

Concentration inequalities for non-causal random fields^{†*}

Rémy Garnier

*CY University - AGM UMR 8088,
2 Bd. Adolphe Chauvin. 95000 Cergy-Pontoise, France*

*CDiscount,
120-126 Quai de Bacalan, 33300 Bordeaux, France
e-mail: remy.garnier@ext.cdiscout.com*

Raphaël Langhendries

*University Paris 1 Panthéon-Sorbonne - SAMM,
90, rue de Tolbiac 75013 PARIS CEDEX 13 - France*

*Safran Aircraft Engines,
77550 Moissy-Cramayel, France
e-mail: r.langhendries@gmail.com*

Abstract: Concentration inequalities are widely used for analyzing machine learning algorithms. However, the current concentration inequalities cannot be applied to some non-causal processes which appear for instance in Natural Language Processing (NLP). This is mainly due to the non-causal nature of such involved data, in the sense that each data point depends on other neighboring data points. In this paper, we establish a framework for modeling non-causal random fields and prove a Hoeffding-type concentration inequality. The proof of this result is based on a local approximation of the non-causal random field by a function of a finite number of i.i.d. random variables.

MSC2020 subject classifications: Primary 60G60, 60G48, 60E15; secondary 68Q32, 62M45.

Keywords and phrases: Non-causal random fields, learning in non-i.i.d scenarios, Hoeffding’s inequality, local dependence, learning theory.

Received January 2021.

Contents

1	Introduction	1682
1.1	Our contribution	1684
1.2	Outline of the paper	1684
2	Model	1685
2.1	Some definitions and notations	1685
2.2	Local non-causal relations	1686

*This work was partly involved in the CY Initiative of Excellence (grant “Investissements d’Avenir” ANR-16-IDEX-0008), Project “EcoDep” PSI-AAP2020-0000000013.

[†]We thank Paul Doukhan, Joseph Rynkiewicz and Xiequan Fan for their advice.

2.2.1	Contraction hypothesis	1686
2.2.2	Coupling hypothesis	1688
2.2.3	Examples	1688
2.3	Function of interest Φ and the statistic $S_{\mathcal{I}}$	1690
3	Main Results	1691
3.1	Concentration inequality within i.i.d. assumption	1691
3.2	Concentration inequalities and expected deviation bounds for non-causal random fields	1692
3.3	Comparison with other results	1694
3.4	Application to learning theory	1696
3.4.1	Application to the completion problem	1696
3.4.2	Model selection	1699
4	Approximation	1702
4.1	Approximation of X_t	1702
4.1.1	Notations	1702
4.1.2	Exact reconstruction	1702
4.1.3	Intuition	1703
4.1.4	Definition	1703
4.1.5	Approximation error	1704
4.2	Concentration inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$	1705
4.2.1	Counting random variables	1705
4.2.2	An extension of McDiarmid's inequality	1706
4.2.3	Difference bound for $\tilde{S}_{\mathcal{I}}^{[d]}$	1707
4.2.4	McDiarmid's inequality	1709
5	Concentration inequalities for $S_{\mathcal{I}}$	1710
5.1	General concentration inequalities	1710
5.2	Concentration inequalities with optimized parameters	1712
5.3	Expected deviation bounds	1713
A	Proof of Corollary 3.3	1715
B	Proof of Lemma 4.1	1717
C	Proof of Lemma 4.5	1718
D	Proof of Lemma 4.6	1719
E	Proof of Theorem 5.2	1721
	References	1723

1. Introduction

Concentration inequalities are widely used in statistical learning. For example, model selection techniques rely heavily on concentration inequalities [28]. They have also been used for high dimensional procedures [1, 7] or for studying various machine learning framework, such as time series forecasting [25], online machine learning [33] or classification problems [19].

Many concentration inequalities have been proposed under different scenarios and assumptions. The simpler case corresponds to the *independent and identically distributed hypothesis*. Interested readers may consult [9] for an overview on concentration inequalities in this case.

However, in many machine learning applications, the i.i.d. assumption does not hold. It is the case for instance for time series which exhibit inherent temporal dependence.

As simple model of dependent times series are Markov chains. A Markov chain is a sequence of random variables (X_t) verifying:

$$X_t = F(X_{t-1}, \dots, X_{t-i}, \varepsilon_t), \text{ for a function } F \text{ and i.i.d. innovations } (\varepsilon_t)_{t \in \mathbb{Z}}. \quad (1.1)$$

Many approaches have been used to establish concentration inequalities for Markov chains. These approaches include renewal technique (see [6]), Marton coupling (see [32]) and martingale decomposition (see [2, 12, 14, 16, 23]).

However, Markov chains are not sufficient to model any type of dependence. Some data may exhibit non-causal dependence. In the uni-dimensional case, it means that one data point does not merely depend on past data points, but also future data.

For example, this situation occurs for textual data. Textual data are a typical example of data generated by a *non-causal* process, because dependence on words goes forward, but also backward. A key task involving textual data is the *completion problem*. It consists in filling blanks in a text using surrounding words. This task is achieved by creating a language model, i.e., a probability distribution of words for a given context (past and future words).

In practice, non-causal models are already used to learn language models. Notable examples of such models are bidirectional neural networks [34]. They have recently received a lot of attention for their performances in Natural Language Processing. In particular, the BERT model [21] has become a staple for a very large range of NLP tasks, such as translation, part-of-speech tagging, sentiment analysis. However, despite their success in practical applications, it lacks a theoretical framework to analyze such non-causal models.

If the dimension of the lattice of random variables increases, we obtained a *random field*. The natural extension of Markov chains to random fields leads to causal random fields: here the dependence is propagated along with preferential directions (see [17] for an example of application). Non-causal random fields appear more naturally in some applications. For instance, it is natural to model the generation of pictures by a non-causal random field defined over a two-dimensional lattice. In this case, the completion problem consists in filling missing pixels using neighboring pixels [4]. It had to use information from each direction (up, down, left, right) to be able to complete the missing pixel. However, contrary to the case of textual data, such pixel completion techniques are not state-of-the-art for image processing. Another application of importance is the case of completing geographical data sets which make sense for an ecology setting (see <http://doukhan.u-cergy.fr/EcoDep.html>), for which applications are of fundamental importance.

1.1. Our contribution

In this article, we extend the Markovian framework presented in Eq. (1.1) to handle non-causal data. In the 1-dimensional case, we assume that the model is a solution of the following equations

$$X_t = F(X_{t-s}, \dots, X_{t-1}, X_{t+1}, \dots, X_{t+s}, \varepsilon_t), \text{ with i.i.d. } (\varepsilon_t)_{t \in \mathbb{Z}}.$$

We straightforwardly generalize this approach to random fields indexed by \mathbb{Z}^κ , i.e., the multi-dimensional case. In this case, data are a solution of the following equation

$$X_t = F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t), \text{ for some neighborhood } \mathcal{B} \text{ and i.i.d. } (\varepsilon_t)_{t \in \mathbb{Z}}.$$

This kind of random field was introduced by [18]. It provides conditions for the existence and uniqueness of the solution of such an equation.

We aim at establishing concentration inequalities for these processes using realistic hypotheses on the data. In particular, we want our hypotheses to be reasonable for text data. These hypotheses are non-causal versions of the hypotheses of [2] or [16].

Our main results are Hoeffding-type inequalities. The main hypothesis for these inequalities is relaxed versions of the contraction assumption used in [2, 14, 16].

Our technique of proof relies on a convenient local approximation of a non-causal random field by a function of a finite number of i.i.d. random variables. We also give some examples of such random fields and compare our framework with other classical dependent frameworks. Finally, we present a simple application of our results to model selection, inspired by [27].

1.2. Outline of the paper

In Section 2, we introduce non-causal random fields proposed in [18]. We also present our hypotheses in this section. We take some examples of models satisfying our framework and define the statistic $S_{\mathcal{I}}$ of interest for which we aim at proving a concentration inequality. Main results may be found in Section 3 as well as some immediate applications to machine learning and comparisons with other concentration inequalities. The rest of the paper is dedicated to the proof of main results.

In the section 4, we introduce an approximation $\tilde{S}_{\mathcal{I}}^{[d]}$ of $S_{\mathcal{I}}$ which only depends on a *finite* number of independent variables. We also establish a result evaluating the quality of this approximation.

Then, in section 5 we prove the main concentration inequality using the McDiarmid's inequality proposed by [13] on $\tilde{S}_{\mathcal{I}}^{[d]}$.

2. Model

In this section, we will present a model inspired by [18] in order to consider non-causal random field. Then, we introduce our hypotheses and the targeted statistic $S_{\mathcal{I}}$.

2.1. Some definitions and notations

From now, all random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Dimension of the random field Let $\kappa \in \mathbb{N}$ denote the *dimension* of the random field of interest.

This *dimension* should not be confused the classical dimension that appears in high-dimensional statistics, i.e., the number of parameters of the model. Here, the number of parameters increases exponentially with the dimension κ .

Probabilistic setting Let \mathcal{X} be a Banach space endowed with a norm $\|\cdot\|$. We define the m -norm $\|X\|_m$ of a random variable X as

$$\|X\|_m = (\mathbb{E}\|X\|^m)^{\frac{1}{m}} \quad \text{with } \|\cdot\| \text{ is a norm on } \mathcal{X}.$$

We also use the uniform norm $\|X\|_\infty = \inf\{C, \|X\| \leq C \text{ a.s.}\} = \lim_{m \rightarrow \infty} \|X\|_m$.

Neighborhoods $\mathcal{V}(\delta, s)$ and \mathcal{B} Let δ be a κ -tuple of non-negative integer, i.e., $\delta = (\delta_1, \dots, \delta_\kappa) \in \mathbb{N}^\kappa$ and s be a κ -tuple of integers, i.e., $s = (s_1, \dots, s_\kappa) \in \mathbb{Z}^\kappa$. The δ -neighborhood of s , $\mathcal{V}(\delta, s)$ is defined as the following κ -orthotope

$$\mathcal{V}(\delta, s) = \{t = (t_1, \dots, t_\kappa) \in (\mathbb{Z})^\kappa / \forall i, s_i - \delta_i \leq t_i \leq s_i + \delta_i\}.$$

Thereafter, we will study a specific neighborhood $\mathcal{B} = \mathcal{V}(\delta, 0) \setminus \{0\}$ for a fixed value of δ .

Moreover we will denote by n_B the cardinal of $\mathcal{V}(\delta, 0) = \mathcal{B} \cup \{0\}$ (thus $n_B = \text{card}(\mathcal{B}) + 1$).

