CONDITIONAL QUANTILE SEQUENTIAL ESTIMATION FOR STOCHASTIC CODES

Tatiana Labopin-Richard, F Gamboa, Aurélien Garivier

CONDITIONAL QUANTILE SEQUENTIAL ESTIMATION FOR STOCHASTIC CODES

T. LABOPIN-RICHARD, F. GAMBOA, AND A. GARIVIER

Abstract. This paper is devoted to the sequential estimation of a conditional quantile. More precisely the quantile of the output of a real stochastic code with inputs in $\mathbb{R}^d$. We introduce a stochastic algorithm based on the Robbins-Monro algorithm with data tailored by the k-nearest neighbours theory. We give conditions on the model in order that the algorithm is convergent. Further, we provide non-asymptotic rate of convergence of the means square error. We also focus on the best tuning parameters of the algorithm.

1. Introduction

In the last decades, computer code experiment problems have been widely studied by statisticians (for example [15], [22], [21], ...).

1.1. **Stochastic code.** In the cumputer code experiment, a stochastic code is a numerical black box model with a random seed inside. Mathematically speaking, we can model it in the following way. Let X (the inputs vector of the code) be a random vector of $\mathbb{R}^d$. Let ϵ (the random seed)be the random vector of $\mathbb{R}^m$. We assume that ϵ and X are independent random vectors. Further let g be a *regular* map from $\mathbb{R}^d \times \mathbb{R}^m$ to $\mathbb{R}$; the output of the stochastic code g is

$$Y = g(X, \epsilon). \tag{1}$$

This black box is said to be stochastic because of the unobserved random seed ϵ. Indeed, contrary to deterministic numerical black box, the code (1) does not, in general, return the same output when we feed with the same input at two different times. Notice that ϵ and g are both unknown but realisations of (X, Y) may be observed. Generally, one run of the code may be very expensive.

In this work, we will propose and study an algorithm to estimate a conditional quantile. For a fixed level $\alpha \in [\frac{1}{2}, 1]$, the target of our algorithm is

$$\theta^*(x) := q_\alpha \left(g(x, \epsilon) \right) \ \ (x \in \mathbb{R}^d)$$

Here, we denote by $q_\alpha(Z)$ the upper quantile of level α of a random variable Z. In other words

$$q_\alpha(Z) = F_Z^{-1}(\alpha),$$

Date: December 11, 2015.

where $F_Z^{-1} := \inf\{x : F_Z(x) \geq u\}$ is the generalized inverse of the cumulative distribution function of a law Z.

1.2. **The algorithm.** If a call to the code is not too expensive then classical methods may be performed to estimate a quantile. Indeed, having at hand a sample $(Y_i^x)_i$ where each Y_i^x is distributed as $g(x, \epsilon)$, we can estimate the quantile with the empirical quantile or with some classical stochastic algorithm (see next paragraph). Here, we are looking for a recursive method which allows to estimate the conditional quantile for different values of x at the same time. We begin by drawing a sample. We first draw a sample of inputs $(X_1, \dots X_n)$ that will feed the code. We then observe a sample of outputs $(Y_1 = g(X_1, \epsilon), \dots, Y_n = g(X_n, \epsilon))$. Using the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ we will iterate a stochastic algorithm allowing the estimation of the conditional quantile on several x at the same time. This algorithm is based both on the classical Robbins Monro algorithm to estimate a quantile and on the k-nearest neighbors methodology. Let us see how the algorithm works.

Robbins and Monro introduced in [18] a general stochastic algorithm to approximate the root of a function $h : \mathbb{R}^d \longrightarrow \mathbb{R}^d$. This algorithms is

$$\begin{cases} \theta_0 \in \mathbb{R}^d \\ \theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Z_{n+1}) \end{cases} \tag{2}$$

where (θ_n) is a $\mathbb{R}^d$-valued sequence, (γ_n) is a deterministic step-size sequence and (Z_n) an i.i.d sample of observations. The function H is related to the function h by the formula

$$\mathbb{E}(H(\theta_n, Z_{n+1})|\mathcal{F}_n) = h(\theta_n),$$

where $\mathcal{F}_n := \sigma(Z_1, \dots Z_n)$ is the past sigma field. This kind of algorithms has been widely studied by many authors. In an asymptotic point of view, Robbins and Monro showed convergence of the mean square error in [18]. The almost sure convergence is proved with differents methods and under different hypothesis by Blum in [5] and Schreck and al. in [23]. Fabian, Ruppert and then Sacks study the asymptotic rate of convergence respectively in [11], [19] and [20]. In [26], Woodroofe investigate the probability of large deviations for (θ_n). From a non-asymptotic point of view, there are several recent results under different assumptions. Frikha and Menozzi give in [12] non-asymptotic concentration bounds under Gaussian concentration assumption. Further Moulines et al. propose in [17] non-asymptotic inequality on the mean square error under convex assumptions.

The quantile is a classical example of target for this algorithm. Indeed the quantile of order α $(0 < \alpha < 1)$ of Z is the root of the function $h(\theta^*) = F_Z(\theta) - \alpha$ where F_Z is the cumulative distribution function of teh distribution Z. So that, to estimate the quantile (in the simple case where we have at hand a sample (Z_n) of independent and identically ditributed copies of Z) the algorithm is then the following

$$\begin{cases} \theta_0 \in \mathbb{R} \\ \theta_{n+1} = \theta_n - \gamma_{n+1}\left(\mathbf{1}_{Z_{n+1}\leq\theta_{n+1}} - \alpha\right) \end{cases} \tag{3}$$

This algorithm is consistent and leads to an estimate having Gaussian asymptotic distribution (we refer to [10] for a sum up of the asymptotic theory on Robbins-Monro algorithm). In [6], Cardot et al. study this algorithm in the median case ($\alpha = \frac{1}{2}$). In this paper, they took $\gamma_n = \frac{1}{n^\gamma}$ with $\frac{1}{2} < \gamma < 1$. Further, they provide non-asymptotic confidence balls and non-asymptotic inequality for the mean square error.

The algorithm would then be useful if we would like to estimate the conditional quantile for only one fixed x. To construct an algorithm converging for every x, we will use in addition the k-nearest neighbors theory. Let us fix an input x. To estimate the conditional quantile at x with the previous algorithm, we need to observe a sample of the output corresponding to the input x. As discused before, we can not afford to obtain a sample of the output for each input we are interested in. We only have at hand a global sample. At time $n+1$, we will use a variation of the classical quantile algorithm in up-dating only when the input falls in a neighborhood of x :

$$||X_{n+1} - x|| \leq ||X - x||_{(k_n,n)}$$

where denotes $Z_{(i,n)}$ the i-th order statistic of a sample $(Z_i)_{i=1\ldots n}$. Finally for a fixed x, our algorithm to estimate the α-quantile of the law $g(x,\epsilon)$ is the following

$$\begin{cases} \theta_0(x) \in \mathbb{R} \\ \theta_{n+1}(x) = \theta_n(x) - \gamma_{n+1} H(\theta_n(x), Y_{n+1}) \mathbf{1}_{X_{n+1}\in kNN_n(x)} \end{cases} \tag{4}$$

where we denote :

- (γ_n) is a deterministic sequence of steps. We will mainly study the case where $\gamma_n = \frac{1}{n^\gamma}$ for $0 < \gamma \leq 1$, $(n \in \mathbb{N}^*$.
- $kNN_n(x)$ is the set of the k_n nearest neighborhood of x fot the Euclidean norm on $\mathbb{R}^d$ that is

$$kNN_n(x) := \{X_i : \ ||X_i - x|| \leq ||X - x||_{(k_n,n)}, \ i = 1 \ldots n\}$$

  We study the case where $k_n = \lfloor n^\beta \rfloor$ for $0 < \beta < 1$.
- The function H, (inspired from the classical Robbins Monro theorem) is

$$H(\theta_n(x), Y_{n+1}) = \mathbf{1}_{Y_{n+1}\leq\theta_n(x)} - \alpha.$$

The idea of considering neighbors of x is classical. It appears for example in the estimation of the conditional mean. Stone in [24] and [25] study this regression problem and propose an estimator of the conditional mean based on k-nearest neighbors. He also gives conditions on k_n for this estimator to converge. Bhattacharya and al. Then use in [4] this idea to introduce estimator of the conditional quantile (non-recursive) in the case where the inputs ly in $\mathbb{R}$. This estimator is built on the generalized inverse of the empirical cumulative distribution function computed on the k_n responses corresponding

to the k_n nearest inputs of X. They study how to tune k_n to achieve optimum balance between bias and random error and show the weak convergence of their algorithm.

Notice that in our problem we have to find conditions both on k_n and γ_n. The paper is organized as follows. In Section 2 we are interested in the a.s convergence of the algorithm. We show that if $\frac{1}{2} < \gamma < \beta < 1$, then the algorithm is strongly consistent. We also prove a non-asymptotic inequality on the mean square error. This leads to the rate of convergence of the alorithm. We discuss the best parameters. Finally, in Section 3, we present some numerical simulations to illustrate our results. The technical points of the proofs are differed to Section 5.

2. Main results

In the previous section, we discuss a general method to build a sequential conditionnal quantile stochastic algorithm. In this section, we explain how to tune the parameters of this algorithm. We also give theoretical garantees of convergenceunder technical hypohesis.

2.1. **Notations and assumptions.**

In this paper, there are lots of constants. For sake of simplicity in the notations, we will split them in three types.

1) Constants denoted (L, U) are lower and upper bound of compact support of random variables. We put as an index something which make understand the support of which random variable we are dealing with.
2) Constants $(N_i)_{i\in\mathbb{N}^*}$ are integer constants which are ranks after which some properties are true.
3) Constants $(C_i)_{i\in\mathbb{N}^*}$ are all other constants.

In the case 2) et 3), when we do not precise any thing else, these constants only depends on the model, that is on g and on the distribution of (ϵ, X). Further, we will denote $C_i(u)$ or $N_i(u)$ for $u \in \mathcal{P}(\{\alpha, x, d\})$ a constant depending of the model, on the probability level α, on the point x and on the dimension d.

We recap on appendix the values of all the constants.

In the sequel, we denote Y^x a random variable with distribution $g(x, \epsilon)$. Moreover, $\mathcal{B}_x$ is the set of the balls of $\mathbb{R}^d$ centered in the point x. For $B \in \mathcal{B}_x$ we denote r_B its radius and when $r_B > 0$, Y^B is a random variable of conditional distribution $\mathcal{L}(Y|X \in B)$.

Remark 2.1. *When the couple (X, Y) has a density $f_{(X,Y)}$, when the random vector X has a density f_X which is positive, we can define the law $\mathcal{L}(Y|X = x)$ thanks to its density function*

$$f_{Y|X=x} = \frac{f_{(X,Y)}(x, .)}{f_X}.$$

In this case, we have when $B = \{x\}$,

$$Y^B \sim Y^x \sim g(x, \epsilon) \sim \mathcal{L}(Y|X = x).$$

In the sequel, we still denote F_Z the cumulative distribution function of a random variable Z.

We will need to suppose two kind of assumptions. The first one is unavoidable, since we deal with k-nearest neighbors. The three others are more technical.

Assumption A1 For all x in the support of X (that we will denote $\text{Supp}(X)$ in the sequel), there exists a constant $M(x)$ such that the following inequality holds

$$\forall B \in \mathcal{B}_x,\ \forall t \in \mathbb{R},\ |F_{Y^B}(t) - F_{Y^x}(t)| \leq M(x) r_B.$$

In other words we assume that our stochastic code is continuous enough : the law of two responses corresponding to two different but close inputs are not completely different. The assumption is clearly required, since we want approximate the law of $g(x, \epsilon)$ by $\mathcal{L}(Y|x \in kNN_n(x))$.

Remark 2.2. *When we do not consider compact support law, we can show that the last assumption holds for example as soon as (X, Y) had a regular density. In all case, it is easier to show this assumption, that the couple (X, Y) has a density. See Subsection 3.1 for an example.*

Assumption A2 The law of inputs has a density and this density is lower-bounded by a constante $C_{input} > 0$ on its support.

This hypothesis is strong. It implies that the law of X has a compact support. Notice that this kind of assumptions are usual in k-nearest neighbors context, (see for example [13]).

Assumption A3 The code function g takes its values in a compact $[L_g, U_g]$.

Lemma 2.1. *Under assumption **A3** and if $\beta \geq \gamma$, for all x, $\theta_n(x) \in [L_g - (1-\alpha), U_g + \alpha]$ a.s. .*

Proof. Imagine $\theta_n(x)$ leaves the compact set $[L_g, U_g]$ by the right at step N_0. At worst, it was in U_g at step $N_0 - 1$. This situation can be resumed in this way :

$$\begin{aligned} \theta_{N_0-1} &= U_g \\ \theta_{N_0} &= U_g + \alpha\gamma_{N_0}. \end{aligned}$$

Then, in the next step, since $\theta_{N_0} > U_g$, we have $Y_{N_0+1} \leq \theta_{N_0}$ and then

$$\theta_{N_0+1} = U_g + \alpha\gamma_{N_0} - (1-\alpha)\gamma_{N_0+1}\mathbf{1}_{X_{N_0+1} \in kNN_{N_0}(x)}.$$

Finally, since $\theta_n > U_g$ the algorithm either does not move (if $X_{n+1} \notin kNN_n(x)$) or comes back in direction of $[L_u, U_g]$ with a step of $(1-\alpha)\gamma_{n+1}$. Then, if

$$\sum_{n \geq 0} \gamma_n \mathbf{1}_{X_{n+1} \in kNN_n(x)} = +\infty \ a.s$$

the algorithm comes back to the compact set $[L_g, U_g]$. Let us then show that if $\beta \geq \gamma$, the previous sum diverges a.s. Let us denote S_n the partial sum. First, the (X_i) are independent and the occurrence or non-occurrence of the events $\{X_{n+1} \in kNN_n(x)\}$ are unchanged by finite permutations of the indices. Then, since S_n is at non negative terms, we know by the Hewit-Savage zero-one law that it converges to a constant or it

diverges to $+\infty$. The same argument gives that V_n also converges to a constant or it diverges to $+\infty$ where

$$V_n = \frac{\mathbb{E}(S_n)}{S_n}.$$

But,

$$\begin{aligned}\mathbb{E}(S_n) &= \sum_{k=1}^{n} \gamma_n \mathbb{P}(X_{k+1} \in kNN_k(x)) \\ &= \sum_{k=1}^{n} \frac{\gamma_n k_n}{(n+1)}\end{aligned}$$

where we used Lemma 6.1 to compute the probability. So, when $\beta \geq \gamma$ the sum $E(S_n)$ diverges to $+\infty$. Finally, either V_n converges to a constant and then S_n diverges to $+\infty$. Or V_n diverges to $+\infty$ and then V_n diverges to $+\infty$ in probability. It implies that for all $\epsilon > 0$, $\mathbb{P}(V_n > \epsilon) \longrightarrow 1$ (because $V_n \geq 0$). Then there is a contradiction with $\mathbb{E}(V_n) = 1$. Finally, V_n almost surely converges to a constant and so S_n diverges to $+\infty$ a.s.

