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 estimation of conditional quantile, more precisely the quantile of the output of a real stochastic code whose inputs are in $\mathbb{R}^d$. In this purpose, we introduce a stochastic algorithm based on Robbins-Monro algorithm and on k-nearest neighbors theory. We propose conditions on the code for that algorithm to be convergent and study the non-asymptotic rate of convergence of the means square error. Finally, we give optimal parameters of the algorithm to obtain the best rate of convergence.

1. Introduction

1.1. **Stochastic code.** A stochastic code is a numerical black box with randomness inside. We can model it in this way. Let X, the inputs vector, be a random vector of $\mathbb{R}^d$, let ϵ, the random seed, be a random vector of $\mathbb{R}^m$ and let g be a map of $\mathbb{R}^d \times \mathbb{R}^m$ to $\mathbb{R}$; the output of the stochastic code g is

$$Y = g(X, \epsilon).$$

This black box is said to be stochastic because of the random seed ϵ. Indeed, contrary to numerical black box, this one does not necessarily return the same output when we provide it the same input at two different times. We note that ϵ and g are both unknown but we can observe the output when we provide the code an input. Nevertheless, each computation is very costly

The goal of this work is to find and study an algorithm which estimates the following conditional quantile : for a level $\alpha \in [\frac{1}{2}, 1]$ fixed, the target of our algorithm is

$$\theta^x(x) := q_\alpha \left(\mathcal{L} \left(g(X, \epsilon) | X = x \right) \right)$$

where we denote $q_\alpha(Z)$ the upper quantile of level α of the law Z that is

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

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

Date: September 16, 2015.

1.2. **The algorithm.** When each call to the code is not very expensive, several methods are well known to estimate quantiles. Indeed, if we have at hand a sample $(Y_i^x)_i$ where each Y_i^x is distributed like $g(X,\epsilon)|X=x$, we can estimate the quantile with the empirical quantile or with the classical stochastic algorithm for a quantile (see next paragraphe). Here, we are looking for a recursive method which allows us to estimate the conditional quantile for different x at the same time. In this purpose we begin by creating an observation sample (of a size limited by our budget) : we chose a sample of inputs $(X_1,\dots X_n)$ that we provide to the code. We then observe a sample of outputs $(Y_1 = g(X_1,\epsilon),\dots,Y_n = g(X_n,\epsilon))$. Thanks to this sample $(X_1,Y_1,\dots X_n,Y_n)$ we iterate a stochastic algorithm which allows us to estimate the conditional quantile of several x in the same time (at each iteration, several estimator of conditional quantile are updated). This algorithm is based on the classical Robbins Monro algorithm to estimate the quantile and use k-nearest neighbors theory. Let us see how this algorithm is constructed.

Robbins and Monro introduced in [17] stochastic algorithms to approximate the root of a function $h:\mathbb{R}^d \longrightarrow \mathbb{R}^d$. These algorithms are in the form

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

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

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

This kind of algorithms have been studied by several authors. In an asymptotic point of view, Robbins and Monro showed convergence of the mean square in [17]; the almost sure convergence is proved with differents methods and under different hypothesis by Blum in [5] and Schreck and al. in [20]; Fabian, Ruppert and then Sacks study the asymptotic rate of convergence in [11], [18] and [19]; in [23] Woodroofe investigate the probability of large deviations. 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 and Moulines et al. propose in [16] non-asymptotic inequality on the mean square error under strictly convex hypothesis.

The quantile is a classical example of target for these algorithms because the quantile is the root of the function $h(\theta^*) = F_X(\theta) - \alpha$ for F_X the cumulative distribution function and α the level of the quantile. To estimate the quantile (in the simple case where we have at hand a sample of a law X) the algorithm is then the following

$$\begin{cases} \theta_0 \in \mathbb{R} \\ \theta_{n+1}(x) = \theta_n(x) - \gamma_{n+1}\left(\mathbf{1}_{X_{n+1}\leq\theta_{n+1}} - \alpha\right) \end{cases} \tag{2}$$

This algorithm satisfies hypothesis of Robbins Monro algorithm to be consistent and normality gaussian (see [10] for a sum up of the asymptotic theory on Robbins-Monro algorithm). Moreover, Cardot and al. studies in [6] this algorithm for the median ($\alpha = \frac{1}{2}$)

when $\gamma_n = \frac{1}{n^\gamma}$ with $\frac{1}{2} < \gamma < 1$. They propose non-asymptotic confidence balls and then non-asymptotic inequality on the mean square error.

This algorithm would then be useful if we would like to estimate the conditional quantile for only an x . To construct an algorithm which converges for every x, we then use in addition the k-nearest neighbors theory. Let us fix an input x. To estimate the conditional quantile in x with the previous algorithm, we need to have at hand a sample of the output corresponding to the input x. But, as we said before, we can't afford to obtain a sample of the output for each input we are interested in. We only have at hand the previous observation sample. Then, for our x, we will use the classical algorithm to estimate the quantile with the sample (Y_n) but in up-dating only when Y_i is not too far from the theorical response (if we would have provided x as input) : in other word, we update the algorithm only when the input X_i we provided to the code is close to x, that is to say when X_i belongs to the k-nearest neighborhood of x :

$$||X_i - x|| \leq ||X - x||_{(k_n,n)}$$

where we denote $Z_{(i,n)}$ the i-th order statistics of a sample (Z_i) of size n. Finally when we fixed an input x , the algorithm proposed to estimate the α-quantile of the law $\mathcal{L}(Y|X = x)$ is the following

$$\begin{cases} \theta_0(x) \in \mathbb{R} \\ \theta_{n+1}(x) = \theta_n(x) - \gamma_{n+1} H(\theta_n(x), Y_{n+1}) \mathbf{1}_{X_{n+1} \in kNN_n(x)} \end{cases} \tag{3}$$

where we denote :

- (γ_n) the deterministic sequel of steps of our stochastic algorithm. We study the case where $\gamma_n = \frac{1}{n^\gamma}$ for $0 < \gamma \leq 1$.
- $kNN_n(x)$ the set of the k_n nearest neighborhood of x whithin the meaning of the euclidiean norm on $\mathbb{R}^d$ that is

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

  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),

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

This idea of considering neighbors of x is not new. It first appeared in the literature to estimate the conditional mean of the same model that the one we consider. Stone in [21] and [22] study this regression problem and propose an estimator based on k-nearest neighbors : the mean of the k_n responses coming from the k_n inputs nearest to X. He also gives conditions on k_n for this estimator to converge (these classical conditions are $k_n \longrightarrow +\infty$ and $\frac{k_n}{n} \longrightarrow 0$). Bhattacharya and al. then use in [4] this idea to introduce estimator of the conditional quantile (non-recursive) in the case where inputs are on dimension 1, as 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 (at rate $\mathcal{O}\left(n^{\frac{2}{5}}\right)$).

