HAL open science

CONDITIONAL QUANTILE SEQUENTIAL ESTIMATION FOR STOCHASTIC CODES

Tatiana Labopin-Richard, F Gamboa, Aurélien Garivier

To cite this version:

Tatiana Labopin-Richard, F Gamboa, Aurélien Garivier. CONDITIONAL QUANTILE SEQUENTIAL ESTIMATION FOR STOCHASTIC CODES. 2015. hal-01187329v5

HAL Id: hal-01187329

https://hal.science/hal-01187329v5

Preprint submitted on 19 May 2016 (v5), last revised 20 Jul 2019 (v7)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 in the context of real stochastic codes with vector-valued inputs. Our algorithm is a combination of k-nearest neighbours and of a Robbins-Monro estimator. We discuss the convergence of the algorithm under some conditions on the stochastic code. We provide non-asymptotic rates of convergence of the mean square error and we discuss the tuning of the algorithm's parameters.

1. Introduction

Computer code experiments have encountered, in the last decades, a growing interest among statisticians in several fields (see [27] and references therein and also [19, 26, 22, 18, 5]...). In the absence of noise, a numerical black box $g : \mathbb{R}^d \to \mathbb{R}$ maps an *input vector* X to $Y = g(X) \in \mathbb{R}$. When the black box does include some randomness, the code is called *stochastic* and the model is as follows: a random vector $\epsilon \in \mathbb{R}^m$, called *random seed*, models the stochasticity of the function, while X is a random vector. The random seed and the input are assumed to be stochastically independent. The map g (which satisfies some regularity assumption specified below) is defined on $\mathbb{R}^d \times \mathbb{R}^m$ and outputs

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

hence yielding possibly different values for the same input X. One observes a sample of (X, Y), without having access to the details of g. However, those observations are often expensive (for example when g has a high computational complexity) and one aims at learning rapidly some properties of interest on g.

We focus in this work on the estimation of the conditional quantile of the output Y given the input X. For a given level $\alpha \in [1/2, 1)$ and for every possible input $x \in \mathbb{R}^d$, the target is

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

where $q_\alpha(Z) := F_Z^{-1}(\alpha)$ is the quantile of level α of the random variable Z and $F_Z^{-1}(u) := \inf\{x : F_Z(x) \geq u\}$ is the generalized inverse of the cumulative distribution function of Z. Moreover, we would like to estimate such a quantile for different values of x.

Date: May 19, 2016.

1.1. **The algorithm.** For a fixed value of x, there are several well-known procedures to estimate the quantile $\theta^*(x)$. Given a sample $(Y_i^x)_{i=1\dots n}$ of $Y^x := g(x,\epsilon)$, the empirical quantile is a solution. For a sequential estimation, one may use a Robbins Monro [23] estimator. This method permits to iteratively approximate the zero of a function $h : \mathbb{R} \to \mathbb{R}$ by a sequence of estimators defined by induction: $\theta_0 \in \mathbb{R}^d$ and for all $n \geq 0$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(\theta_n, Z_{n+1}) \ .$$

Here, (γ_n) is the learning rate (a deterministic step-size sequence), (Z_n) is an i.i.d sample of observations, and H is a noisy version of h. Denoting $\mathcal{F}_n := \sigma(Z_1, \dots Z_n)$, H is such that

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

Classical conditions for the the choice of the step sizes (γ_n) are

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

These conditions ensure the convergence of the estimates under weak assumptions. For example, convergence in mean square is studied in [23], almost sure consistency is considered in [7, 28], asymptotic rate of convergence are given in [13, 24, 25], while large deviations principles are investigated in [31]. There has been a recent interest on non-asymptotic results. Risk bounds under Gaussian concentration assumption (see [14]) and finite time bounds on the mean square error under strong convexity assumptions (see [21, 28] and references therein), have been given. Quantile estimation corresponds to the choice $h : t \mapsto F(t) - \alpha$, where F is the cumulative distribution function of the target distribution. One can show that the estimator

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

is consistent and asymptotically Gaussian (see [12] chapters 1 and 2 for proofs and details). It is important to remind, however, that the lack of strong convexity prevents most non-asymptotic results to be applied directly, except when the density is lower-bounded. We nevertheless mention that Godichon et al. prove in [8, 16] such non-asymptotic results for the adaptation of algorithm (2) to the case where Z is a random variable on an Hilbert space of dimension higher than 2.

Of course, unless x can take a finite but small number of different values, it is not possible to use this algorithm with a sample of Y^x for each x. Even more, when the code has a high computational complexity, the overall number of observations (all values of x included) must remain small, and we need an algorithm using only one limited sample $(X_i, Y_i)_{i=1\dots n}$ of (X, Y). Then, the problem is more difficult. For each value of x, we need to estimate quantile of the conditional distribution given x using a *biased* sample. To address this issue, we propose to embed Algorithm (2) into a non-parametric estimation procedure. For a fixed input x, the new algorithm only takes into account the pairs (X_i, Y_i) for which the input X_i is close to x, and thus (presumably) the law of Y_i close to that of Y^x. To set up this idea, we use the k-nearest neighbours method, introducing the sequential estimator:

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

where

- $kNN_n(x)$ is the set of the k_n nearest neighbours of x for the euclidean norm on $\mathbb{R}^d$. Denoting by $||X - x||_{(i,n)}$ the ith statistic order of a sample $(||X_i - x||)_{i=1\dots n}$ of size n, we have

$$\{X_{n+1} \in kNN_{n+1}(x)\} = \{||X_{n+1} - x|| \leq ||X - x||_{(k_{n+1},n)}\} \ .$$

  In this work, we discuss choices of the form $k_n = \lfloor n^\beta \rfloor$ for $0 < \beta < 1$, $n \in \mathbb{N}^*$.
- (γ_n) is the deterministic steps sequence. We also study the case $\gamma_n = n^{-\gamma}$ for $0 < \gamma \leq 1$, $n \in \mathbb{N}^*$.

The k-nearest neighbours method of localization first appears in [29, 30] for the estimation of conditional expectations. In [6], Bhattacharya et al. apply it to the (non-recursive) estimation of the conditional quantile function for real-valued inputs. If the the number of neighbours k_n is small, then few observations are used and the estimation is highly noisy; on the other hand, if k_n is large, then values of Y_i may be used that have a distribution significantly different from the target. The challenge is thus to tune k_n so as to reach an optimal balance between bias and variance.

In this work, this tuning is combined with the choice of the learning rate. The main objective of this work is thus to optimize the two parameters of Algorithm (3), i.e. the step size γ_n and the number of neighbours k_n. The paper is organized as follows: Section 2 deals with the almost sure convergence of the algorithm. Further, it contains the main result of our paper that is a non-asymptotic inequality on the mean square error from which an optimal choice of parameters is derived. 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

We explain here how to tune the parameters of the algorithm. We also provide conditions allowing theoretical guarantees of convergence. Before that, we start by some notation and technical assumptions.

2.1. Notation and assumptions.

The constants appearing in the sequel are of three different types:

1) (L, U) denote lower- and upper bounds for the support of random variables. They are indexed by the names of those variables;
2) $(N_i)_{i\in\mathbb{N}^*}$ are integers denoting the first ranks after which some properties hold;
3) $(C_i)_{i\in\mathbb{N}^*}$ are positive real numbers used for other purposes.

Without further precision, constants of type 2) and 3) only depend 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 $u \in \mathcal{P}(\{\alpha, x, d\})$ (the power set of a $\{\alpha, x, d\}$) constant depending on the model, on

the probability level α, on the point x and on the dimension d. The values of all the constants are summarized in the Appendix.

For any random variable Z, we denote by F_Z its cumulative distribution function. We denote by $\mathcal{B}_x$ the set of the balls of $\mathbb{R}^d$ centred at x. For $B \in \mathcal{B}_x$, we denote by r_B its radius and for $r_B > 0$, we call Y^B a random variable with distribution $\mathcal{L}(Y|X \in B)$.

Remark 2.1. *If the pair (X, Y) has a density $f_{(X,Y)}$ and if the marginal density $f_X(x)$ is positive, then we can compute the density of $\mathcal{L}(Y|X = x)$ by*

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

and when $B = \{x\}$,

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

We will make four assumptions. The first one is hardly avoidable, since we deal with k-nearest neighbours. The three others are more technical.

Assumption A1 For all x in the support of X (that we will denote $\mathrm{Supp}(X)$ 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 words, we assume that the stochastic code is sufficiently smooth. The law of two responses corresponding to two different but close inputs are not completely different. The assumption is clearly required, since we want to approximate the law $\mathcal{L}(Y^x)$ by the law $\mathcal{L}(Y|X \in kNN_n(x))$.

Remark 2.2. *If we consider random vector supported by $\mathbb{R}^d \times \mathbb{R}$, we can show that Assumption **A1** holds, for example, as soon as (X, Y) had a regular density. In all cases, it is easier to prove this assumption when the couple (X, Y) has a density. See Subsection 3.1 for an example.*

Assumption A2 The law of X has a density and this density is lower-bounded by a constant $C_{input} > 0$ on $\mathrm{Supp}(X)$.

This hypothesis implies in particular that the law of X has a compact support. Notice that this kind of assumptions is usual in k-nearest neighbours context (see for example [15]).

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

Under Assumption **A3** and if $\beta \geq \gamma$, then

$$\sqrt{C_1} := \max\left(U_Y - L_Y + (1 - \alpha), U_Y + \alpha - L_Y\right) = U_Y - L_Y + \alpha ,$$

is a bound of $|\theta_n(x) - \theta^*(x)|$ (see Lemma 5.8 in Appendix).