Additionally, we introduce the notation $\mathcal{B} + s$ which denote the set \mathcal{B} shifted by s , i.e., $\mathcal{B} + s = \mathcal{V}(\delta, s) \setminus \{s\}$.

Innovations $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$. Let $(\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$ be a random field indexed by \mathbb{Z}^κ of *independent and identically distributed* random variables ε_t on a Banach space E . To shorten the notation, we denote by $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$ the whole random field.

Additionally, we define μ_ε the probability distribution of one random variable ε_t and $\mu = \bigotimes_{t \in \mathbb{Z}^\kappa} \mu_\varepsilon$, the product of these distributions on $E^{\mathbb{Z}^\kappa}$. Then, μ is the distribution of $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$.

2.2. Local non-causal relations

In this paper, we will study the κ -dimensional non-causal random field $(X_t)_{t \in \mathbb{Z}^\kappa}$. This random field presents a *local dependence*, for each $t \in \mathbb{Z}^\kappa$, X_t depends on its neighborhood indexed by $\mathcal{V}(\delta, t)$ and on the innovation ε_t .

Formally, we assume it exists a function $F : \mathcal{X}^{\mathcal{B}} \times E \rightarrow \mathcal{X}$ such that $(X_t)_{t \in \mathbb{Z}^\kappa}$ is a stationary solution of the following equation

$$\forall t \in \mathbb{Z}^\kappa, X_t = F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t). \quad (2.1)$$

The reference [18] ensures the existence and uniqueness of such solutions giving contraction hypothesis on F . Here, we just suppose to have one (strongly) stationary random field $(X_t)_{t \in \mathbb{Z}^\kappa}$ verifying this equation and we do not assume the uniqueness.

Thereafter, we denote by μ_X the stationary distribution of one marginal random variable X_t . We assume that laws μ_X and μ_ε are stable by F . It means that, if $(X_t)_{t \in \mathcal{B}}$ is drawn with marginal distribution μ_X and ε is drawn with distribution μ_ε , then the random variable $F((X_s)_{s \in \mathcal{B}}, \varepsilon)$ follows the distribution μ_X .

For $s \in \mathcal{B}$, X_{t+s} depends on X_t , but X_t also depends on X_{t+s} . Therefore, it is no longer possible to describe $(X_t)_{t \in \mathbb{Z}^\kappa}$ as a result of a martingale process. This is why, we call $(X_t)_{t \in \mathbb{Z}^\kappa}$ a non-causal random field.

Remark 2.1. *It is important to note that the evolution of the random field is defined by a local equation. This is similar to settings of auto-regressive times series in the uni-dimensional case. This is why our framework can be seen as a generalization of stationary Markov chains. Indeed when $\delta = \kappa = 1$ and $\mathcal{B} = \{-1\}$, $(X_t)_{t \in \mathbb{Z}}$ is an homogeneous and stationary Markov chain.*

Remark 2.2. *We have to assume that the set \mathcal{B} of index is finite. Indeed the quantity $n_{\mathcal{B}} = \text{card}(\mathcal{B}) + 1$ play an important role in our results. Consequently, models with infinite memory (such as $AR(\infty)$) don't fit into the framework defined by Equation 2.1*

2.2.1. Contraction hypothesis

Numerous concentration inequalities for Markov chains and other martingales rely on contraction hypotheses (e.g. [2, 14, 16]).

Let's consider a \mathcal{X} -valued Markov chain $(X_t)_{t \in \mathbb{Z}}$ given by

$$X_t = F(X_{t-1}, \varepsilon_t), \text{ where } \varepsilon_t \text{ are i.i.d. random variables.}$$

The classical contraction hypothesis requires that it exists $\rho < 1$ such that for all $y, y' \in \mathcal{X}$, and for all $\varepsilon \in E$

$$\|F(y, \varepsilon) - F(y', \varepsilon)\| \leq \rho \|y - y'\|. \quad (2.2)$$

It is relatively easy to extend this condition to a non-causal random field defined by Equation (2.1). It would be the following condition.

Definition 2.1. ABSOLUTE CONTRACTION *It exists $(\lambda_t)_{t \in \mathcal{B}}$, such that $\rho := \sum_{t \in \mathcal{B}} \lambda_t < 1$, and for any \mathcal{X} -valued tuples $\mathcal{Y} = (y_t)_{t \in \mathcal{B}}$ and $\mathcal{Y}' = (y'_t)_{t \in \mathcal{B}}$ indexed by \mathcal{B} and for all $\varepsilon \in E$.*

$$\|F(\mathcal{Y}, \varepsilon) - F(\mathcal{Y}', \varepsilon)\| \leq \sum_{s \in \mathcal{B}} \lambda_s \|y_s - y'_s\|. \quad (2.3)$$

This condition is similar to the condition proposed by [18]. It is strong condition, and it is not satisfied by many usual models.

We therefore want to relax this condition. To that end, we consider that the contraction is only verified for a m order moment. This leads to the following hypothesis.

Definition 2.2. WEAK CONTRACTION HYPOTHESIS (\mathbf{H}_1^m). *Let $m \in \mathbb{N}$. There exists $(\lambda_s)_{s \in \mathcal{B}} \in [0, 1]^{\mathcal{B}}$ such that, for all tuples of \mathcal{X} -valued random variables $(Y_t)_{t \in \mathbb{Z}^{\kappa}}, (Y'_t)_{t \in \mathbb{Z}^{\kappa}}$ with marginal distribution μ_X and a random fields $\varepsilon = (\varepsilon_t)_{t \in \mathbb{Z}^{\kappa}}$ with product distribution μ .*

1. $\forall t \in \mathbb{Z}^{\kappa}, \|Y_t - Y'_t\|$ admit moment at order m .
2. $\forall t \in \mathbb{Z}^{\kappa}, \|F((Y_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) - F((Y'_{t+s})_{s \in \mathcal{B}}, \varepsilon_t)\|_m \leq \sum_{s \in \mathcal{B}} \lambda_s \|Y_{t+s} - Y'_{t+s}\|_m$.
3. $\sum_{s \in \mathcal{B}} \lambda_s < 1$. We denote thereafter $\rho = \sum_{s \in \mathcal{B}} \lambda_s$.

We emphasize that (\mathbf{H}_1^m) depends on m because the contraction concerns only the m -moment and does not require the function F to be contracting (Equation (2.3)). This hypothesis is hard to verify on the new model because ε_t and (Y_t) are in general dependent on one another. To prove this condition, we generally need to verify a stronger contraction hypothesis. However, it should be possible to test this hypothesis, whereas it is impossible for the absolute contraction condition because we do not have access to the function F .

However, the weak contraction hypothesis leads to weaker results than absolute contraction. Therefore we also use the following compromise: When m goes to ∞ , (\mathbf{H}_1^m) becomes

Definition 2.3. UNIFORM CONTRACTION HYPOTHESIS (\mathbf{H}_1^∞). *There exists $(\lambda_s)_{s \in \mathcal{B}} \in [0, 1]^{\mathcal{B}}$ such that, for all \mathcal{X} -valued tuples of random variables $(Y_t)_{t \in \mathbb{Z}^{\kappa}}, (Y'_t)_{t \in \mathbb{Z}^{\kappa}}$ with marginal distribution μ_X and a random fields $\varepsilon = (\varepsilon_t)_{t \in \mathbb{Z}^{\kappa}}$ with product distribution μ .*

1. $\forall t \in \mathbb{Z}^{\kappa}, \|Y_t - Y'_t\|$ admit moment at order ∞
2. $\forall t \in \mathbb{Z}^{\kappa}, \|F((Y_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) - F((Y'_{t+s})_{s \in \mathcal{B}}, \varepsilon_t)\|_\infty \leq \sum_{s \in \mathcal{B}} \lambda_s \|Y_{t+s} - Y'_{t+s}\|_\infty$.
3. $\sum_{s \in \mathcal{B}} \lambda_s < 1$.

Remark 2.3. *If the set \mathcal{X} has a finite diameter, i.e. $\text{diam}(\mathcal{X}) = \sup_{x, x' \in \mathcal{X}} (\|x - x'\|)$*

and if there are no null set under the measure μ_X , then $\forall t \in \mathbb{Z}^{\kappa}, \|Y_t - Y'_t\|_\infty \leq \text{diam}(\mathcal{X})$. Moreover, $F : \mathcal{X}^{\mathcal{B}} \times E \rightarrow \mathcal{X}$ admits also a finite diameter $\text{diam}(F) = \sup_{x, x' \in \mathcal{X}} (\|F(x, \varepsilon) - F(x', \varepsilon)\|_\infty)$.

Let's assume that $\text{diam}(F) < \text{diam}(\mathcal{X})$. In this case, it is tempting to think that hypothesis (\mathbf{H}_1^∞) is satisfied. However, it may exist couplings of $(Y_t)_{t \in \mathbb{Z}^{\kappa}}$,

$(Y'_t)_{t \in \mathbb{Z}^\kappa}$ with marginal distribution μ_X such that $\|Y_t - Y'_t\|_\infty$ is far smaller than $\text{diam}(\mathcal{X})$. Therefore hypothesis (\mathbf{H}_1^∞) is not necessarily satisfied when the diameter of F is smaller than the diameter of \mathcal{X} .

We have the following relation between those hypotheses and the absolute contraction.

Lemma 2.1. *If F verifies the absolute contraction and if the first condition of the weak (respectively uniform) contraction condition holds, then it verifies the weak (resp. uniform) contraction condition.*

The first condition of weak and uniform contraction condition are immediately verified as soon as \mathcal{X} is bounded. However, this is not a necessary condition, in particular for (\mathbf{H}_1^m) . In particular, its first condition is verified as soon as μ_X is short tail.

The relation between (\mathbf{H}_1^m) and (\mathbf{H}_1^∞) is uneasy to establish. If (\mathbf{H}_1^∞) is verified, then, for every coupling, it exists a rank m_0 such that the condition 2 of the hypothesis (\mathbf{H}_1^m) is verified for every $m > m_0$. However, it is not clear if such a rank m_0 exists for every coupling.

2.2.2. Coupling hypothesis

We introduce a coupling hypothesis similar to the condition used in [16]. It controls the moment of the difference between two independent variables following the same distribution μ_X .