In our problem we have to find conditions on k_n for our algorithm to converge but we also have to tune an other parameter which is the step of the deterministic sequence γ of our gradient-descent algorithm. The paper is organized this way. In Section 2 we are interested in the a.s convergence of the algorithm. We show that if $\frac{1}{2} < \gamma < \beta < 1$, the algorithm is strongly consistent. We also propose a non-asymptotic inequality on the mean square error for the algorithm, and then obtain the rate of convergence. This allows us to exhibit best parameters ($\gamma = \frac{1}{1+d}$ and $\beta = \frac{1}{1+d} + \eta$ where η is as small as possible). Finally, in Section 3, we present some numerical simulations to illustrate our results. The technical points of the proofs are differed in Section 5.

2. Main results

In the previous section, we propose a general methode to solve our problem, by introducing a stochastic algorithm. In this section, we propose to explain how to tune parameters of this algorithm. We also give theoretical garantees under technical hypohesis.

2.1. Notations and assumptions.

For the following theorems we need to suppose two kinf of assumptions. The first one is inevitable, since we deal with k-nearest neighbors. The three others are technical hypothesis. Their necessity is debatable but relaxing them constrain to developpe very technical proofs, which is not our focus in this paper.

Let us first introduce the function

$$\mathcal{C}: \begin{cases} (E, d_H) \longrightarrow (\mathcal{M}^1(\mu), d_{VT}) \\ A \longmapsto \mathcal{L}(Y|X \in A) \end{cases}$$

where E is the set of the sets on the metric space $\left(\mathbb{R}^d, ||.||\right)$ (where $||.||$ is the euclidean norm), d_H is the Hausdorff distance defined by

$$d_H(X, Y) = \max\{\sup_{y \in Y} \inf_{x \in X} ||x - y||, \sup_{x \in X} \inf_{y \in Y} ||x - y||\},$$

$\mathcal{M}^1(\mu)$ is the set of the probability measures and d_V is the totale variation distance

$$d_{VT}(P, Q) = \sup_A |P(A) - Q(A)|.$$

In the sequel we will denote Y^A a random variable which is distributed like $g(X, Y)|X \in A$. When $A = \{x\}$ we only denote Y^x to reduce the amount of notations. Then F_{Y^A} denotes the cumulative distribution function of the law $g(X, Y)|X \in A$. We also denote $\theta^*(x)$ the quantile we want to estimate.

Following assumptions will be useful for our main theorems :

Assumption A1 The function $\mathcal{C}$ is M-Lipschitz in terms of :

$$\forall x \in \text{Supp}(X), \forall A \in E_x, \ \forall t \in \mathbb{R}, \ |F_{Y^A}(t) - F_{Y^x}(t)| \leq M \max_{a \in A} ||x - a||. \tag{4}$$

where we denote $\mathrm{Supp}(Z)$ the support of the law of Z and E_x the subset of E of sets containing the point x.

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 $\mathcal{L}(Y^{kNN(x)})$ by $\mathcal{L}(Y^x)$.

Remark 2.1. *When we do not consider compact support law, we can show that this hypothesis is true as soon as the density function on the input law f_X has a bounded derivative and the density of the couple $f_(X,Y)$ has a derivative with respect to the first variable which is bounded.*

Assumption A2 The law of inputs had a density function and this density is lower-bounded by a constante $C_{inputs} > 0$.

This hypothesis is very strong. It implies in particular that the law of inputs has compact support but this kind of hypothesis are usual in k-nearest neighbors theory when you do not want to make technical development, as we can see for example in [13].

Assumption A3 The code function g is at values in a compact $[A, B]$.

Remark 2.2. *This assumptions implies, when $\beta > \gamma$ that for all x, $\theta_n(x)$ is bounded a.s uniformly in ω.*

Indeed, let N_0 be the first step for which θ_{N_0} goes out of $[A, B]$, by the right. At step $N_0 - 1$, the algorithm is, in the worst case in B. Then, at step N_0, we get $\theta_{N_0}(x) = B + \alpha\gamma_{N_0+1}$. At the next step, since $Y_{N_0+2} \leq \theta_{N_0+1}$, the algorithm do not move or comes back in direction of $[A, B]$ by a step of $-\gamma_{N_0+2}(1-\alpha)\alpha$. A classical results of Robbins-Monro is that, in this case, the algorithm comes back to $[A, B]$ because the sum of the deterministic stepwises γ is divergent. Here, we also have to take account of $\mathbf{1}_{X \in kNN(x)}$.

Imagine we fixe a point x. At step n, this point has a probability to be concerned by the new data of $n^\beta/n = n^{\beta-1}$. Then, until step n and at time t, the points has updated $\sum_{k \leq n} k^{\beta-1} \approx n^\beta$ times (and so $n \approx t^{\frac{1}{\beta}}$). Then, the stepwise $\gamma_n = \frac{1}{n^\gamma} = \frac{1}{(t^{\frac{1}{\beta}})^\gamma} = \frac{1}{n^{\frac{\gamma}{\beta}}}$ satisfies $\sum_n \gamma_n = +\infty$ because $\gamma < \beta$.

Finally, the algorithm comes back to the compact and if it goes out later, it will not goes further than $B + \alpha\gamma_{N_0+1}$ because the sequence γ is decreasing.

Then, we have shown that

$$\forall x, \forall n, \theta_n(x) \in [A - (1-\alpha); B + \alpha] := [A', B'].$$

Denoting A_x the minimum of the support of X and B_x its maximum, we then have $\sqrt{R} := \max(B + \alpha - A_x, B_x - A + (1-\alpha))$ is the uniform bound of $\theta_n(x) - \theta^*(x)$.