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

Lemma 2.1. *Denoting* $C_2(x,\alpha) := \min\left(C_g(x), \frac{1-\alpha}{U_Y+\alpha-L_Y},\right)$, *we have thanks to assumption* ***A4****,*

$$\forall n \in \mathbb{N}^*,\ \left[F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right]\left[\theta_n(x)-\theta^*(x)\right] \geq C_2(x,\alpha)\left[\theta_n(x)-\theta^*(x)\right]^2 . \tag{4}$$

Proof. When $\theta_n(x) \in [L_Y, U_Y]$, it is obvious that the inequality (4) holds for $C_2 := C_g(x)$. When $\theta_n(x) \in [L_{\theta_n}, L_Y]$, we have

$$L_{\theta_n} \leq \theta_n(x) \leq L_Y \leq \theta^*(x) ,$$

and then $F_{Y^x}(\theta_n(x)) = 0$. Thus,

$$\begin{aligned}(\theta_n(x)-\theta^*(x))(F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))) &= (\theta_n(x)-\theta^*(x))^2 \frac{(0-\alpha)}{\theta_n(x)-\theta^*} \\ &\geq (\theta_n(x)-\theta^*(x))^2 \frac{-\alpha}{L_Y-(1-\alpha)-U_Y} \\ &\geq C_2(x,\alpha)(\theta_n(x)-\theta^*(x))^2 .\end{aligned}$$

The proof of the last case follows similarly using that $C_2(x,\alpha,d) \geq \frac{1-\alpha}{U_Y+\alpha-L_Y}$. □

This assumption is useful to deal with non-asymptotic inequality for the mean square error. It is the substitute of the strong convexity assumption made in [21] which is not true in the case of the quantile.

2.2. **Almost sure convergence.** The following theorem studies the almost sure convergence of our algorithm.

Theorem 2.1. *Let x be a fixed input. Under Assumptions* ***A1*** *and* ***A2****, Algorithm (3) is almost surely convergent whenever* $\frac{1}{2} < \gamma \leq \beta < 1$.

Sketch of proof : In the sequel, we still denote $\mathcal{F}_n := \sigma(X_1,\dots,X_n,Y_1,\dots,Y_n)$ and $\mathbb{E}_n$ and $\mathbb{P}_n$ the conditional expectation and probability given $\mathcal{F}_n$. For sake of simplicity, we denote

$$H(\theta_n(x), X_{n+1}, Y_{n+1}) := \left(\mathbf{1}_{Y_{n+1}\leq\theta_n(x)} - \alpha\right)\mathbf{1}_{X_{n+1}\in kNN_{n+1}(x)} .$$

The proof is organized in three steps.

1) We decompose $H(\theta_n(x), X_{n+1}, Y_{n+1})$ as a sum of a drift and a martingale increment :

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

Then,

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

is a martingale which is bounded in L^2. So it converges almost surely.

2) We show the almost sure convergence of $(\theta_n)_n$:

a) First, we check that (θ_n) does not diverge to $+\infty$ or $-\infty$.
b) Then, we prove that (θ_n) converges almost surely to a finite limit.

3) We conclude by identifying the limit : $\theta^*(x)$, 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 ensured by Assumptions **A1** and **A2**. The entire proof is available in Section 5.

Comments on parameters. In the Theorem 2.1, we assume that $0 < \beta < 1$. $\beta > 0$ means that the number of neighbours goes to $+\infty$ and then, $||X - x||_{(k_n,n)} \to 0$. The condition $\beta < 1$ allows to apply the Lemma 5.4. It is a technical condition. The condition $\beta \geq \gamma$ can be understood in this way. When considering Algorithm (2), we deal with the *global learning rate* $\gamma_n = n^{-\gamma}$. In Algorithm (3), since for a fixed input x, there is not an update at each step n, we introduce the *effective learning rate* γ_{k_n} in the following way. At step k, $\theta_k(x)$ has a probability of k^β/k to be updated. Then, until time n, the algorithm is updated a number of times equal to

$$N = \sum_{k \leq n} k^{\beta-1} \sim n^\beta \ .$$

Thus, there were $N = n$ updates at time $n^{\frac{1}{\beta}}$. Then, in mean, it is as if the algorithm was defined by

$$\theta_{k_n}(x) = \theta_{k_n-1}(x) + \gamma_{k_n}\left(\mathbf{1}_{Y_{k_n} \geq \theta_{k_n}(x)} - \alpha\right) \ ,$$

with the learning rate

$$\gamma_{k_n} = \frac{1}{\left(n^{\frac{1}{\beta}}\right)^\gamma} = \frac{1}{n^{\frac{\gamma}{\beta}}} \ .$$

This is a well-known fact that this algorithm has a *good* behaviour if, and only if, the sum

$$\sum_n \gamma_{k_n} = \sum_n \frac{1}{n^{\frac{\gamma}{\beta}}} \ ,$$

is divergent. That is if, and only if $\beta > \gamma$. At last, the condition $\frac{1}{2} < \gamma < 1$ is a classical assumption on the Robbins Monro algorithm to be consistent (see for example in [23]). Here, we restrict the condition to $\gamma < 1$ because we need $1 > \beta \geq \gamma$.

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

Theorem 2.2. *Under hypothesis **A1**, **A2**, **A3** and **A4**, the mean square error $a_n(x)$ of the algorithm (3) satisfies the following inequality : $\forall(\gamma, \beta, \epsilon)$ such that $0 < \gamma \leq \beta < 1$ and $1 > \epsilon > 1 - \beta$, $\forall n \geq N_0 + 1$ where $N_0 = 2^{\frac{1}{\epsilon-(1-\beta)}}$,*

$$a_n(x) \leq \exp\left(-2C_2(x,\alpha)(\kappa_n - \kappa_{N_0})\right) C_1 + \sum_{k=N_0+1}^{n} \exp\left(-2C_2(x,\alpha)\left(\kappa_n - \kappa_k\right)\right) d_k$$
$$+ C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) \ ,$$

where for $j \in \mathbb{N}^*$, $\kappa_j = \sum_{i=1}^{j} i^{-\epsilon-\gamma}$ *and*

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

Sketch of proof : The idea of the proof is to establish the recursive inequality on $a_n(x)$ (following [21]), that is for $n \geq N_0$,

$$a_{n+1}(x) \leq a_n(x)(1 - c_{n+1}) + d_{n+1}$$

where for all $n \in \mathbb{N}^*$, $0 < c_n < 1$ and $d_n > 0$. We conclude by using Lemma 5.7. In this purpose we begin by expanding the square

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

Taking the expectation conditionally to $\mathcal{F}_n$, and using the Bayes formula, we get

$$\begin{aligned}\mathbb{E}_n\left((\theta_{n+1}(x) - \theta^*(x))^2\right) \leq \mathbb{E}_n\left((\theta_n(x) - \theta^*(x))^2\right) + \gamma_{n+1}^2 P_n \\ - 2\gamma_{n+1}\left(\theta_n(x) - \theta^*(x)\right) P_n \left[F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] ,\end{aligned} \tag{5}$$

where $P_n = \mathbb{P}_n\left(X_{n+1} \in kNN_{n+1}(x)\right)$ as in Lemma 5.1 and $B_n^{k_{n+1}}(x)$ is the ball of $\mathbb{R}^d$ centred in x and of radius $||X - x||_{(k_{n+1},n)}$. We rewrite this inequality to make appear two different errors :

1) First, the quantity $F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))$ represents the *bias error* (made because the sample is biased). Using **A1**, we get

$$|F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))| \leq M(x)||X - x||_{(k_{n+1},n)} .$$

and by **A3**, $|\theta_n(x) - \theta^*(x)| \leq \sqrt{C_1}$. Thus,

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

2) The second quantity, $F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))$ represents the *on-line learning error* (made by using a stochastic algorithm). 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 the inequality (5),we get

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

This inequality reveals a problem : thanks to Lemmas 5.1 and 5.6 (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 truncated parameter ϵ_n. Instead of writing a recursive inequality on $a_n(x)$ we write such inequality with the quantity $b_n(x) := \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2 \mathbf{1}_{P_n>\epsilon_n}\right]$. Choosing $\epsilon_n = (n+1)^{-\epsilon}$, we have to tune an other parameter but thanks to **A3** and concentration inequalities (see lemma 5.4), it is easy to deduce a recursive inequality on $a_n(x)$ from the one on $b_n(x)$, for $n \geq N_0$.

Comments on the parameters. We choose $0<\beta<1$ for the same reasons as in Theorem 2.1. About γ, the inequality is true on the entire area $0<\gamma<1$ as soon as $\gamma \leq \beta$ (which is unusual, as you can see in [16] for example). We will nevertheless see in the sequel that this is not because the inequality is true that $a_n(x)$ converges to 0. We will discuss later *good* choices for (γ,β).

Compromise between the two errors. 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
$$\exp\left(-2C_2(x,\alpha)(x)\sum_{k=N_0+1}^{n}\frac{1}{k^{\epsilon+\gamma}}\right),$$
of the inequality. This term decreases to 0 if and only if $\gamma+\epsilon<1$ which implies $\beta>\gamma$. Then β has to be chosen not too small.
- The *on-line learning error* gives the term $(k_n/n)^{1/d+1} = n^{(1-\beta)(1+1/d)}$ in the remainder. For the remainder to decrease to 0 with the faster rate, we then need that β is as small as possible compared to 1. Then β has to be chosen not 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 d_n in n to exhibit dominating terms. d_n is the sum of three terms. The exponential one is always negligence as soon as n is big enough because $1>\epsilon$. The two other are powers of n. Comparing their exponent, we can find the dominating term in function of γ and β. Actually, there exists a rank $N_1(x,d)$ and some constants C_5 and $C_6(x,d)$ such that, for $n \geq N_0+1$,

if $\beta \leq 1-d\gamma$, we get
$$d_n \leq C_5 n^{-2\gamma+\beta-1}.$$
if $\beta > 1-d\gamma$, we get
$$d_n \leq C_6(x,d) n^{-\gamma+(1+\frac{1}{d})(\beta-1)}.$$

Copying that in the Theorem 2.2, we deduce the following result.

Corollary 2.1. *Under assumptions of Theorem 2.2, there exists ranks $N_4(x,\alpha,d)$ and constants $C_7(x,\alpha,d)$ and $C_8(x,\alpha)$ such that $\forall n \geq N_4(x,\alpha,d)$,*

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

When $\beta \leq 1 - d\gamma$, and $1 - \epsilon < \min(1 - \beta + \gamma, 1 - \gamma)$,

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

Remark 2.3. *For other values of γ and β, the derived inequalities do not imply the convergence to 0 of $a_n(x)$ this is why we do not present them.*

From this corollary we can derive *optimal* choices for (β, γ), that is parameters for which our upper-bound on the mean square error decreases with the fastest rate.

