. Expectation and variance are considered with respect to $(\xi_m)_{m\geq 1}$ while the input $(Y_i, x_i)_{1\leq i\leq n}$ is considered fixed and we note $\mathbb{E}_\xi$ and Var_ξ to emphasize this. We define

$$\bar{g}_j(x) = \mathbb{E}_\xi[g_j(x, \xi)], \quad x \in [0,1]^p, \ j = 1, \ldots, n,$$

and

$$S = (\bar{g}_i(x_j))_{1\leq i,j\leq n}. \tag{3.2}$$

Note that $\bar{g}_1, \ldots, \bar{g}_n$ are well-defined and in $\mathbb{B}$ because Ξ is finite so that there are no measurability or integrability issues.

Proposition 3.5. *Consider the boosting algorithm (3.1) under Assumption 3.1 and let the input $(Y_i, x_i)_{1\leq i\leq n}$ be fixed.*

i) For $x \in [0,1]^p$ and $m \geq 0$,

$$\mathbb{E}_\xi[\hat{F}_m^\lambda(x)] = \bar{Y}_n + \sum_{i=1}^{n} w_{m,i}^\lambda \bar{g}_i(x), \tag{3.3}$$

where $w_m^\lambda = (w_{m,i}^\lambda)_{1\leq i\leq n}$ is defined by (2.4) with S given by (3.2).

ii) There exists a positive constant K such that, for all $x \in [0,1]^p$, $m \geq 0$ and $\lambda < 1$,

$$\mathrm{Var}_\xi[\hat{F}_m^\lambda(x)] \leq K(m+1)\lambda^2(1+K\lambda)^m n \|\tilde{Y}\|_\infty^2 \left\{1 + (\lambda m K)^2 \mathrm{e}^{2\lambda m \|S\|_\infty}\right\},$$

where $\|\cdot\|_\infty$ denotes here the maximum norm on $\mathbb{R}^n$. We use the same notation for the infinity-norm of $n \times n$ matrices.

As will be clear from the proof, the constant K can be taken as $2M_1 + M_1^2 + (n+1)M_2$, where

$$M_1 = \max_{1\le j\le n+1} \sum_{i=1}^{n} |\bar{g}_i(x_j)| \tag{3.4}$$

and

$$M_2 = \max_{1\le j\le n+1} \sum_{i=1}^{n} \mathrm{Var}_\xi[g_i(x_j)]. \tag{3.5}$$

Corollary 3.6. *For all $t \ge 0$ and $x \in [0,1]^p$, the convergence (1.3) holds in quadratic mean with deterministic limit*

$$\hat{F}_t(x) = \bar{Y}_n + \sum_{i=1}^{n} w_{t,i}\bar{g}_i(x), \tag{3.6}$$

where $w_t = (w_{t,i})_{1\le i\le n}$ is defined by (2.8) with S given by (3.2).

Corollary 3.6 states that the vanishing learning rate limit $\hat{F}_t(x)$ for the stochastic boosting algorithm (3.1) is exactly the same as the one for the deterministic boosting algorithm (1.2) with base learner

$$\begin{aligned}\bar{f}(x;(Y_i,x_i)_{1\le i\le n}) &= \mathbb{E}_\xi[\hat{f}(x;(Y_i,x_i)_{1\le i\le n},\xi)]\\ &= \sum_{j=1}^{n} Y_j\bar{g}_j(x).\end{aligned}$$

In particular, the properties of the limit process $(\hat{F}_t)_{t\ge 0}$ have been studied in Section 2: characterization by a differential equation, behaviour of the training and test error, *etc.*

3.3. Weak convergence in function space

Corollary 3.6 implies that the convergence

$$\hat{F}^\lambda_{[t/\lambda]}(x) \longrightarrow \hat{F}_t(x), \quad \text{as } \lambda \to 0,$$

holds in the sense of finite dimensional distributions, that is joint convergence in distribution of the values of the processes at finitely many points $(t_i,x_i)_{1\le i\le k}$. We strengthen here this convergence into a weak convergence of stochastic processes in the Skorokhod space $\mathbb{D}([0,\infty),\mathbb{B})$ of càd-làg functions with values in $\mathbb{B}$. This stronger notion of convergence allows to consider the asymptotic behavior of functionals that depend on the whole trajectory (such as maximum or integral over compact sets) and not only on finitely many time points. We refer to Billingsley [27, Section 16] for background on the Skorokhod space $\mathbb{D}([0,\infty),\mathbb{R})$ and to Ethier and Kurtz [14, Section 3.5] for the general Skorokhod space $\mathbb{D}([0,\infty),E)$ on a metric space E.

Theorem 3.7. *Consider the boosting algorithm (3.1) under Assumption 3.1 and let the input $(Y_i,x_i)_{1\le i\le n}$ be fixed.*

i) For fixed $\lambda > 0$, the boosting sequence $(\hat{F}^\lambda_m)_{m\ge 0}$ is a time homogeneous Markov chain with values in $\mathbb{B}$.
ii) As $\lambda \to 0$, the convergence in distribution

$$(\hat{F}^\lambda_{[t/\lambda]})_{t\ge 0} \xrightarrow{d} (\hat{F}_t)_{t\ge 0} \tag{3.7}$$

holds in the Skorokhod space $\mathbb{D}([0,\infty),\mathbb{B})$.

Our proof relies on the theory of approximations of Markov chains by diffusions, see *e.g.* [15]. The assumption that Ξ is finite is important because it ensures that $(\hat{F}_m^\lambda)_{m\geq 0}$ remains in the finite dimensional subspace

$$\mathcal{F} = \text{span}(1, g_i(\cdot, \xi); 1 \leq i \leq n, \xi \in \Xi),$$

where span denotes the linear span of vectors in $\mathbb{B}$. In Theorem 3.7, the Markov property and the functional convergence can equivalently be stated with $\mathbb{B}$ replaced by the finite-dimensional space $\mathcal{F}$. This property is crucial in order to use the theory of multidimensional diffusion by [15].

The limit process $(\hat{F}_t)_{t\geq 0}$ is not only càd-làg but continuous with respect to time as it is the solution of the linear differential Equation (2.11), where the functions $(g_i)_{1\leq i\leq n}$ need to be replaced by $(\bar{g}_i)_{1\leq i\leq n}$ in the definition of $\mathcal{L}$. The limit process is even smooth in time as shown by the explicit solution given in Theorem 2.9.

4. Numerical illustration

We study numerically the behavior of the vanishing learning rate limit $\hat{F}_t(\cdot)$ given in Equation (2.7). We consider the experimental design from [11] Section 6.1: the sample $(Y_i, X_i)_{1\leq i\leq n}$ is generated according to

$$\begin{cases} X_i \sim \text{Unif}([-1,1]), \\ \varepsilon_i \sim \mathcal{N}(0, 1/4), \\ Y_i = f(X_i) + \varepsilon_i, \end{cases} \tag{4.1}$$

with regression function

$$f(x) = 1 - \Big|2|x| - 1\Big|, \quad x \in [-1,1].$$

The covariates $(X_i)_{1\leq i\leq n}$ and the errors $(\varepsilon_i)_{1\leq i\leq n}$ are assumed i.i.d. and independent on each other. For the boosting procedure, we use a cubic smoothing spline with 5 degrees of freedom as linear base learners, see Example 2.4 with $r = 2$. Recall that the degrees of freedom, noted df, is equal to the trace of the matrix S defined in Equation (2.3) and reflects the complexity of the base learner.

We first simulate $n = 100$ observations of model (4.1) and compute the limit functions $\hat{F}_t(\cdot)$ for different time values $t = 0$, 1, 10, 100 and 1000. We can see that $t = 10$ produces a fairly good fit while $t = 0$ or 1 produces an underfit and $t = 100$ or 1000 an overfit. Such large values of t are rarely used in practice and this shows that overfitting eventually arises, but very slowly.

To analyse this overfit and the effect of the learning rate λ, we then compare the training and test errors as defined in Section 2.3. In Figure 2 below, we plot the training and test errors of the boosting predictor $\hat{F}^\lambda_{[t/\lambda]}(\cdot)$ as a function of time $t \geq 0$ (in logarithmic scale), for different learning rates $\lambda = 1$, 0.5, 0.1 and for the vanishing learning rate limit $\lambda \to 0$. We can see that the training error is decreasing while the test error decreases until a minimum at $t \approx \exp(1.8) \approx 6$ and starts to increase again. As $\lambda \to 0$, the error functions converge quickly to their limit and $\lambda = 0.1$ can hardly be distinguished from the limit $\lambda = 0$. One can see that convergence is slower for small time values and uniformly fast for large time values.