Finally, we have shown that if $\beta \geq \gamma$ and if the algorithm leaves the compact set $[L_g, R_g]$ by the right, its goes to $U_g + \alpha$ as a maximum and then comes back to the compact. A similar results holds when the algorithm leave the compact set by the left and finally we have shown that

$$\theta_n(x) \in [L_g - (1-\alpha), R_g + \alpha] \text{ a.s.}$$

□

Denoting L_x the minimum of the support of X and R_x its maximum, we then have unde **A3** and if $\beta \geq \gamma$, $\sqrt{C_1} := |\max(L_g + \alpha - L_X, U_X - U_g + (1-\alpha))|$ is a bound of $|\theta_n(x) - \theta^*(x)|$.

Assumption A4 For each x, the law $g(x, \epsilon)$ has a density which is lower-bounded by a constante $C_g(x) > 0$ on its support.

Lemma 2.2. *Denoting* $C_2(x, \alpha) := \min\left(C_g(x), \frac{1-\alpha}{U_{\theta_n} - L_{\theta_n}}, \right)$*, we have thanks to assumption* ***A4****,*

$$\forall \theta_n(x), [F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))]\,[\theta_n(x) - \theta^*(x)] \geq C_2(x)\,[\theta_n(x) - \theta^*(x)]^2\,. \tag{5}$$

Proof. It is obvious when $\theta_n \in \text{Supp}(Y^x)$. When, it is not the case, we know that $\theta_n \in [L_{\theta_n}, U_{\theta_n}]$. Suppose $L_{\theta_n} \leq L_X \leq \theta^* \leq U_X \leq \theta_n \leq U_{\theta_n}$. Then, we have $F(\theta_n) = 1$, $F(\theta^*) = \alpha$ and

$$C_2(x) \leq \frac{1-\alpha}{U_{\theta_n} - L_{\theta_n}} \leq \frac{1-\alpha}{U_{\theta_n} - L_X} \leq \frac{1-\alpha}{\theta_n - \theta^*}$$

so that

$$\begin{aligned}(\theta_n(x)-\theta^*(x))(F_{Y^x}(\theta_n(x))-F_{Y^x}(\theta^*(x))) &= (\theta_n(x)-\theta^*(x))(1-\alpha)\\ &\geq (\theta_n(x)-\theta^*(x))C_2(x,\alpha)(\theta_n(x)-\theta^*(x))\\ &= C_2(x,\alpha)(\theta_n(x)-\theta^*(x))^2.\end{aligned}$$

The proof of the other cases follows similarly. □

This assumption is useful to deal with non-asymptotic inequality for the mean square error. It is the substitute of the convex assumption made in [17] which is not true in the frame of the quantile.

2.2. **A.s convergence.** The following theorem studies the a.s convergence of our algorithm.

Theorem 2.1. *Let x be a fixed input. Under assumptions **A1** and **A2**, the algorithm 4 is a.s convergent if, and only if, $\frac{1}{2} < \gamma < \beta < 1$.*

Sketch of proof : To prove this theorem, we adapt the proof of Blum in [5] of a.s convergence of the Robbins Monro algorithm to estimate a quantile. We decompose the reasoning into 3 parts and use martingale arguments. In the sequel, we still denote $\mathcal{F}_n := \sigma(X_1, \dots X_n, Y_1, \dots, Y_n)$ the past sigma field and $\mathbb{E}_n$ and $\mathbb{P}_n$ the conditional expectation and probability on $\mathcal{F}_n$. For sake of simplicity we denote

$$H(\theta_n(x), X_{n+1}, Y_{n+1}) := H(\theta_n(x), Y_{n+1})\mathbf{1}_{X_{n+1}\in kNN_n(x)}.$$

1) We decompose $H(\theta_n(x), X_{n+1}, Y_{n+1})$ in two terms : a martingale one and a remainder one :

$$h_n(\theta_n) = \mathbb{E}(H(\theta_n, X_{n+1}, Y_{n+1})|\mathcal{F}_n) \text{ and } H(\theta_n, X_{n+1}, Y_{n+1}) = h_n(\theta_n) + \xi_{n+1}.$$

Then

$$T_n = \theta_n(x) + \sum_{j=1}^{n} \gamma_j h_{j-1}(\theta_{j-1}(x))$$

is a martingale bounded in L^2. So it converges a.s.

2) We show the almost sure convergence of $(\theta_n)_n$.
   a) (θ_n) does not diverges to $+\infty$ or $-\infty$.
   b) (θ_n) converges a.s to a finite limit.
3) The limit is $\theta^*(x)$ the conditional quantile.

Steps 2a), 2b) et 3) are shown by contradiction. The key point is that almost surely, after a certain rank, $h_n(\theta_n) > 0$. This property is true thanks to assumptions **A1** and **A2** it is shown in Section 5.

Comments on parameters. In the Theorem 2.1, we assume $\frac{1}{2} < \gamma < 1$ which is a classical assumption on the Robbins Monro algorithm to be consistent (see for example in [18]). Indeed, a stepwise sequence (γ_n) such that

$$\sum_n \gamma_n = \infty \text{ and } \sum_n \gamma_n^2 < +\infty$$

is needed. The number of neighbors is $\lfloor n^\beta \rfloor$ with $0 < \beta < 1$. $\beta < 1$ means that the number a neighbors goes to $+\infty$.This implies the crucial following property (see Lemma 6.4) :

$$||X - x||_{(k_n,n)} \underset{n\to+\infty}{\longrightarrow} 0.$$

2.3. **Non-asymptotic inequality.** Here, we study the rate of converge of the mean square error that denoted by $a_n(x) := \mathbb{E}\left((\theta_n(x) - \theta^*(x))^2\right)$.

Theorem 2.2. *Let x a fixed input. Under hypothesis **A1**, **A2**, **A3** and **A4**, the mean square error $a_n(x)$ of the algorithm 4 at x satisfies the following inequality : for all $0 < \gamma \leq \beta < 1$, $1 > \epsilon > 1 - \beta$, for $n \geq 2^{\frac{1}{\epsilon-(1-\beta)}} := N_0$,*

$$\begin{aligned} a_n(x) \leq & C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + a_0(x) \exp\left(-2C_2(x)\sum_{k=1}^{n} \frac{1}{k^{\gamma+\epsilon}}\right) \\ & + \sum_{k=1}^{n} \exp\left(-2C_2(x)\sum_{i=k}^{n} \frac{1}{i^{\gamma+\epsilon}}\right)\beta_k \end{aligned}$$

where there exists a constant $C_3(d)$ such that

$$\beta_n = C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + 2\sqrt{C_1}M(x)C_3(d)\gamma_{n+1}\left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1} + \gamma_{n+1}^2\frac{k_n}{n+1}.$$

Sketch of proof : The idea of the proof is to establish the recursive inequality on $a_n(x)$ (following [17]) :

$$a_{n+1}(x) \leq a_n(x)(1 - \alpha_n) + \beta_n$$

where $0 < \alpha_n < 1$ and $\beta_n > 0$ and to conclude using Lemma 6.6. In this purpose we begin by expanding the square

$$\begin{aligned} (\theta_{n+1}(x) - \theta^*(x))^2 = & (\theta_n(x) - \theta^*(x))^2 + \gamma_{n+1}^2 \left[(1-2\alpha)\mathbf{1}_{Y_{n+1}\leq\theta_n(x)} + \alpha^2\right]\mathbf{1}_{X_{n+1}\in kNN_n(x)} \\ & - 2\gamma_{n+1}(\theta_n(x) - \theta^*(x))\left(\mathbf{1}_{Y_{n+1}\leq\theta_n(x)} - \alpha\right)\mathbf{1}_{X_{n+1}\in kNN_n(x)} \end{aligned}$$

Taking the expectation conditionally on $\mathcal{F}_n$, and using the Baye's formula, we get

$$\begin{aligned} \mathbb{E}_n\left(\theta_{n+1}(x) - \theta^*(x))^2\right) \leq & \mathbb{E}_n\left((\theta_n(x) - \theta^*(x))^2\right) + \gamma_{n+1}^2 P_n \\ & - 2\gamma_{n+1}\left(\theta_n(x) - \theta^*(x)\right)P_n\left[F_{Y^{B_n^{k_n}(x)}}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] \end{aligned}$$

where $P_n = \mathbb{P}_n\left(X_{n+1} \in kNN_n(x)\right)$ as in Lemma 6.1. Then may rewrite this inequality by the way of two different errors.

1) The first error is $F_{Y B_n^{k_n}(x)}(\theta_n(x)) - F_{Y^x}(\theta_n(x))$. It is the error we make by using the response corresponding to an input close to x instead of x. This is the variance error. Using **A1**,

$$|F_{Y B_n^{k_n}(x)}(\theta_n(x)) - F_{Y^x}(\theta_n(x))| \leq M(x)||X - x||_{(k_n,n)}$$

and by **A3**,

$$|\theta_n(x) - \theta^*(x)| \leq \sqrt{C_1}$$

thus,

$$-2\gamma_{n+1}(\theta_n(x) - \theta^*(x))P_n \left[F_{Y B_n^{k_n}(x)}(\theta_n(x)) - F_{Y^x}(\theta_n(x))\right] \leq 2\gamma_{n+1}\sqrt{C_1}M(x)P_n||X - x||_{(k_n,n)}$$

2) The second term, $F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*)$ is the error we make by approximating θ^* by θ_n. This is a bias error. Thanks to Assumption **A4** we get

$$(\theta_n - \theta^*)\left[F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] \geq C_2(x,\alpha)\left[\theta_n(x) - \theta^*(x)\right]^2.$$

Taking now the expectation of our inequality we get (by using Remark 2.1)

$$\begin{aligned} a_{n+1}(x) \leq a_n(x) &- 2\gamma_{n+1}C_2(x,\alpha)\mathbb{E}\left[(\theta_n(x) - \theta^*(x))^2 P_n\right] + \gamma_{n+1}^2\mathbb{E}(P_n) \\ &+ 2\gamma_{n+1}M(x)\sqrt{C_1}\mathbb{E}(||X - x||_{(k_n,n)}P_n). \end{aligned}$$

This equation inequality a problem : thanks to Lemmas 6.1 and 6.5 (and so thanks to assumption **A2**) we can deal with the two last terms but we are not able to compute $\mathbb{E}\left[(\theta_n(x) - \theta^*(x))^2 P_n\right]$. To solve this problem, we use a truncature parameter ϵ_n : instead of writing a recursive inequality on $a_n(x)$ we write such inequality with $b_n(x)$, which is easier. Chosing $\epsilon_n = \frac{1}{n^\epsilon}$, we have to tune an other parameter but thanks to **A3**) and concentration inequalities (see lemma 6.3), it is easy to deduce a recursive inequality on $a_n(x)$ from the one on $b_n(x)$.

In fact, simulations (see Section 3) seem to show that in practice, the inequality is true relatively soon.

Comments on the parameters. We chose $0 < \beta < 1$ for the same reasons as in Theorem 2.1. About γ, the inequality (2.2) is true on the entier area $0 < \gamma < 1$ as soon as $\gamma \leq \beta$ (which is unusual, as you can see in [14] for example). We will nevertheless see in the sequel that this is not because the inequality is true that it implies a fast convergence to 0 of the mean square error.

Compromise between bias and variance. We can easily see the compromise we have to do on β to deal with the two previous errors. Indeed

- The bias error gives the term $\exp\left(-2C_2(x,\alpha)(x)\sum_{k=1}^{n}\frac{1}{k^{\epsilon+\gamma}}\right)$ of the inequality. This term decreases to 0 if and only if $\gamma + \epsilon < 1$ which implies $\beta > \gamma$. Then β must not be too small.

- The variance error gives the term $\left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1}$ in the remainder. For the remainder to decrease to 0, we then need that $\beta < 1$ and then we can not choose β too big.

From this theorem, we can get the rate of convergence of the mean square error. In that purpose, we have to study the order of the remainder β_n in n to exhibit dominating terms. It is sum of three terms. The exponential one is always negligeable as soon as n is big enough because $1 > \epsilon$. the two other are power of n. Comparing their exponent, we can exhibit the dominating term. Indeed, there exists a rank $N_1(x,d)$ such that, for $n \geq N_1(x,d)$,

If $\beta \leq 1 - d\gamma$, we get

$$\beta_n \leq C_5 n^{-2\gamma+\beta-1}$$

If $\beta > 1 - d\gamma$, we get

$$\beta_n \leq C_6(x,d) n^{-\gamma+(1+\frac{1}{d})(\beta-1)}.$$

We notice that N_0 and $N_1(x,d)$ are not the same kind of rank. In fact, $N_1(x,d)$ is reasonably small whatever the model parameters are, because it is only the rank after which exponential term and power of n term with big exponent are bigger than a power term with small exponent. Even if the constants in front of the terms can increase $N_1(x,d)$, it stays reasonably small. N_0 is not so nice, because, it increases exponentially when ϵ is close to $1-\beta$ (and we will see in Corollary 2.2 that optimal parameters is $\epsilon = 1 - \beta + \eta_1$ with η_1 small).