Corollary 2.2. *Under the same assumptions than in Theorem 2.2, the optimal parameters are $\gamma = \frac{1}{1+d}$ and $\beta = \gamma + \eta_\beta$ where $\eta_\beta > 0$ is as small as possible. With these parameters, there exists a constant $C_9(x, \alpha, d)$ such that $\forall n \geq N_4(x, \alpha, d)$,*

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

where $\eta = \frac{\eta_\epsilon}{2} + \eta_\beta$ and $\eta_\epsilon = 1 - \beta - \epsilon$.

Comments on the constant $\mathbf{C_9(x, \alpha, d)}$. Like all the other constants of this paper, we know the explicit expression of $C_9(x, \alpha, d)$. An example of values of this constant is given in Subsection 3.1.

We can notice that the constant $C_9(x, \alpha, d)$ depends on x only through $C_g(x)$ and $M(x)$. Nevertheless, often in practice, $C_g(x)$ and $M(x)$ do not really depend 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 do not depend on x), our result is uniform in x. Then, it is easy to deal with 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. The bound gets worst. We can easily understand this phenomenon because when α is big, we have a small probability to sample on the right of the quantile, and the algorithm is then less powerful.

Let us now comment the dependency on the dimension d. The constant $C_9(x, d, alpha)$ decreases when the dimension d increases. Nevertheless, this decreasing is too small to balance the behaviour of the rate of convergence which is in $n^{\frac{-1}{1+d}}$. This is an example of the curse of dimensionality.

Comment on the rank $\mathbf{N_4(x, \alpha, d)}$. This rank is the maximum of four ranks. There are two kinds of ranks. The ranks $(N_i)_{i \neq 0}$ depend on constants of the problem but are reasonably small, because the largest of them is the rank after which exponential terms are smaller than power of n terms, or smaller power of n terms are smaller than bigger power of n terms. They are then often inferior to N_0 in practice. We only need this rank to find optimal parameters (and at this stage our reasoning is no more non-asymptotic).

The rank N_0 is completely different. It was introduced in the first theorem because we could not deal with $a_n(x)$ directly. In fact it is the rank after which the deviation

inequality, allowing us to use $b_n(x)$, is true. It depends on the gap between ϵ and $1-\beta$. The optimal ϵ to obtain the rate of convergence of the previous corollary is $\epsilon = 1-\beta+\eta_\epsilon$ with η_ϵ as small as possible. The constant η_ϵ appears on the rank N_0 and also on the rate of convergence (let us suppose that $N_4 = N_0$ which is the case most of time)

$$\forall n \geq N_0 = \exp\left(2\eta_\epsilon^{-1}\right),\ a_n(x) = \mathcal{O}\left(n^{\frac{-1}{1+d}+\frac{\eta_\epsilon}{2}+\eta_\beta}\right) .$$

Then the smaller is η_ϵ, the faster is the rate of convergence, but also the larger is the rank after which inequalities are true.

Let us give an example. For a budget of $N = 1000$ calls to the code, one may choose $\eta_\epsilon = 0.3$ for the inequality to be theoretically true for $n = N$. The Table 1 gives the theoretical precision for different values of d and compares it with the ideal case where $\eta_\epsilon = 0$.

d	1	2	3
η_ϵ=0.3	0.088	0.28	0.5
η_ϵ=0	0.031	0.1	0.17

TABLE 1. Expected precision for the MSE when $N = 1000$

We can observe that, when $\eta_\epsilon > 0$, the precision increases with the dimension faster than when η_ϵ. Moreover, as soon as $\frac{1}{1+d} < \eta_\epsilon/2$ ($d = 6$ for our previous example), the result does not allow to conclude that a_n decreases to 0 with this choice of η_ϵ.

Nonetheless, simulations (see next part) seem 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 exist when we implement the algorithm). In practice, the optimal rate of convergence for optimal parameters is reached early (see Section 3).

3. Numerical simulations

In this part we present some numerical simulations to illustrate our results. We consider simplistic examples so as to be able to evaluate clearly the strengths ans the weaknesses of our algorithm. 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

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

where $X \sim \mathcal{U}([0,1])$ and $\epsilon \sim \mathcal{U}([-0.5, 0.5])$. We try to estimate the quantile of level $\alpha = 0.95$ for $x = 0.5$ and initialize our algorithm to $\theta_1 = 0.3$. Let us check that our assumptions are fulfilled in this case. We have $\mathcal{L}(g(x,\epsilon)) = \mathcal{U}([-\frac{1}{2}+x^2; \frac{1}{2}+x^2])$. Then

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

Moreover, the code function g takes its values in the compact set $[L_Y, U_Y] = [-\frac{1}{2}; \frac{3}{2}]$. Let us study assumption **A1**. Let B be an interval containing x, denoted $B = [x-a, x+b]$

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

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

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

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

There are again two different 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 $(t + 1/2)^{1/2}$ relative to $(x - a)$ is not always the same. Then, if $t \in [-\frac{1}{2}; -\frac{1}{2}(x-a)^2]$, we get

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

as $0 < a < 1$. Finally, in this case, **A1** is true with $M(x) = 2/3$. We can compute exactly in the same way for the other cases and we always find an $M(x) \leq 2/3$. The assumption **A2** is also satisfied, taking $C_{input} = 1$. We have already explained that assumption **A3** is true for $[L_Y, U_Y] = [-1/2, 3/2]$. Finally assumption **A4** is also satisfied with $C_g(x) = 1$ and $C_2(x, \alpha) = 0.02$.

3.1.1. *Almost sure convergence.* Let us first deal with the almost sure convergence. We plot in Figure 1, for $(\beta, \gamma) \in [0,1]^2$, the relative error of the algorithm. Best parameters are clearly in the area $\beta > \gamma \geq 1/2$. We can even observe that for $\beta \approx 1$, $\beta \leq \gamma$ or $\gamma < 1/2$, the algorithm does not converge almost surely (or very slowly). This is in accordance with our theoretical results. Nevertheless, we can observe a kind of

continuity for γ around $1/2$: in practice, the convergence becomes really slow only when γ is significantly far away from $1/2$.

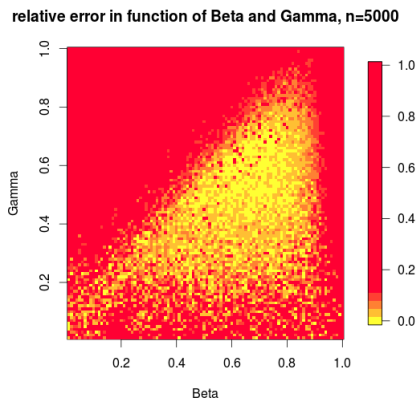

FIGURE 1. Relative error for $n = 5000$ in fonction of β and γ

3.1.2. *Mean Square Error (MSE).* Let us study the best choice of β et γ in terms of L^2-convergence. We plot in Figures 2, the mean square error in function of γ and β (we estimate the MSE by a Monte Carlo method of 100 iterations).

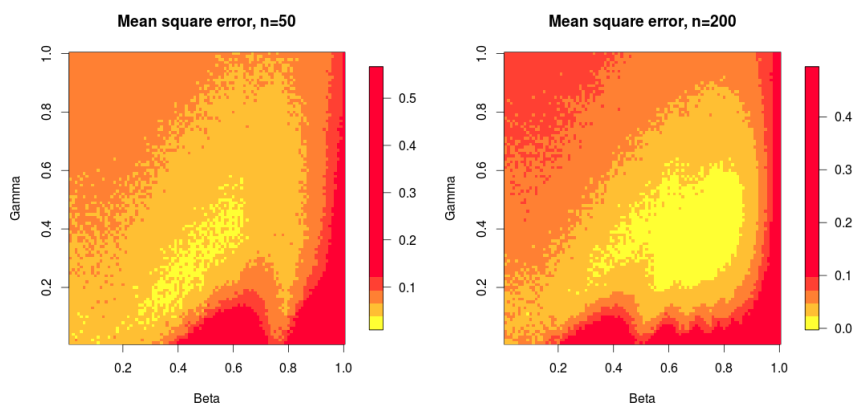

FIGURE 2. Mean square error in function of β and γ for the square function

Simulations confirm that the theoretical optimal area $\gamma = 0.5$ and $\beta = \gamma + \eta_\beta$ gives the smallest MSE. Nevertheless, it seems that in practice we can relax the condition that *the gap η_β between β and γ is as small as possible.* Indeed, when η_β is reasonably big, simulations show that we are still in the optimal area.

3.1.3. *Theoretical bound.* In this case, we have at hand all the parameters to compute the theoretical bound of our theorems. In particular, in corollary 2.2, we get

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

Table 2 summarizes the value of the constants needed to compute the theoretical bound in this case.

Constant	α	$M(x)$	C_{input}	$C_g(x)$	$C_2(x,\alpha)$	$U_Y - L_Y$
Value	0.95	$\frac{2}{3}$	1	1	0.02	2
Constant	$\sqrt{C_1}$	$C_3(d)$	$C_4(d)$	$C_5(x,d)$	$C_6(x,d)$	$C_9(x,d,\alpha)$
Value	2.95	7.39	2	1.95	12	180

TABLE 2. Constant values

For $N = 1000$, we obtain the bound $a_N(x) \leq 5.8$ which is over-pessimistic compared to the practical results. We can then 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 $[L_Y, U_Y]$. This happens only with a very small probability. If we do not take this case into account, we have $C_2(x,\alpha) = 1$. Then $C_9(x,\alpha,d) \approx 3.7$ and then, for $N = 1000$, the bound is 0.11. Practical results are still better (we can observe that for $n = 50$ only, we have a MSE inferior to 0.05 !), but the gap is less important.