Importantly, training error decreases slowly than expected, in contrast with the exponential rate of convergence announced by the theory. The convergence to 0 of the training error cannot be observed on Figure 2 where t ranges from 0 to $\exp(4) \approx 50$. This phenomenon is explained by the order of magnitude of the eigenvalues of the base learner. The eigenvalues of S quickly decrease to very small values as seen in Figure 3 below, where the 60 largest eigenvalues are plotted in logarithmic scale (some numerical instability arises for smaller eigenvalues). The rate of convergence to 0 of the training error is exponential with rate $\mathrm{e}^{-t\mu_{100}}$, where μ_{100} denotes the smallest eigenvalues. Here we already have $\mu_{60} \approx 6 \cdot 10^{-7}$ explaining the slow rate of decrease of the training error and the fact that convergence to zero is not observed in practice on usual time range.

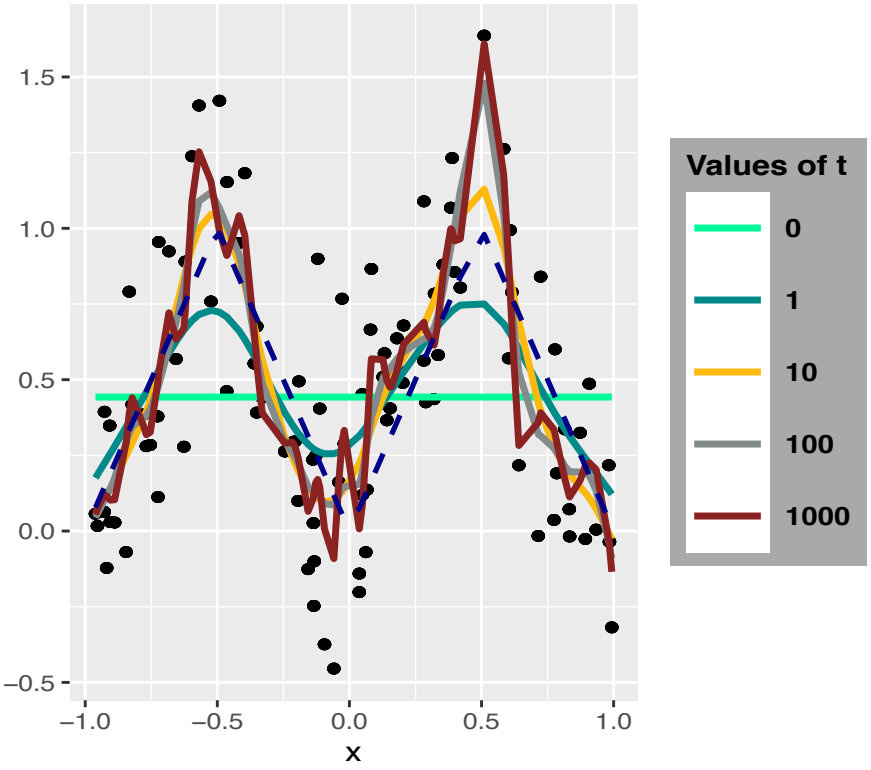

FIGURE 1. Output of the L^2-Boosting algorithm in the vanishing learning rate asymptotic with input $(Y_i, X_i)_{1\le i\le 100}$ generated from model (4.1) at different values of t. The black dots represent the observed data set $(Y_i, X_i)_{1\le i\le 100}$ and the dashed blue line represents the regression function f.

We next discuss what this behaviour of eigenvalues imply for the boosting predictor $\hat{F}_t(\cdot)$. When considering prediction at $\mathbf{x} = (x_i)_{1\le i\le n}$, we can write $\hat{F}_t(\mathbf{x}) = (\hat{F}_t(x_i))_{1\le i\le n}$ as

$$\hat{F}_t(\mathbf{x}) = u_1 u_1^\top Y + \sum_{i=2}^{n} (1 - \mathrm{e}^{-\mu_i t}) u_i u_i^\top Y \tag{4.2}$$

where $(\mu_i)_{1\le i\le n}$ and $(u_i)_{1\le i\le n}$ are the eigenvalues and eigenvectors of S and $Y = (Y_i)_{1\le i\le n}$. The rank one matrix $u_i u_i^\top$ is the matrix of the orthogonal projection on the eigenspace $\mathbb{R}u_i$. We interpret Equation (4.2) as a smoothed projection: for $\mu_i t \gg 1$, $1 - \mathrm{e}^{-\mu_i t} \approx 1$ and the boosting operator acts as the projection on this dimension; on the opposite, for $\mu_i t \ll 1$, $1 - \mathrm{e}^{-\mu_i t} \approx 0$ and the boosting operator acts as a filter on this dimension. The fact that the eigenvalues are spread out on multiple orders of magnitude implies that most of the coefficients are close to 0 or 1, whence the name smoothed projection. This is illustrated on Figure 4 where the coefficients from Equation (4.2) are plotted for various values of t. The sum of the coefficient is equal to the trace of the linear boosting operator, that is its number of degrees of freedom

$$\mathrm{df}(t) = 1 + \sum_{i=2}^{n} (1 - \mathrm{e}^{-\mu_i t}),$$

which is an increasing function of time. As we can see, boosting acts as a smoothed projection on the linear spaced spanned by the first eigenvectors. The larger the time, the larger the number of degrees of freedom of the smoothed projection, that is the larger the projection space.

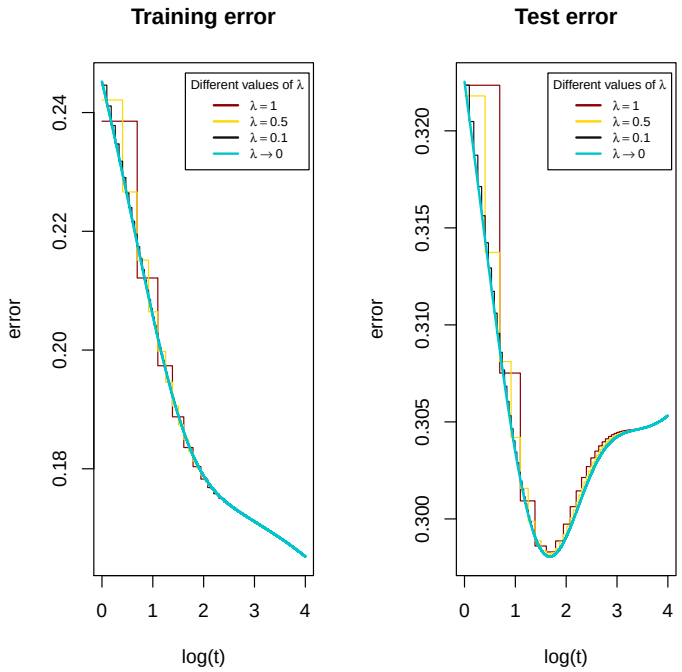

FIGURE 2. Training and test error of the predictors $\hat{F}_t(\cdot)$ and $\hat{F}^{\lambda}_{[t/\lambda]}(\cdot)$ for different values of λ.

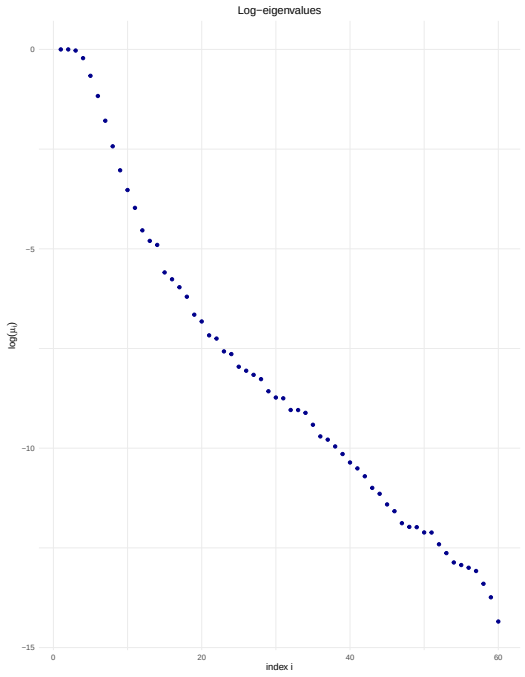

FIGURE 3. Decay of the base learner eigenvalues in logarithmic scale – only the 60 largest eigenvalues are plotted due to numerical instability for smaller ones.