Assumption A4 For each x, the law $g(X, \epsilon)|X = x$ has a density which is lower-bounded by a constante $D(x) > 0$.

Then, denoting $D_{code}(x) := \min(D(x), \frac{1-\alpha}{B'-A'}, \frac{\alpha}{B'-A'})$ (with notations of previous paragraph), we have

$$\forall \theta_n, \left[F_{Y^x}(\theta_n) - F_{Y^x}(\theta^*(x))\right]\left[\theta_n - \theta^*(x)\right] \geq D_{code}(x)\left[\theta_n - \theta^*(x)\right]. \tag{5}$$

Indeed, it is obvious when $\theta_n \in \text{Supp}(Y^x)$. When, it is not the case, we know that $\theta_n \in [A', B']$. Imagine, that $A' \leq A_x \leq \theta^* \leq B_x \leq \theta_n \leq B'$. Then, we have $F(\theta_n) = 1$, $F(\theta^*) = \alpha$ and

$$D_{code}(x) \leq \frac{1-\alpha}{B'-A'} \leq \frac{1-\alpha}{B'-A^x} \leq \frac{1-\alpha}{\theta_n - \theta^*}$$

this is why

$$(\theta_n - \theta^*)(F(\theta_n) - F(\theta^*)) = (\theta_n - \theta^x)(1-\alpha) \geq (\theta_n - \theta^*)D_{code}(x)(\theta_n - \theta^*) = D_{code}(x)(\theta_n - \theta^*)^2.$$

The same proof allow to study other cases.

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 [16] which is not true in the case 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**, if $\frac{1}{2} < \gamma < \beta < 1$, then the algorithm 3 at x is a.s convergent.*

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.

1) We decompose H into a a martingale term and a remainder term by setting

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

Then

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

is bounded in L^2 martingale and so 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 conditionnal quantile we want to estimate.

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** as you can see in section 5.

Comments on parameters. In this theorem, we have $\frac{1}{2} < \gamma < 1$ which is a classical assumption for Robbins Monro algorithm to be consistent as you can see in [17] because we need a stepwise sequence γ_n such that

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

The number of neighbors is $\lfloor n^\beta \rfloor$ with $0 < \beta < 1$. $\beta > 0$ means that we consider at least one neighbor. $\beta < 1$ means that we want a number a neighbors that goes to $+\infty$ which implies the crucial following property (see Lemma 6.4)

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

Finally we need to choose $\beta > \gamma$ as we have seen before.

2.3. **Non-asymptotic inequality.** We want to study the rate of converge of the mean square error that we denote $a_n(x) := \mathbb{E}\left(\left(\theta_n(x) - \theta^*(x)\right)^2\right)$ where $\theta^*(x)$ is our target, in other words the quantile of the law $\mathcal{L}(Y|X = x)$.

Theorem 2.2. *Let x be an fixed input. Under hypothesis **A1**, **A2**, **A3** and **A4**, the mean square error $a_n(x)$ of the algorithm 3 at x satisfies the following inequality : for all $0 < \gamma < 1$, $0 < \beta < 1$ and $1 > \epsilon > 1 - \beta$, for $n \geq 2^{\frac{1}{\epsilon-(1-\beta)}} := N_0$,*

$$a_n(x) \leq R\exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + a_0(x)\exp\left(-2D_{code}(x)\sum_{k=1}^{n}\frac{1}{k^{\gamma+\epsilon}}\right) + \sum_{k=1}^{n}\exp\left(-2D_{code}(x)\sum_{i=k}^{n}\frac{1}{i^{\gamma+\epsilon}}\right)\beta_k$$

where $\beta_n = R\exp\left(-\frac{3n^{1-\epsilon}}{8}\right) + 2\sqrt{R}MD(d)\gamma_{n+1}\left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1} + \gamma_{n+1}^2\frac{k_n}{n+1}$, R is defined in remark 2.2, $D(d) = \sqrt[d]{2}\left(1 + \frac{8}{3d} + \frac{1}{\sqrt[d]{C_{input}H(d)}}\right)$ and $H(d) = \frac{\pi^{\frac{5}{2}}}{\Gamma(\frac{d}{2})+1}$.

Sketch of proof : The idea of the proof is to establish a recursive inequality on $a_n(x)$ (an idea from [16]) of the form

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

and to conclude with Lemma 6.6. In this purpose we begin by developing 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 to $\mathcal{F}_n$, and using the Baye's formula, we get

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

where $P_n = \mathbb{P}_n\left(X_{n+1} \in kNN_n(x)\right)$ as in lemma 6.1. Then we make appear the two errors we need to deal with.

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

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

and by **A3**,

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

thus,

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

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 the bias error. Thanks to Assumption **A4** we get

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

Taking now the expectation of our inequality we get (by using remark 2.2)

$$a_{n+1}(x) \leq a_n(x) - 2\gamma_{n+1}D_{code}(x)\mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2P_n\right] + \gamma_{n+1}^2\mathbb{E}(P_n) + 2\gamma_{n+1}M\sqrt{R}\mathbb{E}(||X-x||_{(k_n,n)}P_n$$

This equation reveals 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 $\mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2P_n\right]$. 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. For γ, we have $0 < \gamma < 1$ and then we can explore what happens when $\gamma \leq \frac{1}{2}$ which is unusual (for exampl, this case is not studied in [14]). It can be explains by the fact that our second parameter β can compensate a small choice of γ. Finally, we have to take $\beta > \gamma$ for the same reason than in previous theorem.