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 for which the differentiability fails). Assumptions can be checked as above. Since the almost surely convergence is true and gives really same kind of plots than the previous case, we only study the convergence of the MSE. In that purpose, we plot in Figure 3 the MSE (estimated by 100 iterations of Monte Carlo simulations) in function of γ and β, for n=300 (the discontinuity constraints us to make more iterations to have a sufficient precision) and $\theta_1 = 0.3$. Conclusions are the same than in the previous example concerning the best parameters. Nevertheless, we can observe that the lack of smoothness implies some strange behaviour around $\gamma = 1$.

3.3. **Dimensions 2 and 3.** In dimension d, we showed 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.

3.3.1. *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 choose $n = 400$ and want to study the quantile in the input point $x = (0,0)$ and initialize our algorithm in $\theta_1 = 0.3$. In Figure 4, we can see that $\beta = 1$ and $\gamma = 1$ are still really

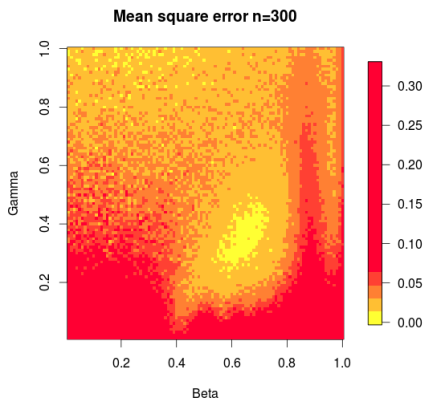

FIGURE 3. MSE in function of β and γ for absolut value function

bad parameters. As in our theoretical results, $\gamma = \frac{1}{1+d} = \frac{1}{3}$ 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 $n = 400$ iterations, in this case, the MSE, in the optimal parameters case reaches 0.06.

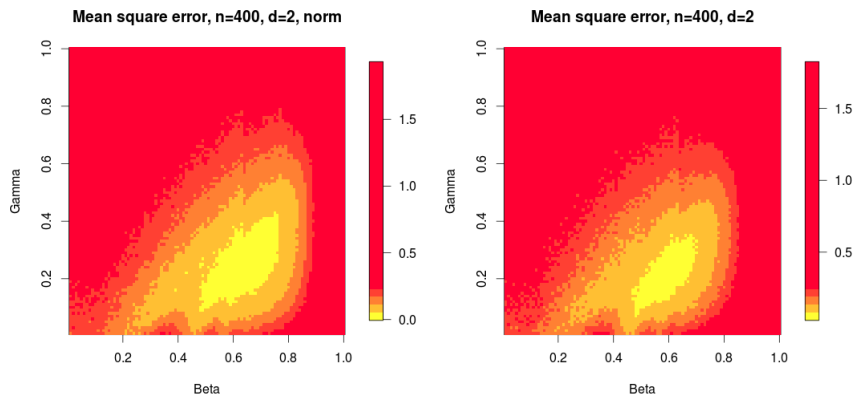

FIGURE 4. Mean square error in function of β and γ

3.3.2. *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 5 are the same than in dimension 2. The scale is not the same, the convergence is slower again but with $n = 500$ we nevertheless obtain a MSE of 0.10.

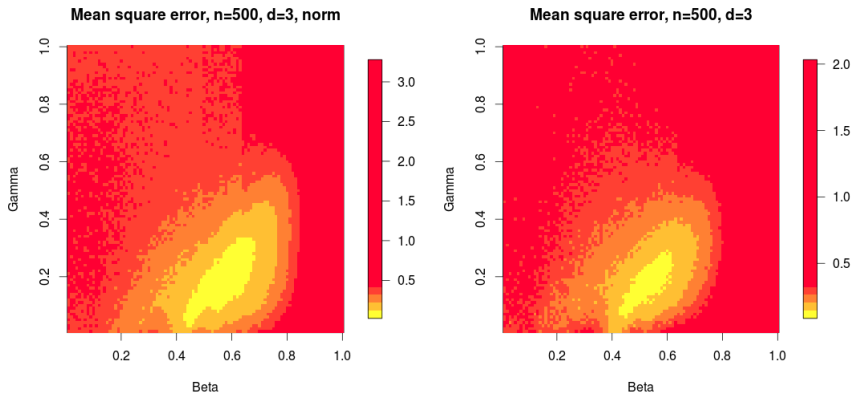

FIGURE 5. Mean square error in function of β and γ

4. Conclusion and perspectives

In this paper, we proposed a sequential method for the estimation of a conditional quantile of the output of a stochastic code where inputs lie in $\mathbb{R}^d$. We introduced a combination of k-nearest neighbours and Robins-Monro estimator. The algorithm thus elaborated had then two parameters to tune : the number of neighbours $k_n = \lfloor n^\beta$ and the learning rate $\gamma_n = n^{-\gamma}$. Obtaining a bias-variance decomposition of the risk, we showed that our algorithm is convergent for $\frac{1}{2} < \gamma < \beta < 1$ and we studied its mean square error non-asymptotic rate of convergence. Moreover, we proved that we have to choose $\gamma = \frac{1}{1+d}$ and $\beta = \gamma + \eta_\beta$ ($\eta_\beta > 0$) to get the best rate of convergence. Numerical simulations have showed 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. Relaxing the conditions will be the object of a future work. Moreover, the proof that we propose constrained us to use an artefact parameter ϵ which implies that the non-asymptotic inequality is theoretically true for big n, even if simulations confirm that this problem does not exist in practice. A second perspective is then to find a better way to prove this inequality for smaller n.

Finally, it is a very interesting future work to write non-asymptotic lower-bound for the mean square error of our algorithms.

5. Appendix 1 : Technical lemmas and proofs

5.1. **Technical lemmas and notation.** For sake of completeness, we start by recall and prove some well-known facts on order statistics.

Lemma 5.1. *When X has a density, denoting $P_n = \mathbb{P}(X \in kNN_{n+1}(x)|X_1, \dots X_n)$, we have the following properties*

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

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

Proof. Conditionally to $X_1, \dots, X_n$, the event $\{X \in kNN_{n+1}(x)\}$ is equivalent to the event $\{||X-x|| \leq ||X-x||_{(k_{n+1},n)}\}$. Then,

$$\begin{aligned} P_n &= \mathbb{P}(X \in kNN_{n+1}(x) | X_1 \dots X_n) \\ &= \mathbb{P}_X\left(||X-x|| \leq ||X-x||_{(k_{n+1},n)} | X_1 \dots X_n\right) \\ &= F_{||X-x||}\left(||X-x||_{(k_{n+1},n)}\right) . \end{aligned}$$

Since X has a density, the cumulative distribution function $F_{||X-x||}$ is continuous. Indeed, using the sequential characterization 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 [9]), we get

$$P_n = F_{||X-x||}\left(||X-x||_{(k_{n+1},n)}\right) \sim U_{(k_{n+1},n)} \sim \beta(k_{n+1}, n-k_{n+1}+1) ,$$

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

Let us know recall some deviation results.

Lemma 5.2. *We denote $\mathcal{B}(n,p)$ the binomial distribution of parameters n and p, for $n \geq 1$ and $p \in [0,1]$. Then, if $Z \sim \mathcal{B}(n,p)$, we get*

$$\begin{aligned} \mathbb{P}\left(\frac{Z}{n} < \frac{p}{2}\right) &\leq \exp\left(-\frac{3np}{32}\right) , \\ \mathbb{P}\left(\frac{Z}{n} > 2p\right) &\leq \exp\left(-\frac{3np}{8}\right) . \end{aligned}$$

Proof. Let (Z_i) be an independent sample of Bernoulli of parameter p and let

$$Z = \frac{1}{n} \sum_{k=1}^{n} Z_i .$$

We apply the Bernstein's inequality (see Theorem 8.2 of [11]) 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. □

We now give some technical lemma useful to prove our main results.

Lemma 5.3. *Suppose $\beta\geq\gamma$. Then, for $C\geq 3$, we get*

$$\mathbb{P}\left(\sum_n \gamma_n \mathbf{1}_{X_n\in kNN_n(x)}\leq C\right)=0 .$$

Proof. First, it is a well known result (see [9]) that if $U\sim\mathcal{U}([0,1])$, then $X\overset{\mathcal{L}}{=}F^{-1}(U)$. Since F is non-decreasing, we get

$$\mathbf{1}_{U_n\in kNN_n(x)}=\mathbf{1}_{F^{-1}(U_n)\in kNN_n(F(x))}\ a.s.$$

So that, it is enough to show the result for $X\sim\mathcal{U}\left([0,1]\right)$.

Let x be a real number in $[0,1]$. Let ϵ be a positive real number. Let n_0 be an integer such that

$$\sum_{n\geq n_0}\exp\left(-\frac{3k_n}{16}\right)\leq\epsilon . \tag{6}$$

Let n_1^x be the integer such that if $x\in\{0,1\}$, $n_1^x=1$ and if $x\in]0,1[$, for $n\geq n_1^x$,

$$\begin{cases}\dfrac{k_n}{2n}+x\leq 1 ,\\ x-\dfrac{k_n}{2n}\geq 0 .\end{cases}$$

We denote $N:=\max(n_0,n_1^x)$. We set

$$\Omega:=\left\{\forall n\geq N,\ \sum_{j=1}^n \mathbf{1}_{|X_j-x|\leq\frac{k_n}{4n}}\leq k_n\right\} .$$