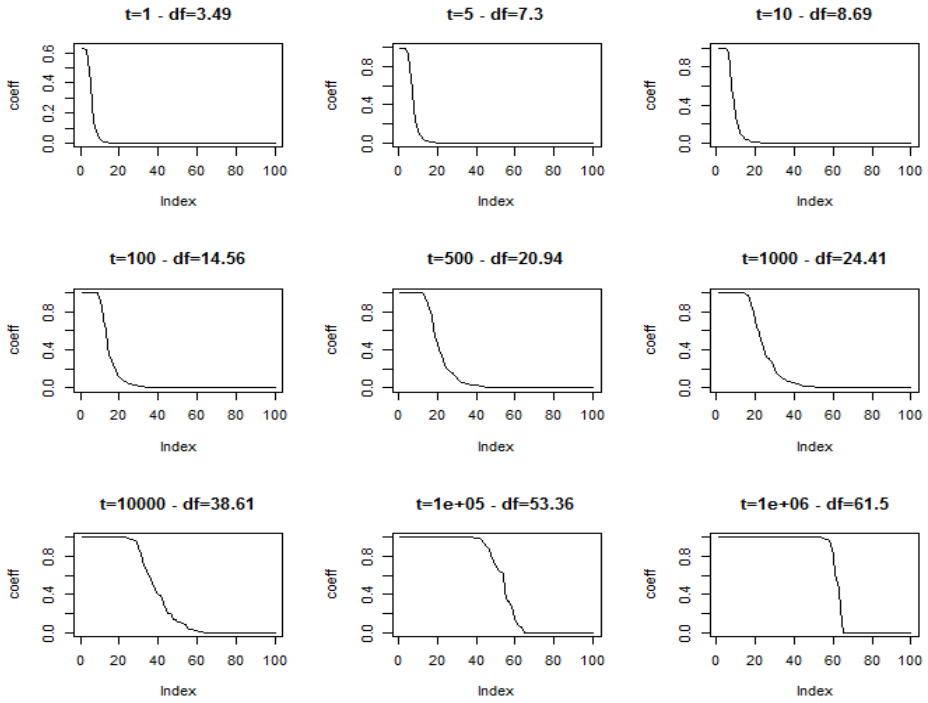

FIGURE 4. Coefficient appearing in Equation (4.2), the interpretation of the boosting linear operator as a smoothed projection.

We finally investigate the behaviour of the eigenvectors of base learner. Figures 5 and 6 respectively show the first and last eigenvectors. Quite strikingly, we can see that the first eigenvectors contains a well structured signal (akin to a polynomial basis) while the last eigenvectors mostly contains noise. The interpretation of linear boosting as a smoothed projection is thus meaningful as projection is performed on small dimensions containing signal, while higher dimensions associated to noise are filtered out.

To illustrate our results using real data, we have selected the "Communities and Crime" dataset (refer to [28] for further details). This dataset provides information on various socio-economic factors and crime rates in diverse communities across the United States. Typically, the dataset includes variables such as demographic characteristics (*e.g.*, population size, education level), economic indicators (*e.g.*, income levels, employment rates), and crime statistics (*e.g.*, rates of various types of crimes such as murder, assault, burglary). Here, we have chosen to model the variable "ViolentCrimesPerPop," which quantifies the rate of violent crimes per capita within a given community. This metric serves as a crucial measure for evaluating the prevalence and severity of violent crimes across different communities.

We compare the performances of predictors $\hat{F}_t$ and $\hat{F}^{\lambda}_{[t/\lambda]}$ with different values of λ based on cubic smoothing spline with 5 degrees of freedom. To achieve this, we divided the dataset into a training set (80% of the observations, *i.e.*, 1596 observations), on which we calibrated our models, and a test set (20% of the dataset rows, *i.e.*, 398 observations). The time t for each predictor was selected through cross-validation. Table 1 summarizes the test error of each model. It can be observed that larger values of the learning rate ($\lambda = 1$ or 0.5) lead to suboptimal test errors. On the other hand, infinitesimal gradient boosting ($\lambda \to 0$) achieves the smallest test error. We can furthermore see that the test error stabilizes for small learning rates ($\lambda = 0.1$, 0.05 and 0.01). This justifies the common practice to fix a small learning rate ($\lambda = 0.05$ or $\lambda = 0.01$) so that the vanishing learning rate asymptotic applies and avoids the selection of this hyperparameter.

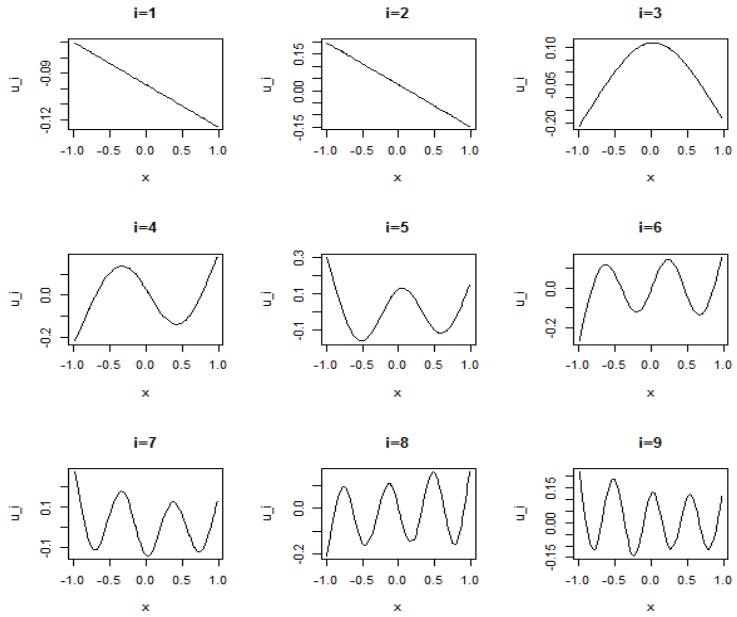

FIGURE 5. First eigenvectors of the linear base learner interpreted as signal.

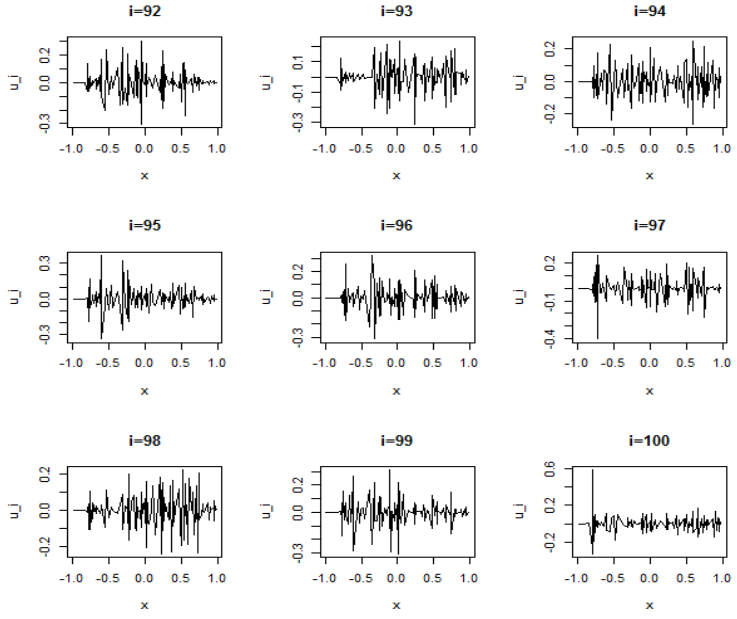

FIGURE 6. Last eigenvectors of the linear base learner interpreted as noise.

TABLE 1. Test errors of standard and infinitesimal gradient boosting procedures on the "Communities and Crime" dataset.

λ	1	0.5	0.1	0.05	0.01	$\to 0$
Test error	0.22956	0.22927	0.21737	0.21733	0.21732	0.21728

5. Conclusion

Gradient boosting is a powerful and widely used machine learning algorithm due to its high predictive accuracy and robustness to overfitting. This ensemble machine technique comes with several hyperparameters that require careful calibration to achieve optimal model performance. In particular, the learning rate which controls the step size at each iteration during the gradient boosting process. This parameter balances the trade-off between training speed and model accuracy. A smaller learning rate often leads to more accurate models but requires more iterations. Gradient boosting with a shrinkage parameter approaching zero is typically more convenient in certain scenarios. This improves the model generalization in the sense that each individual base learner added to the ensemble has a very small impact on the overall model. This leads to a more gradual and stable learning process.

In this work, we studied the asymptotic behavior of gradient boosting algorithms when the learning parameter goes to zero (infinitesimal gradient boosting). We mainly considered L^2-boosting for regression with a linear base learner. We proved a deterministic limit and characterized it as the unique solution of a linear differential equation in an infinite-dimensional function space. This discovery has at least two advantages: the ease of implementing the algorithm and the circumvention of the problem of choosing the learning rate. The result process in this case is written as a weighted sum of the functions $(g_i)_{i=1,\dots,n}$ and easily implemented in practice.

The existence of the vanishing learning rate limit justifies the tradeoff observed in practice between learning rate and number of iterations. Roughly speaking, for small learning rates, dividing the learning rate by a factor 2 and multiplying the number of iterations by a factor 2 will lead to similar performances. This justifies the common practice to fix a small learning rate (typically $\lambda = 0.01$) and then to look for the optimal number of iterations with cross-validation. This is usually done by monitoring the prediction error on a validation set as the number of iterations increases.