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

- The bias error gives the term $\exp\left(-2D_{code}(x)\sum_{k=1}^{n}\frac{1}{k^{\epsilon+\gamma}}\right)$ 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 $1 > \epsilon$. Let us sudy the two others. They are power of n, then we have to compare their exposant, to exhibit the dominating one. First, we denote N_1 the rank after which when $\beta \leq 1 - d\gamma$,

$$\max\left(2\sqrt{R}MD(d)\gamma_{n+1}\left(\frac{k_n}{n+1}\right)^{1+\frac{1}{d}}, R\exp\left(-\frac{3}{8}n^{1-\epsilon}\right)\right) \leq n^{-2\gamma+\beta-1}, \text{ and}$$

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

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

Finally, when $n \geq \max(N_0, N_1)$ and $\beta \leq 1 - d\gamma$ we get

$$\beta_n \leq 3n^{-2\gamma+\beta-1}$$

and when $n \geq \max(N_0, N_2)$ and $\beta > 1 - d\gamma$, we get

$$\beta_n \leq 4\sqrt{R}MD(d)n^{-\gamma+(1+\frac{1}{d})(\beta-1)}.$$

We notice that N_0 and N_1 are not the same kind of rank. In fact, N_1 is reasonably small whatever the parametersare , because it is only the rank after which an exponential term is smaller than a power term. In fact, it depends on the problem constants ($D_{code}(x)$, M, ...). N_0 is not so nice, because, it increase 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). We then understand that ranks like N_1 don't make any problem : if an equality is true after N_1, it is nevertheless a non-asymptotic inequality, because the inequality is true when n is reasonably small. The difficulty will mostly be in N_0. For this reason and to simplify notations, in the sequel, we still denote N_1 ranks which are not huge but allows us to exhibit dominating terms. In fact, N_1 is simply the maximum of all these ranks which are reasonably small compared to N_0.

Corollary 2.1. *$a_n(x)$ decreases to 0 with the following rate : $\forall n \geq \max(N_0, N_1)$, when $\beta > 1 - d\gamma$ and $1 - \beta < \epsilon < \min\left(1-\gamma, \left(1+\frac{1}{d}\right)(1-\beta)\right)$,*

$$a_n(x) \leq \frac{C_1}{n^{-\epsilon+\left(1+\frac{1}{d}\right)(1-\beta)}}$$

where $C_1 = \frac{2^{\gamma+3}\sqrt{R}MD(d)}{D_{code}(x)}$, and when $\beta \leq 1 - d\gamma$ and $1 - \eta < \min\left(1-\beta+\gamma, 1-\gamma\right)$

$$a_n(x) \leq \frac{C_2}{n^{\gamma-\beta+1-\epsilon}},$$

where $C_2 = \frac{3\times 2^{2+\gamma}}{D_{code}(x)}$.

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* $\beta = \gamma + \eta$ *where* $\eta > 0$ *is as small as possible. In this case, we get for* $n \geq \max(N_0, N_1)$*, when* $\gamma \geq \frac{1}{1+d}$

$$a_n(x) \leq \frac{C_1}{n^{\frac{1}{d}(1-\gamma)-\eta'}},$$

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

$$a_n(x) \leq \frac{C_2}{n^{\gamma-\eta'}}$$

where in the two cases $\eta' = \frac{\eta}{d} - \eta_1$ *and* $\eta_1 = \epsilon - (1-\beta)$.

Comparison with others results. When they study the MSE 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.

To 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 $\beta = \frac{1}{2} + \eta$ in dimension 1, a rate of convergence of the MSE in $n^{\frac{1}{2}}$ which is then a slower. Nevertheless, our result is non-asymptotic and our algorithm is easier to compute that their estimator which necessitate 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 more rapidly when parameters are* $\gamma = \frac{1}{1+d}$ *and* $\beta = \gamma + \eta$ *where* $\eta > 0$ *is as small as possible. We indeed obtain with these parameters, for* $n \geq \max(N_0, N_1)$

$$a_n(x) \leq \frac{C_1}{n^{\frac{1}{1+d}-\eta'}}$$

where η' *is the the same than in corollary 2.1 and* $C_1 = 2^{3+\frac{1}{1+d}} \frac{\sqrt{R}MD(d)}{D_{code}(x)}$.

Sketch of proof : It is easy optimization.

Comment on the rank N_0. As we saw before the rank N_1 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 $\epsilon = 1-\beta+\eta_1$ 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 $N_0 = \exp\left(2\eta_1^{-1}\right)$ the rate of convergence is on $\mathcal{O}\left(n^{\frac{-1}{1+d}+\frac{\eta}{d}+\eta_1}\right)$. Then the more η_1 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 $\eta_1 = 2(\ln(10000))^{-1} \approx 0.217$. In this case, we can theoretically obtain a risk of $N^{\frac{-1}{1+d}+\frac{2}{\ln(N)}}$ (where we forgot the term $d\eta^{-1}$ which is very small compared to the two others terms). It means that in dimension 1, the MSE 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, very regular is the 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 know that $\mathcal{L}(Y|X=x) = \mathcal{U}([-\frac{1}{2}+x^2; \frac{1}{2}+x^2])$. Then we have

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

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

$(a > 0,\ b > 0)$, then

$$|F_{Y^A}(t) - F_{Y^x}(t)| \leq \left| \frac{\int_{-\infty}^{t} \int_A f_{(X,Y)}(z,y)dydz}{\int_A f_X(z)dz} - \int_{-\infty}^{t} f_{(X,Y)}(x,y)dy \right|$$
$$\leq \frac{\int_{-\frac{1}{2}}^{t} \int_{x-a}^{x+b} \left| \mathbf{1}_{[-\frac{1}{2}+z^2;\frac{1}{2}+z^2]} - \mathbf{1}_{[-\frac{1}{2}+z^2;\frac{1}{2}+z^2]} \right| (y)dzdy}{\mu(A)}$$

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 :

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

From now, there are again two cases. We always have $(t+\frac{1}{2})^{\frac{1}{2}} \leq x$ since $t \in [-\frac{1}{2}; x^2 - \frac{1}{2}]$. 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

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

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

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

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 $n = 50000$. 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 $\frac{1}{2}$.

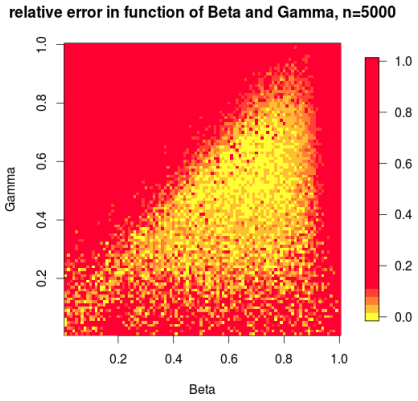

FIGURE 1. Relative error in fonction of β and γ

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.

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 risk in function of β to check that the risk 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 risk 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 mean square error (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 more the decreasing is fast.

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.