Corollary 2.1. *$a_n(x)$ decreases to 0 with the following rate*
$\forall n \geq \max(N_0, N_1(x,d), N_2, N_3(x,\alpha,d))$, when $\beta > 1 - d\gamma$ and
$1 - \beta < \epsilon < \min\left(1-\gamma, \left(1+\frac{1}{d}\right)(1-\beta)\right)$, there exists a constant $C_7(x,\alpha,d)$ such that

$$a_n(x) \leq \frac{C_7(d,x,\alpha,\epsilon,\gamma)}{n^{-\epsilon+\left(1+\frac{1}{d}\right)(1-\beta)}}$$

and when $\beta \leq 1 - d\gamma$ and $1 - \eta < \min(1-\beta+\gamma, 1-\gamma)$ there exists a constant $C_8(x,\alpha,\epsilon,\gamma)$ such that

$$a_n(x) \leq \frac{C_8(x,\alpha,\epsilon,\gamma)}{n^{\gamma-\beta+1-\epsilon}},$$

where we make appears dependence in ϵ and γ in the constants, just like the dependence on x, α and x. In the other cases, the inequality of Theorem 2.2, does not allow to show that $a_n(x)$ decreases to 0.

Sketch of proof : The proof consists in studying each term with comparison between sums and integrales and to exhibit dominating terms and their order in n.

Corollary 2.2. *Under the same hypothesis than in Theorem 2.2, when γ is fixed, the choice of β giving the best rate of convergence of the mean square error is $\beta = \gamma + \eta_\beta$ where $\eta_\beta > 0$ is as small as possible. In this case, we get for $n \geq \max(N_0, N_1(x,d), N_2, N_3(x,\alpha,d))$, when $\gamma \geq \frac{1}{1+d}$*

$$a_n(x) \leq \frac{C_7(x,\alpha,d,\epsilon,\gamma)}{n^{\frac{1}{d}(1-\gamma)-\eta}},$$

and when $\gamma < \frac{1}{1+d}$

$$a_n(x) \leq \frac{C_8(x,\alpha,\epsilon,\gamma)}{n^{\gamma-\eta}}$$

where in the two cases $\eta = \frac{\eta_\beta}{d} - \eta_\epsilon$ *and* $\eta_\epsilon = \epsilon - (1-\beta)$.

Comparison with others results. When they study the mean square error for the classical stochastic algorithm to estimate the quantile, Godichon et al. show in [14] that non-asymptotic rate of convergence is in $\mathcal{O}\left(n^{-\gamma}\right)$ for $\frac{1}{2} < \gamma < 1$. Our study shows a rate of convergence of $\mathcal{O}\left(n^{-\gamma+1+\eta}\right)$ for these γ. Our rate is lower but it is logical because we have a second level of approximation since we only have at hand a sample of bias laws. Moreover, we are able to give the rate of convergence for $0 < \gamma \leq \frac{1}{2}$ also.

Let us compare our results to classical result on k-nearest neighbors. Bhattacharya and al. in [4] show that, to estimate conditional quantile with the generalized inverse of empirical cumulative function, the best number of neighbors is for $\beta = \frac{4}{5}$ when inputs are in $\mathbb{R}$. With this parameter, they show the weak convergence of their estimator at speed $\mathcal{O}\left(n^{\frac{2}{5}}\right)$. Our result gives for optimal $\beta = \frac{1}{2} + \eta_\beta$ in dimension 1, a rate of convergence of the mean square error in $n^{\frac{1}{2}}$ which is then slower. Nevertheless, our result is non-asymptotic and our algorithm is easier to compute that their estimator which require to calculate a generalized inverse. Moreover, our inequality is true whatever the dimension d of the input space.

Corollary 2.3. *Under the same assumptions than in Theorem 2.2, the mean square error decreases faster when parameters are* $\gamma = \frac{1}{1+d}$ *and* $\beta = \gamma + \eta_\beta$ *where* $\eta_\beta > 0$ *is as small as possible. We indeed obtain with these parameters, for* $n \geq \max(N_0, N_1(x,d), N_2, N_3(x,\alpha,d))$ *there exists a constant* $C_9(x,\alpha,d)$ *such that*

$$a_n(x) \leq \frac{C_9(x,\alpha,d)}{n^{\frac{1}{1+d}-\eta}}$$

where η *is the the same than in corollary 2.1.*

Comments on the constant C_9 **:** As you can see in Appendix, the constant $C_9(x,\alpha,d)$ is the minimum between $C_7(x,\alpha,d,1,\frac{1}{1+d})$ and $C_8(x,\alpha,d,1,\frac{1}{1+d})$. The explicit form of these two constants show that this minimum is often $C_8(x,\alpha,d,1,\frac{1}{1+d})$. Indeed, it is true as soon as $0.5 \leq \sqrt{C_5}M(x)C_3(d)$ where $C_3(d)$ decrease when d increases but is always biger than 1. Then, the minimum is C_8 as soon as $M(x)\sqrt{C_1}$ is biger than 0.5. You can see an example of values of these constants in Subsection 3.1.

We also can notice that the constant $C_9(x,\alpha,d)$ depends on x only by the dependancy on x of $C_g(x)$ and $M(x)$. In practice, there are lots of case where these two constants does not depends on x (see for example Subsection 3.1). In these cases (or when we

can easily find a bound of $C_g(x)$ and $M(x)$ which does not depend on x), our result is uniform in x and we can consider the integrated mean square error and conclude that

$$\int_X a_n(x) f_X(x) dx \leq \frac{C_9(\alpha, d)}{n^{\frac{1}{1+d}-\eta}}.$$

When α increases to 1, we try to estimate extremal quantile. $C_2(x, \alpha)$ becomes smaller and then $C_9(x, \alpha, d)$ increases. Finally the bound gets worst. We can easily understand this phenomonon because when α is big, we have a small probability to sample on the right of the quantile, and the algorithm is less powerfull.

Let us now comment the dependancy in the dimension d. As we saw before, it is more usual that $C_9 = C_8$. In this case, C_9 derease when the dimension d increases in $2^{\frac{1}{1+d}}$. Nevertheless, this dereasing is smaller to balance the lost caused in the rate of convergence which is in $n^{\frac{-1}{1+d}}$. This is by the way why we chose best parameter γ by optimizing the rate of convergence and not the constant in front of it.

Sketch of proof : It is easy optimization.

Comment on the rank N_0. As we saw before the rank $(N_i)_{i\neq 0}$ depends on constants of the problem but is reasonably small. This is not the case of the rank N_0 which depends on the gap between ϵ and $1-\beta$. The problem comes from the fact that optimal ϵ to obtain rate of convergence of the two previous corollaries is $\epsilon = 1 - \beta + \eta_\epsilon$ with η_1 as small as possible. But, $\eta_1 = \epsilon - (1-\beta)$ appears on the rank N_0 but also on the rate of convergence : after the rank $N_0 = \exp\left(2\eta_\epsilon^{-1}\right)$ the rate of convergence is on $\mathcal{O}\left(n^{\frac{-1}{1+d}+\eta}\right)$. Then the more η si small, the more the rate of convergence is fast but the more the rate is true for big n. Our results are non-asymptotic but nevertheless true when n is large.

Imagine, you have a budget of 10000 calls to the code. Then if you want your inequality to be theoretically true for $N = 10000$, we have to take $\eta_\epsilon = 2(\ln(10000))^{-1} \approx 0.217$. In this case, we can theoretically obtain a risk of $N^{\frac{-1}{1+d}+\frac{2}{\ln(N)}}$ (where we forgot the term $d\eta_\beta^{-1}$ which is very small compared to the two others terms). It means that in dimension 1, the mean square error decreases theoretically to 6%, which is acceptable. But in dimension $d > 1$ is not very good : we obtain 30% in dimension 2 and 63% in dimension 3.

Nonetheless, simulations (see next part) seems to show that this difficulty is only an artifice of our proof (we needed to introduce ϵ_n because we do not know how to compute $\mathbb{E}((\theta_n - \theta^*)P_n)$, but it does not really exists when we compute the algorithm). Our simulations show that the optimal rate of convergence when we choose optimal parameters is fast reached (see Section 3).

3. Numerical simulations

In this part we present some numerical simulations to illustrate our theorems. To begin with, we deal with dimension 1. We study two stochastic codes.

3.1. **Dimension 1- square function.** The first example, is the very regular code caracterized by the function

$$g(X, \epsilon) = X^2 + \epsilon$$

where $X \sim \mathcal{U}([0,1])$ and $\epsilon \sim \mathcal{U}([-0.5, 0.5])$. We try to estimate the quantile for $x = 0.5$ and initialize our algorithm to $\theta_1 = 0.3$. Let us show that our assumptions are fullfilled in this case. We have $\mathcal{L}(g(x,\epsilon)) = \mathcal{U}([-\frac{1}{2} + x^2; \frac{1}{2} + x^2])$. Then e

$$f_{(X,Y)}(u,v) = \mathbf{1}_{[-\frac{1}{2}+u^2, \frac{1}{2}+u^2]}(v).$$

Moreover, the code function g is at values in the compact set $[L_g, U_g] = [-\frac{1}{2}; \frac{3}{2}]$. Let us study assumption **A1**. Let A be an interval containing x, denoted $B = [x-a, x+b]$ ($a > 0$, $b > 0$), then

$$\begin{aligned} |F_{Y^B}(t) - F_{Y^x}(t)| &\leq \left| \frac{\int_{-\infty}^{t} \int_B f_{(X,Y)}(z,y) dy dz}{\int_B f_X(z) dz} - \int_{-\infty}^{t} f_{(X,Y)}(x,y) dy \right| \\ &\leq \frac{\int_{-\frac{1}{2}}^{t} \int_{x-a}^{x+b} \left| \mathbf{1}_{[-\frac{1}{2}+z^2; \frac{1}{2}+z^2]} - \mathbf{1}_{[-\frac{1}{2}+z^2; \frac{1}{2}+z^2]} \right| (y) dz dy}{\mu(B)} \end{aligned}$$

Now, we have to distinguish the cases in function of the localization of t. There are lots of cases, but computations are nearly the same. That is why we will develop only one case here.

If $t \in [-\frac{1}{2}; x^2 - \frac{1}{2}]$, we have :

$$\begin{aligned} |F_{Y^B}(t) - F_{Y^x}(t)| &\leq \frac{\int_{x-a}^{x+b} \int_{-\frac{1}{2}}^{t} \left| \mathbf{1}_{[-\frac{1}{2}+z^2; \frac{1}{2}+z^2]} - \mathbf{1}_{[-\frac{1}{2}+z^2; \frac{1}{2}+z^2]} \right| (y)}{a+b} \\ &= \frac{\leq \int_{x-a}^{x+b} \left(\mathbf{1}_{z \geq x}(0) + \mathbf{1}_{z \leq x}(t - z^2 + \frac{1}{2}) \mathbf{1}_{z \geq \sqrt{t+\frac{1}{2}}} \right) dz}{a+b} \\ &= \frac{\leq \int_{x-a}^{x} (t + \frac{1}{2} - z^2) dz}{b+a} \end{aligned}$$

From now, there are again two cases.

Since $t \in [-\frac{1}{2}; x^2 - \frac{1}{2}]$, we always have $(t + \frac{1}{2})^{\frac{1}{2}} \leq x$. But the position of $\sqrt{t + \frac{1}{2}}$ in relation to $(x - a)$ is not always the same. Then, if $t \in [-\frac{1}{2}; -\frac{1}{2}(x-a)^2]$, we get

$$
\begin{aligned}
|F_{Y^B}(t) - F_{Y^x}(t)| &\leq \frac{\int_{x-a}^{x+b}(t - z^2 + \frac{1}{2})dz}{b+a} \\
&\leq (t + \frac{1}{2})a - \frac{x^3}{3} + \frac{(x-a)^3}{3} \\
&\leq (x-a)^2 a - x^2 a + a^2 x - \frac{a^3}{3} \\
&\leq -a^2 x + \frac{2a^3}{3} \\
&\leq 0 + r_B \times 1^2 \times \frac{2}{3}
\end{aligned} \tag{6}
$$

where we use that $0 < a < 1$.

Finally, in this case **A1** is true with $M(x) = \frac{2}{3}$. We can compute exactly in the same way for the other cases and we always find an $M(x) \leq \frac{2}{3}$. The assumption **A2** is also satisfies, taking $C_{input} = 1$. We have already explained that assumption **A3** is true for $[L_g, U_g] = [-\frac{1}{2}, \frac{3}{2}]$. Finally assumption **A4** is also satisfies with $C_g(x) = 1$ and $C_2(x, \alpha) = \frac{1-\alpha}{3}$.

3.1.1. *a.s convergence.* Let us first deal with the almost sure convergence.

To check the convergence when $0 < \gamma < \beta < 1$, we plot in Figure 1 the relative error of the algorithm in function of γ and β when $n = 5000$. Best parameters are clearly $\beta > \gamma = \frac{1}{2}$. We can even observe that for $\beta \approx 1$ or $\beta \leq \gamma$, the algorithm does not converge almost surely (or very slowly). This is in accordance with our theoretical results. Since we also have plotted the relative error fo $\gamma < \frac{1}{2}$, we can check that the behaviour of our algorithm in this area is not good. Nevertheless, we can observe a kind of ocntinuity : in practice, the convergence becomes really slow only when γ is significally far from $\frac{1}{2}$.

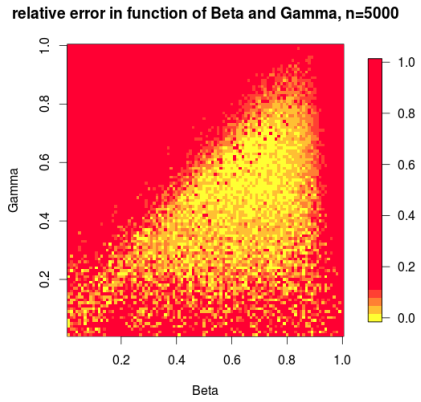

FIGURE 1. Relative error in fonction of β and γ

To complete this observations, we plot in Figure 2 evolution of iterations of the re-centered algorithm $(\theta_n - \theta^*)$ for different parameters. Conclusions are the same.