For a correct and rigorous use of infinitesimal gradient boosting in practice, it is necessary to choose the number of iterations carefully in order to avoid overfitting problems. One of the most effective approaches to determine the optimal number of iterations is to use cross-validation. Split your dataset into training and validation subsets. Train the gradient boosting model with a range of iteration values, and for each value, evaluate the model's performance on the validation set. Select the number of iterations that results in the best validation performance.

It's important to note that there is no one-size-fits-all answer for the number of iterations, and the optimal value can vary from one problem to another. Therefore, experimenting with different values, using cross-validation, and monitoring model performance are essential steps in finding the right number of iterations.

6. Proofs

6.1. Proofs for Section 2

Proof of Proposition 2.2. A straightforward induction based on the recursive Equation (2.3) yields the first formula for w_m^λ. Furthermore, the identity

$$(I - \lambda S) \sum_{j=0}^{m-1} (I - \lambda S)^j = \sum_{j=0}^{m} (I - \lambda S)^j - I$$

implies

$$\lambda S \sum_{j=0}^{m-1} (I - \lambda S)^j = I - (I - \lambda S)^m.$$

When S is invertible, we deduce

$$\lambda \sum_{j=0}^{m-1} (I - \lambda S)^j = S^{-1}[I - (I - \lambda S)^m]$$

and the second formula for w_m^λ follows. □

Proof of Proposition 2.6. Using the first formula for w_m^λ from Proposition 2.2 and the binomial theorem, we get

$$w_m^\lambda = \lambda \sum_{j=0}^{m-1} \sum_{k=0}^{j} \binom{j}{k} (-\lambda S)^k \tilde{Y} = \lambda \sum_{k=0}^{m-1} (-\lambda S)^k \sum_{j=k}^{m-1} \binom{j}{k} \tilde{Y}.$$

By the Hockey-stick identity,

$$w_m^\lambda = \lambda \sum_{k=0}^{m-1} \binom{m}{k+1} (-\lambda S)^k \tilde{Y} = -\sum_{j=1}^{m} \binom{m}{j} (-\lambda)^j S^{j-1} \tilde{Y}. \tag{6.1}$$

Let $T > 0$. In view of (2.2) and (2.7),

$$\sup_{t\in[0,T]} \sup_{x\in[0,1]^p} \left| \hat{F}^\lambda_{[t/\lambda]}(x) - \hat{F}_t(x) \right| \le \sup_{t\in[0,T]} \sum_{i=1}^{n} \left| w^\lambda_{[t/\lambda],i} - w_{t,i} \right| \sup_{x\in[0,1]^p} |g_i(x)| .$$

The functions $g_1, \dots, g_n$ are locally bounded so that

$$\sup_{t\in[0,T]} \sup_{x\in[0,1]^p} \left| \hat{F}^\lambda_{[t/\lambda]}(x) - \hat{F}_t(x) \right| \le M \sup_{t\in[0,T]} \left\| w^\lambda_{[t/\lambda]} - w_t \right\| \tag{6.2}$$

for some $M > 0$. Here, $\|\cdot\|$ denotes any norm on $\mathbb{R}^n$ and we use below the same notation for the induced operator norm of $n \times n$ matrix. We need to prove for $0 < \zeta < 1$ that

$$\lambda^{-\zeta} \| w^\lambda_{[t/\lambda]} - w_t \| \to 0 \quad \text{uniformly for } t \in [0, T]. \tag{6.3}$$

Equation (6.1) implies

$$w^\lambda_{[t/\lambda]} - w_t = \sum_{j=1}^{[t/\lambda]} \left(\frac{(-t)^j}{j!} - \binom{[t/\lambda]}{j} (-\lambda)^j \right) S^{j-1} \tilde{Y} + \sum_{j>[t/\lambda]} \frac{(-t)^j}{j!} S^{j-1} \tilde{Y},$$

whence we deduce $\lambda^{-\zeta} \| w^\lambda_{[t/\lambda]} - w_t \| \le \mathrm{I} + \mathrm{II}$ with

$$\mathrm{I} = \lambda^{-\zeta} \sum_{j=1}^{[t/\lambda]} \left| \frac{t^j}{j!} - \binom{[t/\lambda]}{j} \lambda^j \right| \|S\|^{j-1} \|\tilde{Y}\|,$$

$$\mathrm{II} = \lambda^{-\zeta} \sum_{j>[t/\lambda]} \frac{t^j}{j!} \|S\|^{j-1} \|\tilde{Y}\|.$$

Consider the first term. For $t \in [0,T]$, $\lambda > 0$ and $1 \le j \le [t/\lambda]$,

$$\binom{[t/\lambda]}{j}\lambda^j \ge \frac{1}{j!}\left(\lambda\,[t/\lambda] - \lambda(j-1)\right)^j \ge \frac{1}{j!}(t-\lambda j)^j,$$

which entails

$$\left|\frac{t^j}{j!} - \binom{[t/\lambda]}{j}\lambda^j\right| \le \frac{1}{j!}(t^j - (t-\lambda j)^j) \le \frac{1}{j!}(T^j - (T-\lambda j)^j)$$

where the last inequality uses the fact that $t \mapsto t^j - (t-\lambda j)^j$ is increasing on $[\lambda j, \infty)$. We deduce

$$\mathrm{I} \le \lambda^{-\zeta}\|\tilde{Y}\| \sum_{j=1}^{[t/\lambda]} \frac{1}{j!}(T^j - (T-\lambda j)^j)\|S\|^{j-1}.$$

Thanks to the inequality

$$T^j - (T-\lambda j)^j = \lambda j \sum_{k=0}^{j-1} T^{j-k-1}(T-\lambda j)^k \le \lambda j^2 T^{j-1},$$

we get the upper bound

$$\mathrm{I} \le \lambda^{1-\zeta}\|\tilde{Y}\| \sum_{j=1}^{\infty} \frac{1}{j!} j^2 (T\|S\|)^{j-1}.$$

Since the last series converges and $t \in [0,T]$ is arbitrary, we deduce that $\mathrm{I} = O(\lambda^{1-\zeta})$ as $\lambda \to 0$, uniformly for $t \in [0,T]$.

Observe now that for $t \in (0,T]$ (the case $t = 0$ is obvious), the second term satisfies

$$\mathrm{II} \le \lambda^{-\zeta}\|\tilde{Y}\|\,([t/\lambda]+1)^{-1}\,t \sum_{j > [t/\lambda]} \frac{(t\|S\|)^{j-1}}{(j-1)!} \le \lambda^{1-\zeta}\|\tilde{Y}\| \exp(T\|S\|).$$

We deduce that $\mathrm{II} \to 0$ as $\lambda \to 0$ uniformly in $t \in [0,T]$. This proves Equation (6.3) and, in view of Equation (6.2), the convergence $\lambda^{-\zeta}(\hat{F}^{\lambda}_{[t/\lambda]}(x) - \hat{F}_t(x)) \to 0$ uniformly on $[0,T] \times [0,1]^p$.

Finally, in the case when S is invertible, we have

$$\begin{aligned} w_t &= -\sum_{j \ge 1} \frac{(-t)^j}{j!} S^{j-1}\tilde{Y} = -S^{-1}\sum_{j \ge 1} \frac{(-t)^j}{j!} S^j \tilde{Y} \\ &= -S^{-1}(\mathrm{e}^{-tS} - I)\tilde{Y}. \end{aligned}$$

This proves Equation (2.9). □

Proof of Corollary 2.7. When S is symmetric, it can be written as the sum of rank 1 orthogonal projections

$$S = \sum_{j=1}^{n} \mu_j u_j u_j^{\top}.$$

Then we have

$$\sum_{k\geq 1}\frac{(-t)^k}{k!}S^{k-1}=\sum_{j=1}^{n}\left(\sum_{k\geq 1}\frac{(-t)^k}{k!}\mu_j^{k-1}\right)u_ju_j^\top$$
$$=\sum_{j=1}^{n}\frac{\mathrm{e}^{-\mu_j t}-1}{\mu_j}u_ju_j^\top$$

where the series are normally convergent and extension by continuity is used in the last equality when $\mu_j=0$. We deduce from Proposition 2.6 that

$$w_t=\sum_{j=1}^{n}\frac{1-\mathrm{e}^{-\mu_j t}}{\mu_j}u_ju_j^\top\tilde{Y}$$

and Equation (2.10) follows from Equation (2.7) since $w_{t,i}=v_i^\top w_t$, $1\leq i\leq n$. □

Proof of Theorem 2.9. We first check the claim that $\mathcal{L}$ is a bounded linear operator on $\mathbb{B}$. We denote by $\|\cdot\|_\infty$ the norm in $\mathbb{B}$. For all $x\in[0,1]^p$,

$$|\mathcal{L}(Z)(x)|\leq\sum_{i=1}^{n}|Z(x_i)||g_i(x)|\leq\Big\|\sum_{i=1}^{n}|g_i|\Big\|_\infty\|Z\|_\infty,$$

whence we deduce, taking the supremum over $x\in[0,1]^p$,

$$\|\mathcal{L}(Z)\|_\infty\leq\Big\|\sum_{i=1}^{n}|g_i|\Big\|_\infty\|Z\|_\infty.$$

This proves that the linear operator $\mathcal{L}$ is bounded.