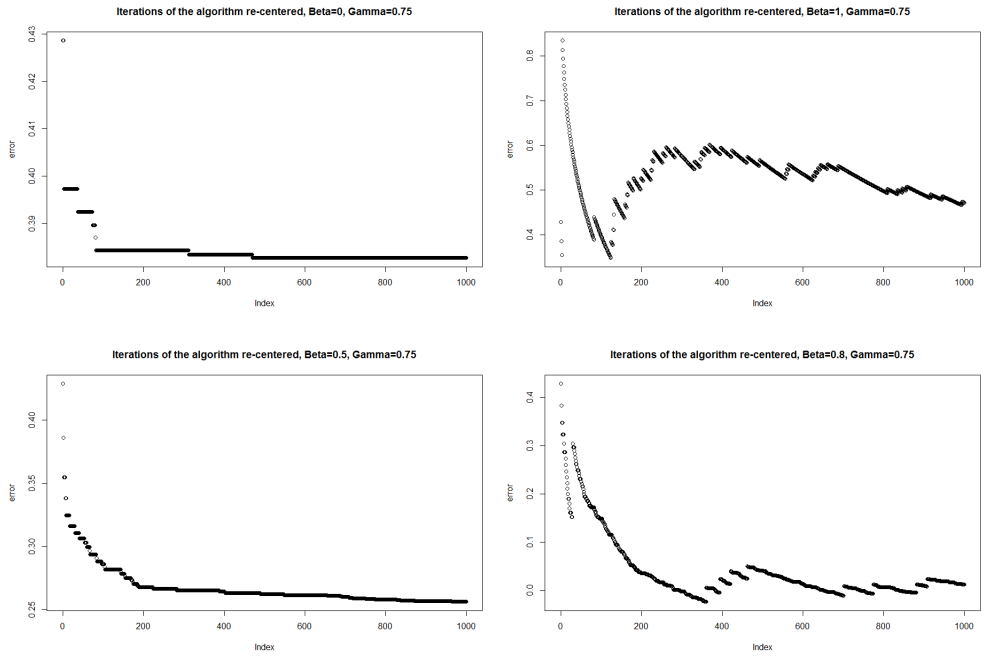

FIGURE 2. Convergence a.s of the algorithm

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 :

$$a_n(x) \leq \frac{C_1}{n^{\frac{1}{1+d}-\eta'}}.$$

Since $R = 4$, $d = 1$, $C_{input} = 1$, $D_{code}(x) = 0.017$, $M = \frac{2}{3}$, we get $H(d) \approx 6$ and $D(d) \approx 7.5$. Finally, $C_1 \approx 6000$. Then we obtain a bound of 420 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 $D_{code}(x)$ 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 $[A, B]$. This clearly can happen with a very small probability. If we do not take account of this case, we have $D_{code}(x) = 2$. Then $C_1 \approx 110$ and then, for $n = 200$, the bound is 7.7. 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.

3.2. **Dimension 1 - absolute value function.** Let us see what happens when the function g is not continuous 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).

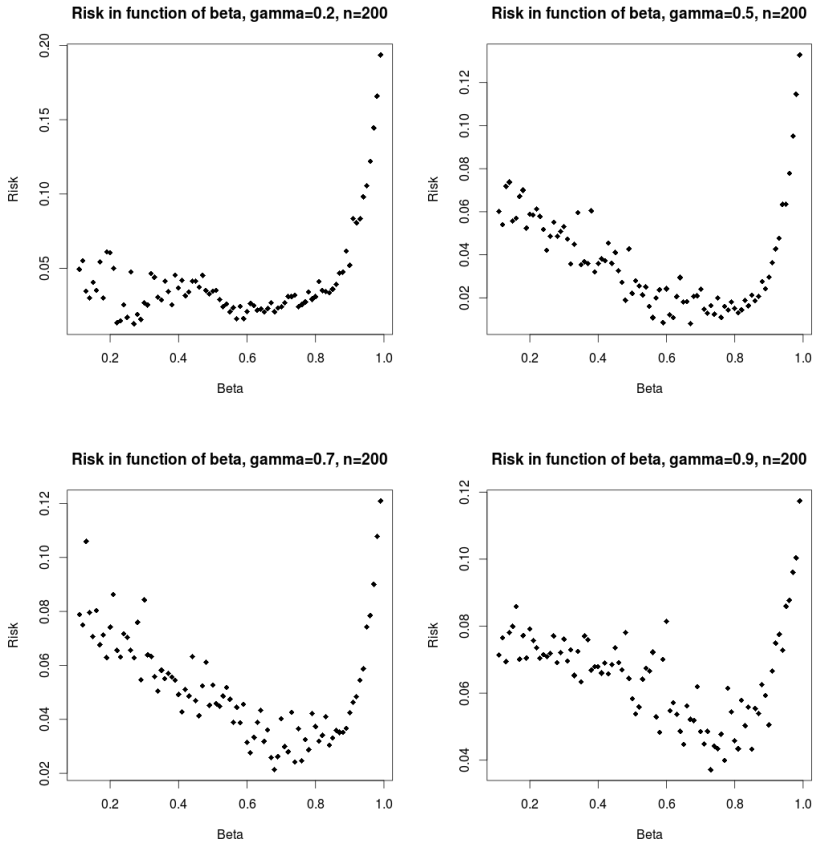

FIGURE 3. Choice of β when γ is fixed

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 code :

$$g_1(X, \epsilon) = ||X||^2 + \epsilon \text{ and } g_2(X, \epsilon) = x_1^2 + x_2 + \epsilon,$$

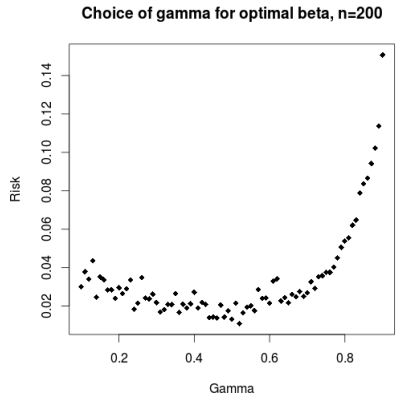

FIGURE 4. Choice of γ when β is optimal

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 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% !

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 $\epsilon \sim \mathcal{U}([-0.5, 0.5])$. 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%.