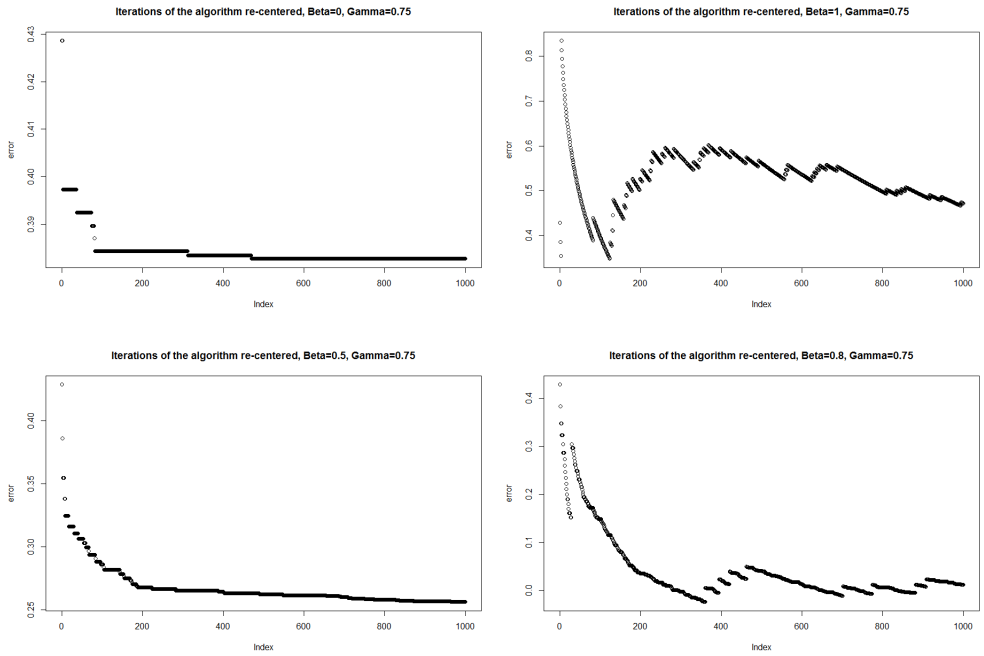

FIGURE 2. Convergence a.s of the algorithm

3.1.2. *Mean Square Error (MSE).* Let us study the best choice of β when γ is fixed (illustrations of corollary 2.2). For this simulations (Figure 3), we estimate the MSE by Monte Carlo with 100 realisations, when γ is fixed, for β between 0 and 1 and $n = 200$. We plot the MSE in function of β to check that it is smaller when β is just superior to γ. Simulations are good illustrations of the corollary except when γ is too close to 1.

Let us now illustrate the choice of γ when β is optimal (illustrations of corollary 3.1.2). In this part (Figure 4) , we study the influence of γ when β is "optimal", that is for β just superior to γ. First, we plot the MSE estimated by a Monte Carlo method with 100 iterations in function of γ. We can see that the best choice of γ is then $\frac{1}{2}$.

Then in Figure 5, we plot in logarithmical scale the convergence of the MSE (still with Monte Carlo of 100 realisations) for different values of γ. It appears that the more close to $\frac{1}{2}$ we are, the faster is the decreasing.

Finally, let us sum up all and find the optimal parameters. We plot in Figures 6, the mean square error in function of γ and β (still estimate by Monte Carlo of 100 iterations).

We can see that best parameters are $\gamma = \frac{1}{2}$ and β superior to γ.

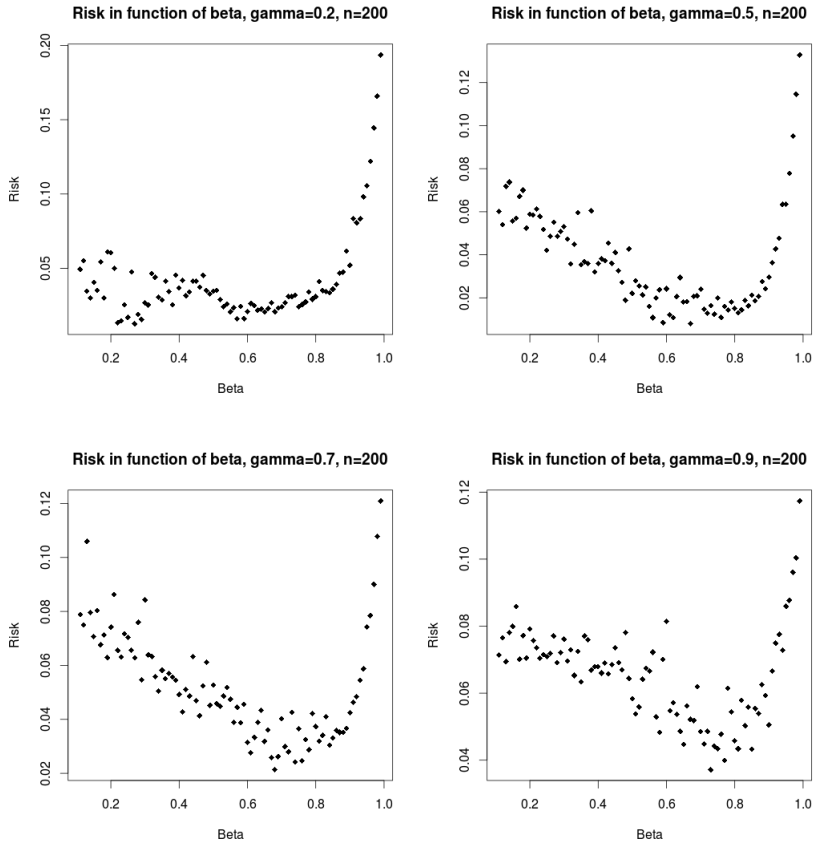

FIGURE 3. Choice of β when γ is fixed

3.1.3. *Theorical bound.* In this case, we have at hand all the parameters, to compute the theoretical bound, obtained in our theorems. In particular, in corollary 2.3, we get :

$$a_n(x) \leq \frac{C_9(x,d,\alpha)}{n^{\frac{1}{1+d}-\eta}}.$$

Since $\alpha = 0.95$, $C_1 = 6$, $d = 1$, $C_{input} = 1$, $C_2(x,\alpha) = 0.017$, $M(x) = \frac{2}{3}$, we get $C_3(d) = 7.39$ because $C_4(d) = 2$. Then $C_5 = 2$ and $C_6(x,d) = 48$. So $C_9(x,\alpha,d) = 330$ and we obtain a bound of 23 for $n = 200$ which is very far away from the practical results we got. Our bounds are clearly not optimal, but they allow us to find optimal parameters. We can think to a way to improve this bound. First of all, the constant $C_2(x,\alpha)$ is in fact not so small. Indeed, we have to take a margin in the proof, for the case where θ_n goes out of $[L_g, U_g]$. This clearly can happen with a very small probability. If we do not take account of this case, we have $C_2(x,\alpha) = 1$. Then $C_9(x,\alpha,d) \approx 5.65$ and then, for $n = 200$, the bound is 0.39. Practical results are still better (we can observe that for $n = 50$ only, we have a MSE inferior to 0.5% !), but the gap is less important. Theorically, we need a budget of 12770 calls to the code to get the precision of 5%.

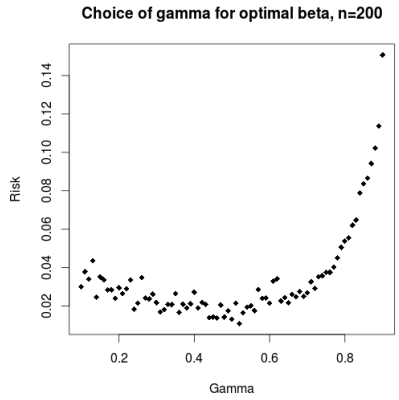

FIGURE 4. Choice of γ when β is optimal

3.2. **Dimension 1 - absolute value function.** Let us see what happens when the function g is less smooth with respect to the first variable. We study the code

$$g(X, \epsilon) = |X| + \epsilon$$

where $X \sim \mathcal{U}([-1, 1])$ and $\epsilon \sim \mathcal{U}([-0.5, 0.5])$. We want to study the conditional quantile in $x = 0$ (the point in which the continuity fails).

We do not try to check our assumptions, because computations are nearly the same than in previous case, but they are true. Since the a.s convergence is true and gives really same kind of plots than previous case, we only study the convergence of the MSE. To deal with the MSE, we also check that best parameters are the theoretical one in practice. In that purpose, we plot in Figure 7 the MSE (estimated by 100 iterations of Monte Carlo simulations) in function of γ and β, for n=300 (the discontinuity constrains us to make more iterations to have a sufficient precision) and $\theta_1 = 0.3$. Conclusions are the same than in previous example concerning the best parameters. Nevertheless, we can observe that the lack of continuity implies some strange behaviour around $\gamma = 1$.

3.3. **Dimensions 2 and 3.** In dimension $d > 1$, our theorems give that theoretical optimal parameters are $\gamma = \frac{1}{1+d}$ and $\beta = \gamma + \eta$. To see what happens in practice, we still plot Monte Carlo estimations (200 iterations) of the MSE in function of γ and β.

3.4. **Dimension 2.** In dimension 2, we study two codes :

$$g_1(X, \epsilon) = ||X||^2 + \epsilon \text{ and } g_2(X, \epsilon) = X_1^2 + X_2 + \epsilon,$$

where $X = (X_1, X_2) \sim \mathcal{U}\left([-1, 1]^2\right)$ and $\epsilon \sim \mathcal{U}([-0.5, 0.5])$. In each case, we chose $n = 400$ and want to study the quantile in the input point $x = (0, 0)$ and initialize our algorithm in $\theta_1 = 0.3$. In Figure 8, we can see that $\beta = 1$ and $\gamma = 1$ are still really bad parameters. As in theoretical point of view, $\gamma = \frac{1}{1+d} = \frac{1}{3}$ seems to be the best choice. Nevertheless, even if it is clear that $\beta < \gamma$ is a bad choice, the experiments seems to show that best parameter β is strictly superior to γ, more superior than in theoretical

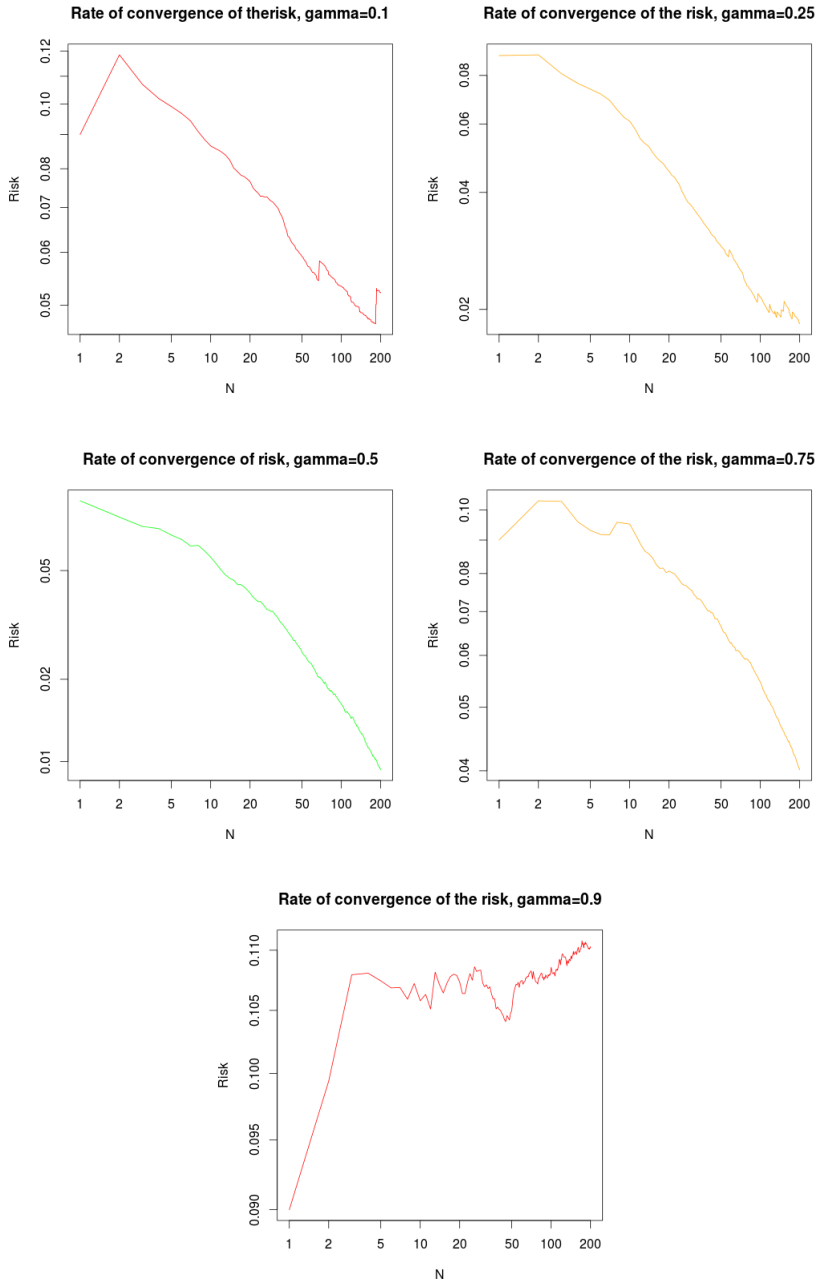

FIGURE 5. Convergence of the mean square error in logarithm scale

case, where we take β as close as possible of γ. As we said before, in practice, N_0 seems not to be the true limit rank. Indeed, with only $, = 400$ iterations, in this case, the MSE, in the opitmal parameters case reach 6% !