Definition 2.4. COUPLING HYPOTHESIS (\mathbf{H}_2^m) . *Let $m \in \mathbb{N} \cup \{\infty\}$. We suppose that for two \mathcal{X} -valued independent random variables X, Y with marginal distribution μ_X , it exists a constant \mathbb{V}_m such that*

$$\|Y - X\|_m \leq \mathbb{V}_m.$$

This hypothesis is immediately verified as soon as the diameter of \mathcal{X} is finite. Moreover, if (\mathbf{H}_1^m) is verified, so is (\mathbf{H}_2^m) . Nevertheless, we point out that the quantity \mathbb{V}_m plays a important role in our concentration inequality result and may be significantly smaller than the diameter of \mathcal{X} . Consequently, even when the diameter is finite, it may be advantageous to use hypothesis (\mathbf{H}_1^m) and (\mathbf{H}_2^m) .

Remark 2.4. *If $m, m' \in \mathbb{N} \cup \{\infty\}$ and $m' \leq m$, then $(\mathbf{H}_2^m) \implies (\mathbf{H}_2^{m'})$*

2.2.3. Examples

We provide below some examples of non-causal random fields.

Non-causal linear fields Under our framework, the simplest possible non-causal random fields is the bidirectional linear model. In this case $\kappa = 1$, and we have α_{-1}, α_1 such that

$$F(X_{t-1}, X_{t+1}, \varepsilon_t) = \alpha_{-1}X_{t-1} + \alpha_1X_{t+1} + \varepsilon_t.$$

Where the ε_t are a Gaussian white noise of variance σ^2 . In this setting, absolute contraction is satisfied if $\alpha_{-1} + \alpha_1 < 1$.

Remark 2.5. *This is equivalent, when $\alpha_1 \neq 0$ to a classical AR(2) model.*

Bidirectional linear field can be generalize to linear random fields

$$X_t = F((X_{t+i})_{i \in \mathcal{B}}, \varepsilon_t) = \sum_{s \in \mathcal{B}} \alpha_s X_{t+s} + \varepsilon_t.$$

In any case, absolute contraction is satisfied as soon as $\sum_{s \in \mathcal{B}} \alpha_s < 1$.

Finite LARCH random fields Finite LARCH(n) random fields [35] are defined by

$$X_t = F((X_{t-i})_{i \in [1,n]}, \varepsilon_t) = \varepsilon_t \left(\alpha_0 + \sum_{i=1}^n \alpha_i X_{t-i} \right).$$

They can be generalized to non-causal LARCH(n) fields defined by

$$X_t = F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) = \varepsilon_t \left(\alpha_0 + \sum_{s \in \mathcal{B}} \alpha_s X_{t+s} \right).$$

In this case, a sufficient condition to fulfill (\mathbf{H}_1^m) is: $\|\varepsilon_t\|_\infty \sum_{s \in \mathcal{B}} \alpha_s < 1$.

Finite ARCH random fields ARCH models are widely used in econometric and can be easily extended to the non-causal case. Here, we propose a Bi-ARCH(1,1) model defined by

$$\begin{aligned} X_t &= F(X_{t-1}, X_{t+1}, \varepsilon_t) \\ &= \varepsilon_t \sqrt{\alpha_{-1}X_{t-1} + \alpha_1X_{t+1} + \beta}, \text{ where } \varepsilon \text{ is a real random variable.} \end{aligned}$$

In this case, (\mathbf{H}_1^m) is satisfied providing $\|\varepsilon\|_\infty(\alpha_1 + \alpha_{-1}) < 1$. (\mathbf{H}_2^m) is satisfied if ε is bounded and the process stationary. Those processes can be extend to the multidimensional case, such processes verify

$$X_t = F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) = \varepsilon_t \sqrt{\alpha_0 + \sum_{s \in \mathcal{B}} \alpha_s X_{t+s}^2}.$$

Bidirectional RNN Bidirectional recurrent neural networks (BRNN) have been used in Natural Language Processing. They have many applications from text translation to part of speech tagging and speech recognition. Here, we present the formal version of a single-layer bidirectional neural network with a white noise ε_t .

$$F(\mathcal{X}_t, \varepsilon_t) = f(A\mathcal{X}_t) + \varepsilon_t.$$

Where A is a $p \times 2k$ matrix, f is an activation function and \mathcal{X}_t is the $2k$ vector

$$\mathcal{X}_t = (X_{t-k}, \dots, X_{t-1}, X_{t+1}, \dots, X_{t+k}).$$

We suppose that the activation function is 1-Lipschitz. This is the case for most activation function (sigmoid, RELU, softmax). There is also an operator norm $\|\cdot\|_{op,m}$ associated with the norm $\|\cdot\|_m$.

With this condition, the contraction condition (\mathbf{H}_1^m) is verified as soon as

$$\|A\|_{op,m} < 1.$$

If the white noise is bounded, the condition (\mathbf{H}_2^∞) is verified. Instead, if it is subgaussian, we only have (\mathbf{H}_2^m).

2.3. Function of interest Φ and the statistic $S_{\mathcal{I}}$

Throughout this article, we focus on a function $\Phi : \mathcal{X}^{\bar{\mathcal{B}}} \mapsto \mathbb{R}$ defined on a small neighborhood

$$\bar{\mathcal{B}} = \mathcal{V}(\bar{\delta}, 0) = \prod_{i=1}^{\kappa} [-\bar{\delta}_i, \bar{\delta}_i]. \quad (2.4)$$

And we denote by $n_{\bar{\mathcal{B}}}$ the cardinal of $\bar{\mathcal{B}}$. It is important to clarify that definitions of \mathcal{B} and $\bar{\mathcal{B}}$ are slightly different, the point $\{0\}$ is included in the set $\bar{\mathcal{B}}$ and not in \mathcal{B} . Whereas definitions of $n_{\mathcal{B}}$ and $n_{\bar{\mathcal{B}}}$ are comparable, indeed $n_{\mathcal{B}} = \text{card}(\mathcal{B}) + 1$ and $n_{\bar{\mathcal{B}}} = \text{card}(\bar{\mathcal{B}})$.

Then, for a given subset \mathcal{I} of indices, we define the statistic $S_{\mathcal{I}}$

$$S_{\mathcal{I}} = \sum_{s \in \mathcal{I}} \Phi((X_{s+t})_{t \in \bar{\mathcal{B}}}). \quad (2.5)$$

We first recall that the process is *strongly* stationary, thus $\mathbb{E}[\Phi((X_{s+t})_{t \in \bar{\mathcal{B}}})] = \mathbb{E}[S_{\mathcal{I}}]$ is well defined. Our goal is to control the difference between $S_{\mathcal{I}}$ and $\mathbb{E}[S_{\mathcal{I}}]$, which is the deviation of $S_{\mathcal{I}}$. We introduce below some hypotheses concerning this function Φ .

Lipschitz separability hypothesis (\mathbf{H}_3) We assume that Φ is L -Lipschitz separable, i.e., there exists a constant $L > 0$ such that

$$\begin{aligned} &\text{For all } (u_t)_{t \in \bar{\mathcal{B}}}, (v_t)_{t \in \bar{\mathcal{B}}} \text{ tuples of } \mathcal{X}^{\bar{\mathcal{B}}}, \\ &\|\Phi((u_t)_{t \in \bar{\mathcal{B}}}) - \Phi((v_t)_{t \in \bar{\mathcal{B}}})\| \leq L \sum_{t \in \bar{\mathcal{B}}} \|u_t - v_t\|. \end{aligned} \quad (2.6)$$

This hypothesis is close to the condition proposed by [16] in the causal dependent case. Moreover, recent works (see [36]) suggest that such hypothesis is suitable to deal with deep learning models and provide algorithm to estimate the Lipschitz constant.

To simplify, we assume throughout the article that $L = 1$ but setting $L \neq 1$ would only add a multiplicative factor in our results.

Bound for Φ (\mathbf{H}_4) *We assume that Φ is bounded.*

$$\forall x \in \mathcal{X}^{\mathcal{B}}, |\Phi(x)| \leq M. \quad (2.7)$$

This hypothesis is immediately verified when \mathcal{X} is bounded. The case of unbounded \mathcal{X} will be discussed in section 3.4.

3. Main Results

3.1. Concentration inequality within i.i.d. assumption

Numerous concentration inequalities have been established to control the deviation of a random variables. Hoeffding's and McDiarmid's inequalities, proposed respectively in [20] and [30], are among the most widely used in machine learning. Their classical versions are based on the i.i.d. assumption.

Below, we recall the Hoeffding's inequality.

Theorem 3.1 (Hoeffding's inequality). *Let X_1, \dots, X_n be independent real random variables and $S_{\mathcal{I}} = \sum_{i=1}^n X_i$.*

We suppose that there exist two tuples (a_1, \dots, a_n) and (b_1, \dots, b_n) such that with probability 1,

$$\forall i \in [1, n], a_i \leq X_i \leq b_i. \quad (3.1)$$

Then

$$\mathbb{P}(S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}] \geq \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (3.2)$$

In the field of statistical learning, Hoeffding's and McDiarmid's have lots of applications ([9, 28] give examples).

In this article, we aim to prove a Hoeffding-type inequality in the non-causal setting defined in Section 2.

Remark 3.1. *Other types of concentration inequalities (Bernstein, von Bahr-Esseen) could have been proved in the same context using the same approach. Indeed, it should also be possible, as in the i.i.d. case, to derive additional inequalities from the exponential inequality given in Lemma 4.8.*

Nevertheless, we focus on Hoeffding's inequality because of its simplicity and its use in many applications.

3.2. Concentration inequalities and expected deviation bounds for non-causal random fields

We state below simplified versions of our results. Full theorems may be found in Section 5.2.

Theorem 3.2 (Concentration inequality, uniform contraction case). *Let $n_B = \text{Card}(\mathcal{B}) + 1$ and $n_{\bar{B}} = \text{Card}(\bar{\mathcal{B}})$. If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) are verified, there is a constant A such that, for $\varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty$,*

$$\mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq 2 \exp\left(\frac{-2(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty)^2}{(n_{\bar{B}}\mathbb{V}_\infty)^2(1 + An_{\bar{B}}n_B^3\kappa!^2 \lceil \ln(n) \rceil^\kappa n)}\right).$$

With A such that $\Upsilon(d)^2 d^\kappa \leq A(\kappa!)^2 \lceil \ln(n) \rceil^\kappa$, where Υ is the function defined in Lemma 4.6 and can be bounded independently of $n_{\bar{B}}$, m or n .