5. Conclusion and perspectives

In this paper we aimed at estimating a conditional quantile of the output of a stochastic code where inputs are 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 the 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 $\beta = \gamma + \eta$ and $\gamma = \frac{1}{2}$. Numerical

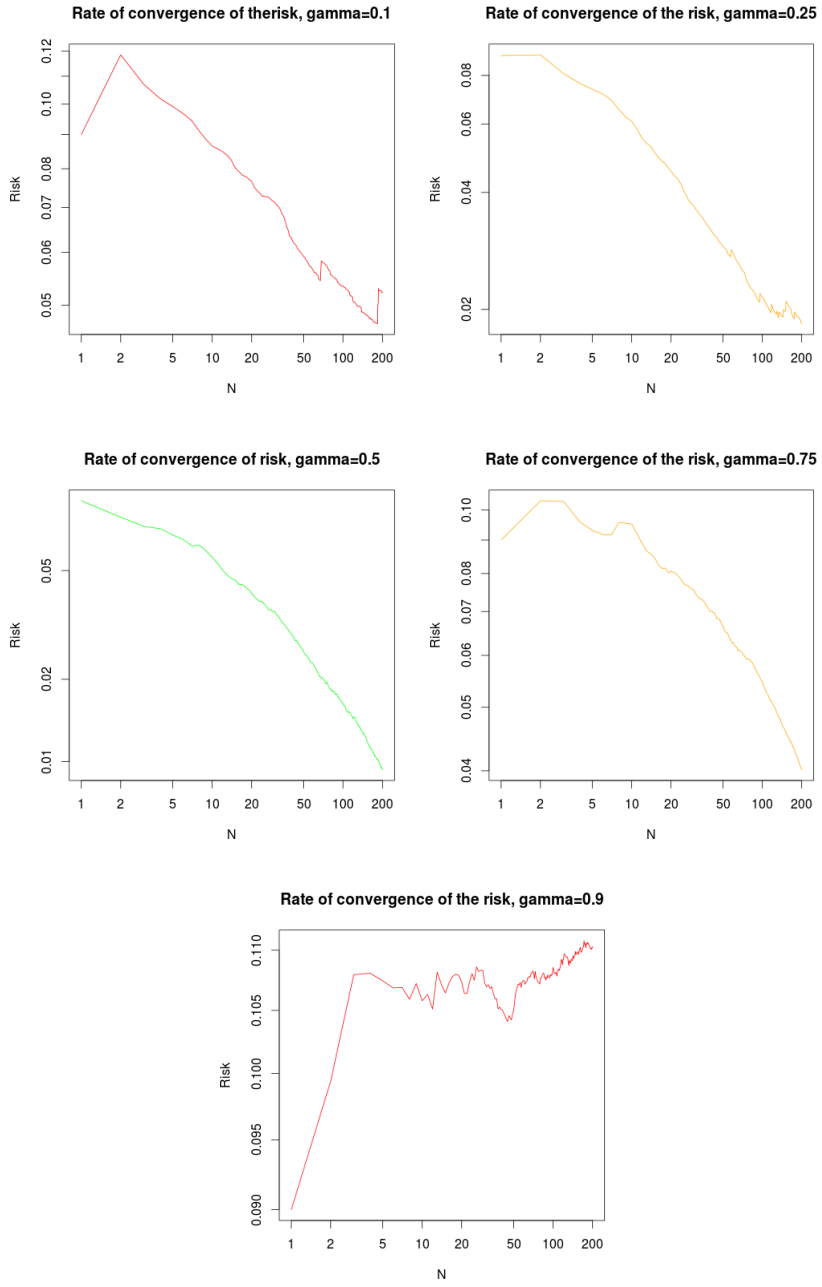

FIGURE 5. Convergence of the mean square error in logarithm scale

simulations that we made show that our algorithm with theoretical optimal parameters is really powerful to estimate a conditional quantile, even in dimension $d > 1$.

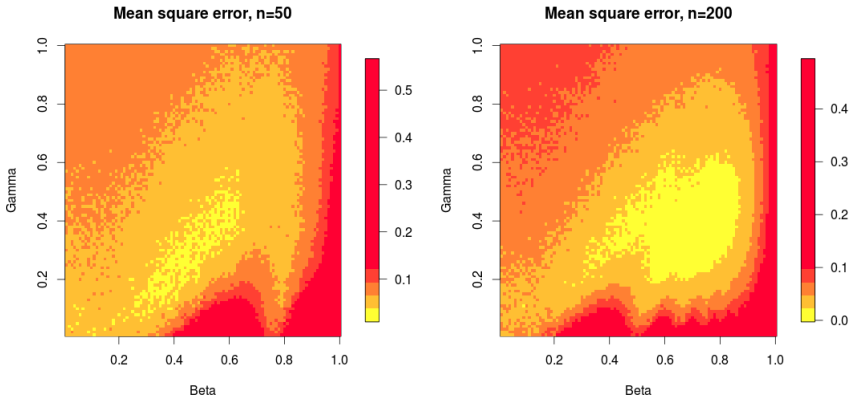

FIGURE 6. Mean square error in function of β and γ for the square function

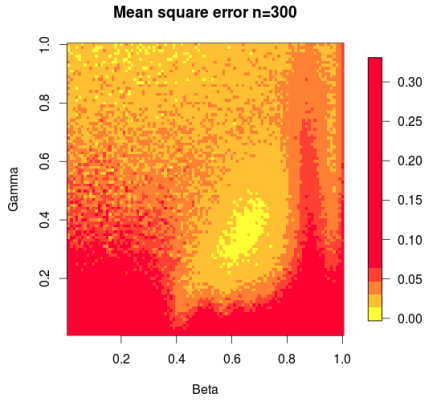

FIGURE 7. MSE in function of β and γ for absolut value function

Even if the theoretical guarantees are shown under strong technical assumptions, 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.**