On this event, for every $n\geq N$, there are at most k_n elements X_i such that $|X_i-x|$ is inferior to $\frac{k_n}{4n}$. Thus, if an element satisfies $|X_j-x|\leq\frac{k_n}{4n}$, it belongs to the k_n-nearest neighbours of x. Then,

$$
\begin{aligned}
\mathbb{P}\left(\overline{\Omega}\right) &\leq \sum_{n\geq N} \mathbb{P}\left(\sum_{j=1}^{n} \mathbf{1}_{|X_j - x| \leq \frac{k_n}{4n}} > k_n\right) \\
&=: \sum_{n\geq N} \mathbb{P}\left(Z_n > k_n\right) \\
&= \sum_{n\geq N} \mathbb{P}\left(\frac{\mathcal{B}(n,p)}{n} > \frac{k_n}{n}\right) .
\end{aligned} \tag{7}
$$

where, since $n \geq n_1^x$,

$$
\begin{aligned}
p &= \mathbb{P}\left(|X - x| \leq \frac{k_n}{4n}\right) \\
&= \begin{cases} \mathbb{P}\left(-\frac{k_n}{n} + x \leq X \leq \frac{k_n}{4n} + x\right) & \text{if } x \in]0,1[\\ \mathbb{P}\left(X \leq \frac{k_n}{4n}\right) & \text{if } x = 0 \\ \mathbb{P}\left(X \leq 1 - \frac{k_n}{4n}\right) & \text{if } x = 1 \end{cases} \\
&= \begin{cases} \frac{k_n}{2n} & \text{if } x \in]0,1[\\ \frac{k_n}{4n} & \text{otherwise} \end{cases} \\
&\leq \frac{k_n}{2n} .
\end{aligned}
$$

Then, Equation (7) gives

$$
\begin{aligned}
\mathbb{P}\left(\overline{\Omega}\right) &\leq \sum_{n\geq N} \mathbb{P}\left(\sum_{j=1}^{n} \mathbf{1}_{|X_j - x| \leq \frac{k_n}{4n}} > k_n\right) \\
&\leq \mathbb{P}\left(\frac{\mathcal{B}\left(n, \frac{k_n}{2n}\right)}{n} > \frac{k_n}{n}\right) \\
&\leq \exp\left(-\frac{3k_n}{16}\right) \leq \epsilon .
\end{aligned} \tag{8}
$$

where we used the second inequality of Lemma 5.2 and the Equation (6). But, as we noticed above, on the event Ω, we have

$$
\mathbf{1}_{X_n \in kNN_n(x)} \geq \mathbf{1}_{|X_n - x| \leq \frac{k_n}{4n}} .
$$

Finally,

$$\mathbb{P}\left(\Omega \cap \sum_{n\geq N} \gamma_n \mathbf{1}_{X_n \in kNN_n(x)} \leq C\right) \leq P\left(\sum_{n\geq N} \gamma_n \mathbf{1}_{|X_n - x| \leq \frac{k_n}{4n}} \leq C\right) . \tag{9}$$

Let now $(I_k)_k$ be a partition of $[|N, +\infty|]$ such that

$$\forall k \geq 1, \sum_{n\in I_k} \gamma_n \frac{k_n}{4n} \in [2C, 2C+1] .$$

Such a partition exists since, as $\beta \geq \gamma$, the sum $\sum_n \gamma_n \frac{k_n}{n}$ is divergent. Then,

$$\mathrm{Var}\left[\sum_{n\in I_k} \gamma_n \mathbf{1}_{|X_k - x| \leq \frac{k_n}{4n}}\right] \leq \mathbb{E}\left[\sum_{n\in I_k} \gamma_n \mathbf{1}_{|X_k - x| \leq \frac{k_n}{4n}}\right] \leq 2C+1 .$$

The Chebyshev's inequality gives

$$\mathbb{P}\left(\sum_{n\in I_k} \gamma_n \mathbf{1}_{|X_k - x| \leq \frac{k_n}{4n}} \leq C\right) \leq \frac{2C+1}{C^2} \leq \frac{7}{9} < 1 ,$$

since $C \geq 3$.

$$\mathbb{P}\left(\bigcap_k \left\{\sum_{n\in I_k} \gamma_n \mathbf{1}_{|X_n - x| \leq \frac{k_n}{4n}}\right\} \leq C\right) = 0 .$$

$$\mathbb{P}\left(\sum_{n\geq N} \gamma_n \mathbf{1}_{|X_n - x| \leq \frac{k_n}{4n}} \leq C\right) = 0 . \tag{10}$$

Thanks to (7), (9) and (10), we get

$$\mathbb{P}\left(\sum_{n\geq N} \gamma_n \mathbf{1}_{X_n \in kNN_n(x)} \leq C\right) \leq \mathbb{P}(\overline{\Omega}) + 0 \leq \epsilon ,$$

which holds for all $\epsilon > 0$. □

Lemma 5.4. *Denoting A_n the event $\{X_1, \dots X_n \mid P_n > \epsilon_n\}$ where $\epsilon_n = \frac{1}{(n+1)^\epsilon}$ and the parameter ϵ satisfies $1 > \epsilon > 1 - \beta$, we have for $n \geq 1$,*

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

Proof. Thanks to the Lemma 5.1, we obtain

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

where we denote I_ϵ the incomplete β function. A classical result (see [1]) allows us to write this quantity in terms of the binomial distribution

$$\mathbb{P}(A_n^C) = \mathbb{P}(\mathcal{B}(n, \epsilon_n) \geq k_{n+1}) \ .$$

Thanks to Lemma 5.2, we know that

$$\mathbb{P}(\mathcal{B}(n+1, \epsilon_n) \geq k_{n+1}) \leq \exp\left(-\frac{3(n+1)\epsilon_{n+1}}{8}\right) \leq \exp\left(-\frac{3(n+1)^{1-\epsilon}}{8}\right) \ ,$$

as soon as $k_{n+1}/(n+1) \geq 2\epsilon_n$, which is true as soon as $n \geq 2^{1/(\epsilon-(1-\beta))}$ because $\epsilon > 1-\beta$. □

Lemma 5.5. *Under hypothesis of Theorem 2.1, $||X-x||_{(k_{n+1},n)}$ converges almost surely to 0.*

Proof. Let u be a positive number.

$$\begin{aligned} p_u := \mathbb{P}(X \in \mathcal{B}(x,u)) &= \int_{\mathcal{B}(x,u)} f(t)dt \\ &\geq \mu_X\left(\mathcal{B}(x,u)\right) = C_1 \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \\ &= C_{input} C_4(d) u^d =: q_u \ . \end{aligned} \tag{11}$$

Let Z be a random variable of law $\mathcal{B}(n, p_u)$. Since $||X-x||_{(k_{n+1},n)} > u$ implies that there are at the most k_{n+1} elements of the sample which satisfy $X \in \mathcal{B}(x, q_u)$, we get :

$$\mathbb{P}(||X-x||_{(k_{n+1})} > u) = \mathbb{P}(Z < k_{n+1}) \ .$$

Thanks to equation (11), and denoting $\tilde{Z}$ a random variable of law $\mathcal{B}(n, q_u)$, we have

$$\mathbb{P}(||X-x||_{(k_{n+1})} > u) \leq \mathbb{P}(\tilde{Z} < k_{n+1}) \ .$$

Lemma 5.2 implies that $\mathbb{P}(||X-x||_{(k_{n+1})} > u)$ is the general term of a convergent sum. Indeed, when n is large enough, then $k_{n+1}/n < q_u/2$ because k_{n+1}/n converges to 0 ($\beta < 1$). The Borel-Cantelli Lemma then implies that $||X-x||_{(k_{n+1},n)}$ converges almost surely to 0. □

Lemma 5.6. *With the same notation as above,*

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

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

$$\begin{aligned} \mathbb{E}(||X-x||_{(k_{n+1},n)} P_n) &= \mathbb{E}\left(||X-x||_{(k_{n+1},n)} \tilde{F}\left(||X-x||_{(k_{n+1},n)}\right)\right) \\ &= \int y \tilde{F}(y) f_{||X-x||_{(k_{n+1},n)}}(y) dy \ , \end{aligned}$$

with

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

Then we get

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

We denote $U_{|.|}$ the upper bound of the support of $||X-x||$, and write

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

Using same arguments that in Lemma 2.1, denoting $C_{10}(d) = \sqrt[d]{\frac{2(k_{n+1}+1)}{(n+1)C_{input}C_4(d)}}$,we get

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

where we use Lemma 5.2 in the second integral because $u > C_{10}(d)$ implies $\frac{k_{n+1}+1}{n+1} < \frac{q_u}{2}$. Then, we obtain

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

because for $n \geq 1$, we get $k_n \geq 1$. □

Lemma 5.7. *Let (b_n) be a a real sequence. If there exist sequences $(c_n)_{n\geq 1} \in [0,1]^{\mathbb{N}}$ and $(d_n)_{n\geq 1} \in]0,+\infty[^{\mathbb{N}}$ such that*

$$\forall n \geq N_0,\ b_{n+1} \leq b_n(1 - c_{n+1}) + d_{n+1} \ ,$$

then for all $n \geq N_0 + 1$,

$$\forall n,\ b_n \leq \exp\left(-\sum_{k=1}^{n} N_0 + 1c_k\right) b_{N_0} + \sum_{k=N_0+1}^{n} \exp\left(-\left(\sum_{j=1}^{n} c_j - \sum_{j=1}^{k} c_j\right)\right) d_k \ .$$

Proof. This inequality appears in [21] and references therein. It can be proved by induction using that $\forall x \in]0,+\infty[,\ \exp(x) \geq 1 + x$. □

Let us first prove the following consequence of Assumption **A3**.

Lemma 5.8. *Under assumption **A3**, if $\beta \geq \gamma$, then for all x and for all $n \geq 1$,*

$$\theta_n(x) \in [L_Y - (1-\alpha), U_Y + \alpha], a.s.$$

Proof. Suppose that $\theta_n(x)$ leaves the compact set $[L_Y, U_Y]$ by the right at step N_0. By definition, $\theta_{N_0-1} \leq U_Y$ and consequently $\theta_{N_0} \leq U_Y + \alpha\gamma_{N_0}$. At next step, since $\theta_{N_0} > U_Y$, we have $Y_{N_0+1} \leq \theta_{N_0}$ and then

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

Then, the algorithm either does not move (if $X_{N_0+1} \notin kNN_{N_0+1}(x)$) or comes back in direction of $[L_Y, U_Y]$ with a step of $(1-\alpha)\gamma_{N_0+1}$. Then, if

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

the algorithm almost surely comes back to the compact set $[L_Y, U_Y]$. Thanks to Lemma 5.3, we know that, since $\beta \geq \gamma$, the previous sum diverges almost surely. A similar result holds when the algorithm leaves the compact set by the left and finally we have shown that almost surely,

$$\theta_n(x) \in [L_Y - (1-\alpha), U_Y + \alpha] =: [L_{\theta_n}, U_{\theta_n}] \ .$$

□

5.2. **Proof of Theorem 2.1 : almost sure convergence.** To prove this theorem, we adapt the classical analysis of the Robbins-Monro algorithm (see [7]). In the sequel we do not write $\theta_n(x)$ but θ_n to make the notation less cluttered.