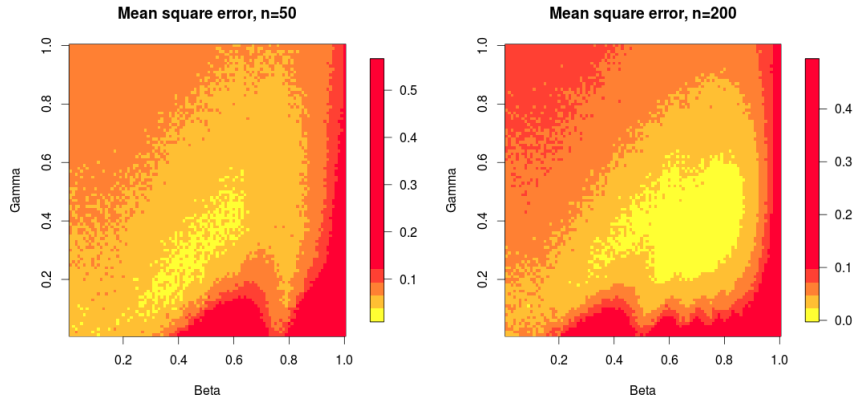

FIGURE 6. Mean square error in function of β and γ for the square function

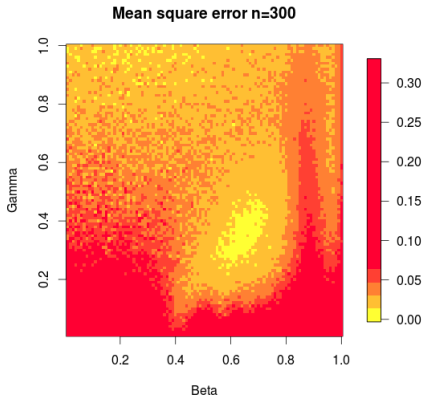

FIGURE 7. MSE in function of β and γ for absolut value function

4. DIMENSION 3

In dimension 3, we study the two codes :

$$g_1(X,\epsilon) = ||X||^2 + \epsilon \text{ and } g_2(X,\epsilon) = X_1^2 + X_2 + \frac{X_3^3}{2} + \epsilon,$$

where $X = (X_1, X_2, X_3) \sim \mathcal{U}\left([-1,1]^3\right)$ and $\epsilon \sim \mathcal{U}\left([-0.5, 0.5]\right)$. In each case, we choose $n = 500$ and want to study the quantile in the input point $(0,0,0)$. The interpretation of Figure 9 are the same than in dimension 2. The scale is still not the same, the decrease is again more slow but with $n = 500$ we nevertheless obtain a MSE of 10%.

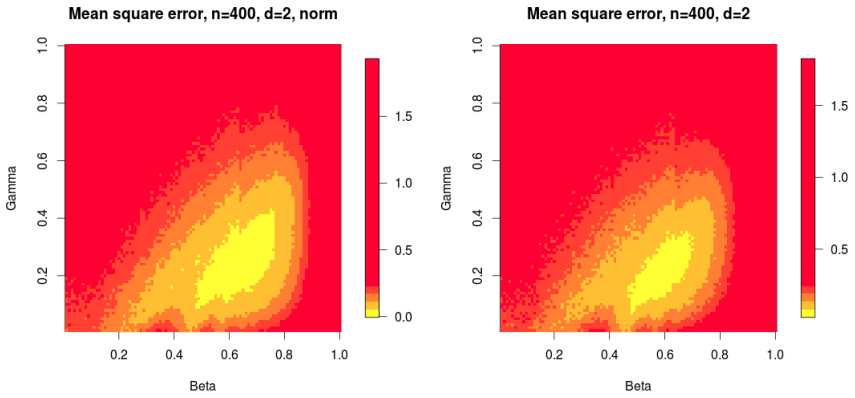

FIGURE 8. Mean square error in function of β and γ

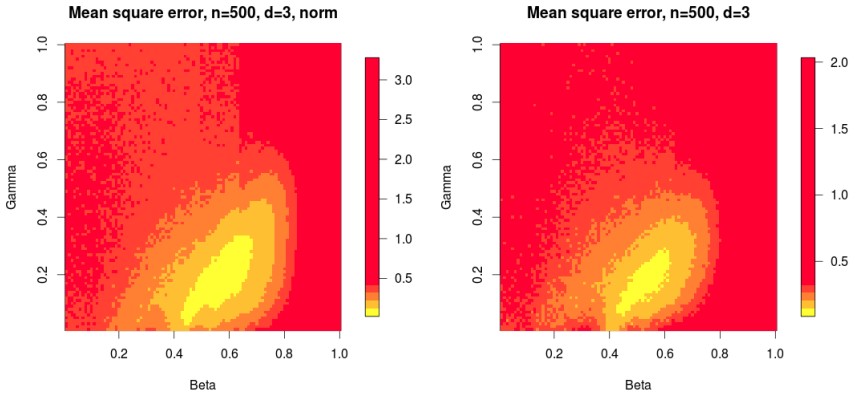

FIGURE 9. Mean square error in function of β and γ

5. Conclusion and perspectives

In this paper we aimed at estimating a conditional quantile of the output of a stochastic code where inputs ly in $\mathbb{R}^d$. In this purpose we introduced a new stochastic algorithm using k-nearest neighbors theory. Dealing with the two erros made by this approximation, we show that our algorithm is convergent for $\frac{1}{2} < \gamma < \beta < 1$ and study its non-asymptotic rate of convergence of the mean square error. Moreover, we show that to get the best rate of convergence, we have to chose $\beta = \gamma + \eta_\beta$ and $\gamma = \frac{1}{2}$. Numerical simulations show that our algorithm with theoretical optimal parameters is really powerful to estimate a conditional quantile, even in dimension $d > 1$.

The theoretical guarantees are shown under strong technical assumptions, but our algorithm is a general methodology to solve the problem. A futur work can consist in trying to relax this technical assumptions. Moreover, the proof we propose constrained

us to use an artifact parameter ϵ which implies that the non-asymptotic inequality is theoretically true for big n, even if simulations confirm that this problem do not exists in practice. A second perspective is then to find a better way to prove this inequality for smaller n. Finally, it could be interesting to find a way to chose the new input at each step. Maybe we could build a criteria which allows us to chose the best new input to provide to the code, to reduce the error made by our algorithm.

6. Annexes : technical lemmas and proofs

6.1. Technical lemmas and notations.

Lemma 6.1. *Denoting $P_n = \mathbb{P}(X \in kNN_n(x)|X_1, \dots X_n)$, we have the following properties*

1) $P_n = F_{||X-x||}\left(||X-x||_{(k_n,n)}\right)$
2) $P_n \sim \beta(k_n, n-k_n+1)$
3) $\mathbb{E}(P_n) = \frac{k_n}{n+1}$.

where you denote $F_{||X-x||}$ the cumulative distribution function of the law $||X-x||$, $||X-x||_{(k_n,n)}$ the k_n order statistic of the sample $(||X_1-x||, \dots ||X_n-x||)$ and $\beta(k_n, n-k_n+1)$ the beta distribution of parameters k_n and $n-k_n+1$.

Proof. Conditionnally to $X_1, \dots X_n$, "X is in the set $kNN_n(x)$" is equivalent to "X satisfies $||X-x|| \leq ||X-x||_{(k_n,n)}$". Then

$$\begin{aligned} P_n &= \mathbb{P}(X \in kNN_n(x)|X_1 \dots X_n) \\ &= \mathbb{P}_X\left(||X-x|| \leq ||X-x||_{(k_n,n)}|X_1 \dots X_n\right) \\ &= F_{||X-x||}\left(||X-x||_{(k_n,n)}\right) \end{aligned}$$

Since X is at density, the cumulative distribution function $F_{||X-x||}$ is continuous. Indeed, using the sequential caracterization we get for a sequence (t_n) converging to t

$$\begin{aligned} F_{||X-x||}(t_n) &= \mathbb{P}(X \in B_d(x, t_n)) \\ &= \int_{\mathbb{R}^d} f(z)\mathbf{1}_{B_d(x,t_n)}(z). \end{aligned}$$

Since f is integrable, the Lebesgue theorem allows us to conclude that

$$\lim_n \int_{\mathbb{R}^d} f(z)\mathbf{1}_{B_d(x,t_n)}(z) = \int_{\mathbb{R}^d} \lim_n f(z)\mathbf{1}_{B_d(x,t_n)}(z) = \mathbb{P}(X \in B_d(x,t)),$$

so the cumulative distribution function is continuous.

Then thanks to classical result on statistics order and quantile transform (see [7]), we get

$$\begin{aligned} P_n &= F_{||X-x||}\left(||X-x||_{(k_n,n)}\right) \\ &\sim U_{(k_n,n)} \\ &\sim \beta(k_n, n-k_n+1) \end{aligned}$$

where we denoted $U_{(k_n,n)}$ the k_n statistic order of a independant sample of size n distributed like a uniform law on $[0,1]$.

□

Lemma 6.2. *Denoting $\mathcal{B}(n,p)$ the binomial distribution of parameters n and p, we have*

$$\mathbb{P}\left(\frac{\mathcal{B}(n,p)}{n} < \frac{p}{2}\right) \leq \exp\left(-\frac{3np}{32}\right)$$
$$\mathbb{P}\left(\frac{\mathcal{B}(n,p)}{n} > 2p\right) \leq \exp\left(-\frac{3np}{8}\right)$$

Proof. Let us prove the first inequality. By noticing that

$$Z \underset{\mathcal{L}}{=} \frac{1}{n}\sum_{k=1}^{n} Z_i$$

where $(Z_n)_n$ is an independant sample of $\mathcal{B}(p)$ (Bernoulli law of paramater p), we apply the Bernstein's inequality (see Theorem 8.2 of [9]) to conclude that

$$\mathbb{P}(Z - p < -\epsilon p) \leq \exp\left(-\frac{3np\epsilon^2}{8}\right)$$
$$\mathbb{P}(Z - p > \epsilon p) \leq \exp\left(-\frac{3np\epsilon^2}{8}\right)$$

The results follow by taking $\epsilon = \frac{1}{2}$ in the first case and $\epsilon = 1$ in the second case. □

Lemma 6.3. *Denoting A_n the event $\{X_1, \dots X_n \mid P_n > \epsilon_n\}$ where $\epsilon_n = \frac{1}{n^\epsilon}$ and $1 > \epsilon > 1 - \beta$, we have for $n \geq 1$,*

$$\mathbb{P}(A_n^C) \leq \exp\left(-\frac{3n^{1-\epsilon}}{8}\right)$$

Proof. Thanks to the Lemma 6.1, we obtain

$$\begin{aligned}\mathbb{P}(A_n^C) &= \mathbb{P}(\beta(k_n, n - k_n + 1) \geq \epsilon_n) \\ &= I_{\epsilon_n}(k_n, n - k_n + 1)\end{aligned}$$

where we denote I_ϵ the incomplet β function. A classical result (see [1]) allow us to exprim this quantity in function avec the binomial distribution. Then

$$\mathbb{P}(A_n^C) = \mathbb{P}(\mathcal{B}(n, \epsilon_n) \geq k_n)$$

Thanks to Lemma 6.2, we know that

$$\mathbb{P}(\mathcal{B}(n, \epsilon_n) \geq k_n) \leq \exp\left(-\frac{3n\epsilon_n}{8}\right)$$

as soon as

$$\frac{k_n}{n} \geq 2\epsilon_n$$

which is true as soon as $n \geq 2^{\frac{1}{\epsilon-(1-\beta)}}$ because $\epsilon > 1 - \beta$. We now use the Cramer's method to study the deviation of the Binomial distribution (see [8])

$$\mathbb{P}\left(\frac{\mathcal{B}(n,\epsilon_n)}{n} \geq \frac{k_n}{n}\right) \leq \exp\left(-n\left(\frac{k_n}{n}\log\left(\frac{k_n}{n\epsilon_n}\right) + \left(1-\frac{k_n}{n}\right)\log\left(\frac{1-\frac{k_n}{n}}{1-\epsilon_n}\right)\right)\right)$$
$$= \left(\frac{n\epsilon_n}{k_n}\right)^{k_n}\left(\frac{1-\epsilon}{1-\frac{k_n}{n}}\right)^{n-k_n}$$

Then, since $\epsilon_n = n^{-\epsilon}$ and $k_n \sim n^{\beta}$, computations give us

$$\log\left(\mathbb{P}(A_n^C)\right) = n^{\beta}\left(1-\epsilon+\beta\right)\log(n) + \left(n - n^{\beta}\log\left(1-\frac{n^{\beta-1}-n^{-\epsilon}}{1-n^{\beta-1}}\right)\right.$$
$$\leq n^{\beta}\left(1-\epsilon+\beta\right)\log(n) + n^{\beta} - n^{-\epsilon+1}$$
$$= \mathcal{O}\left(-n^{\beta}\log(n)\right)$$

and the result follows.