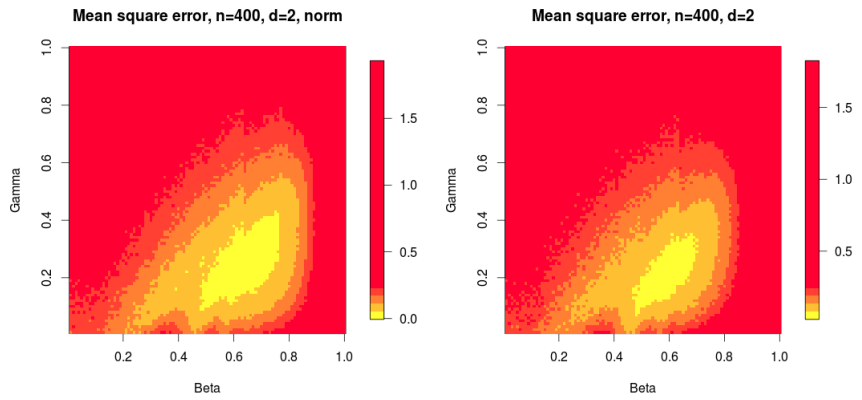

FIGURE 8. Mean square error in function of β and γ

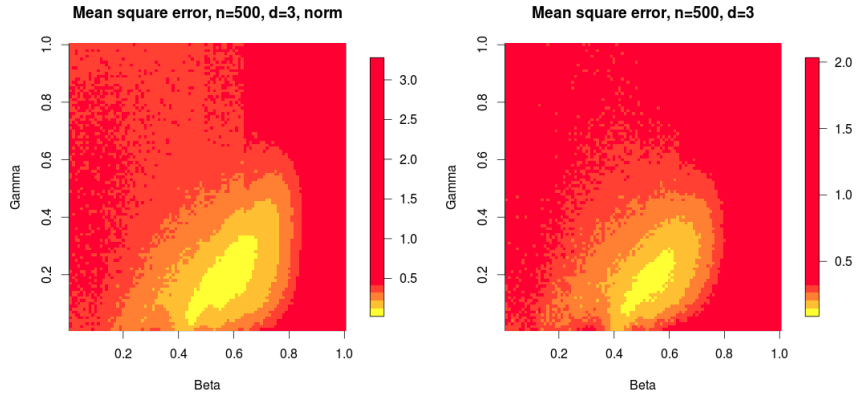

FIGURE 9. Mean square error in function of β and γ

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* $||X-x||$ *where* X *is the input law,* $||X-x||_{(k_n,n)}$ *the* k_n *order statistic of the sample* $(||X_1 - x||, \dots ||X_n - x||)$ *of size* n *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

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

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

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

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

□

Lemma 6.2. *Denoting $\mathcal{B}(n,p)$ the Binomiale law 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 Binomiale 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])

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

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

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

and the result follows.

□

Definition 6.1. *Let $B_n^{k_n}(x)$ be the set 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.

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

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,*

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

where $D(d) = \sqrt[d]{\frac{4}{C_{input}H(d)}}\left(1+\frac{8}{3d}\right)$, $H(d) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$, *and* C_{input} *is defined by assumption* ***A2*** .

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]), denoting A the upper-bound of the support of the denisty of the inputs law :

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

Using same arguments that in Lemma 2.1, we get

$$\begin{aligned} I := \int_0^A \mathbb{P}(||X-x||_{(k_n+1,n+1)} > u)du &= \int_0^{\sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}H(d)}}} \mathbb{P}(\mathcal{B}(n+1,q_u) < k_n+1)du \\ &\quad + \int_{\sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}H(d)}}}^A \mathbb{P}(\mathcal{B}(n+1,q_u) < k_n+1)du \\ &\leq \int_0^{\sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}H(d)}}} 1du + \int_{\sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}H(d)}}}^A \exp\left(-\frac{3(n+1)C_{input}H(d)u^d}{32}\right)du \end{aligned}$$

where we use Lemma 6.2 in the second integrale because $u > \sqrt[d]{\frac{2(k_n+1)}{(n+1)CH(d)}}$ implies $\frac{k_n+1}{n+1} < \frac{q_u}{2}$. Then, denoting $B = \sqrt[d]{\frac{2(k_n+1)}{(n+1)C_{input}H(d)}}$ we get

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

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. *Creation of a martingale.* Let us 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 $\mathcal{F}_n = \sigma(X_1, \dots X_n, Y_1, \dots Y_n)$. In the sequel we still denote $\mathbb{P}_n$ ad $\mathbb{E}_n$ 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 A such that for n large enough, $\theta_n(\omega) > A > \theta^*$. Then

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

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

$$F_{Y^{B_n^{k_n}(x)}}(A) - F_{Y^x}(A) \geq -M \max_{a\in B_n^{k_n}(x)} ||x-a|| = -M||X-x||_{(k_n,n)}.$$

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 $n \geq N$

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

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 a and b satisfying

$$a > \theta^*,\ \liminf \theta_n < a < b < \limsup \theta_n.$$

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

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

We the choose m and n so that

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

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

$$\alpha\gamma_n \leq \frac{\alpha(b-a)}{3(1-\alpha)} < (b-a)$$

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

$$(1-\alpha)\gamma_n \leq \frac{1}{3}(b-a) < (b-a).$$

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

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

because $n < j < m$, we get $\theta^* < a < \theta_j$ 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

$$\theta_n - \theta_n \leq \frac{b-a}{3}$$

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

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

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 $\theta_n > \epsilon_1 > \theta^*$,

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

Then, when n is large enough to have

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

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

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 quadratic risk $\mathbb{E}((\theta_n(x) - \theta^*(x))^2) := a_n(x)$. 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 quadratic risk on this event.

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

where R is the constant of the Remark 2.2. Lemma 6.3 gives the quantity $\mathbb{P}(A_n^C)$. We finally obtain

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

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 conditionnaly to $\mathcal{F}_n$, we get then

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

Using the Baye's formula, we get

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

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 $\mathcal{L}(Y|X=x)$ 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**,

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

and by **A3**

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

thus,

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

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

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

Let us come back to equation (12).