Constant A does not depend on n_B , $n_{\bar{B}}$ or n and its explicit value it may be found in Theorem 5.1. The dominant term in the denominator is $A\mathbb{V}_\infty^2 n_B^3 n_{\bar{B}}^3 \kappa!^2 \lceil \ln n \rceil^\kappa n$. It is $\mathcal{O}(n(\ln n)^\kappa)$, whereas classical i.i.d. Hoeffding’s inequality are $\mathcal{O}(n)$.

Moreover, this term is strongly affected by the dimension of the random field κ . However, in practice, this constraint is not a problem, because, in most practical cases, $\kappa \in \{1, 2\}$.

Other important factors are parametric dimensions of our model, represented by n_B and $n_{\bar{B}}$. It is logical that the quality of the inequality decreases when the number of variables to control increases. Our inequality is very sensitive to these factors and is therefore not a suitable result for high dimension estimation, i.e., when n is smaller than n_B and $n_{\bar{B}}$. We recall that for a simple Markov chain both of these terms would be equal to 2.

Finally, the condition $\varepsilon \geq 2n_{\bar{B}}\mathbb{V}_\infty$ is not restrictive for applications as we will show in subsection 3.4.

We now give a simplified version of our concentration inequality for the weaker assumptions (\mathbf{H}_1^m) and (\mathbf{H}_2^m) .

Theorem 3.3 (Concentration inequality, weak contraction case). *Let $n_B = \text{Card}(\mathcal{B}) + 1$ and $n_{\bar{B}} = \text{Card}(\bar{\mathcal{B}})$. If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) are verified, there are constants A, B, C, D, E, F, H such that, for $\varepsilon \geq 2F(n_{\bar{B}}n_B)^3 \kappa! \lceil \ln(n) \rceil^{2\kappa} n^{\frac{2}{m}}$.*

$$\begin{aligned} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq & 2 \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - F(n_{\bar{B}}n_B)^3 \kappa! \lceil \ln(n) \rceil^{2\kappa} n^{\frac{2}{m}}\right)^2}{\left(Hn_{\bar{B}}n^{\frac{2}{m}}\right)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}\right) \\ & + \frac{\rho^m}{n} \left(2n_B n_{\bar{B}} C \lceil \ln(n) \rceil^\kappa + \left(\frac{D}{n_B^3 n_{\bar{B}}^2 \Upsilon(d) \ln(n)^{2\kappa}}\right)^m\right). \end{aligned}$$

These constants are

- $A = \frac{1 + \frac{\ln(\rho^{-1})}{\kappa! n_B} + \ln(\rho^{-1}) \frac{\left(\frac{\kappa-1}{e}\right)^{\kappa-1}}{(\kappa-1)!}}{\ln(\rho^{-1})^\kappa}$, then $\Upsilon(d) \leq A\kappa!$, where Υ is the function defined in Lemma 4.6 and can be bounded independently of $n_{\bar{B}}$, m or n .
- $B = \frac{1}{\rho^2} + \frac{1}{\rho(n_{\bar{B}}n_B)^{\frac{1}{m}}} + \frac{2M}{\Upsilon(d)\mathbb{V}_m(n_{\bar{B}}n_B)^2}$.
- $C = \left[\frac{1 - \frac{1}{m}}{\ln(\rho^{-1})} \right]^\kappa$, then $d^\kappa \leq C \lceil \ln(n) \rceil^\kappa$.
- $D = \frac{\rho \ln(\rho^{-1})^{2\kappa}}{2\left(1 - \frac{1}{m}\right)^{2\kappa}}$.
- E such that $\Upsilon(d)^2 d^\kappa \leq E(\kappa!)^2 \lceil \ln(n) \rceil^\kappa$.
- $F = 2BC^2 A \mathbb{V}_m$.
- $H = \frac{\mathbb{V}_m}{\rho}$.

Similar remarks can be made about the dimension κ and on parameters n_B , $n_{\bar{B}}$ as about the previous theorem.

The denominator in the exponential is $\mathcal{O}\left(nn^{\frac{4}{m}}(\ln(n))^\kappa\right)$. Compared to the i.i.d case, there is an additional term in $n^{\frac{4}{m}}(\ln(n))^\kappa$ that we intend to explain. The term $n^{\frac{4}{m}}$ comes from the (\mathbf{H}_1^m) which is less strong than the classical contraction hypothesis. As in the strong contraction case, the term $(\ln(n))^\kappa$ comes from the dependence, and is obtained by similar article for unidirectional dependence. [16].

(\mathbf{H}_1^m) and (\mathbf{H}_2^m) are rather weaker assumptions than (\mathbf{H}_1^∞) and (\mathbf{H}_2^∞) and thus lead to deteriorated concentration inequalities. In fact, in the exponential term, the denominator is asymptotically dominated by $\mathcal{O}\left(n^{1+\frac{4}{m}}(\ln(n))^\kappa\right)$ instead of $\mathcal{O}(n(\ln(n))^\kappa)$ under assumptions (\mathbf{H}_1^∞) and (\mathbf{H}_2^∞) and $\mathcal{O}(n)$ under i.i.d. assumption.

We point out that this theorem only gives interesting results when $m > 4$ (the reason for this becomes clearer in Theorem 3.4).

The two extra additive terms decrease faster than the main term and are not dominant for applications as we will show in subsection 3.4.

These two theorems lead to the next corollaries for the expected deviation.

Corollary 3.1 (Simplified expected deviation bound, uniform contraction case). *We assume (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) . Let A be the constant defined in Theorem 3.2. It holds*

$$\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \leq n_{\bar{B}} \mathbb{V}_\infty \left(2 + \sqrt{\frac{\pi}{2} (1 + An_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)} \right).$$

Corollary 3.2 (Simplified expected deviation bound, weak contraction case). *We assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) . Let A, B, C, D, E, F, H be the constants defined in Theorem 3.3. It holds*

$$\begin{aligned} & \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \\ & \leq 2n_{\bar{B}} H n^{\frac{2}{m}} \sqrt{\frac{\pi}{2} (1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)} + 2F (n_{\bar{B}}n_B)^3 \kappa! \lceil \ln(n) \rceil^{2\kappa} n^{\frac{2}{m}} \\ & + 2\rho^m M \left(2n_B n_{\bar{B}} C \lceil \ln(n) \rceil^\kappa + \left(\frac{D}{n_B^3 n_{\bar{B}}^2 \Upsilon(d) \ln(n)^{2\kappa}} \right)^m \right). \end{aligned}$$

Remark 3.2. *To show how constants occurring in Corollaries 3.2 and 3.3 can be explicitly computed, we provide below a simple example.*

We assume that it exists α_{-1} and α_{+1} such that

$$\forall t \in \mathbb{Z}, X_t = F(X_{t-1}, X_{t+1}, \varepsilon_t) = \alpha_{-1}X_{t-1} + \alpha_{+1}X_{t+1} + \varepsilon_t.$$

This framework corresponds to the most simple bi-linear random field.

Additionally, we introduce a function ϕ and a statistic $S_{\mathcal{I}}$ such that

$$S_{\mathcal{I}} = \sum_{t \in \mathcal{I}} \phi(X_{t-1}, X_{t+1}).$$

In this simple case, $n_B = n_{\bar{B}} = 3$, $\kappa = 1$ and $\rho = \alpha_{-1} + \alpha_{+1}$. We now assume that (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) hold.

A is defined in Theorem 3.2 as a constant such that $\Upsilon(d)^2 d^\kappa \leq A(\kappa!)^2 [\ln(n)]^\kappa$.

We recall from Theorem 4.1 that $\Upsilon(d) \leq v = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}$ and from Theorem 5.1 that $d = \left\lceil \frac{\ln(n)}{\ln(\rho^{-1})} \right\rceil$. Assuming that $n \geq 2$, it follows that

$$\begin{aligned} A &\leq \frac{\left(\frac{1}{3} + \frac{1}{\ln(\rho^{-1})} + 1 \right)^2}{\ln(\rho^{-1})} \left(\frac{1}{\ln(\rho^{-1})} + \frac{1}{\ln(2)} \right) \\ &= \left(\frac{1}{\ln(\rho^{-1})} \right)^4 \left(1 + \frac{4}{3} \ln(\rho^{-1}) \right) \left(1 + \frac{\ln(\rho^{-1})}{\ln(2)} \right). \end{aligned}$$

Therefore, we get the following bound for $\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|]$.

$$\begin{aligned} \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] &\leq G(\rho, \mathbb{V}_\infty, n) \\ &\underset{n \rightarrow \infty}{\sim} \frac{27\mathbb{V}_\infty}{\ln(\rho)^2} \sqrt{\frac{\pi}{2} \left(1 + \frac{4}{3} \ln(\rho^{-1}) \right) \left(1 + \frac{\ln(\rho^{-1})}{\ln(2)} \right) n \ln(n)}. \end{aligned}$$

When hypotheses (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) hold, then the constants occurring in Corollary 3.2 can be similarly explicitly computed. See Corollary 5.3 for instance.

Note that $S_{\mathcal{I}}$ is defined as a sum, and not an average of random variables. Therefore the expected deviation $\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|]$ does not go to 0.

3.3. Comparison with other results

In this section, we compare our hypotheses and results with other frameworks and concentration inequalities in the literature.

Independent and identically distributed case The first concentration inequalities were proved for i.i.d. random variables. We have already presented the classical Hoeffding’s inequality. McDiarmid’s is similar, but somewhat more

general, since it does not refer to a sum of random variables $\sum_{i=1}^n \phi(X_i)$, but to a general random function of random variables $\phi(X_1, \dots, X_n)$.

The requirements for Hoeffding's inequality (see Equation (3.1)) or its equivalent for McDiarmid's inequality, are *uniform* in the sense of [24] (uniform bounded difference assumption). However, some attempts [13, 24] have been made to relax this condition. Instead of assuming such conditions with probability 1, they may only be required with probability $1 - \rho$ (with ρ small). However, relaxing these conditions degrades the quality of the bound.

This can be compared to the difference between the conditions (\mathbf{H}_1^∞) and (\mathbf{H}_1^m) . Indeed, the moment condition implies that the application F is a contracting mapping with high probability.

Markov chain The causal dependence can be easily described using a Markov chain, where the values X_t verify an equation of type $X_t = F(X_{t-1}, \varepsilon_t)$ for an (i.i.d) innovation ε_t . We have borrowed a lot from this approach. Our non-causal master Equation (2.1) has the same form as a classical Markov chain causal equation.