□

Definition 6.1. *Let $B_n^{k_n}(x)$ be the ball centered in x such that*

$$\mathbb{P}(X \in kNN_n(x)|X_1 \dots X_n) = \mathbb{P}(X \in B_n^{k_n}(x)),$$

in fact

$$B_n^{k_n}(x) = B_{||.||_d}(x, ||X-x||_{(k_n,n)}).$$

Lemma 6.4. *Under hypothesis of theorem 2.1, $||X-x||_{(k,n)}$ converges to 0 a.s.*

Proof. Let u be a strictly non-negative number.

$$\begin{aligned} p_u :&= \mathbb{P}(X \in \mathcal{B}(x,u)) = \int_{\mathcal{B}(x,u)} f(t)dt \\ &\geq \mu_X\left(\mathcal{B}(x,u)\right) = C_1\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \\ &= C_{input}C_4(d)u^d := q_u \end{aligned} \tag{7}$$

Since $\{||X-x||_{(k_n,n)} > u\} \subset \{$there are at the most k_n elements of the sample which satisfy $X \in \mathcal{B}(x,u)\}$, we get, by denoting $Z \sim \mathcal{B}(n,p_u)$,

$$\mathbb{P}(||X-x||_{(k_n,n)} > u) = \mathbb{P}(Z < k_n)$$

Thanks to equation (7), we get, by denoting $\tilde{Z} \sim \mathcal{B}(n,q_u)$,

$$\mathbb{P}(||X-x||_{(k_n,n)} > u) \leq \mathbb{P}(\tilde{Z} < k_n)$$

Thanks to Lemma 6.2, we then know that $\mathbb{P}(||X-x||_{(k_n,n)} > u)$ is the general term of a convergent sum. Indeed, for n large enough, $\frac{k_n}{n} < \frac{q_u}{2}$ because $\frac{k_n}{n}$ converges to 0 ($\beta < 1$). The Borel-Cantelli Lemma (see for example Proposition 5.1.2 of [3]) then implies that $||X-x||_{(k_n,n)}$ converges to 0 a.s. □

Lemma 6.5. *With de forcoming notations,*

$$\mathbb{E}(||X-x||_{(k_n,n)}P_n) \leq C_3(d)\left(\frac{k_n}{n+1}\right)^{1+\frac{1}{d}}$$

Proof. Let us denote $\tilde{F}$ and $\tilde{f}$ the cumulative and density distribution function of the law of $||X-x||$.

$$\begin{aligned}\mathbb{E}(||X-x||_{(k_n,n)}P_n) &= \mathbb{E}\left(||X-x||_{(k_n,n)}\tilde{F}\left(||X-x||_{(k_n,n)}\right)\right)\\ &= \int y\tilde{F}(y)f_{||X-x||_{(k_n,n)}}(y)dy\end{aligned}$$

with

$$f_{|X-x|_{(k_n,n)}}(y) = \frac{n!}{(k_n-1)!(n-k_n)!}\tilde{F}(y)^{k_n-1}\left(1-\tilde{F}(y)\right)^{n-k_n}\tilde{f}(y)$$

Then we get

$$\begin{aligned}\mathbb{E}(||X-x||_{(k_n,n)}P_n) &= \int y\tilde{F}(y)^{k_n}(1-\tilde{F}(y))^{n-k_n}\tilde{f}(y)\frac{n!}{(k_n-1)!(n-k_n)!}\\ &= \frac{k_n}{n+1}\mathbb{E}\left(||X-x||_{(k_n+1,n+1)}\right)\end{aligned}$$

Let us now use a classical inequality between expectancy and probability (see for example Proposition 3.4.8 of [3]). Let us denote $U_{|.|}$ the upper bound of the support of $||X-x||$,

$$\mathbb{E}(||X-x||_{(k_{n+1},n+1)}) \leq \int_0^{U_{|.|}} \mathbb{P}(||X-x||_{(k_{n+1},n+1)} > u)du.$$

Using same arguments that in Lemma 2.1, we get denoting $C_{11}(d) = \sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}C_4(d)}}$

$$\begin{aligned}I := \int_0^{U_{|.|}} \mathbb{P}(||X-x||_{(k_n+1,n+1)} > u)du &= \int_0^{C_{10}(d)} \mathbb{P}(\mathcal{B}(n+1,q_u) < k_n+1)du\\ &\quad + \int_{C_{10}(d)}^{U_{|.|}} \mathbb{P}(\mathcal{B}(n+1,q_u) < k_n+1)du\\ &\leq \int_0^{C_{10}(d)} 1du + \int_{C_{10}(d)}^{U_{|.|}} \exp\left(-\frac{3(n+1)C_{input}C_4(d)u^d}{32}\right)du\end{aligned}$$

where we use Lemma 6.2 in the second integrale because $u > C_{10}(d)$ implies $\frac{k_n+1}{n+1} < \frac{q_u}{2}$. Then, we get

$$
\begin{aligned}
I = &\leq C_{10}(d) + \int_{C_{11}(d)}^{+\infty} \exp\left(-\frac{3(n+1)C_{input}C_4(d)u^d}{32}\right) du \\
&\leq C_{10}(d) + \int_0^{+\infty} \frac{u^{d-1}}{C_{10}(d)^{d-1}} \exp\left(-\frac{3(n+1)C_{input}C_4(d)u^d}{32}\right) du \\
&= C_{10}(d) + \frac{C_{10}(d)}{C_{10}(d)^d} \frac{32}{3(n+1)dC_{input}C_4(d)} \int_0^{+\infty} \frac{3(n+1)dC_{input}C_4(d)u^{d-1}}{32} \exp\left(-\frac{3(n+1)C_{input}C_4(d)u^d}{32}\right) du \\
&= C_{10}(d) + \frac{C_{11}(d)}{C_{10}(d)^d} \frac{32}{3(n+1)dC_{input}C_4(d)} \left[-\exp\left(-\frac{3(n+1)C_{input}C_4(d)u^d}{32}\right)\right]_0^{+\infty} \\
&= C_{10}(d)\left(1 + \frac{3(n+1)dC_{input}C_4(d)}{32C_{10}(d)^d}\right) \\
&= \sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}C_4(d)}}\left(1 + \frac{16}{3d(k_n+1)}\right) \\
&= \sqrt[d]{\frac{k_n}{n+1}}\left[\sqrt[d]{\frac{2}{C_{input}C_4(d)}}\sqrt[d]{\frac{k_n+1}{k_n}}\left(1 + \frac{16}{3d(k_n+1)}\right)\right] \\
&= \sqrt[d]{\frac{k_n}{n+1}}\sqrt[d]{\frac{4}{C_{input}C_4(d)}}\left(1 + \frac{8}{3d}\right) \\
&:= C_3(d)\sqrt[d]{\frac{k_n}{n+1}}
\end{aligned}
$$

where we use in the last inqequality that for $n \geq 1$, $k_n \geq 1$.

□

Lemma 6.6. *Let (b_n) be a deterministic sequel such that there exists a constant C and a sequence $(\alpha_n)_n$ such that*

$$\forall n,\ b_{n+1} \leq b_n(1 - 2C\alpha_n) + \beta_n$$

then

$$\forall n,\ b_n \leq \exp\left(-2C\sum_{k=1}^{n}\alpha^k\right)b_0 + \sum_{k=1}^{n}\exp\left(-2C\sum_{j=k}^{n}\alpha_j\right)\beta_k.$$

Proof. Proof by induction. □

6.2. **Proof of Theorem 2.1 : a.s convergence of the algorithm.** To proove this theorem, we adapt the proof of Robbins-Monro in the classical case (see [5]). In the sequel we don't write $\theta_n(x)$ but θ_n to make the notation less cluttered.

6.2.1. *Introduction of a martingale.* Let us recall that we denote $H(\theta_n, X_{n+1}, Y_{n+1}) := \left(\mathbf{1}_{Y_{n+1}\leq\theta_n-\alpha}\right)\mathbf{1}_{X_{n+1}\in kNN_n(x)}$ and $\mathcal{F}_n = \sigma(X_1, \dots X_n, Y_1, \dots Y_n)$ and $\mathbb{P}_n$ and $\mathbb{E}_n$ the probability and expectancy conditionnaly to $\mathcal{F}_n$. Let us denote

$$\begin{aligned} h_n(\theta_n) &:= \mathbb{E}(H(\theta_n, X_{n+1}, Y_{n+1})|\mathcal{F}_n) \\ &= \mathbb{P}_n(Y_{n+1} \leq \theta_n \cap X_{n+1} \in kNN_n(x)) - \alpha \mathbb{P}_n(X_{n+1} \in kNN_n(x)) \\ &= P_n\left[(F_{Y^{kNN_n(x)}}(\theta_n) - F_{Y^x}(\theta^*)\right] \end{aligned}$$

where we use the notations $P_n := \mathbb{P}(X \in kNN_n(x)|X_1, \dots X_n)$ as in Lemma 6.1. We then have exhibited a martingale T_n

$$\begin{aligned} T_n &= \theta_n + \sum_{j=1}^{n} \gamma_j h_{j-1}(\theta_{j-1}) \\ &= \theta_0(x) - \sum_{j=1}^{n} \gamma_j \xi_j \end{aligned}$$

with $\xi_j = H(\theta_{j-1}, X_j, Y_j) - h_{j-1}(\theta_{j-1})$. This martingale is bounded in $\mathbb{L}^2$. Indeed, as

$$\sup_n |\xi_n| \leq \alpha + (1+\alpha) = 1 + 2\alpha$$

the Burkholder inequality gives the existence of a constant C such that

$$\begin{aligned} \mathbb{E}(|T_n|^2) &\leq \mathbb{E}\left(\left(\sum_{j=1}^{n} \gamma_j \xi_j\right)^2\right) \\ &\leq C\mathbb{E}\left(|\sum_{j=1}^{n} (\gamma_j \xi_j)^2|^2\right) \\ &\leq C(1+2\alpha)\sum_{j=1}^{n} \gamma_j^2 \end{aligned}$$

which allows us to conclude because $\sum_{n\geq 0} \gamma_n^2 < +\infty$.

6.2.2. *The sequel* (θ_n) *converges a.s.* First, let us show that

$$\mathbb{P}(\theta_n = +\infty) + \mathbb{P}(\theta_n = -\infty) = 0. \tag{8}$$

Let us suppose that this probability is positive (we name Ω_1 the non-negligeable set where $\theta_n(\omega)$ diverges to $+\infty$ and the same arguments would show the result when the limit is $-\infty$). Let ω be in Ω_1. We have $\theta_n(\omega) \leq \theta^*$ for only a finit number of n. Let us then show that for n large enough, $h_n(\theta_n(\omega)) > 0$. First, we know that P_n follows a Beta distribution. This is why

$$\mathbb{P}(P_n = 0) = 0 \ \forall n$$

and then the Borel-Cantelli Lemma gives that

$$\mathbb{P}(\exists N \ \forall n \geq N \ P_n > 0) = 1.$$

As we suppose Ω_1 has a strictly non-negativ measure, we know that there exists Ω_2 of strictly non-negative measure such that $\forall \omega \in \Omega_2$, $\theta_n(\omega) \to +\infty$ and for all n large enough, $P_n(\omega) > 0$. Since

$$h_n(\theta_n(\omega)) = P_n \left(F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - \alpha \right),$$

we have now to show that

$$F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - \alpha > 0.$$

As $\theta_n(\omega)$ diverges to $+\infty$, we can find D such that for n large enough, $\theta_n(\omega) > D > \theta^*$. Then

$$\begin{aligned} F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - \alpha &= F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - F_{Y^x}(\theta^*) \\ &= F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - F_{Y^{B_n^{k_n}(x)}}(D) + F_{Y^{B_n^{k_n}(x)}}(D) - F_{Y^x}(D) \\ &+ F_{Y^x}(D) - F_{Y^x}(\theta^*) \end{aligned}$$

First, $F_{Y^{B_n^{k_n}(x)}}(\theta_n(\omega)) - F_{Y^{B_n^{k_n}(x)}}(D) \geq 0$ because a cumulative distribution function is non-decreasing. Then, we set $\eta = F_{Y^x}(D) - F_{Y^x}(\theta^*)$ which is a finite value. To deal with the last term, we use our assumption **A1**.

$$F_{Y^{B_n^{k_n}(x)}}(D) - F_{Y^x}(D) \geq -M(x)||X - x||_{(k_n,n)}.$$

but we know, thanks to lemma 6.4 that $||X - x||_{(k_n,n)}$ converges to 0 p.s. Like so there exists a set Ω_3 of probability strictly non-negative such that $\forall \omega \in \Omega_3$, the previous reasonning is true and for $\epsilon < \frac{\eta}{L}$, there exists rank $N(\omega)$ such that if $n \geq N$

$$F_{Y^{B_n^{k_n}(x)}}(D) - F_{Y^x}(D) \geq 0 + -L\epsilon + \eta > 0. \tag{9}$$

Finally for $\omega \in \Omega_3$ (set of strictly non-negative measure), we have shown that

$$\lim_n \left[\theta_n(\omega) + \sum_{j=1}^{n} \gamma_{j-1} h(\theta_{j-1}(\omega)) \right] = +\infty$$

which is a absurde because of the previous part : T_n is almost surely convergent. Then θ_n does not diverges to $+\infty$ or $-\infty$.

Now, we will show that (θ_n) converges a.s. In all the sequel of the proof we will make reasonning ω by ω like in previous part. To make the reading more easy, we do not write ω and Ω any more. Thanks to equation 8 and previous subsection, we know that, with probability strictly non-negative, there exists a sequel (θ_n) such that

$$\begin{cases} (a)\ \theta_n + \sum_{j=1}^{n} \gamma_{j-1} h(\theta_{j-1}) \text{converges to a finite limit} \\ (b)\ \liminf \theta_n < \limsup \theta_n \end{cases}$$

Let us suppose that $\limsup \theta_n > \theta^*$ (we will find a contradiction and the same argument would allow us to conclude in the other case). Let us choose c and d satisfying

$$c > \theta^*,\ \liminf \theta_n < c < d < \limsup \theta_n.$$

As the sequel (γ_n) converges to 0, and (T_n) is a Cauchy sequence , we can find a deterministic rank N and two entires n and m such that $N \leq n < m$ implies

$$\begin{cases} (a)\ \gamma_n \leq \dfrac{(d-c)}{3(1-\alpha)} \\ (b)\ \left|\theta_m - \theta_n - \displaystyle\sum_{j=n}^{m-1} \gamma_j h(\theta_{j-1})\right| \leq \dfrac{d-c}{3} \end{cases}$$

We the choose m and n so that

$$\begin{cases} (a)\ N \leq n < m \\ (b)\ \theta_n < c,\ \theta_m > d \\ (c)\ n < j < m \Rightarrow c \leq \theta_j \leq d \end{cases} \tag{10}$$

This is possible since beyond N, the distance between two iterations will be either

$$\alpha\gamma_n \leq \frac{\alpha(d-c)}{3(1-\alpha)} < (d-c)$$

because $\alpha < \frac{3}{5}$ or

$$(1-\alpha)\gamma_n \leq \frac{1}{3}(d-c) < (d-c).$$

Moreover, since c and d are chosen to have an iteration inferior to c and an iteration superior to b, the algorithm will necessarely go through the segment $[c, d]$. We the take n and m the times of enter and exit of the segment. Now,

$$\begin{aligned} \theta_m - \theta_n &\leq \frac{d-c}{3} + \sum_{j=n}^{m-1} \gamma_{j+1} h_j(\theta_j) \\ &\leq \frac{d-c}{3} + \gamma_{n+1} h_n(\theta_n) \end{aligned}$$

because $n < j < m$, we get $\theta^* < c < \theta_j$ and we have already shown that in this case, $h_j(\theta_j) > 0$. We then only have to deal with the terme θ_n. If $\theta_n > \theta^*$, we can apply the same result and then

$$\theta_n - \theta_n \leq \frac{d-c}{3}$$

which is in contradiction with (b) of equation (10). When $\theta < \theta^*$,

$$\begin{aligned} \theta_m - \theta_n &\leq \frac{d-c}{3} + \gamma_n h(\theta_{n-1}) \\ &\leq \frac{d-c}{3} + \gamma_n(1-\alpha) \\ &\leq \frac{d-c}{3} + \frac{d-c}{3} < (d-c) \end{aligned}$$

which is still a contradiction with (b) of (10).
We have shown that the algorithm converges a.s.