Point i) In the Banach space $\mathbb{B}$, the differential equation (2.11) is linear of first order with constant bounded linear operator $\mathcal{L}$ and hence it follows from the general theory (see [29]) that it admits a unique solution starting from any point Z_0. We check that $Z(t)$ defined by Equation (2.12) is this solution. For $t=0$, $\mathrm{e}^{-t\mathcal{L}}=\mathrm{Id}$ so that Equation (2.12) yields $Z(0)=Z_0$. On the other hand, differentiating Equation (2.12) thanks to the relations $(\mathrm{e}^{-t\mathcal{L}})'=-\mathcal{L}\mathrm{e}^{-t\mathcal{L}}$, we obtain

$$\begin{aligned}Z'(t)&=-\mathcal{L}\mathrm{e}^{-t\mathcal{L}}Z(0)+\mathcal{L}\mathrm{e}^{-t\mathcal{L}}\mathcal{Y}\\&=-\mathcal{L}(Z(t)-(\mathrm{Id}-\mathrm{e}^{-t\mathcal{L}})\mathcal{Y})+\mathcal{L}\mathrm{e}^{-t\mathcal{L}}\mathcal{Y}\\&=-\mathcal{L}Z(t)+\mathcal{L}\mathcal{Y}\\&=-\mathcal{L}Z(t)+G.\end{aligned}$$

In the last equality, we use the fact that $\mathcal{Y}$ is such that $\mathcal{L}(\mathcal{Y})=G$. This proves that $Z(t)$ is the solution of (2.11) with initial condition $Z(0)=Z_0$.

Point ii) We finally check that $(\hat{F}_t)_{t\geq 0}$ is the solution of (2.11) with initial condition $\bar{Y}_n$. By construction, we have $\hat{F}_0=\bar{Y}_n$. Furthermore, differentiating the relation

$$\hat{F}_t=\bar{Y}_n+\sum_{i=1}^{n}w_{t,i}g_i$$

yields

$$\hat{F}_t' = \sum_{i=1}^{n} w_{t,i}' g_i$$

where $w_{t,i}'$ is the i-th component of

$$w_t' = \sum_{j\geq 1} \frac{(-t)^{j-1}}{(j-1)!} S^{j-1}\tilde{Y} = \mathrm{e}^{-tS}\tilde{Y}.$$

The derivative of w_t is obtained by differenting the power series defining w_t and we can see that w_t satisfies the differential equation

$$w_t' = -Sw_t + \tilde{Y}, \quad t \geq 0.$$

As a consequence, for $t \geq 0$,

$$\begin{aligned}
\hat{F}_t' &= \sum_{i=1}^{n} \Big(-\sum_{j=1}^{n} S_{i,j} w_{t,j} + \tilde{Y}_i \Big) g_i \\
&= -\sum_{i=1}^{n} \Big(\bar{Y}_n + \sum_{j=1}^{n} w_{t,j} g_j(x_i) \Big) g_i + \sum_{i=1}^{n} Y_i g_i \\
&= -\sum_{i=1}^{n} \hat{F}_t(x_i) g_i + \sum_{i=1}^{n} Y_i g_i \\
&= -\mathcal{L}(\hat{F}_t) + G.
\end{aligned}$$

This proves that $(\hat{F}_t)_{t\geq 0}$ is the unique solution of Equation (2.11) starting from $\bar{Y}_n$ and its explicit form follows from point *ii*). □

Proof of Proposition 2.13. Using the linear independence of $g_1, \ldots, g_n$ together with Equation (2.7), we can see that the output $\hat{F}_t$ remains bounded in $\mathbb{B}$ as $t \to \infty$ if and only if the weight w_t remains bounded in $\mathbb{R}^n$ as $t \to \infty$. Using the explicit formula (2.8) for w_t and the Jordan decomposition of S, we prove that $(w_t)_{t\geq 0}$ remains bounded if and only if, for all Jordan block B of S,

$$\sum_{k\geq 1} \frac{(-t)^k}{k!} B^{k-1} \quad \text{remains bounded as } t \to \infty. \tag{6.4}$$

Indeed, the assumption $\sum_{i=1}^{n} g_i(x) = 1$ implies that 1 is an eigenvalue of S associated to the constant eigenvector 1_n. It follows that the centered input $\tilde{Y}$ in the definition of (2.8) can provide a contribution related to any other Jordan block.

Finally, we characterize the property (6.4). Write the Jordan block B of size s in the form $B = \mu I_s + N$ where μ is an eigenvalue of S and N a nilpotent matrix of order $s-1$. A standard discussion, akin to the criterion for the stability of linear systems of differential equations (see [30]), reveals that (6.4) holds if and only if μ has positive real part or if $s = 1$ and μ has a null real part. □

Proof of Proposition 2.14. Point i). The convergence

$$\mathrm{err}_{train}(t) = \frac{1}{n}\|\mathrm{e}^{-tS}\tilde{Y}\| \longrightarrow 0, \quad \text{as } t \to \infty,$$

for all possible input is equivalent to the matrix convergence

$$\mathrm{e}^{-tS} \longrightarrow 0, \quad \text{as } t \to \infty.$$

Note indeed that the centered input $\tilde{Y}$ belongs to the space orthogonal to 1_n but the direction 1_n is well-controlled since Assumption 2.12 implies $S1_n = 1_n$ and hence $\mathrm{e}^{-tS}1_n = \mathrm{e}^{-t}1_n \to 0$. Finally, the convergence $\mathrm{e}^{-tS} \longrightarrow 0$ is equivalent to the fact that all the (complex) eigenvalues of S have a positive real part (see for example [31] and the references therein).

Point ii). The relation

$$\begin{aligned}\mathbb{E}[\mathrm{err}_{train}(t)] &= \frac{1}{n}\mathbb{E}[\|R_t\|^2] \\ &= \frac{1}{n}\|\mathbb{E}[R_t]\|^2 + \frac{1}{n}\mathrm{Trace}(\mathrm{Var}(R_t))\end{aligned} \tag{6.5}$$

yields the decomposition into squared bias and variance. The vector of residuals is $R_t = \mathrm{e}^{-tS}\tilde{Y}$ where $\tilde{Y} = (Y_i - \bar{Y}_n)_{1\le i\le n}$ has expectation and variance

$$\begin{aligned}\mathbb{E}[\tilde{Y}] &= f - \bar{f}1_n = \tilde{f}, \\ \mathrm{Var}[\tilde{Y}] &= \sigma^2\left(I - \frac{1}{n}1_n 1_n^\top\right) = \sigma^2 J.\end{aligned}$$

We deduce the squared bias and variance

$$\begin{aligned}\mathrm{bias}^2(t) &= \frac{1}{n}\left\|\mathbb{E}[R_t]\right\|^2 = \frac{1}{n}\left\|\mathrm{e}^{-tS}\tilde{f}\right\|^2, \\ \mathrm{var}_{train}(t) &= \frac{1}{n}\mathrm{Trace}(\mathrm{Var}[R_t]) = \frac{\sigma^2}{n}\mathrm{Trace}\left(\mathrm{e}^{-tS}J\mathrm{e}^{-tS^\top}\right)\end{aligned}$$

and Equation (2.16) follows from Equation (6.5).