In order to establish concentration inequalities on Markov chains, further assumptions about the behavior of the function F are required. For example, classical hypotheses include conditions on the mixing time and spectral gap [32]. However, these conditions are difficult to translate in a non-causal and multidimensional case. Therefore, we use an approach similar to [2, 16] and introduce a contraction hypothesis. However, as far as we know, our contraction hypotheses (\mathbf{H}_1^∞) and (\mathbf{H}_1^m) are weaker than the absolute contraction hypothesis used in previous works (e.g. [16]). That is why the convergence rate obtained by [16] is better than our own. They obtain a convergence rate in $\mathcal{O}(\sqrt{n})$ instead of the $\mathcal{O}(\sqrt{\ln(n)n})$ of our Corollary 3.1.

Simultaneous auto regressive scheme The simultaneous autoregressive scheme models the evolution of a lattice through time. At each time step, values on the lattice are updated using the previous values. Causal and non-causal versions of such models were studied in [11]. Our hypothesis (\mathbf{H}_1^m) is close to the Stochastic Lipschitz Continuity hypothesis from this article. Our work differs in that we are interested in non-asymptotic results.

Weak dependence for causal time series Another possibility is to use weak dependence (see [15]) to model dependence in time series. It can be applied to more general processes because it accounts for long-range dependence. However, weak dependence conditions are non-local, and therefore more difficult to verify. In the causal case, they lead to a similar rate of convergence as our results under the uniform contraction hypothesis. For example, they are used by [3] for model selection to obtain a convergence rate asymptotically dominated by $\mathcal{O}(\sqrt{\ln^{\frac{5}{2}}(n)n})$.

Weak dependence for non-causal time series For non-causal time series, many results have been obtained using the notion of weak dependence [15]. This

notion replaces our contraction condition (\mathbf{H}_1^∞) or (\mathbf{H}_1^m) . However, it is difficult to compare our setting with these results because their concentration inequalities are not Hoeffding or Mc-Diarmid type inequalities. In particular, they introduce an autocovariance term that does not appear in our concentration inequalities.

Markovian random field Few results are similar to ours for general random fields. The usual framework for treating non-causal random fields is the Markovian random field framework, where the random variables are distributed on a graph, instead of a lattice. To our knowledge, there are no similar concentration results that apply to general Markovian random fields.

An interesting sub-case of Markovian random fields is the Ising model, which is commonly used in physics and imagery [29]. Ising models can be viewed as a special case of our model, where $\mathcal{X} = \{-1, 1\}$ and $\bar{\mathcal{B}} = \{-1, 1\}^\kappa$. The transition is governed by the following equation

$$F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) = 2\mathbb{1} \left(\varepsilon > \frac{\exp(\beta \sum_{s \in \mathcal{B}} X_{t+s})}{\exp(\beta \sum_{s \in \mathcal{B}} X_{t+s}) + \exp(-\beta \sum_{s \in \mathcal{B}} X_{t+s})} \right) - 1,$$

where ε_t follow an uniform distribution in $(0, 1)$. Some concentration results have been established with coupling conditions [10], it leads to polynomial (rather exponential) concentration inequalities. However, it is hard to compare them with our results, because, even if (\mathbf{H}_1^∞) is not satisfied for the Ising model, (\mathbf{H}_1^m) seems reasonable but hard to verify.

3.4. Application to learning theory

In this subsection, we provide an application of our results to learning theory. We adapt the approach of [5], [26] and [27] with our framework, to get an oracle bound for the model selection problem.

3.4.1. Application to the completion problem

A classical problem in learning theory is the completion problem (both in regression or classification). The objective is to predict X_t using its neighbours on a κ dimensional lattice $(X_t)_{t \in \bar{\mathcal{B}}}$. The prediction is given by a model \hat{f}

$$\hat{x}_s = \hat{f}((x_{s+t})_{t \in \bar{\mathcal{B}} \setminus \{0\}}).$$

This completion problem may seem trivial, but it is actually the backbone of many NLP tasks. Indeed, predicting words is often used as a primary task to train the encoder in an encoder-decoder scheme. It creates a language model, i.e., a distribution of probability of a word in a text. This language model is then used for more complex task (translation, text segmentation, question answering etc.).

We also introduce a cost function $c : \mathcal{X}^2 \mapsto \mathbb{R}$. $c(\hat{x}_s, x_s)$ quantifies the error made when the considered algorithm predict \hat{x}_s instead of x_s . In this case, the function of interest Φ introduced earlier corresponds to the cost of \hat{f} on a sample

$$\Phi((X_{s+t})_{t \in \bar{\mathcal{B}}}) = c(\hat{f}((X_{s+t})_{t \in \bar{\mathcal{B}} \setminus \{0\}}), X_s).$$

And

$$S_{\mathcal{I}} = \sum_{s \in \mathcal{I}} \Phi((X_{s+t})_{t \in \bar{\mathcal{B}}}) = \sum_{s \in \mathcal{I}} c(\hat{f}((X_{s+t})_{t \in \bar{\mathcal{B}} \setminus \{0\}}), X_s).$$

Remark 3.3. We emphasize that \mathcal{B} and $\bar{\mathcal{B}}$ may be different in practice. Indeed, in practice, we can not know the size of \mathcal{B} . We may also want to try a simpler model (with $n_{\bar{\mathcal{B}}} < n_{\mathcal{B}}$) to reduce overfitting.

Remark 3.4. The cost function $c : \mathcal{X}^2 \mapsto \mathbb{R}$ can be chosen by the user. Classical choices include cross entropy, Kullback-Leibler divergence, etc. When \mathcal{X} is unbounded, some classical cost function such as quadratic distance must be truncated to satisfy (\mathbf{H}_4) .

Remark 3.5. If (\mathbf{H}_4) is verified with a given upper bound M , it is always possible to bound this cost function by a constant and then to divide this function by this constant.

Therefore, we could assume that $M = 1$ in hypothesis (\mathbf{H}_4) without loss of generality.

On Figure 1, the red neighborhood $\mathcal{V}(\bar{\delta}, t)$ is used to compute $\hat{x}_t = \hat{f}((x_{t+s})_{s \in \bar{\mathcal{B}}})$. However, to compute this in practice, we must know the values of all the x_t in the neighborhood $\mathcal{V}(\bar{\delta}, t)$. It implies that the training set (the blue set) is only a subset of the set of all known values (the grey one)

In the completion problem, we want to control the gap between the theoretical risk of an estimator \hat{f} , denoted by $R(\hat{f})$,

$$R(\hat{f}) = \mathbb{E}[c(\hat{f}((X_{s+t})_{t \in \bar{\mathcal{B}} \setminus \{0\}}), X_s)], \tag{3.3}$$

and its empirical risk,

$$\mathcal{R}_{emp}(\hat{f}) = \frac{1}{n} \sum_{s \in \mathcal{I}} c(\hat{f}((X_{s+t})_{t \in \bar{\mathcal{B}} \setminus \{0\}}), X_s). \tag{3.4}$$

Immediate application of our results gives tail and moment inequalities for $|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|$ under hypotheses (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_1^m) , (\mathbf{H}_2^m) .

Theorem 3.4 (Tail inequalities). Let $n_{\mathcal{B}} = \text{Card}(\mathcal{B}) + 1$ and $n_{\bar{\mathcal{B}}} = \text{Card}(\bar{\mathcal{B}})$.

- If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) are verified, there are constants such that, for $\varepsilon > \frac{2n_{\bar{\mathcal{B}}}\mathbb{V}_\infty}{n}$,

$$\begin{aligned} & \mathbb{P} \left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \geq \varepsilon \right) \\ & \leq 2 \exp \left(\frac{-2n^2 \left(\varepsilon - \frac{2n_{\bar{\mathcal{B}}}\mathbb{V}_\infty}{n} \right)^2}{(n_{\bar{\mathcal{B}}}\mathbb{V}_\infty)^2 (1 + An_{\bar{\mathcal{B}}}n_{\mathcal{B}}^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)} \right). \end{aligned}$$

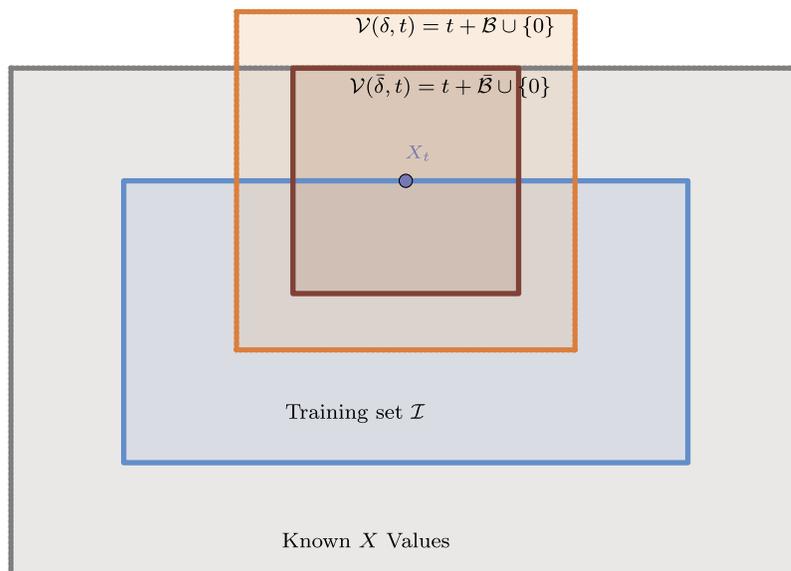


FIG 1. Learning diagram

- If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) are verified, there are constants such that, for $\varepsilon \geq \frac{2L_1(n)}{n^{1-\frac{2}{m}}}$,

$$\begin{aligned} & \mathbb{P}\left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \geq \varepsilon\right) \\ & \leq 2 \exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\varepsilon}{2} - \frac{L_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{(Hn_{\bar{B}})^2(1 + En_{\bar{B}}n_{\bar{B}}^2(\kappa!)^2[\ln(n)]^\kappa n)}\right) + \frac{\rho^m}{n}L_2(n), \end{aligned}$$

with $L_1(n) = F(n_{\bar{B}}n_B)^3 \kappa! [\ln(n)]^{2\kappa}$ and $L_2(n) = 2n_B n_{\bar{B}} C [\ln(n)]^\kappa + \left(\frac{D}{n_B^3 n_{\bar{B}}^2 \Upsilon(d) \ln(n)^{2\kappa}}\right)^m$.

Proof. It is a direct application of Theorems 3.2 and 3.3 with $S_{\mathcal{I}} = \sum_{s \in \mathcal{I}} c(\hat{f}((X_{s+t})_{t \in \bar{B} \setminus \{0\}}, X_s))$. Moreover, the constants involved are the same as in Theorem 3.2 and 3.3. \square

If we assume that $\mathbb{E}[\exp(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)]$ exists, we get the following useful result.