6.2.3. *The algorithm converges a.s to θ^*.* Again we reason by contradiction. Let us name θ the limit such that $\mathbb{P}(\theta \neq \theta^*) > 0$. With probability strictly non-negative, we can find a sequel (θ_n) which converges to θ such that

$$\begin{cases} (a)\ \theta^* < \epsilon_1 < \epsilon_2 < \infty \\ (b)\ \epsilon_1 < \theta < \epsilon_2 \end{cases}$$

(or $-\infty < \epsilon_1 < \epsilon_2 < \theta^*$ but arguments are the same). Then, for n large enough, we get

$$\epsilon_1 < \theta_n < \epsilon_2.$$

In the first hand, (T_n) and (θ_n) are convergent, we also know that $\sum_{j=1}^{n} \gamma_{j+1} h(\theta_j)$ converges a.s.

But, in the second hand, let us show that $h_n(\theta_n) = P_n \left(F_{Y^{B_n^{k_n}(x)}}(\theta_n) - \alpha \right)$ is lower bounded. First we know thanks to Lemma 6.3, that for $1 < \epsilon < 1 - \beta$ and $\epsilon_n = \frac{1}{n^\epsilon}$

$$\mathbb{P}(P_n \leq \epsilon_n) \exp\left(-\frac{3n\epsilon_n}{8} \right).$$

This the general term of a convergent sum, the Borel-Cantelli Lemma gives

$$\mathbb{P}(\exists N\ \forall n \geq N\ P_n > \epsilon_n) = 1.$$

Moreover, as we have already seen in equation (9), since $\theta_n > \epsilon_1 > \theta^*$,

$$F_{Y^{B_n^{k_n}(x)}}(\theta_n) - \alpha \geq 0 - M(x)||X - x||_{(k_n,n)} + F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)$$

Then, when n is large enough to have

$$||X - x||_{(k_n,n)} \leq \frac{F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)}{2M(x)},$$

$$F_{Y^{B_n^{k_n}(x)}}(\theta_n) - \alpha \geq \frac{F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)}{2}.$$

Finally there exists a set Ω of probability strictly non-negative such that, $\forall \omega \in \Omega$

$$\sum_{k=1}^{n} \gamma_{k+1} h_k(\theta_k) \geq \frac{F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)}{2} \sum_{k=1}^{n} \gamma_k P_k \geq \sum_{k=1}^{n} \frac{1}{n^{\gamma+\epsilon}}$$

which is a contradiction (with the first hand point) because the sum is divergent $(\gamma + \epsilon < 1)$.

6.3. **Proof of Theorem 2.2 : Non-asymptotic inequality on the mean square error.** Let x be fixed in $[0,1]$. We want to study the means square error $a_n(x)$. In that purpose, let us establish an inductive inequality to conclude with Lemma 6.6. In the sequel, we will need to study $\theta_n(x)$ on the event A_n of the Lemma 6.3. Then, we begin to find a link between the quadratic risk and the mean square error on this event.

$$\begin{aligned} a_n(x) &= \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2 \mathbf{1}_{A_n}\right] + \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2 \mathbf{1}_{A_n^C}\right] \\ &\leq \mathbb{E}\left[(\theta_n-\theta^*)^2 \mathbf{1}_{A_n}\right] + C_1 \mathbb{P}(A_n^C) \end{aligned}$$

where R is the constant of the Remark 2.1. Lemma 6.3 gives the quantity $\mathbb{P}(A_n^C)$. We finally obtain

$$\mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2\right] \leq \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2 \mathbf{1}_{A_n}\right] + C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) \tag{11}$$

Let us now study the sequence $b_n(x) := \mathbb{E}\left[(\theta_n(x)-\theta^*)^2 \mathbf{1}_{A_n}\right]$.
First,

$$b_{n+1}(x) \leq \mathbb{E}\left[(\theta_{n+1}(x)-\theta^*(x))^2\right].$$

But

$$\begin{aligned} (\theta_{n+1}(x)-\theta^*(x))^2 = (\theta_n(x)-\theta^*(x))^2 &+ \gamma_{n+1}^2\left[(1-2\alpha)\mathbf{1}_{Y_{n+1}\leq\theta_n(x)}+\alpha^2\right]\mathbf{1}_{X_{n+1}\in kNN_n(x)} \\ &- 2\gamma_{n+1}(\theta_n(x)-\theta^*(x))\left(\mathbf{1}_{Y_{n+1}\leq\theta_n(x)}-\alpha\right)\mathbf{1}_{X_{n+1}\in kNN_n(x)} \end{aligned}$$

Taking the expectation conditionaly to $\mathcal{F}_n$, we get then

$$\begin{aligned} \mathbb{E}_n\left((\theta_{n+1}(x)-\theta^*(x))^2\right) \leq \mathbb{E}_n\left((\theta_n(x)-\theta^*(x))^2\right) &+ \gamma_{n+1}^2 \mathbb{P}_n\left(X_{n+1}\in kNN_n(x)\right) \\ &- 2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right)\left[\mathbb{P}_n\left(Y_{n+1}\leq\theta_n(x)\cap X_{n+1}\in kNN_n(x)\right)\right. \\ &\left.\times\, \mathbb{P}_n\left(X_{n+1}\in kNN_n(x)\right)F_{Y^x}(\theta^*)\right] \end{aligned}$$

Using the Baye's formula, we get

$$\begin{aligned} \mathbb{E}_n\left(\theta_{n+1}(x)-\theta^*(x))^2\right) \leq \mathbb{E}_n\left((\theta_n(x)-\theta^*(x))^2\right) &+ \gamma_{n+1}^2 P_n \\ &- 2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right)P_n\left[F_{Y^{B_n^{k_n}(x)}}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] \end{aligned}$$

where $P_n = \mathbb{P}_n\left(X_{n+1}\in kNN_n(x)\right)$ as in lemma 6.1. Let us split the double product term into two terms representing the two errors we made by iterating our algorithm. We still denote F_{Y^x} and $F_{Y^{B_n^{k_n}(x)}}$ the cumulative functions of the laws $g(x,\epsilon)$ and $\mathcal{L}(Y|X\in kNN_n(x))$.

$$\begin{aligned}\mathbb{E}_n\left(\theta_{n+1}(x)-\theta^*(x))^2\right) \leq & \left(\theta_n(x)-\theta^*(x)\right)^2+\gamma_{n+1}^2 P_n \\ & -2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right) P_n\left[F_{Y^{B_n^{k_n}(x)}}(\theta_n(x))-F_{Y^x}(\theta_n(x))\right] \\ & -2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right) P_n\left[F_{Y^x}(\theta_n(x))-F_{Y^x}(\theta^*(x))\right]\end{aligned} \tag{12}$$

We now use our hypothesis. By **A1**,

$$|F_{Y^{B_n^{k_n}(x)}}(\theta_n(x))-F_{Y^x}(\theta_n(x))| \geq M(x)||X-x||_{(k_n,n)}$$

and by **A3**

$$|\theta_n(x)-\theta^*(x)| \leq \sqrt{C_1}$$

thus,

$$-2\gamma_{n+1}(\theta_n(x)-\theta^*(x))P_n\left[F_{Y^{B_n^{k_n}(x)}}(\theta_n(x))-F_{Y^x}(\theta_n(x))\right] \leq 2\gamma_{n+1}\sqrt{C_1}M(x)P_n||X-x||_{(k_n,n)}$$

On the other hand, thanks to **A4** we know that,

$$\left(\theta_n-\theta^*\right)\left[F_{Y^x}(\theta_n(x))-F_{Y^x}(\theta^*(x))\right] \geq C_2(x,\alpha)\left[\theta_n(x)-\theta^*(x)\right]^2.$$

Let us come back to equation (12).

$$\begin{aligned}\mathbb{E}_n\left(\theta_{n+1}(x)-\theta^*(x))^2\right) \leq & \left(\theta_n(x)-\theta^*(x)\right)^2\left(\mathbf{1}_{A_n}+\mathbf{1}_{\bar{A_n}}\right)+\gamma_{n+1}^2 P_n \\ & -2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right)^2 C_2(x,\alpha)P_n+2\gamma_{n+1}M(x)\sqrt{C_1}||X-x||_{(k_n,n)}P_n\end{aligned}$$

where we used again Remark 2.1. To conclude, we take the expectation

$$\begin{aligned}b_{n+1}(x) \leq & C_1\mathbb{P}(A_n^C)+b_n(x)-2\gamma_{n+1}C_2(x,\alpha)\mathbb{E}\left[P_n\left(\theta_n(x)-\theta^*\right)^2\right] \\ & +\gamma_{n+1}^2\mathbb{E}(P_n)+2\gamma_{n+1}\sqrt{C_1}M(x)\mathbb{E}\left[P_n||X-x||_{(k_n,n)}\right]\end{aligned}$$

We have to compute the two expectancies. Thanks to Lemma 6.5, we first know that for n large enough,

$$\mathbb{E}(||X-x||_{(k_n,n)}P_n) \leq \left(\frac{k_n}{n+1}\right)^{1+\frac{1}{d}} C_3(d),$$

The second one is more difficult to compute. This is why we need the event A_n. By definition of A_n

$$\begin{aligned}-2\gamma_{n+1}C_2(x,\alpha)\mathbb{E}\left[P_n\left(\theta_n(x)-\theta^*\right)^2\right] & \leq -\gamma_{n+1}\epsilon_n C_2(x,\alpha)\mathbb{E}\left[\left(\theta_n(x)-\theta^*(x)\right)^2\mathbf{1}_{A_n}\right] \\ & = -2\gamma_{n+1}\epsilon_n C_2(x,\alpha)b_n(x)\end{aligned}$$

We obtain for $n \geq 1$

$$b_{n+1}(x) \leq b_n(x)\left(1-2C_2(x,\alpha)n^{-\gamma-\epsilon}\right)+\beta_n$$

with

$$\beta_n = C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + 2C_2(x,\alpha)M(x)C_3(d)\gamma_{n+1}\left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1} + \gamma_{n+1}^2\frac{k_n}{n+1},$$

Let us now use Lemma 6.6,

$$b_n(x) \leq \exp\left(-2C_2(x,\alpha)\sum_{k=1}^{n} k^{-\gamma-\epsilon}\right) b_0(x) + \sum_{k=1}^{n} \exp\left(-2C_2(x,\alpha)\sum_{j=k}^{n} j^{-\epsilon-\gamma}\right)\beta_k$$

To conclude, we reinject equation 6.3 in Equation 11 and obtain

$$a_n(x) \leq \exp\left(-2C_2(x,\alpha)\sum_{k=1}^{n} k^{-\gamma-\epsilon}\right) b_0(x) + \sum_{k=1}^{n} \exp\left(-2C_2(x,\alpha)\sum_{j=k}^{n} j^{-\epsilon-\gamma}\right)\beta_k + C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right).$$

6.4. **Proof of Corollary 2.1 : Rate of convergence.** In this part, we will denote

$$T_n^1 := \exp\left(-2C_2(x,\alpha)\sum_{k=1}^{n} k^{-\gamma-\epsilon}\right)$$

and

$$T_n^2 := \sum_{k=1}^{n} \exp\left(-\sum_{j=k}^{n} j^{-\epsilon-\gamma}\right)\beta_k.$$

We will find their order in n to conclude. When γ is fixed, our inequality shows thanks to T_n^1 that $a_n(x)$ can converges to 0 only when the sum

$$\sum_{k\geq 1} \frac{1}{k^{\gamma+\epsilon}} = +\infty.$$

This is why we must first consider $\epsilon \leq 1-\gamma$. As $\epsilon < 1-\beta$, we have to take $\beta > \gamma$.

Remark 6.1. *The case where $\epsilon = 1-\gamma$ is possible but its study shows that it is a less interessant case than for $\epsilon < 1-\gamma$ (there is a dependency in the value of $C_2(x,\alpha)$ but the optimal rate is the same as the one in the case we study). The case $\epsilon > 1-\gamma$ show that $a_n(x)$ is bounded, but we already know it. In the sequel, we then only consider $\epsilon < 1-\gamma$.*

$$\begin{aligned} T_n^1 &= \exp\left(-2C_2(x,\alpha)\sum_{k=1}^{n}\frac{1}{k^{\epsilon}}\right) \\ &\leq \exp\left(-2C_2(x,\alpha)\int_1^{n+1}\frac{1}{t^{\epsilon+\gamma}}dt\right) \\ &\leq \exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\epsilon-\gamma}-1}{(1-\epsilon-\gamma)}\right) \end{aligned}$$

To deal with the second term T_n^2 we fisrt study the order in n of β_n. Comparing exponant we get that there exists a rank $N_1(x,d)$ and constants C_5 and $C_6(d)$ (see appendix for explicit forms) such that if $\beta \leq 1-d\gamma$

$$\beta_n \leq C_5 n^{-2\gamma+\beta-1},$$

and if $\beta > 1 - d\gamma$,

$$\beta_n \leq C_6(x,d) n^{-\gamma+(1+\frac{1}{d})(\beta-1)}$$

.

We have to distinguish the two cases in the sequel.