Point iii). When S is symmetric, we use the decomposition $S = \sum_{i=1}^n \mu_i u_i u_i^\top$ which implies

$$\mathrm{e}^{-tS}\tilde{f} = \sum_{i=1}^n \mathrm{e}^{-t\mu_i}(u_i^\top \tilde{f})u_i$$

and

$$\mathrm{bias}^2(t) = \frac{1}{n}\|\mathrm{e}^{-tS}\tilde{f}\|^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{e}^{-2t\mu_i}(u_i^\top \tilde{f})^2.$$

On the other hand, using furthermore the relations $S^\top = S$, $\mathrm{Trace}(AB) = \mathrm{Trace}(BA)$ and $J^2 = J$, we get

$$\begin{aligned}\mathrm{var}_{train}(t) &= \frac{\sigma^2}{n}\mathrm{Trace}\Big(\mathrm{e}^{-tS}J\mathrm{e}^{-tS^\top}\Big) = \frac{\sigma^2}{n}\mathrm{Trace}\Big(J\mathrm{e}^{-2tS}J^T\Big)\\ &= \frac{\sigma^2}{n}\sum_{i=1}^n \mathrm{e}^{-2t\mu_i}\mathrm{Trace}\Big(Ju_iu_i^\top J^\top\Big) = \frac{\sigma^2}{n}\sum_{i=1}^n \mathrm{e}^{-2t\mu_i}\|Ju_i\|^2.\end{aligned}$$

□

Proof of Proposition 2.15. Point i). The proof is similar to the proof of point *ii)* in Proposition 2.14 and we give the main lines only. Equation (2.15) is equivalent to

$$\mathrm{err}_{test}(t) = \frac{1}{n}\big\|R'_t\big\|^2 \quad \text{with } R'_t = (Y'_i - \hat{F}_t(x_i))_{1\le i\le n},$$

so that $\mathbb{E}[\mathrm{err}_{test}(t)] = \mathrm{bias}^2(t) + \mathrm{var}_{test}(t)$ with

$$\begin{aligned}\mathrm{bias}^2(t) &= \frac{1}{n}\big\|\mathbb{E}[R'_t]\big\|^2,\\ \mathrm{var}_{test}(t) &= \frac{1}{n}\mathrm{Trace}(\mathrm{Var}(R'_t)).\end{aligned}$$

Since $\mathbb{E}[R'_t] = \mathbb{E}[R_t] = \mathrm{e}^{-tS}\tilde{f}$, the squared bias is the same as in Proposition 2.14. Finally, the formula for the variance follows from the relation

$$R'_t = Y' + \bar{Y}_n 1_n + (\mathrm{e}^{-tS} - I)\tilde{Y}$$

where $Y' = (Y'_i)_{1\le i\le n}$, $\bar{Y}_n 1_n$ and $(\mathrm{e}^{-tS} - I)\tilde{Y}$ are uncorrelated with variance $\sigma^2 I$, $\sigma^2 n^{-1}1_n 1_n^\top$ and $\sigma^2(I - \mathrm{e}^{-tS})J(I - \mathrm{e}^{-tS})^\top$ respectively.

Point ii). The formula are proved in the same way as the formulas of point *iii)* in Proposition 2.14 and we omit the proof for the sake of brevity. The claimed properties of the squared bias and variance are straightforward. To prove the monotonicity of the expected test error near the origin and at infinity, it is enough to compute the derivative and prove that it is negative near 0 and positive near infinity. The limit $2\sigma^2$ relies on the fact that $\sum_{i=1}^n \|Ju_i\|^2 = n - 1$ because J is an orthogonal projection of rank $n - 1$ and $(u_i)_{1\le i\le n}$ and orthonormal basis. Details are left to the reader. □

Proof of Proposition 2.17. Conditionally on $X' = x'$, we have the decomposition

$$\begin{aligned}\mathbb{E}[(Y' - \hat{F}_t(X'))^2 \mid X' = x'] &= \sigma^2 + \mathbb{E}[(f(x') - \hat{F}_t(x'))^2]\\ &= \sigma^2 + (f(x') - \mathbb{E}[F_t(x')])^2 + \mathrm{Var}[F_t(x')].\end{aligned}$$

By Proposition 2.6,

$$\hat{F}_t(x') = \bar{Y}_n + g(x')^\top S^{-1}\left(I_n - \mathrm{e}^{-tS}\right)\tilde{Y},$$

with expectation and variance given by

$$\begin{aligned}\mathbb{E}[\hat{F}_t(x')] &= \bar{f} + g(x')^\top S^{-1}\left(I_n - \mathrm{e}^{-tS}\right)\tilde{f}\\ \mathrm{Var}[\hat{F}_t(x')] &= \frac{\sigma^2}{n} + g(x')^\top S^{-1}\left(I_n - \mathrm{e}^{-tS}\right)J_n\left(I_n - \mathrm{e}^{-tS}\right)S^{-1}g(x').\end{aligned}$$

We deduce

$$\mathbb{E}[(Y' - \hat{F}_t(X'))^2 \mid X' = x'] = \sigma^2 + (f(x') - \bar{f} - \tilde{f}^\top S^{-1}\left(I_n - \mathrm{e}^{-tS}\right) g(x'))^2 + \frac{\sigma^2}{n} + g(x')^\top S^{-1}\left(I_n - \mathrm{e}^{-tS}\right) J_n \left(I_n - \mathrm{e}^{-tS}\right) S^{-1} g(x').$$

Integrating with respect to x', we obtain the announced value of the test error. □

6.2. Proofs for Section 3

Proof of Proposition 3.5. As a preliminary, we state a Markov property of the stochastic boosting algorithm. For $x \in [0,1]^p$, we note $\mathbf{x} = (x_i)_{1\le i\le n+1}$ with the convention $x_{n+1} = x$ and also $\hat{F}_m^\lambda(\mathbf{x}) = (\hat{F}_m^\lambda(x_i))_{1\le i\le n+1}$. We observe that $(\hat{F}_m^\lambda(\mathbf{x}))_{m\ge 0}$ is a time homogeneous Markov chain with values in $\mathbb{R}^{n+1}$. Indeed, the recursive relation (3.1) implies that $\hat{F}_{m+1}^\lambda(\mathbf{x})$ depends only of $(\hat{F}_m^\lambda(x_i))_{1\le i\le n}$, and ξ_{m+1}. The time homogeneous Markov property follows since $\hat{F}_m^\lambda(\mathbf{x})$ contains $(\hat{F}_m^\lambda(x_i))_{1\le i\le n}$ in its first n components and ξ_{m+1} is independent on the past $\hat{F}_0^\lambda(\mathbf{x}), \dots, \hat{F}_m^\lambda(\mathbf{x})$.

Point i). Taking conditional expectation, Equation (3.1) implies

$$\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(\mathbf{x}) \mid \hat{F}_m^\lambda(\mathbf{x})] = \hat{F}_m^\lambda(\mathbf{x}) + \lambda \sum_{i=1}^n (Y_i - \hat{F}_m^\lambda(x_i))\bar{g}_i(\mathbf{x}), \quad m \ge 0, \tag{6.6}$$

with $\bar{g}_j(\mathbf{x}) = (\bar{g}_j(x_i))_{1\le i\le n+1}$. We deduce

$$\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(\mathbf{x})] = \mathbb{E}_\xi[\hat{F}_m^\lambda(\mathbf{x})] + \lambda \sum_{i=1}^n (Y_i - \mathbb{E}_\xi[\hat{F}_m^\lambda(x_i)])\bar{g}_i(\mathbf{x}), \quad m \ge 0.$$

Considering component $n+1$, we see that the functions $x \mapsto \mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(x)]$ satisfy the recursive relation (1.2) where the linear base learner is given by (2.1) with g_j replaced by $\bar{g}_j$. Proposition 2.2 then yields the explicit form for $\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(\mathbf{x})]$ stated in Equation (3.3).

Point ii). In order to compute the variance of $\hat{F}_m^\lambda(x_j)$, $1 \le j \le n+1$, we use the recursive relation (1.2) together with the variance decomposition

$$\mathrm{Var}_\xi[\hat{F}_{m+1}^\lambda(x_j)] = \mathrm{Var}_\xi[\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(x_j) \mid \hat{F}_m^\lambda(\mathbf{x})]] + \mathbb{E}_\xi[\mathrm{Var}_\xi[\hat{F}_{m+1}^\lambda(x_j) \mid \hat{F}_m^\lambda(\mathbf{x})]].$$

In view of Equation (6.6), the first term satisfies

$$\begin{aligned}
&\mathrm{Var}_\xi[\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda(x_j) \mid \hat{F}_m^\lambda(\mathbf{x})]] \\
&= \mathrm{Var}_\xi\Big[\hat{F}_m^\lambda(x_j) + \lambda \sum_{i=1}^n (Y_i - \hat{F}_m^\lambda(x_i))\bar{g}_i(x_j)\Big] \\
&= \mathrm{Var}_\xi[\hat{F}_m^\lambda(x_j)] + \lambda^2 \sum_{1\le i,k\le n} \bar{g}_i(x_j)\bar{g}_k(x_j)\mathrm{Cov}_\xi[\hat{F}_m^\lambda(x_i), \hat{F}_m^\lambda(x_k)] \\
&\quad - 2\lambda \sum_{i=1}^n \bar{g}_i(x_j)\mathrm{Cov}_\xi[\hat{F}_m^\lambda(x_j), \hat{F}_m^\lambda(x_i)] \\
&\le (1 + \lambda M_1)^2 \max_{1\le i\le n+1} \mathrm{Var}_\xi[\hat{F}_m^\lambda(x_i)],
\end{aligned}$$

where M_1 is given by Equation (3.4). In the last inequality, we use the Cauchy-Schwartz inequality to upper bound the covariances. We next provide an upper bound for the second term in the variance decomposition. We have