5.2.1. *Martingale decomposition.* In this sequel, we still denote $H(\theta_n, X_{n+1}, Y_{n+1}) := \left(\mathbf{1}_{Y_{n+1}\leq \theta_n - \alpha}\right) \mathbf{1}_{X_{n+1}\in kNN_{n+1}(x)}$, $\mathcal{F}_n = \sigma(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ and $\mathbb{P}_n$ and $\mathbb{E}_n$ the probability and expectation conditionally to $\mathcal{F}_n$. We introduce

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

Then,

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

with $\xi_j = H(\theta_{j-1}, X_j, Y_j) - h_{j-1}(\theta_{j-1})$ is a martingale. It is bounded in $\mathbb{L}^2(\mathbb{R})$. Since

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

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

$$\mathbb{E}(|T_n|^2) \leq \mathbb{E}\left(\left(\sum_{j=1}^n \gamma_j \xi_j\right)^2\right) \leq C\mathbb{E}\left(\left|\sum_{j=1}^n (\gamma_j \xi_j)^2\right|^2\right) \leq C(1+2\alpha)\sum_{j=1}^n \gamma_j^2 < \infty \ .$$

5.2.2. *The sequence (θ_n) converges almost surely.* First, let us prove that

$$\mathbb{P}(\theta_n \to \infty) + \mathbb{P}(\theta_n \to -\infty) = 0. \tag{12}$$

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 finite number of n.

Let us show that on an event $\Omega \subset \Omega_1$ with positive measure, for n large enough, $h_n(\theta_n(\omega)) > 0$. First, we know that P_n follows a Beta distribution. This is why $\forall n,\ \mathbb{P}(P_n = 0) = 0$. Then, the Borel-Cantelli Lemma gives that

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

As Ω_1 has a positive measure, we know that there exists $\Omega_2 \subset \Omega_1$ with positive measure such that $\forall \omega \in \Omega_2$, $\theta_n(\omega) \to +\infty$ and for all n large enough, $P_n(\omega) > 0$. Since

$$h_n(\theta_n(\omega)) = P_n\left(F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - \alpha\right)\ ,$$

we have now to show that on $\Omega \subset \Omega_2$ of positive measure,

$$F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - \alpha > 0\ .$$

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

$$\begin{aligned} F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - \alpha &= F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - F_{Y^x}(\theta^*) \\ &= F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - F_{Y^{B_n^{k_{n+1}}(x)}}(D) + F_{Y^{B_n^{k_{n+1}}(x)}}(D) - F_{Y^x}(D) \\ &+ F_{Y^x}(D) - F_{Y^x}(\theta^*)\ . \end{aligned}$$

First, $F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(\omega)) - F_{Y^{B_n^{k_{n+1}}(x)}}(D) \geq 0$ 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**.

$$F_{Y^{B_n^{k_{n+1}}(x)}}(D) - F_{Y^x}(D) \geq -M(x)||X - x||_{(k_{n+1},n)}\ .$$

We know, thanks to Lemma 5.5, that $||X - x||_{(k_{n+1},n)}$ converges almost surely to 0. Then, there exists a set $\Omega_3 \subset \Omega_1$ of probability strictly non-negative such that forall ω in Ω_3, the previous reasoning is true. And for $\epsilon < \frac{\eta}{L}$, there exists rank $N(\omega)$ such that if $n \geq N$,

$$F_{Y^{B_n^{k_{n+1}}(x)}}(D) - F_{Y^x}(D) \geq 0 - L\epsilon + \eta > 0\ . \tag{13}$$

Finally, for $\omega \in \Omega_3$ (set of strictly non-negative measure), we have shown that after a certain rank, $h_n(\theta_n(\omega)) > 0$. This implies that on Ω_3 of positive measure,

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

which is absurd because in the previous part we proved that T_n is almost surely convergent. Then θ_n does not diverge to $+\infty$ or $-\infty$.

Now, we will show that (θ_n) converges almost surely. In all the sequel of the proof, we reason ω by ω like in the previous part. To make the reading more easy, we do not write ω and Ω any more. Thanks to Equation (12) and to the previous subsection, we know that, with probability positive, there exists a sequence (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, the same argument would allow us to conclude in the other case). Let us choose c and d satisfying $c > \theta^*$ and $\liminf \theta_n < c < d < \limsup \theta_n$. Since the sequence (γ_n) converges to 0, and since (T_n) is a Cauchy sequence, we can find a deterministic rank N and two integers n and m such that $N \leq n < m$ implies

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

We choose m and n so that

$$\begin{cases} (a)\ N \leq n < m \\ (b)\ \theta_n < c,\ \theta_m > d \\ (c)\ n < j < m \Rightarrow c \leq \theta_j \leq d \ . \end{cases} \tag{14}$$

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 necessarily go through the segment $[c, d]$. We then 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 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 (14). 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 (14). We have shown that the algorithm converges almost surely.

5.2.3. *The algorithm converges almost surely to θ^*.* Again we reason by contradiction. Let us name θ the limit such that $\mathbb{P}(\theta \neq \theta^*) > 0$. With positive probability, 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 in this case). Then, for n large enough, we get

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

Finally, on the one hand, (T_n) and (θ_n) are convergent, and we also know that the sum $\sum \gamma_{j+1} h(\theta_j)$ converges almost surely. Let us then show that on the other hand, $h_n(\theta_n) = P_n(F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n) - \alpha)$ is lower bounded. First we know thanks to Lemma 5.4, that for $1 < \epsilon < 1 - \beta$ and $\epsilon_n = \frac{1}{(n+1)^\epsilon}$,

$$\mathbb{P}(P_n \leq \epsilon_n) \leq \exp\left(-\frac{3(n+1)^{1-\epsilon}}{8}\right) \ .$$

This is the general term of a convergent sum. Therefore, 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 (13), since $\theta_n > \epsilon_1 > \theta^*$,

$$F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n) - \alpha \geq 0 - M(x)||X - x||_{(k_{n+1},n)} + F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*) \ .$$

Then, when n is large enough so that

$$||X - x||_{(k_{n+1},n)} \leq \frac{F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)}{2M(x)}$$

holds, we have

$$F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n) - \alpha \geq \frac{F_{Y^x}(\epsilon_1) - F_{Y^x}(\theta^*)}{2} \ .$$

Finally there exists a set Ω of positive probability such that, $\forall \omega \in \Omega$

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

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

5.3. **Proof of Theorem 2.2 : Non-asymptotic inequality on the mean square error.** Let x be fixed in $[0,1]$. We want to find an upper-bound for the mean square error $a_n(x)$ using Lemma 5.7. In the sequel, we will need to study $\theta_n(x)$ on the event A_n of the Lemma 5.4. Then, we begin to find a link between $a_n(x)$ and the mean square error on this event.

$$\begin{aligned} a_n(x) &= \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2\,\mathbf{1}_{A_n}\right] + \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2\,\mathbf{1}_{A_n^C}\right] \\ &\le \mathbb{E}\left[(\theta_n-\theta^*)^2\,\mathbf{1}_{A_n}\right] + C_1\mathbb{P}(A_n^C) \\ &\le \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2\,\mathbf{1}_{A_n}\right] + C_1\exp\left(-\frac{3(n+1)^{1-\epsilon}}{8}\right) , \\ &\le \mathbb{E}\left[(\theta_n(x)-\theta^*(x))^2\,\mathbf{1}_{A_n}\right] + C_1\exp\left(-\frac{3n^{1-\epsilon}}{8}\right) , \end{aligned} \tag{15}$$

thanks to Lemma 5.4 and for $n \ge N_0$.

Let us now study the sequence $b_n(x) := \mathbb{E}\left[(\theta_n(x)-\theta^*)^2\,\mathbf{1}_{A_n}\right]$. First, for $n \ge 0$,

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

But,

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

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

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

Using the Bayes formula, we get

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

Let us split the double product into two terms representing the two errors we made by iterating our algorithm.

$$\begin{aligned} \mathbb{E}_n\left(\theta_{n+1}(x)-\theta^*(x))^2\right) &\le \left(\theta_n(x)-\theta^*(x)\right)^2 + \gamma_{n+1}^2 P_{n+1} \\ &\quad - 2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right)P_{n+1}\left[F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))\right] \\ &\quad - 2\gamma_{n+1}\left(\theta_n(x)-\theta^*(x)\right)P_n\left[F_{Y^x}(\theta_n(x)) - F_{Y^x}(\theta^*(x))\right] . \end{aligned} \tag{16}$$

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

$$|F_{Y^{B_n^{k_{n+1}}(x)}}(\theta_n(x)) - F_{Y^x}(\theta_n(x))| \geq M(x)||X-x||_{(k_{n+1},n)} \ ,$$

and by **A3**,

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

Thus,

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

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

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

Coming back to Equation (16), we get

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

To conclude, we take the expectation

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

But, by definition of A_n,we get

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

Finally,

$$b_{n+1}(x) \leq b_n(x)\left(1-2C_2(x,\alpha)\gamma_{n+1}\epsilon_n\right) + e_{n+1} \ ,$$

with

$$e_{n+1} := C_1\mathbb{P}(A_n^C) + \gamma_{n+1}^2\mathbb{E}(P_n) + 2\gamma_{n+1}\sqrt{C_1}M(x)\mathbb{E}\left[P_n||X-x||_{(k_{n+1},n)}\right] \ .$$

Now using Lemmas 5.6, 5.4 and 5.1 we get for $n \geq N_0$ with

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

The conclusion holds thanks to Lemma 5.7, for $n \geq N_0+1$,

(17)

$$b_n(x) \leq \exp\left(-2C_2(x,\alpha)(\kappa_n-\kappa_{N_0})\right)b_{N_0}(x) + \sum_{k=N_0+1}^{n}\exp\left(-2C_2(x,\alpha)\left(\kappa_n-\kappa_k\right)\right)d_k \ .$$