Corollary 3.3.

- If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) are verified. For each $s > 0$,

$$\mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)]$$

$$\begin{aligned} &\leq \exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n} + \frac{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2\lceil\ln(n)\rceil^{\kappa}n)}{8n^2}\right) \\ &\quad \times \left(1 + \frac{n_{\bar{B}}\mathbb{V}_{\infty}s}{n}\sqrt{2\pi(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2\lceil\ln(n)\rceil^{\kappa}n)}\right). \end{aligned} \tag{3.5}$$

- If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) are verified. For each $s > 0$,

$$\begin{aligned} &\mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)] \\ &\leq \exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4sL(n) + (Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_{\bar{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^{\kappa}n)\right)\right) \\ &\quad \times \left(1 + \frac{\exp(Ms)\rho^m L_2(n)}{n} + \frac{2Hn_{\bar{B}}s}{n^{1-\frac{2}{m}}}\sqrt{2\pi(1 + En_{\bar{B}}n_{\bar{B}}^3(\kappa!)^2\lceil\ln(n)\rceil^{\kappa}n)}\right). \end{aligned} \tag{3.6}$$

The proof can be found in the appendix A.

3.4.2. Model selection

In this subsection, we present a model selection result for a finite set of models. It is an extension of well known results ([5]) within our framework. It should also be possible to obtain results for an infinite set of learning function using the Vapnik-Chervonenkis dimension [8].

We consider a finite set \mathcal{F} of models \hat{f} and assume that we have an algorithm that can find the model with the lowest empirical risk. Our goal is to bound the difference between the theoretical risk of the selected model and the theoretical risk of the better model in the set.

We assume here that all models \hat{f} are independent of the dataset used to compute the empirical risk. This hypothesis has been relaxed in previous work [31]. However, in order to provide a simple example, we retain this hypothesis.

Formally, we define: $\tilde{f} = \operatorname{argmin}_{\hat{f} \in \mathcal{F}} (\mathcal{R}_{emp}(\hat{f}))$ and $f^* = \operatorname{argmin}_{\hat{f} \in \mathcal{F}} (R(\hat{f}))$ and we aim to bound $R(\tilde{f}) - R(f^*)$.

Corollary 3.4 (Model selection on dependent non-causal field, tail inequality). *If (\mathbf{H}_1^{∞}) , (\mathbf{H}_2^{∞}) and (\mathbf{H}_3) hold. For $\varepsilon > \frac{4n_{\bar{B}}\mathbb{V}_{\infty}}{n}$,*

$$\mathbb{P}\left(R(\tilde{f}) - R(f^*) > \varepsilon\right) \leq 2N \exp\left(\frac{-2n^2\left(\frac{\varepsilon}{2} - \frac{2n_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)^2}{(n_{\bar{B}}\mathbb{V}_{\infty})^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2\lceil\ln(n)\rceil^{\kappa}n)}\right). \tag{3.7}$$

If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) , hold. For $\varepsilon \geq \frac{4L_1(n)}{n^{1-\frac{2}{m}}}$,

$$\mathbb{P}\left(R(\tilde{f}) - R(f^*) > \varepsilon\right)$$

$$\leq N \left(2 \exp \left(\frac{-2n^{2-\frac{4}{m}} \left(\frac{\varepsilon}{4} - \frac{L_1(n)}{n^{1-\frac{2}{m}}} \right)^2}{(Hn_{\bar{B}})^2 (1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)} \right) + \frac{\rho^m}{n} L_2(n) \right). \quad (3.8)$$

Proof.

$$\begin{aligned} R(\tilde{f}) - R(f^*) &= R(\tilde{f}) - \mathcal{R}_{emp}(\tilde{f}) + \mathcal{R}_{emp}(\tilde{f}) - \mathcal{R}_{emp}(f^*) + \mathcal{R}_{emp}(f^*) - R(f^*) \\ &\leq \left(R(\tilde{f}) - \mathcal{R}_{emp}(\tilde{f}) \right) + \left(\mathcal{R}_{emp}(\tilde{f}) - \mathcal{R}_{emp}(f^*) \right) \\ &\quad + \left(\mathcal{R}_{emp}(f^*) - R(f^*) \right) \\ &\leq 2 \sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| + \left(\mathcal{R}_{emp}(\tilde{f}) - \mathcal{R}_{emp}(f^*) \right) \\ &\leq 2 \sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \quad \text{because } \left(\mathcal{R}_{emp}(\tilde{f}) - \mathcal{R}_{emp}(f^*) \right) \leq 0. \end{aligned}$$

Thus, for all ε ,

$$\begin{aligned} \mathbb{P} \left(R(\tilde{f}) - R(f^*) > \varepsilon \right) &\leq \mathbb{P} \left(\sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| > \frac{\varepsilon}{2} \right) \\ &\leq \sum_{\hat{f} \in \mathcal{F}} \mathbb{P} \left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| > \frac{\varepsilon}{2} \right) \quad \text{by union bound} \\ &\leq N \mathbb{P} \left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| > \frac{\varepsilon}{2} \right). \end{aligned}$$

Then, Theorem 3.4 yields result for both hypotheses. □

Bound for the expectation can also be obtained. We focus here on the case where hypotheses (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) hold.

Corollary 3.5 (Model selection on dependent non-causal field, expectation inequality). *If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) hold,*

$$\begin{aligned} &\mathbb{E}[R(\tilde{f}) - R(f^*)] \\ &\leq \frac{2n_{\bar{B}}\mathbb{V}_\infty}{n} \left(2 + \sqrt{(1 + An_{\bar{B}}n_B^3\kappa!^2 \lceil \ln(n) \rceil^\kappa n)} \left(\sqrt{\frac{\ln(N)}{2}} + \sqrt{2\pi} \right) \right). \end{aligned}$$

We can also provide an asymptotic equivalent,

$$\begin{aligned} \mathbb{E}[R(\tilde{f}) - R(f^*)] &\leq H(n_{\bar{B}}, n_B, \kappa, \rho, N, n) \\ &\underset{n \rightarrow \infty}{\sim} 2\mathbb{V}_\infty \left(\frac{\sqrt{\ln(N)}}{2} + \sqrt{2\pi} \right) n_{\bar{B}}n_B\kappa! \sqrt{An_{\bar{B}}n_B} \sqrt{\frac{\ln(n)^\kappa}{n}}. \end{aligned}$$

Proof. We showed in previous corollary that $R(\tilde{f}) - R(f^*) \leq 2 \sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|$.

Moreover, for all $s > 0$,

$$\begin{aligned} & \mathbb{E}[\sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|] \\ &= \frac{1}{s} \mathbb{E} \left[\ln \left(\sup_{\hat{f} \in \mathcal{F}} \exp \left(s |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \right) \right) \right] \\ &\leq \frac{1}{s} \ln \left(\mathbb{E} \left[\sup_{\hat{f} \in \mathcal{F}} \exp \left(s |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \right) \right] \right) \quad \text{using Jensen inequality} \\ &\leq \frac{1}{s} \ln \left(\mathbb{E} \left[\sum_{\hat{f} \in \mathcal{F}} \exp \left(s |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \right) \right] \right) \\ &= \frac{1}{s} \ln \left(N \mathbb{E} \left[\exp \left(s |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \right) \right] \right). \end{aligned}$$

If (\mathbf{H}_1^∞) , (\mathbf{H}_2) and (\mathbf{H}_3) hold, applying Equation (3.5) from Corollary 3.3. We get

$$\begin{aligned} \mathbb{E}[R(\tilde{f}) - R(f^*)] &\leq \frac{2 \ln(N)}{s} + \frac{4n_{\bar{B}} \mathbb{V}_\infty}{n} + \frac{s(n_{\bar{B}} \mathbb{V}_\infty)^2 (1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)}{4n^2} \\ &\quad + \frac{2n_{\bar{B}} \mathbb{V}_\infty}{n} \sqrt{2\pi (1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)}, \end{aligned}$$

using $\ln(1+x) \leq x$.

Then choosing $s = \frac{2n}{n_{\bar{B}} \mathbb{V}_\infty} \sqrt{\frac{2 \ln(N)}{1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n}}$, we get

$$\begin{aligned} & \mathbb{E}[R(\tilde{f}) - R(f^*)] \\ &\leq \frac{2n_{\bar{B}} \mathbb{V}_\infty}{n} \left(\sqrt{\frac{\ln(N)}{2} (1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)} + 2 \right. \\ &\quad \left. + \sqrt{2\pi (1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)} \right) \\ &\leq \frac{2n_{\bar{B}} \mathbb{V}_\infty}{n} \left(2 + \sqrt{(1 + An_{\bar{B}} n_B^3 \kappa!^2 \lceil \ln(n) \rceil^\kappa n)} \left(\sqrt{\frac{\ln(N)}{2}} + \sqrt{2\pi} \right) \right). \quad \square \end{aligned}$$

Remark 3.6. In the case where $\mathbb{E} \left[\exp \left(s |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \right) \right]$ does not exist, or if we only have (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) , it is possible to use a simpler bound. It may be obtained noticing

$$\mathbb{E}[R(\tilde{f}) - R(f^*)] \leq \mathbb{E}[\sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|] \leq N \mathbb{E}[|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|]$$

and using Corollaries 3.1 or 3.2. If we do so, learning bound increase with N instead of $\sqrt{\ln(N)}$. If we assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) , we then get

asymptotic bound

$$\mathbb{E}[R(\tilde{f}) - R(f^*)] \leq H(n_{\bar{B}}, n_{\mathcal{B}}, \kappa, \rho, N, n) \underset{n \rightarrow \infty}{\sim} \frac{H\kappa! \sqrt{2E\pi(n_{\bar{B}}n_{\mathcal{B}})^3 \ln(n)^\kappa}}{n^{\frac{1}{2} - \frac{2}{m}}}.$$

Where H and E are constants defined in Corollary 3.2.

Remark 3.7. In [27], a bound is provided for the expected maximal deviation in the i.i.d case,

$$\mathbb{E}[\sup_{\hat{f} \in \mathcal{F}} |R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|] \leq \sqrt{\frac{\ln(2N)}{2n}}.$$

Under our setting, with hypotheses (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) , we obtained comparable bound (Corollary 3.5) with an extra term $\sqrt{\ln(n)^\kappa}$.