$$\begin{aligned}
&\mathrm{Var}_\xi[\hat{F}^\lambda_{m+1}(x_j) \mid \hat{F}^\lambda_m(\mathbf{x})] \\
&= \lambda^2 \mathrm{Var}_\xi\Big[\sum_{i=1}^n (Y_i - \hat{F}^\lambda_m(x_i)) g_i(x_j) \mid \hat{F}^\lambda_m(\mathbf{x})\Big] \\
&= \lambda^2 \sum_{1\le i,k\le n} (Y_i - \hat{F}^\lambda_m(x_i))(Y_k - \hat{F}^\lambda_m(x_k)) \mathrm{Cov}_\xi[g_i(x_j), g_k(x_j)] \\
&\le \lambda^2 M_2 \sum_{i=1}^n (Y_i - \hat{F}^\lambda_m(x_i))^2,
\end{aligned}$$

where M_2 is defined in (3.5). The last line relies on the inequality $u^\top \Sigma u \le \rho(\Sigma)\|u\|^2$, where $u \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$ is a non negative symmetric matrix and $\rho(\Sigma)$ denotes its spectral radius, *i.e.* its largest eigenvalue. We apply this inequality with $u = (Y_i - \hat{F}^\lambda_m(x_i))_{1\le i\le n}$ and $\Sigma = (\mathrm{Cov}_\xi[g_i(x_j), g_k(x_j)])$ and we use the fact that $\rho(\Sigma) \le \mathrm{Trace}(\Sigma)$. We deduce that the second term in the variance decomposition is upper bounded by

$$\begin{aligned}
&\mathbb{E}_\xi[\mathrm{Var}_\xi[\hat{F}^\lambda_{m+1}(x_j) \mid \hat{F}^\lambda_m(\mathbf{x})]] \\
&\le \lambda^2 M_2 \sum_{i=1}^n \mathbb{E}_\xi[(Y_i - \hat{F}^\lambda_m(x_i))^2] \\
&\le n\lambda^2 M_2 \max_{1\le i\le n+1} \mathrm{Var}_\xi[\hat{F}^\lambda_m(x_i)] + \lambda^2 M_2 \sum_{i=1}^n (Y_i - \mathbb{E}_\xi[\hat{F}^\lambda_m(x_i)])^2.
\end{aligned}$$

Collecting the two terms of the variance decomposition, we get

$$\mathrm{Var}_\xi[\hat{F}^\lambda_{m+1}(x_j)] \le (1+\alpha) a_m + \beta b_m$$

with $\alpha = 2\lambda M_1 + \lambda^2 M_1^2 + n\lambda^2 M_2$, $\beta = \lambda^2 M_2$, $a_m = \max_{1\le i\le n+1} \mathrm{Var}_\xi[\hat{F}^\lambda_m(x_i)]$ and $b_m = \sum_{i=1}^n (Y_i - \mathbb{E}_\xi[\hat{F}^\lambda_m(x_i)])^2$. Taking the maximum over $j = 1, \ldots, n+1$, note that the sequence $(a_m)_{m\ge 0}$ satisfies

$$a_{m+1} \le (1+\alpha) a_m + \beta b_m, \quad m \ge 0.$$

By the discrete Gronwall lemma or a straightforward induction, we deduce

$$a_m \le (1+\alpha)^{m+1} a_0 + \beta (1+\alpha)^m \sum_{k=0}^m b_k, \quad m \ge 0.$$

We use now Equations (3.3) and (6.1) to show that

$$\begin{aligned}
b_k &\le 2\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 + 2nM_1^2 \left(\max_{1\le i\le n} |w^\lambda_{k,i}|\right)^2 \\
&\le 2n\|\tilde{Y}\|^2_\infty + 2nM_1^2 \left(\sum_{j=1}^k \binom{k}{j} \lambda^j \|S\|^{j-1}_\infty \|\tilde{Y}\|_\infty\right)^2
\end{aligned}$$

$$\leq 2n\|\tilde{Y}\|_\infty^2 + 2nM_1^2\|\tilde{Y}\|_\infty^2 \left(\lambda k \sum_{j=1}^{k} \frac{(\lambda k\|S\|_\infty)^{j-1}}{(j-1)!}\right)^2$$
$$\leq 2n\|\tilde{Y}\|_\infty^2 \left\{1 + (\lambda k M_1)^2 \mathrm{e}^{2\lambda k\|S\|_\infty}\right\}.$$

This implies that

$$\sum_{k=0}^{m} b_k \leq 2(m+1)n\|\tilde{Y}\|_\infty^2 \left\{1 + (\lambda m M_1)^2 \mathrm{e}^{2\lambda m\|S\|_\infty}\right\}.$$

Therefore, as $a_0 = 0$, it follows that

$$\max_{1\leq i\leq n+1} \mathrm{Var}_\xi[\hat{F}_m^\lambda(x_i)] \leq 2(m+1)\beta(1+\alpha)^m n\|\tilde{Y}\|_\infty^2 \left\{1 + (\lambda m M_1)^2 \mathrm{e}^{2\lambda m\|S\|_\infty}\right\}.$$

Since, for $\lambda < 1$, $\alpha \leq (2M_1 + M_1^2 + (n+1)M_2)\lambda$ and $2\beta \leq (2M_1 + M_1^2 + (n+1)M_2)\lambda^2$, the expected upper bound for the variance of the stochastic boosting output $\hat{F}_m^\lambda(x)$ is obtained. □

Proof of Corollary 3.6. First observe that

$$\lim_{\lambda\to 0} \mathbb{E}_\xi[\hat{F}_{[t/\lambda]}^\lambda(x)] = \bar{Y}_n + \sum_{i=1}^{n} w_{t,i}\bar{g}_i(x) = \hat{F}_t(x).$$

This is a straightforward consequence of Equation (3.3) and of the weight convergence $w_{[t/\lambda]}^\lambda \to w_t$ stated in Proposition 2.6. Together with the convergence of the variance $\mathrm{Var}_\xi[\hat{F}_{[t/\lambda]}^\lambda(x)] \to 0$ deduced from point *ii)* of Proposition 3.5, this yields the convergence in quadratic mean $\hat{F}_{[t/\lambda]}^\lambda(x) \to \hat{F}_t(x)$ as $\lambda \to 0$. □

Proof of Theorem 3.7. Point i). The recursive relation (3.1) can be rewritten as

$$\hat{F}_{m+1}^\lambda = T(\hat{F}_m^\lambda, \xi_{m+1}), \quad m \geq 0,$$

with $T : \mathbb{B} \times \Xi \to \mathbb{B}$ defined by

$$T(F, \xi) = F(\cdot) + \lambda \sum_{i=1}^{n} (Y_i - F(x_i)) g_i(\cdot, \xi).$$

Since $(\xi_m)_{m\geq 1}$ is i.i.d. and independent of $(x_i)_{1\leq i\leq n}$, $(Y_i)_{1\leq i\leq n}$ and $\hat{F}_0^\lambda = \bar{Y}_n$, this implies that $(\hat{F}_m^\lambda)_{m\geq 0}$ is a time homogeneous Markov chain.

Point ii). Note that the Markov chain $(\hat{F}_m^\lambda)_{m\geq 0}$ remains in the finite dimensional subspace

$$\mathcal{F} = \mathrm{span}(1, g_i(\cdot, \xi); 1 \leq i \leq n, \xi \in \Xi) \subset \mathbb{B}$$

and that the Markov property stated in point *i)* remains true if we replace $\mathbb{B}$ by the subspace $\mathcal{F}$. We apply Theorem 11.2.3 in [15], page 272, to the Markov chain $(\hat{F}_m^\lambda)_{m\geq 0}$ on $\mathcal{F}$. Let $f \in \mathcal{F}$ and consider the local drift and volatility of the Markov chain at f defined respectively by

$$b_\lambda(f) = \lambda^{-1}\mathbb{E}_\xi[\hat{F}_{m+1}^\lambda - \hat{F}_m^\lambda \mid \hat{F}_m^\lambda = f]$$

$$a_\lambda(f) = \lambda^{-1}\mathbb{E}_\xi[(\hat{F}^\lambda_{m+1} - \hat{F}^\lambda_m)(\hat{F}^\lambda_{m+1} - \hat{F}^\lambda_m)^\top \mid \hat{F}^\lambda_m = f]$$

where f is implicitly identified with its vector of coordinates in some basis so that the product $ff^\top$ makes sense.

Given $\hat{F}^\lambda_m = f$, we have

$$\hat{F}^\lambda_{m+1} - \hat{F}^\lambda_m = \lambda \sum_{i=1}^n (Y_i - f(x_i)) g_i(\cdot, \xi_{m+1}).$$

We deduce that the local drift is given by

$$b_\lambda(f) = \sum_{i=1}^n (Y_i - f(x_i))\bar{g}_i$$

and does not depend on λ, *i.e.* $b_\lambda(f) = b(f)$.