But thanks to Assumption **A3**, we have already shown that $b_{N_0}(x) \leq a_{N_0}(x) \leq C_1$. To conclude, we re-inject Equation (17) in Equation (15) and obtain for $n \geq N_0 + 1$,

$$a_n(x) \leq \exp\left(-2C_2(x,\alpha)(\kappa_n - \kappa_{N_0})\right) C_1 + \sum_{k=N_0+1}^{n} \exp\left(-2C_2(x,\alpha)\left(\kappa_n - \kappa_k\right)\right) d_k$$
$$+ C_1 \exp\left(-\frac{3n^{1-\epsilon}}{8}\right) .$$

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

$$T_n^0 := C_1 \exp\left(\frac{-3n^{1-\epsilon}}{8}\right) \ , \ T_n^1 := \exp\left(-2C_2(x,\alpha)(\kappa_n - \kappa_{N_0})\right)$$

and

$$T_n^2 := \sum_{k=N_0+1}^{n} \exp\left(-2C_2(x,\alpha)\left(\kappa_n - \kappa_k\right)\right) d_k \ .$$

We want to find a simpler expression for those terms to better see their order in n. First, considering T_n^1 we see that $a_n(x)$ can converge 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 5.1. *The frontier case $\epsilon = 1-\gamma$ is possible but the analysis shows that it is a less interesting choice than $\epsilon < 1-\gamma$ (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). In the sequel, we only consider $\epsilon < 1-\gamma$.*

Let us upper-bound T_n^1. As $x \mapsto 1/x^{\epsilon+\gamma}$ is decreasing, we get

$$\begin{aligned} T_n^1 &= \exp\left(-2C_2(x,\alpha) \sum_{k=N_0+1}^{n} \frac{1}{k^{\epsilon+\gamma}}\right) \\ &\leq \exp\left(-2C_2(x,\alpha) \int_{N_0+1}^{n+1} \frac{1}{t^{\epsilon+\gamma}} dt\right) \\ &\leq \exp\left(-2C_2(x,\alpha) \frac{(n+1)^{1-\epsilon-\gamma} - (N_0+1)^{1-\epsilon-\gamma}}{(1-\epsilon-\gamma)}\right) . \end{aligned}$$

Then, T_n^1 (just like T_n^0) is exponentially small when n grows up. To deal with the second term T_n^2 we first study the order in n of d_n. d_n is composed of three terms :

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

The first one is negligeable (exponentially decreasing). Let us compare the two others which are powers of n. Comparing their exponents, we get that there exists constants C_5 and $C_6(d)$ (their explicit form is given in the Appendix) such that

if $\beta \leq 1 - d\gamma$, then for $n \geq N_0 + 1$,

$$d_n \leq C_5(x,d) n^{-2\gamma+\beta-1} ,$$

if $\beta > 1 - d\gamma$, then for $n \geq N_0 + 1$,

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

Remark 5.2. *Let us detail how one can find C_5 (it is the same reasoning for C_6). If $\beta \leq 1-d\gamma$, we know that when n will be big enough, the dominating term of d_n will be the one in $n^{-2\gamma+\beta-1}$. Then, it is logical to search a constant $C_5(x,d)$ such that $\forall n \geq N_0+1$,*

$$d_n \leq \frac{C_5(x,d)}{n^{2\gamma-\beta+1}} .$$

Such a constant has to satisfy, for all $n \geq N_0 + 1$,

$$C_5(x,d) \geq C_1 \exp\left(-\frac{3}{8} n^{1-\epsilon}\right) n^{2\gamma-\beta+1} + \frac{2\sqrt{C_1} M(x) C_3(d)}{n^{-\gamma+(1-\beta)/d}} + 1 .$$

Since $\beta \leq 1 - d\gamma$, the map $x \mapsto \frac{2\sqrt{C_1}M(x)C_3(d)}{n^{-\gamma+(1-\beta)/d}}$ is positive and decreasing. Then its maximum is reached for $n = N_0 + 1$. Moreover, the map $x \mapsto C_1 \exp\left(-\frac{3}{8}n^{1-\epsilon}\right) n^{2\gamma-\beta+1}$ is also positive and is decreasing on an $[A, +\infty[$. It also has a maximum. The previous inequality is then true for

$$C_5(x,d) := \max_{n \geq N_0+1} C_1 \exp\left(-\frac{3}{8} n^{1-\epsilon}\right) n^{2\gamma-\beta+1} + \frac{2\sqrt{C_1} M(x) C_3(d)}{(N_0+1)^{-\gamma+(1-\beta)/d}} + 1 .$$

Let us study the two previous cases.

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

To upper-bound these sums, we use arguments from [8], which studies the stochastic algorithm to estimate the median on an Hilbert space. The main arguments are comparisons between sums and integrals. Indeed, for $n \geq N_0 + 2$ and $n \geq N_3$ where N_3 is such that

$$\forall n \geq N_3, \ \lfloor \frac{n}{2} \rfloor \geq N_0 + 1 ,$$

$$\begin{aligned}
T_n^2 &= C_6(x,d) \sum_{k=N_0+1}^{n-1} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} + \frac{C_6(x,d)}{n^{\gamma+(1+\frac{1}{d})(1-\beta)}} \\
&= C_6(x,d) \sum_{k=N_0+1}^{\lfloor \frac{n}{2} \rfloor} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} \\
&+ C_6(x,d) \sum_{k=\lfloor \frac{n}{2} \rfloor+1}^{n-1} \exp\left(-2C_2(x,\alpha) \sum_{j=k+1}^{n} \frac{a}{j^{\epsilon+\gamma}}\right) \frac{1}{k^{\gamma+\left(1+\frac{1}{d}\right)(1-\beta)}} + \frac{C_6(x,d)}{n^{\gamma+(1+\frac{1}{d})(1-\beta)}} \\
&=: S_1 + S_2 + S_3 .
\end{aligned}$$

First, the function $x \mapsto x^{-\epsilon-\gamma}$ is decreasing on $]0, +\infty[$ then

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

Then, taking, $1-\beta < \epsilon < \min((1-d\gamma), \left(1+\frac{1}{d}\right)(1-\beta))$, we have since $k \geq \lfloor \frac{n}{2} \rfloor + 1$

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

Now, since for $k \geq 1$,

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

we get

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

Since the function $x \mapsto \exp\left(2C_2(x,\alpha)\frac{n^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)$ is decreasing on $\left[\frac{2C_2(x,\alpha)}{\gamma+\epsilon}, +\infty\right[$, we also define the integer $N_1(x,\alpha)$ the rank such that

$$
\forall n \geq N_1(x,\alpha),\ \lfloor \frac{n}{2} \rfloor + 1 \geq \frac{2C_2(x,\alpha)}{\epsilon+\gamma} \ .
$$

For $n \geq N_1(x,\alpha)$ we get

$$
\begin{aligned}
S_2 &\leq C_6(x,d)\exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\gamma-\epsilon}}{1-\gamma-\epsilon}\right)\frac{2^{(1+\frac{1}{d})(1-\beta)+\gamma}}{n^{(1+\frac{1}{d})(1-\beta)-\epsilon}} \\
&\times \sum_{k=\lfloor\frac{n}{2}\rfloor+1}^{n-1}\int_{\lfloor\frac{n}{2}\rfloor+2}^{n}\exp\left(-2C_2(x,\alpha)\frac{x^{1-\gamma-\epsilon}}{1-\gamma-\epsilon}\right)\frac{1}{x^{\gamma+\epsilon}}dx \\
&\leq \frac{C_6(x,d)}{2C_2(x,\alpha)}\exp\left(-2C_2(x,\alpha)\frac{(n+1)^{1-\gamma-\epsilon}}{1-\gamma-\epsilon}\right)\frac{2^{(1+\frac{1}{d})(1-\beta)+\gamma}}{n^{(1+\frac{1}{d})(1-\beta)-\epsilon}} \\
&\times \left[\exp\left(2C_2(x,\alpha)\frac{n^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)-\exp\left(2C_2(x,\alpha)\frac{(\lfloor\frac{n}{2}\rfloor+2)^{1-\epsilon-\gamma}}{1-\epsilon-\gamma}\right)\right] \\
&\leq \frac{C_6(x,d)}{2C_2(x,\alpha)}\frac{2^{(1+\frac{1}{d})(1-\beta)+\gamma}}{n^{(1+\frac{1}{d})(1-\beta)-\epsilon}} =: \frac{C_7(x,d,\alpha)}{2}\frac{1}{n^{-\epsilon+(1+\frac{1}{d})(1-\beta)}} \ .
\end{aligned}
$$

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,

$$
\begin{aligned}
S_1 &= C_6(x,d)\sum_{k=N_0+1}^{\lfloor\frac{n}{2}\rfloor}\exp\left(-2C_2(x,\alpha)\sum_{j=k+1}^{n}\frac{a}{j^{\epsilon+\gamma}}\right)\frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}} \\
&\leq C_6(x,d)\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\exp\left(-C_2(x,\alpha)n^{1-\epsilon-\gamma}\right)\frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}} \\
&\leq C_6(x,d)\exp\left(-C_2(x,\alpha)n^{1-\epsilon-\gamma}\right)\sum_{k=1}^{\lfloor\frac{n}{2}\rfloor}\frac{1}{k^{\gamma+(1-\beta)\left(1+\frac{1}{d}\right)}} \ .
\end{aligned}
$$

Thanks to the exponential term, S_1 is insignificant compared to S_2 whatever is the behaviour of the sum $\sum_k k^{-\gamma-(1-\beta)\left(1+\frac{1}{d}\right)}$, and so is T_1^n. Then, denoting $N_2(d,x)$ the rank after which we have

$$
S_3 + S_1 + T_n^1 + T_n^0 \leq \frac{C_7(x,\alpha,d)}{2n^{(1+\frac{1}{d})(1-\beta)-\epsilon}} \ ,
$$

we get, in the case where $\beta > 1-\gamma$ and $1-\beta < \epsilon < \min((1-\gamma), \left(1+\frac{1}{d}\right)(1-\beta))$, for $n \geq \max\left(N_0, N_1(x,\alpha), N_2(d,x)\right)$