4. Approximation

In Sections 4 and 5, we will prove the results of section 3. To this end, we introduce a useful approximation of non-causal random fields. [18] has already shown that a non-causal solution (X_t) of (2.1) can be expressed as a function of an infinite number of i.i.d. random variables. We present an approximation of (X_t) by a function of a finite number of i.i.d. random variables.

4.1. Approximation of X_t

4.1.1. Notations

Let's first recall and introduce some notations to handle random fields.

- We already introduce the notation $\boldsymbol{\varepsilon} = (\varepsilon_t)_{t \in \mathbb{Z}^\kappa}$. Similarly, we use the notation $\boldsymbol{\varepsilon}' = (\varepsilon'_t)_{t \in \mathbb{Z}^\kappa}$ where ε'_t are also random variables $\Omega \mapsto E$.
- For $s \in \mathbb{Z}^\kappa$, θ_s will be the s -shift operator, i.e. for $s \in \mathbb{Z}^\kappa$, $\theta_s((\varepsilon_t)_{t \in \mathbb{Z}^\kappa}) = (\varepsilon_{s+t})_{t \in \mathbb{Z}^\kappa}$.
- For $s \in \mathbb{Z}^\kappa$, we denote by $\boldsymbol{\varepsilon}_s$ the s -shifted random field, ie $\boldsymbol{\varepsilon}_s = \theta_s(\boldsymbol{\varepsilon}) = \theta_s((\varepsilon_t)_{t \in \mathbb{Z}^\kappa}) = (\varepsilon_{s+t})_{t \in \mathbb{Z}^\kappa}$.

4.1.2. Exact reconstruction

Theorem 1 from [18] ensures, under suitable conditions on F , the existence and uniqueness of a function H such that, for each $t \in \mathbb{Z}^\kappa$

$$X_t = H(\boldsymbol{\varepsilon}_t). \tag{4.1}$$

We make two comments on this result.

- This theorem provides an expression of each X_t according to an infinite number of i.i.d. random variables (the whole random field $\boldsymbol{\varepsilon}$). However many concentration inequalities involve only a finite number of random variable.
- This theorem relies on an absolute contraction hypothesis on F (similar to Equation 2.3) which is a stronger assumption than (\mathbf{H}_1^∞) and (\mathbf{H}_1^m) .

For those two reasons, we cannot use this *exact* reconstruction of the solution X_t of Equation (2.1).

4.1.3. Intuition

The idea is to approximate each X_t by another random variable \tilde{X}_t which, similarly to $H(\boldsymbol{\varepsilon})$, depends on the innovation $\boldsymbol{\varepsilon}$. However, unlike $H(\boldsymbol{\varepsilon})$, we will only use a finite number of random variable ε_t . The ones that are located in a finite neighborhood surrounding X_t .

4.1.4. Definition

Let's recall some notations and formally define the function $\tilde{H}^{[d]}$.

- In the equation $\forall t \in \mathbb{Z}^\kappa, X_t = F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t)$ (Equation (2.1)), \mathcal{B} is a κ -orthotope defined by $\mathcal{B} = \mathcal{V}(\delta, 0) \setminus \{0\} = \prod_{i=1}^\kappa [-\delta_i, \delta_i] \setminus \{0\}$.
- We define the dilatation by d of the κ -orthotope $\mathcal{V}(\delta, 0)$ centered on s . With our notations, it corresponds to the κ -orthotope $\mathcal{V}(d\delta, s)$ defined for all d in \mathbb{N} and s in \mathbb{Z}^κ by

$$\mathcal{V}(d\delta, s) = \{t = (t_1, \dots, t_\kappa) \in (\mathbb{Z})^\kappa / \forall i, s_i - d\delta_i \leq t_i \leq s_i + d\delta_i\}. \quad (4.2)$$

- We define recursively the function $\tilde{H}^{[d]}$ by

$$\tilde{H}^{[d]}(X, (\varepsilon_t)_{t \in \mathcal{V}(d\delta, s)}) = \begin{cases} X & \text{if } d = 0, \\ F\left(\left(\tilde{H}^{[d-1]}(X, (\varepsilon_u)_{u \in \mathcal{V}(\delta(d-1), t+s)})\right)_{s \in \mathcal{B}}, \varepsilon_s\right) & \text{else.} \end{cases} \quad (4.3)$$

- Eventually, we are able to define the approximation.

$$\forall d \in \mathbb{N}, \forall t \in \mathbb{Z}^\kappa, \tilde{X}_t^{[d]} = \tilde{H}^{[d]}(\bar{X}, (\varepsilon_s)_{s \in \mathcal{V}(d\delta, t)}). \quad (4.4)$$

Where \bar{X} is an independant random variable draw sampled with distribution μ_X .

We can reformulate Equation (4.3) using the notation $\tilde{X}_t^{[d]}$.

$$\forall d \in \mathbb{N}, \forall t \in \mathbb{Z}^\kappa, \tilde{X}_t^{[d]} = \begin{cases} \bar{X} & \text{if } d = 0, \\ F\left(\left(\tilde{X}_{t+s}^{[d-1]}\right)_{s \in \mathcal{B}}, \varepsilon_t\right) & \text{else.} \end{cases}$$

In this way, $\tilde{X}_t^{[d]}$ is an approximation of X_t involving random variables ε_t which belong to the finite neighborhood $\mathcal{V}(d\delta, t)$. Outside, of this neighborhood, we complete the approximation with a random variable \tilde{X} draw from the law μ_X and independent from $(X_t)_{t \in \mathbb{Z}^k}$ and ε .

Remark 4.1. *We emphasize that if the function H from Theorem 1 of [18] exists and is unique, then $\lim_{d \rightarrow \infty} \tilde{H}^{[d]}(X, (\varepsilon_t)_{t \in \mathcal{V}(d\delta, t)}) = H(\varepsilon)$.*

Nevertheless this limit might not exist and not be unique, but even in this case, for all finite d , the approximation $\tilde{H}^{[d]}(X, (\varepsilon_t)_{t \in \mathcal{V}(d\delta, t)})$ is always defined.

4.1.5. Approximation error

The approximation $\tilde{X}_t^{[d]}$ is useful only if we are able to control the approximation error. That is the goal of the two following Lemma.

Lemma 4.1. *Let $m \in \mathbb{N} \cup \{\infty\}$. If (\mathbf{H}_1^m) and (\mathbf{H}_2^m) are verified, then*

$$\forall t \in \mathbb{Z}^k, \forall d \in \mathbb{N}, \|X_t - \tilde{X}_t^{[d]}\|_m \leq \rho^d \mathbb{V}_m.$$

The proof can be found in the appendix B. The lemma is the key to control the quality of our approximation.

Earlier, we define the statistic $S_{\mathcal{I}}$ (see Equation (2.5)), which depends on random variables $(X_t)_{t \in \mathcal{I}}$. We now introduce an approximation of the statistic $S_{\mathcal{I}}$ that we call $\tilde{S}_{\mathcal{I}}^{[d]}$ relying on approximations $\tilde{X}_t^{[d]}$ for t in \mathcal{I} .

We recall that

$$S_{\mathcal{I}} = \sum_{t \in \mathcal{I}} \Phi((X_{t+s})_{s \in \bar{\mathcal{B}}}). \tag{4.5}$$

Similarly, we define

$$\tilde{S}_{\mathcal{I}}^{[d]} = \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}}). \tag{4.6}$$

Using Lemma 4.1, we are able to control the difference between $S_{\mathcal{I}}$ and $\tilde{S}_{\mathcal{I}}^{[d]}$. This is the purpose of the following corollary.

Corollary 4.1 (Moment inequality for the approximation error). *Let $m \in \mathbb{N} \cup \{\infty\}$. We assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) and (\mathbf{H}_3) . We recall that we denote $\text{Card}(I)$ by n and $\text{Card}(\bar{\mathcal{B}})$ by $n_{\bar{\mathcal{B}}}$. Then it holds:*

$$\|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq nn_{\bar{\mathcal{B}}}\rho^d \mathbb{V}_m. \tag{4.7}$$

Proof. By definition of $S_{\mathcal{I}}$ and $\tilde{S}_{\mathcal{I}}^{[d]}$ (Equations (4.5) and (4.6)), using (\mathbf{H}_3) , it holds:

$$\|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq \sum_{t \in \mathcal{I}} \|\Phi((X_{t+s})_{s \in \bar{\mathcal{B}}}) - \Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}})\|_m \leq \sum_{t \in \mathcal{I}} \sum_{s \in \bar{\mathcal{B}}} \|X_{t+s} - \tilde{X}_{t+s}^{[d]}\|_m.$$

Then, using Lemma 4.1

$$\|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq nn_{\bar{\mathcal{B}}}\rho^d \mathbb{V}_m. \tag{4.8} \quad \square$$

4.2. Concentration inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$

In this subsection, we will establish a concentration inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$. Indeed, $\tilde{S}_{\mathcal{I}}^{[d]}$ may be seen as a function of a finite number of random variables, thus it is possible to establish a McDiarmid’s inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$.

More precisely, we aim to show the following theorem.

Theorem 4.1 (Tail inequality for the approximation). *Let $d \in \mathbb{N}$. If we assume (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3)*

$$\forall \varepsilon > 0, \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E} [\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) \leq 2 \exp \left(\frac{-2\varepsilon^2}{(n_{\bar{B}} \mathbb{V}_\infty)^2 \left(n^2 \rho^{2d} + N_2 (n_B \Upsilon(d))^2 \right)} \right). \tag{4.8}$$

If we assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_2) and (\mathbf{H}_3) , $\forall (t_1, t_2) > 0$,

$$\forall \varepsilon > 2npM + p\bar{c},$$

$$\mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E} [\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) \leq 2 \left(p + \exp \left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2} \right) \right), \tag{4.9}$$

with

$$p \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}}n_B \mathbb{V}_m \Upsilon(d)}{t_2} \right)^m \text{ and } \bar{c} = t_1 + N_2 t_2,$$

and

$$\Upsilon(d) \text{ such that } \Upsilon(d) \leq v = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa \left(\frac{\kappa - 1}{\ln(\rho^{-1})e} \right)^{\kappa-1} \underset{\kappa \rightarrow \infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^\kappa}.$$

This theorem is a direct application of a McDiarmid-type inequality. This type of inequality holds if the random variable we want to control is a function of a finite number of *independent* random variables, which is the case here. Indeed, $\tilde{S}_{\mathcal{I}}^{[d]}$ can be expressed as a function of a finite number of random variables $\tilde{X}_t^{[d]}$, and each of these variables $\tilde{X}_t^{[d]}$ is itself a function of a finite number of innovations ε_t and the independent random variable \bar{X} .