To deal with the local volatility $a_\lambda(f)$, note first that there exists a constant $C > 0$ such that the matrix $ff^\top$ has all its coefficients bounded by $C\|f\|^2_\infty$. This is a consequence of the equivalence of norms on the finite dimensional space $\mathcal{F}$: the norm $\|\cdot\|_\infty$ is equivalent to the norm of the vector representing f in the basis implicitly used for computing $ff^\top$. We can thus bound the coefficients of the local volatility matrix by

$$\begin{aligned}
&\lambda C \mathbb{E}_\xi \Big\| \sum_{i=1}^n (Y_i - f(x_i)) g_i(\cdot, \xi_{m+1}) \Big\|^2_\infty \\
&\leq \lambda C M^2 \Big(\max_{1 \leq i \leq n} |Y_i - f(x_i)| \Big)^2,
\end{aligned}$$

where $M = \max_{\xi \in \Xi} \| \sum_{i=1}^n |g_i(\cdot, \xi)| \|_\infty$ is finite because Ξ is finite. We deduce

$$a^\lambda(f) \longrightarrow a(f) \equiv 0 \quad \text{uniformly on compact sets as } \lambda \to 0.$$

The limit functions a and b are continuous. Since $a \equiv 0$ and b is an affine function, the associated martingale problem has exactly one solution starting from any point, see [15] Lemma 6.1.4 page 140 or Theorem 6.3.4 page 152. In fact, because the limit volatility $a \equiv 0$ is vanishing, the solution of the martingale problem is the solution of the differential equation on $\mathcal{F}$

$$f'(t) = b(f(t)) = \sum_{i=1}^n (Y_i - f(x_i))\bar{g}_i, \quad t \geq 0.$$

This is exactly the differential Equation (2.11) and we have proved in Theorem 2.9 that it has a unique solution with initial condition $f(0) = \bar{Y}_n$. Then, Theorem 11.2.3 in [15] implies that the continuous processes defined by interpolation

$$\tilde{F}^\lambda_t = (1 - \{t/\lambda\})\hat{F}^\lambda_{[t/\lambda]} + \{t/\lambda\}\hat{F}^\lambda_{[t/\lambda]+1}, \quad t \geq 0,$$

converge in distribution in the space of continuous functions $\mathbb{C}([0, \infty), \mathcal{F})$

$$(\tilde{F}^\lambda_t)_{t \geq 0} \xrightarrow{d} (\hat{F}_t)_{t \geq 0}. \tag{6.7}$$

The notation $\{u\} = u - [u]$ stands for the fractional part of a real number. Finally, the convergence in distribution (3.7) in the Skorokhod space $\mathbb{D}([0,\infty),\mathcal{F})$ follows by a standard discretization argument. It holds

$$(\hat{F}^{\lambda}_{[t/\lambda]})_{t\geq 0} = \Psi_\lambda((\tilde{F}^{\lambda}_t)_{t\geq 0})$$

where $\Psi_\lambda : \mathbb{C}([0,\infty),\mathcal{F}) \longrightarrow \mathbb{D}([0,\infty),\mathcal{F})$ is the discretization functional

$$\Psi_\lambda((f_t)_{t\geq 0}) = (f_{\lambda[t/\lambda]})_{t\geq 0}.$$

The functional Ψ_λ satisfies the following property: for all converging sequence $(f^\lambda_t) \to (f_t)$ in $\mathbb{C}([0,\infty),\mathcal{F})$ as $\lambda \to 0$, it holds $\Psi_\lambda((f^\lambda_t)) \to (f_t)$ in $\mathbb{D}([0,\infty),\mathcal{F})$ as $\lambda \to 0$. Together with the convergence (6.7), this implies the convergence (3.7) by the generalized continuous mapping theorem ([27], Thm. 2.7). □

Acknowledgements. The authors sincerely thank the Editor and the two anonymous referees for their valuable comments. The first author acknowledges the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX project). The second author acknowledges the support of the ANR project "Efficient inference for large and high-frequency data" (ANR-21-CE40-0021), the ANR project "Méthodes numériques pour l'aide à la décision: préférences dynamiques et risques multivariés" (ANR-21-CE46-0002) and the chair "Efficience et Sobriété Numériques", a joint initiative by Le Mans University and EREN Groupe under the aegis of the Institut Louis Bachelier.

Data availability statement. The real data analyzed in this work (see Section 4) is available on the UCI Machine Learning Repository *via* the following DOI: https://doi.org/10.24432/C53W3X [32].

References

[1] Y. Freund and R. Schapire, Adaptive Game Playing Using Multiplicative Weights, Vol. 29 (1999) 79–103.
[2] S. Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed, Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica* **12** (2002) 111–139.
[3] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, Aggregate features and adaboost for music classification. *Mach. Learn.* **65** (2006) 473–484.
[4] J. Friedman, T. Hastie and R. Tibshirani, Additive logistic regression: a statistical view of boosting. *Ann. Statist.* **28** (2000) 337–407.
[5] J.H. Friedman, Greedy function approximation: a gradient boosting machine. *Ann. Statist.* (2001) 1189–1232.
[6] G. Ridgeway, Generalized boosting models: a guide to the gbm package. (2007). URL https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.
[7] T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California, USA, 2016*. ACM (2016) 785–794. ISBN 978-1-4503-4232-2.
[8] R.E. Schapire and Y. Freund, Boosting: Foundations and Algorithms. Cambridge University Press (2012).
[9] G. Biau and B. Cadre, Optimization by gradient boosting (supplementary material), in *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan, edited by A. Daouia and A. Ruiz-Gazen. Springer, Cham (2001) 23–44*.
[10] L. Breiman, Population theory for boosting ensembles. *Ann. Statist.* **32** (2004) 1–11.
[11] T. Zhang and B. Yu, Boosting with early stopping: convergence and consistency. *Ann. Statist.* **33** (2005) 1538–1579.
[12] P.L. Bartlett and M. Traskin, AdaBoost is consistent. *J. Mach. Learn. Res.* **8** (2007) 2347–2368.
[13] P. Bühlmann and B. Yu, Boosting with the L_2 loss: regression and classification. *J. Am. Statist. Assoc.* **98** (2003) 324–339.
[14] S.N. Ethier and T.G. Kurtz, Markov Processes. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York (1986).
[15] D.W. Stroock and S.R.S. Varadhan, Multidimensional Diffusion Processes. Classics in Mathematics. Springer-Verlag, Berlin (2006). Reprint of the 1997 edition.
[16] A. Dieuleveut, Stochastic approximation in Hilbert spaces. Université Paris sciences et lettres. (2017). English. NNT:2017PSLEE059. tel-01705522v2.
[17] H. Maennel, O. Bousquet and S. Gelly, Gradient Descent Quantizes ReLU Network Features. (2018).
[18] K. Lyu and J. Li, Gradient descent maximizes the margin of homogeneous neural networks, in *International Conference on Learning Representations 2020* (2020).

[19] S.L. Smith, B. Dherin, D.G.T. Barrett and S. De, On the origin of implicit regularization in stochastic gradient descent, in *International Conference on Learning Representations 2021* (2021).
[20] P.-A. Cornillon, N.W. Hengartner and E. Matzner-Løber, Recursive bias estimation for multivariate regression smoothers. *ESAIM: PS* **18** (2014) 483–502.
[21] E. A. Nadaraya, On estimating regression. Theory Proba. Appl. **9** (1964) 141–142.
[22] G.S. Watson, Smooth regression analysis. *Sankhya* **26** (1964) 359–372.
[23] G. Wahba, Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (1990).
[24] L. Györfi, M. Kohler, A. Krzyżak and H. Walk, A Distribution-free Theory of Nonparametric Regression. Springer Series in Statistics, Springer-Verlag, New York (2002).
[25] R. Horn and C. Johnson, Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge (2013).
[26] J.H. Friedman, Stochastic gradient boosting. *Computat. Statist. Data Anal.* **38** (2002) 367–378.
[27] P. Billingsley, Convergence of Probability Measures. Wiley Series in Probability and Statistics: Probability and Statistics, 2nd edn. John Wiley & Sons, Inc., New York (1999).
[28] M. Redmond, Communities and Crime. UCI Machine Learning Repository. (2009).
[29] T. Apostol, Calculus. Vol. II: Multi-variable Calculus and Linear Algebra, with Applications to Differential Equations and Probability. Blaisdell International Textbook Series. Xerox College Publ. (1969).
[30] R. Bellman, Stability Theory of Differential Equations. Dover Books on Intermediate and Advanced Mathematics. Dover Publications (1969).
[31] R. Bellman, Introduction to Matrix Analysis: Second Edition. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (1997).
[32] UCI, Machine Learning Repository DOI: https://doi.org/10.24432/C53W3X