Study of T_n^2 when $\beta > 1 - d\gamma$:

To deal with these terms, we will use arguments from [6].

$$\begin{aligned}
T_n^2 &= C_6(x,d) \sum_{k=1}^{n} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&= C_6(x,d) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor - 1} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&+ C_6(x,d) \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&:= S_1 + S_2
\end{aligned}$$

If we take $1 - \beta < \epsilon < \min((1 - d\gamma), \left(1 + \frac{1}{1+d}\right)(1 - \beta))$, we have

$$\begin{aligned}
S_2 &\leq \left(\frac{1}{\lfloor \frac{n}{2} \rfloor}\right)^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon} C_6(x,d) \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{1}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\epsilon+\gamma}} \\
&\leq \frac{C_6(x,d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(-2C_2(x,\alpha)(x) \frac{(n+1)^{1-\epsilon-\gamma} - (k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{k^{\epsilon+\gamma}} \\
&\leq \frac{C_6(x,d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2C_2(x,\alpha) \frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(2C_2(x,\alpha) \frac{(k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{k^{\epsilon+\gamma}}
\end{aligned}$$

Let us introduce N_2 the rank after which $\forall k \geq \lfloor \frac{n}{2} \rfloor$,

$$\frac{1}{k^{\epsilon+\gamma}} \leq \left(\frac{2}{k+1}\right)^{\epsilon+\gamma}.$$

For $n \geq N_2$,

$$S_2 \leq \frac{C_6(x,d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(2C_2(x,\alpha)(x)\frac{(k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{(k+1)^{\epsilon+\gamma}}$$
$$\leq \frac{C_6(x,d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \int_{\lfloor \frac{n}{2} \rfloor}^{n+1} \exp\left(2C_2(x,\alpha)\frac{(t+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{(t+1)^{\epsilon+\gamma}} dt$$
$$\leq \frac{C_6(x,d)}{2C_2(x,\alpha)n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \exp\left(\frac{2C_2(x,\alpha)}{1-\epsilon-\gamma}(n+1)^{1-\epsilon-\gamma}\right)$$

then for n large enough, there exists a constant $C_7(x,d,\alpha)$ such that

$$S_2 \leq \frac{C_7(x,d,\alpha)}{2n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}}$$

Let us now deal with the term S_1. As $k \leq \lfloor \frac{n}{2} \rfloor$, we have

$$\sum_{j=k+1}^{n} \frac{1}{j^{\epsilon+\gamma}} \geq \frac{n}{2}\frac{1}{n^{\epsilon+\gamma}}$$

then

$$S_1 = C_6(x,d) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}}$$
$$\leq C_6(x,d) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \exp\left(-C_2(x,\alpha)n^{1-\epsilon-\gamma}\right) \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}}$$
$$\leq C_6(x,d) \exp\left(-C_2(x,\alpha)n^{1-\epsilon-\gamma}\right) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}}$$

thanks to the exponentiel, S_1 is insignificant compared to S_2 whatever the behaviour of $\sum k^{-\gamma-(1-\beta)\left(1+\frac{1}{d}\right)}$, and so is T_1^n. Then, denoting $N_3(d,x)$ the rank after which we have

$$\max(S_1, T_n^1) \leq \frac{C_7(x,\alpha,d)}{4n^{(1+\frac{1}{d})(1-\beta)-\epsilon}}$$

we get, in the case where $\beta > 1-\gamma$ and $1-\beta < \epsilon < \min((1-\gamma), \left(1+\frac{1}{1+d}\right)(1-\beta))$, for $n \geq \max\left(N_0, N_1(x,d), N_3, N_3(d,x)\right)$

$$a_n(x) \leq \frac{C_7(x,\alpha,d)}{n^{-\epsilon+\left(1+\frac{1}{d}\right)(1-\beta)}}$$

Study of T_n^2 when $\beta \leq 1 - d\gamma$:

It is the same arguments and we conclude that for $1-\beta < \epsilon < \min(1-\beta+\gamma, 1-\gamma)$ and n large enough ($n \geq \max(N_0, N_1(x,d), N_2, N_3(x,\alpha,d))$),

$$a_n(x) \leq \frac{C_8(x,\alpha,d)}{n^{\gamma-\beta+1-\epsilon}}$$

6.5. **Proof of Corollary 2.2 : choice of parameter β when γ is fixed.** Let us now optimize the rate of convergence by choosing the best parameters. When $\gamma \geq \frac{1}{1+d}$ then $\gamma \geq 1-d\gamma$. So the condition $\beta > \gamma$ implies, $\beta > 1-d\gamma$ and we are in the first case. The rate of convergence is then $n^{\epsilon-\left(1+\frac{1}{d}\right)(1-\beta)}$ for $1-\beta < \epsilon < \min(1-d\gamma, \left(1+\frac{1}{d}\right)(1-\beta))$. To have the greatest rate of convergence, the best choice is then to take ϵ the smallest as possible : $\epsilon = 1-\gamma-\eta+\eta_\epsilon$ with $\eta_\epsilon > 0$. The rate is then in $n^{-(1-\beta)\frac{1}{d}+\eta_\epsilon}$. We then chose β the smallest as possible, that is to say $\beta = \gamma+\eta_\beta$ where $\eta_\beta > 0$. We obtain the rate of convergence $n^{-\frac{1}{d}(1-\gamma)+\eta}$ with $\eta = \eta_\epsilon + \frac{\eta_\beta}{d}$ which conclude the corollary.

When $\gamma < \frac{1}{1+d}$, the two cases are possible. If we take $\beta > 1-d\gamma$, we are in case 1 and in the same way than before, the rate of convergence is in $n^{\epsilon-\left(1+\frac{1}{d}\right)(1-\beta)}$. We take β and ϵ the smallest as possible as in the previous part. But the constraints $\beta > 1-d\gamma$ implies that the smallest β is $1-d\gamma+\eta_\beta$. Then, we choose $\epsilon = d\gamma-\eta_\beta+\eta_\epsilon$ and we obtain the rate of convergence $n^{-\gamma+\eta}$. In the second case, if we take $\gamma < \beta < 1-d\gamma$, we have, for $1-\beta < \epsilon < \min(1-\gamma, 1-\beta+\gamma)$, the rate of convergence $n^{-\gamma+\beta-1+\epsilon}$. In the same way, we take ϵ as small as possible : $\epsilon = 1-\beta-\eta_\epsilon$. This leads to the rate $n^{-\gamma+\eta}$. The choice of β does not matter. When then chose abitrary $\beta = \gamma+\eta_\beta$ to find back the result of the previous part. The two sub-cases given the same result, we choose the first which is the same that first result and the corollary is proved.

6.6. **Proof of corollary 2.3 : choice of parameters γ and β.** When gamma $\gamma \geq \frac{1}{1+d}$ we obtained the rate $n^{-\frac{1}{d}(1-\gamma)+\eta}$, this is why we have to chose γ as small as possible which means $\gamma = \frac{1}{1+d}$; to have the faster convergence. The rate of convergence is then $n^{-\frac{1}{1+d}+\eta}$. When $\gamma < \frac{1}{1+d}$, the rate of convergence is $n^{-\gamma+\eta}$ and the best choice is to take γ near $\frac{1}{1+d}$ and the rate is then $n^{-\frac{1}{1+d}+\eta}$. To conclude, best choices are $\gamma = \frac{1}{1+d}$, $\beta = \gamma+\eta_\beta$ and with these parameters we have shown that

$$a_n(x) \leq \frac{C_9(x,\alpha,d)}{n^{\frac{1}{1+d}-\eta}}$$

where the constant is the minimal constante between $C_7(x,\alpha,d,1,\frac{1}{1+d})$ and $C_8(x,\alpha,d,1,\frac{1}{1+d})$ (because $\epsilon < 1$ and optimal $\gamma = \frac{1}{1+d}$).

7. Appendix

Let us sum up all the constants we need in this paper.

7.1. **Constants of the model.** We denote :

- $M(x)$ the constant of continuity of the model, that is

$$\forall B \in \mathcal{B}_x,\ \forall t \in \mathbb{R},\ |F_{Y^B}(t) - F_{Y^x}t)| \leq M(x) r_B.$$

- C_{input} is the positive lower bound of the density of the inputs f_X.
- $C_g(x)$ is the positive lower bound of the density of the law $g(x, \epsilon)$.

7.2. **Compact support.** We denote :

- $[L_g, U_g]$ the compact in which are included the values of g.
- $[L_X, U_X]$ the compact in which is included the support of the distribution of X.
- $[L_{\theta_n}, U_{\theta_n}] := [L_g - (1-\alpha), U_g + \alpha]$ the sement in which θ_n can take its values $(\forall x)$.
- $U_{|.|}$ the upper bound of the compact support of the distribution of $||X - x||$ $(\forall x)$.

7.3. **Real constants.** We denote :

- $\sqrt{C_1} := \max\left(L_g + \alpha - L_X, U_X - U_g + (1-\alpha)\right)$. C_1 is the uniform in ω bound of $\left(\theta_n(x) - \theta^*(x)\right)^2$.
- $C_2(x, \alpha) := \min\left(C_g(x), \frac{1-\alpha}{U_{\theta_n} - L_{\theta_n}}\right)$. It is the constant such that

$$\left[F_{Y^x}\left(\theta_n(x)\right) - F_{Y^x}\left(\theta^*(x)\right)\right]\left[\theta_n(x) - \theta^*(x)\right] \geq C_2(x, \alpha)\left(\theta_n(x) - \theta^*(x)\right)^2.$$

- $C_3(d) := \sqrt[d]{2}\left(1 + \frac{8}{3d} + \frac{1}{\sqrt[d]{C_{input}} C_4(d)}\right)$.
- $C_4(d) := \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)+1}$.
- $C_5 := 2$.
- $C_6(x, d) := 4\sqrt{C_1} M(x) C_3(d)$.
- $C_7(x, \alpha, d, \epsilon, \gamma) = \frac{2^{\epsilon+\gamma} C_6(d)}{C_2(x,\alpha)}$.
- $C_8(x, \alpha, \epsilon, \gamma) := \frac{2^{\epsilon+\gamma} C_5}{C_2(x,\alpha)}$.
- $C_9(x, \alpha, d) := \min\left(\frac{2^{\frac{1}{1+d}+1} C_6(d)}{C_2(x,\alpha)}, \frac{2^{\frac{1}{1+d}+1} C_5)}{C_2(x,\alpha)}\right)$.
- $C_{10}(d) := \sqrt[d]{\frac{2(k_n+1)}{(n+1) C_{input} C_4(d)}}$.

7.4. **Integer constants.** We denote :

- $N_0 := 2^{\frac{1}{\epsilon - (1-\beta)}}$.
- $N_1(d, x)$ is the interger such that for $n \geq N_1(x, d)$,
  a) If $\beta \leq 1 - d\gamma$,

$$\max\left(2\sqrt{C_1} M(x) C_3(d) \gamma_{n+1} \left(\frac{k_n}{n+1}\right)^{1+\frac{1}{d}}, C_1 \exp\left(-\frac{3}{8} n^{1-\epsilon}\right)\right) \leq \frac{n^{-2\gamma+\beta-1}}{2}.$$

b) If $\beta > 1 - d\gamma$,

$$\max\left(n^{-2\gamma+\beta-1}, C_1 \exp\left(-\frac{3}{8}n^{1-\epsilon}\right)\right) \leq \sqrt{C_1}M(x)C_3(d)n^{-\gamma+\left(1+\frac{1}{d}\right)(\beta-1)}.$$

- N_2 is the rank such that $n \geq N_2$ implies

$$\forall k \geq \left\lfloor \frac{n}{2} \right\rfloor, \ \frac{1}{k^{\epsilon+\gamma}} \leq \left(\frac{2}{k+1}\right)^{\epsilon+\gamma}.$$

- $N_3(x, \alpha, d)$ is the interger such that $\forall n \geq N_3(x, \alpha, d)$,
  a) If $\beta \leq 1 - d\gamma$,

$$\max\left(S_1, T_n^1\right) \leq \frac{C_7(x, \alpha, d)}{4n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}}$$

  b) If $\beta > 1 - d\gamma$,

$$\max\left(S_1, T_n^1\right) \leq \frac{C_8(x, \alpha, d)}{4n^{\gamma-\beta+1-\epsilon}}.$$

References

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

[2] Andrieu, Moulines, and Priouret. Stability of stochastic approximation under verifiable conditions. *The Annals of Applied Probability*, 16(3):1462–1505, 2006.

[3] Philippe Barbe and Michel Ledoux. *Probabilit*. Collection Enseignement sup. EDP Sciences, Les Ulis, 2007. dition corrige de l'ouvrage paru en 1998 chez Belin.

[4] PK Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[5] Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pages 382–386, 1954.

[6] Hervé Cardot, Peggy Cénac, and Antoine Godichon. Online estimation of the geometric median in hilbert spaces: non asymptotic confidence balls. *arXiv preprint arXiv:1501.06930*, 2015.

[7] David and Nagaraja. *Order Statistics*. Wiley, 2003.

[8] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*. Applications of mathematics. Springer, New York, Berlin, Heidelberg, 1998.

[9] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[10] Marie Duflo and Stephen S Wilson. *Random iterative models*, volume 22. Springer Berlin, 1997.

[11] Vaclav Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

[12] Noufel Frikha, Stéphane Menozzi, et al. Concentration bounds for stochastic approximations. *Electron. Commun. Probab*, 17(47):1–15, 2012.

[13] Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification with the nearest neighbor rule in general finite dimensional spaces: necessary and sufficient conditions. *arXiv preprint arXiv:1411.0894*, 2014.

[14] Antoine Godichon. Estimating the geometric median in hilbert spaces with stochastic gradient algorithms. *arXiv preprint arXiv:1504.02267*, 2015.

[15] Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[16] Don O Loftsgaarden, Charles P Quesenberry, et al. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

[17] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[18] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[19] Ruppert. *Handbook of sequential analysis*. CRC Press, 1991.

[20] Jerome Sacks. Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, pages 373–405, 1958.

[21] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

[22] Thomas J Santner, Brian J Williams, and William I Notz. *The design and analysis of computer experiments*. Springer Science & Business Media, 2013.

[23] Amandine Schreck, Gersende Fort, Eric Moulines, and Matti Vihola. Convergence of Markovian Stochastic Approximation with discontinuous dynamics. March 2014.

[24] Charles J Stone. Nearest neighbour estimators of a nonlinear regression function. *Proc. Comp. Sci. Statis. 8th Annual Symposium on the Interface*, pages 413–418, 1976.

[25] Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

[26] Michael Woodroofe. Normal approximation and large deviations for the robbins-monro process. *Probability Theory and Related Fields*, 21(4):329–338, 1972.

FG AG and TLR are with the Institut de Mathématiques de Toulouse (CNRS UMR 5219). Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France.