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

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

Using the same arguments, we conclude that for $1-\beta < \epsilon < \min(1-\beta+\gamma, 1-\gamma)$ and $n \geq \max(N_0, N_1(x,\alpha), N_2(d,x))$ (see Appendix for precise definitions of these ranks), there exists a constant $C_8(x,\alpha,d)$ such that the mean square error satisfies

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

5.5. **Proof of Corollary 2.2 : choice of best parameters β and γ.** Let us now optimize the rate of convergence obtained in previous theorem. When $\beta \geq \gamma$ and $\beta \leq 1-d\gamma$, the rate of convergence is of order $n^{-\gamma+\beta-1+\epsilon}$. To optimize it, we have to choose ϵ as small as possible. Then, we take $\epsilon = 1-\beta+\eta_\epsilon$. The rate becomes $n^{-\gamma+\eta_\epsilon}$. Then, we have also to choose γ as small as possible. In this area, there is only one point in which γ is the smallest, this is the point $(\gamma,\beta) = (\frac{1}{1+d}, \frac{1}{1+d})$. Since we have to take $\beta > \gamma$, the best couple of parameters, in this area, is $(\frac{1}{1+d}, \frac{1}{1+d} + \eta_\beta)$. These parameters follow a rate of convergence of $n^{\frac{-1}{1+d}+\eta}$.

When we are in the second area, the same kind of arguments allows us to conclude to the same optimal point with the same rate of convergence.

In Figure 6, we use the numerical simulations of Section 3 to illustrate the previous discussion.

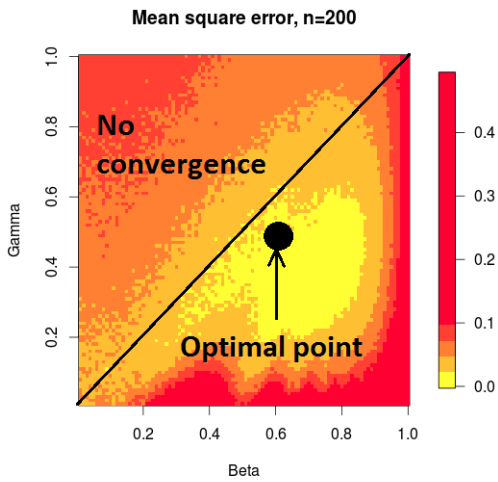

FIGURE 6. Theoretical behaviour of the MSE in function of β and γ

We have finally shown that

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

where the constant is the minimal constant between $C_7(x,\alpha,d)$ and $C_8(x,\alpha,d)$ computed with optimal parameters (γ,β,ϵ).

6. Appendix 2 : Recap of the constants

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

6.1. 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 law f_X.
- $C_g(x)$ is the positive lower bound of the density of the law of $g(x, \epsilon)$.

6.2. Compact support.

We denote :

- $[L_Y, U_Y]$ 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.
- $[L_{\theta_n}, U_{\theta_n}] := [L_Y - (1-\alpha), U_Y + \alpha]$ the segment in which θ_n can take its values $(\forall x)$.
- $U_{|.|}$ the upper bound of the compact support of the distribution of $||X - x||$ $(\forall x)$.

6.3. Real constants.

We denote :

- $\sqrt{C_1} := U_Y + \alpha - L_Y$. C_1 is the uniform in ω and x bound of $(\theta_n(x) - \theta^*(x))^2$.
- $C_2(x, \alpha) := \min\left(C_g(x), \frac{1-\alpha}{U_Y+\alpha-L_Y}\right)$ is the constant such that

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

- $C_3(d) := \sqrt[d]{2}\left(1 + \frac{8}{3d} + \frac{1}{\sqrt[d]{C_{input}} C_4(d)}\right)$.
- $C_4(d) := \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})+1}$.
- $C_5(x,d) := \max\limits_{n \geq N_0+1} C_1 \exp\left(-\frac{3}{8} n^{1-\epsilon}\right) n^{2\gamma-\beta+1} + \frac{2\sqrt{C_1} M(x) C_3(d)}{(N_0+1)^{-\gamma+(1-\beta)/d}} + 1.$
- $C_6(x,d) := \max\limits_{n \geq N_0+1} C_1 \exp\left(-\frac{3}{8} n^{1-\epsilon}\right) n^{\gamma+(1+\frac{1}{d})(1-\beta)} + 2\sqrt{C_1} M(x) C_3(d) + \frac{1}{(N_0+1)^{\gamma-\frac{1}{d}(1-\beta)}}.$
- $C_5^{optim} := \max\limits_{n \geq N_0+1} C_1 \exp\left(-\frac{3}{8} n^{\left(\frac{1}{1+d}+\eta_\beta\right)-\eta_\epsilon}\right)(N_0+1)^{\frac{1}{1+d}-\eta_\beta+1} + 1 + \frac{1}{(N_0+1)^{-\frac{1}{1+d}+\frac{1}{d}(1-\frac{1}{1+d}-\eta_\beta)}}.$
- $C_6^{optim}(x,d) := \max\limits_{n \geq N_0+1} C_1 \exp\left(-\frac{3}{8} n^{\left(\frac{1}{1+d}+\eta_\beta\right)-\eta_\epsilon}\right) n^{\left(1+\frac{1}{d}\right)-\frac{1}{d(1+d)}-\eta_\beta\left(1+\frac{1}{d}\right)} + 2\sqrt{C_1} M(x) C_3(d) + \frac{1}{(N_0+1)^{-\frac{1}{d}+\frac{1}{d(1+d)}+\frac{1}{1+d}+\frac{\eta_\beta}{d}}}.$
- $C_7(x,\alpha,d) := \frac{2^{(1+\frac{1}{d})(1-\beta)+\gamma} C_6(x,d)}{C_2(x,\alpha)}$.
- $C_8(x,\alpha) := \frac{2^{2\gamma-\beta+1} C_5(x,d)}{C_2(x,\alpha)}$.
- $C_9(x,\alpha,d) := \min\left(\frac{2^{1+\frac{1}{d}-\frac{1}{d(1+d)}-\eta_\beta(1+\frac{1}{d})} C_5^{optim}(x,d)}{C_2(x,\alpha)}, \frac{2^{\frac{1}{1+d}-\eta_\beta+1} C_6^{optim}(x,d)}{C_2(x,\alpha)}\right).$
- $C_{10}(d) := \sqrt[d]{\frac{2(k_n+1)}{(n+1) C_{input} C_4(d)}}$.

6.4. **Integer constants.** We denote :

- $N_0 := 2^{\frac{1}{\epsilon-(1-\beta)}}$.
- $N_1(x,\alpha)$ is the rank such that $n \geq N_1(x,\alpha)$ implies

$$\lfloor \frac{n}{2} \rfloor + 1 \geq \frac{2C_2(x,\alpha)}{\epsilon+\gamma}.$$

- $N_2(x,\alpha,d)$ is the integer such that $\forall n \geq N_2(x,\alpha,d)$,
  a) If $\beta \leq 1 - d\gamma$,

$$S_3 + S_1 + T_n^1 + T_n^0 \leq \frac{C_7(x,\alpha,d)}{2n^{(1+\frac{1}{d})(1-\beta)-\epsilon}} \ ,$$

where $T_n^1 := \exp\left(-2C_2(x,\alpha)\sum_{k=N_0+1}^{n} k^{-\gamma-\epsilon}\right)$, $T_n^0 := C_1 \exp\left(\frac{-3n^{1-\epsilon}}{8}\right)$,

$S_3 := \frac{C_6(x,d)}{n^{\gamma+(1+\frac{1}{d})(1-\beta)}}$ and $S_1 := C_6(x,d)\exp(-2C_2(x,\alpha)n^{1-\epsilon-\gamma})\sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} k^{-\gamma-(1-\beta)(1+1/d)}$.

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

$$S_3 + S_1 + T_n^1 + T_n^0 \leq \frac{C_8(x,\alpha,d)}{2n^{\gamma-\beta+1-\epsilon}} \ ,$$

where $T_n^1 := \exp\left(-2C_2(x,\alpha)\sum_{k=N_0+1}^{n} k^{-\gamma-\epsilon}\right)$, $T_n^0 := C_1 \exp\left(\frac{-3n^{1-\epsilon}}{8}\right)$,

$S_3 := \frac{C_5}{n^{2\gamma-\beta+1)}}$ and $S_1 := C_5 \exp(-2C_2(x,\alpha)n^{1-\epsilon-\gamma})\sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} k^{-\gamma-(1-\beta)(1+1/d)}$.

- N_3 is the rank such that $\forall n \geq N_3$, $\lfloor \frac{n}{2} \rfloor \geq N_0 + 1$.
- $N_4(x,\alpha,d) := \max(N_0+2, N_1(x,\alpha), N_2(x,\alpha,d), N_3)$.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

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

[3] Aurélie Arnaud, Julien Bect, Mathieu Couplet, Alberto Pasanisi, and Emmanuel Vazquez. Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.

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

[5] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, 2012.

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

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

[8] 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.

[9] Herbert A. David and Haikady N. Nagaraja. *Order Statistics*. Wiley, 2003.

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

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

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

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

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

[15] 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.

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

[17] Marjorie Jala, Céline Lévy-Leduc, Eric Moulines, Emmanuelle Conil, and Joe Wiart. Sequential design of computer experiments for parameter estimation with application to numerical dosimetry. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 909–913. IEEE, 2012.

[18] Marjorie Jala, Céline Lévy-Leduc, Éric Moulines, Emmanuelle Conil, and Joe Wiart. Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. *Technometrics*, (just-accepted):00–00, 2014.

[19] 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.

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

[21] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning.

[22] Jeremy Oakley. Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):83–93, 2004.

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

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

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

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

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

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

[29] 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.

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

[31] Micheal 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.