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

where we used again Remark 2.2. To conclude, we take the expectation

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

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

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

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

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

We obtain for $n \geq 1$

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

with

$$\beta_n = R\exp\left(-\frac{3n^{1-\epsilon}}{8}\right)+2RMD(d)\gamma_{n+1}\left(\frac{k_n}{n+1}\right)^{\frac{1}{d}+1}+\gamma_{n+1}^2\frac{k_n}{n+1},$$

Let us now use Lemma 6.6,

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

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

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

6.4. **Proof of Corollary 2.2 : Choice of β when γ is fixed.** In this part, we will denote

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

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 $D_{code}(x)$ 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 $\epsilon < 1-\gamma$.*

$$T_n^1 = \exp\left(-2D_{code}(x)\sum_{k=1}^{n}\frac{1}{k^{\epsilon}}\right) \leq \exp\left(-2D\int_1^{n+1}\frac{1}{t^{\epsilon+\gamma}}dt\right) \leq \exp\left(-2D\frac{(n+1)^{1-\epsilon-\gamma}-1}{(1-\epsilon-\gamma)}\right)$$

To deal with the second term T_n^2 we fisrt study the order in n of β_n. There are two cases. When $\beta \leq 1-d\gamma$, we have for n large enough,

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

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

$$\beta_n \leq 4RMD(d)n^{-\gamma+(1+\frac{1}{d})(\beta-1)}$$

.

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].

$$\begin{aligned}
T_n^2 &= 4RMD(d)\sum_{k=1}^{n}\exp\left(-2D_{code}(x)\sum_{j=k+1}^{n}\frac{a}{j^{\epsilon+\gamma}}\right)\frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&= 4RMD(d)\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor-1}\exp\left(-2D_{code}(x)\sum_{j=k+1}^{n}\frac{a}{j^{\epsilon+\gamma}}\right)\frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&+ 4RMD(d)\sum_{k=\lfloor\frac{n}{2}\rfloor}^{n}\exp\left(-2D_{code}(x)\sum_{j=k+1}^{n}\frac{a}{j^{\epsilon+\gamma}}\right)\frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&:= S_1 + S_2
\end{aligned}$$

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

$$\begin{aligned}
S_2 &\leq \left(\frac{1}{\lfloor\frac{n}{2}\rfloor}\right)^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon} 4RMD(d)\sum_{k=\lfloor\frac{n}{2}\rfloor}^{n}\exp\left(-2D_{code}(x)\sum_{j=k+1}^{n}\frac{1}{j^{\epsilon+\gamma}}\right)\frac{1}{k^{\epsilon+\gamma}} \\
&\leq \frac{4RMD(d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}}\sum_{k=\lfloor\frac{n}{2}\rfloor}^{n}\exp\left(-2D_{code}(x)\frac{(n+1)^{1-\epsilon-\gamma}-(k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)\frac{1}{k^{\epsilon+\gamma}} \\
&\leq \frac{4RMD(d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}}\exp\left(-2D_{code}(x)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)\sum_{k=\lfloor\frac{n}{2}\rfloor}^{n}\exp\left(2D_{code}(x)\frac{(k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)\frac{1}{k^{\epsilon+\gamma}}
\end{aligned}$$

Now, for n large enough, we have $\frac{1}{k^{1-\epsilon-\gamma}} \leq \frac{2}{(k+1)^{1-\epsilon-\gamma}}$ and then

$$
\begin{aligned}
S_2 &\leq \frac{4RMD(d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2D_{code}(x)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \exp\left(2D_{code}(x)\frac{(k+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{(k+1)^{\epsilon+\gamma}} \\
&\leq \frac{4RMD(d)}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2D_{code}(x)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \int_{\lfloor \frac{n}{2} \rfloor}^{n+1} \exp\left(2D_{code}(x)\frac{(t+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) \frac{1}{(t+1)^{\epsilon+\gamma}} dt \\
&\leq \frac{4RMD(d)}{2D_{code(x)}n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}} \exp\left(-2D_{code}(x)\frac{(n+1)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right) 2^{\epsilon+\gamma} \exp\left(\frac{2D_{code}(x)}{1-\epsilon-\gamma}(n+1)^{1-\epsilon-\gamma}\right)
\end{aligned}
$$

then for n large enough, there exists a constant $C := 2^{\epsilon+\gamma+1}\frac{RMD(d)}{D_{code}(x)}$ such that

$$S_2 \leq \frac{C}{n^{\left(1+\frac{1}{d}\right)(1-\beta)-\epsilon}}$$

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

$$
\begin{aligned}
S_1 &= 4RMD(d) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \exp\left(-2D_{code}(x) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}} \\
&\leq 4RMD(d) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \exp\left(-D_{code}(x)n^{1-\epsilon-\gamma}\right) \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}} \\
&\leq 4RMD(d) \exp\left(-Dn^{1-\epsilon-\gamma}\right) \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}}
\end{aligned}
$$

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. Let N_1 denote the rank after which we get,

$$\max(S_1, T_n^1) \leq \frac{C}{2n^{(1+\frac{1}{d}(1-\beta)-\epsilon}}.$$

Finally, in the case where $\beta > 1-\gamma$ and $1-\beta < \epsilon < \min((1-\gamma), \left(1+\frac{1}{1+d}\right)(1-\beta))$, for $n \geq \max(N_0, N_1, M_1)$, we get

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

for $C' = 2^{\epsilon+\gamma+2}\frac{RMD(d)}{D_{code}(x)} \leq C_1$ because $\epsilon < 1$.

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,

$$S_2 \leq \frac{c}{n^{\gamma-\beta+1-\epsilon}}$$

with $C = 3 \times \frac{2^{1+\epsilon+\gamma}}{D_{code}(x)} \leq C_2$ since $\epsilon < 1$ and this is the slowest term.

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 β and ϵ the smallest as possible : $\beta = \gamma + \eta$ and $\epsilon = 1 - \gamma - \eta + \eta_1$ with η and η_1 strictly non-negative. We obtain the rate of convergence $n^{-\frac{1}{d}(1-\gamma)+\eta'}$ with $\eta' = \eta_1 + \frac{\eta}{d}$ 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. But the constraints $\beta > 1 - d\gamma$ implies that the smallest β is $1 - d\gamma + \eta$. Then, we choose $\epsilon = d\gamma - \eta + \eta_1$ 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 β and ϵ as small as possible : $\beta = \gamma + \eta$ and $\epsilon = 1 - \gamma - \eta + \eta'$. This leads to the rate $n^{-\gamma+\eta'}$. The two sub-cases given the same result, we choose the first which is the same that first result and the corollary is proved.

6.5. **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 $\gamma = \frac{1}{1+d}$, $\beta = \gamma + \eta$.

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] 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.

[16] 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.

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

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

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

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

[21] 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.

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

[23] 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.