We need to present two things necessary to apply a McDiarmid inequality:

- The necessary counting element to count the number of independent random variables that appears in $\tilde{S}_{\mathcal{I}}^{[d]}$.
- The precise version of the McDiarmid inequality we are going to use.

4.2.1. Counting random variables

The rest of this section will require to count the number of random variables that occur in $\tilde{S}_{\mathcal{I}}^{[d]}$. There are two types of random variables we want to count. The first is the number of approximate random variables $\tilde{X}_t^{[d]}$, the second is the number of innovations ε_t . For this purpose, we introduce the following cardinal number:

- $n = \text{Card}(\mathcal{I})$, the number of variables X_t appearing in $S_{\mathcal{I}}$ (see Section 2.3).
- $n_B = \text{Card}(\mathcal{B} \cup \{0\}) = \text{Card}(\mathcal{B}) + 1$, the number of random variables in the κ -orthotope \mathcal{B} and the point $\{0\}$.
- $n_{\bar{B}} = \text{Card}(\bar{\mathcal{B}})$ the number of components of Φ (see Section 2.3).
- $n_d = \text{Card}(\mathcal{V}(d\delta, t))$ the number of innovations ε_t of $\boldsymbol{\varepsilon}$ used in the approximation $\tilde{X}_t^{[d]}$ (see Equation 4.4).
- $N_1 = \text{Card}(\bigcup_{t \in \mathcal{I}} (\mathcal{V}(\bar{\delta}, t))) = \text{Card}(\bigcup_{t \in \mathcal{I}} (\bar{\mathcal{B}} + t))$ the number of approximate random variables $\tilde{X}_t^{[d]}$ appearing in $\tilde{S}_{\mathcal{I}}^{[d]}$.
- $N_2 = \text{Card}(\bigcup_{t \in \mathcal{I}} (\bigcup_{s \in \bar{\mathcal{B}}+t} (\mathcal{V}(d\bar{\delta}, t))))$ number of innovations ε_t appearing in $\tilde{S}_{\mathcal{I}}^{[d]}$.

There are some relations between these numbers.

Lemma 4.2.

- $n \leq N_1 \leq nn_{\bar{B}}$ and $N_1 \leq N_2 \leq N_1 n_d$.
- $n_d = \text{Card}(\mathcal{V}(d\delta, t)) = \prod_{i=1}^{\kappa} (2d\delta_i + 1) \leq d^{\kappa} \prod_{i=1}^{\kappa} (2\delta_i + \frac{1}{d}) \leq d^{\kappa} n_B$.

Higher bounds for N_1 and N_2 are reached when the above unions involve pairwise disjoint sets. Nevertheless, it is not often the case in practice. For example, in machine learning settings, the training and validation sets are usually connected spaces. Therefore, adding more hypotheses about the topology of \mathcal{I} should improve these bounds.

4.2.2. An extension of McDiarmid's inequality

In order to obtain a concentration inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$, we need a special McDiarmid's inequality. Indeed, if hypotheses (\mathbf{H}_1^m) and (\mathbf{H}_2^m) are verified without absolute contraction, the uniform difference-bound hypothesis (as defined in [24]) is not satisfied, thus classical McDiarmid's inequality [30] does not hold.

Therefore, we need an extended version of McDiarmid's inequality that holds even if the bounded difference hypothesis is verified only with high probability. There are several results of this type ([13, 22, 24, 37]). Here, we have chosen to use the extended McDiarmid's inequality from [13].

We are under the “ A -difference bounded” assumption which corresponds to Assumption 1.2. from [13] and the extension of McDiarmid's inequality (Theorem 2.1. from [13]).

Definition 4.1 (A -difference-bound). $f : \prod_{i=1}^N \Omega_i \rightarrow \mathbb{R}$ is p -difference bounded by $(c_j)_{j \in [1, N]}$ if and only if.

- It exists $A \subset \prod_{i=1}^N \Omega_i$.
- For $(w, w') \in A^2$ such that w and w' differ only in the j -th coordinate, $|f(w) - f(w')| \leq c_j$.

Lemma 4.3 (*p*-difference-bound McDiarmid, Theorem 2.1. from [13]). *If f is A-difference bounded by $(c_j)_{j \in [1, N]}$.*

$\forall t > 0$,

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)|A]| \geq t) \leq 2 \left(p + \exp \left(\frac{-2(t - p\bar{c})^2}{\sum_{j=1}^N c_j^2} \right) \right),$$

with $\bar{c} = \sum_{j=1}^N c_j$ and $p = 1 - \mathbb{P}(A)$.

4.2.3. Difference bound for $\tilde{S}_{\mathcal{I}}^{[d]}$

To check these assumptions (strong difference bound or A-difference bound), we need to bound the difference between the statistic $\tilde{S}_{\mathcal{I}}^{[d]}$ and the statistic $\tilde{S}_{\mathcal{I}}^{\prime [d]}$ when one of its composing random variables is replaced by an i.i.d copy. To deal with this case, we need to introduce further notations.

We recall that

$$\tilde{S}_{\mathcal{I}}^{[d]} = \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}}) = \sum_{t \in \mathcal{I}} \Phi(\tilde{H}^{[d]}(\bar{X}, (\varepsilon_s)_{s \in \mathcal{V}(d\delta, t)})).$$

Two types of random variables are involved in $\tilde{S}_{\mathcal{I}}^{[d]}$.

- The marginal variables ε_s (for $s \in \bigcup_{t \in \mathcal{I}} \mathcal{V}(d\delta, t)$).
- The (unique) variable \bar{X} .

We introduce the following notations.

- For all t in \mathbb{Z}^κ , $\tilde{X}_t^{[d]'} = \tilde{H}^{[d]}(\bar{X}', (\varepsilon_s)_{s \in \mathcal{V}(d\delta, t)})$ where \bar{X}' is drawn from the law μ_X and independent from $(X_t)_{t \in \mathbb{Z}^\kappa}$, $\boldsymbol{\varepsilon}$ and \bar{X} .
- For all t in \mathbb{Z}^κ and for all i in $\mathcal{V}(d\bar{\delta}, t)$, $\tilde{X}_t^{[d, i]} = \tilde{H}^{[d]}(\bar{X}, (\varepsilon_s)_{s \in \mathcal{V}(d\delta, t) \setminus \{i\}} \cup \varepsilon'_i)$ where ε'_i is drawn from the law μ_ε and independent from $(X_t)_{t \in \mathbb{Z}^\kappa}$, $\boldsymbol{\varepsilon}$ and \bar{X} .

For all i in \mathbb{Z}^κ not in $\mathcal{V}(d\bar{\delta}, t)$, we define $\tilde{X}_t^{[d, i]} = \tilde{X}_t^{[d]}$.

- For all t in \mathbb{Z}^κ , $\tilde{S}_{\mathcal{I}}^{[d]'} = \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]'})_{s \in \bar{\mathcal{B}}})$.
- For all t in \mathbb{Z}^κ and for all i in $\bigcup_{t \in \mathcal{I}} \mathcal{V}(d\bar{\delta}, t)$, $\tilde{S}_{\mathcal{I}}^{[d, i]} = \sum_{t \in \mathcal{I}} \Phi(\tilde{X}_t^{[d, i]})$.

Lemma 4.4. *Let $m \in \mathbb{N} \cup \{\infty\}$. If (\mathbf{H}_1^m) and (\mathbf{H}_2^m) are verified,*

- For all t in \mathbb{Z}^κ , $\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d]'}\|_m \leq \rho^d \mathbb{V}_m$.
- For all t and i in \mathbb{Z}^κ .
 - If $i \notin \mathcal{V}(d\bar{\delta}, t)$, $\tilde{X}_t^{[d]} = \tilde{X}_t^{[d, i]}$.
 - Otherwise, there is a unique $c \in [0, d]$ such that $i \in \mathcal{V}(c\delta, t)$ and $i \notin \mathcal{V}((c-1)d\delta, t)$ and $\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d, i]}\|_m \leq \rho^c \mathbb{V}_m$.

Proof. We assume (\mathbf{H}_1^m) and (\mathbf{H}_2^m) .

- The demonstration of the first point is the same as Lemma 4.1.

- Firstly, if $i \notin \mathcal{V}(d\delta, t)$, by definition of $\tilde{X}_t^{[d,i]}$, $\tilde{X}_t^{[d,i]} = \tilde{X}_t^{[d]}$.
 Secondly, if $i \in \mathcal{V}(d\delta, t)$, we can write $\mathcal{V}(d\delta, t) = \bigcup_{c=0}^d \mathcal{V}(c\delta, t) \setminus \mathcal{V}((c-1)\delta, t)$ with $\mathcal{V}(0, t) = \{t\}$ and $\mathcal{V}(-1, t) = \emptyset$.
 Moreover, for all $c_1 \neq c_2$, $\mathcal{V}(c_1\delta, t) \setminus \mathcal{V}((c_1 - 1)\delta, t) \cap \mathcal{V}(c_2\delta, t) \setminus \mathcal{V}((c_2 - 1)\delta, t) = \emptyset$.
 Consequently, it exists an unique c in $[0, d]$ such that $i \in \mathcal{V}(c\delta, t)$ and $i \notin \mathcal{V}((c - 1)\delta, t)$. Then with the same argument as Lemma 4.1, we can show that $\|\tilde{X}_t^{[d]} - \tilde{X}_t^{[d,i]}\|_m \leq \rho^c \mathbb{V}_m$. \square

We can now bound the difference between $\tilde{S}_{\mathcal{I}}^{[d]}$ and $\tilde{S}_{\mathcal{I}'}^{[d]}$ and $\tilde{S}_{\mathcal{I}}^{[d]}$ and $\tilde{S}_{\mathcal{I}}^{[d,i]}$.

Lemma 4.5 (Difference bound for $\tilde{S}_{\mathcal{I}}^{[d]}$). *Let $m \in \mathbb{N} \cup \{\infty\}$. If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) and (\mathbf{H}_3) , then*

$$\|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}'}^{[d]}\|_m \leq nn_{\bar{B}}\rho^d \mathbb{V}_m.$$

And,

$$\|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d,i]}\|_m \leq n_{\bar{B}} \mathbb{V}_m \left(1 + n_B \kappa \sum_{c=1}^d c^{\kappa-1} \rho^c \right).$$

The proof of this lemma can be found in the appendix C.

In previous Lemma, the quantity $\sum_{c=1}^d c^{\kappa-1} \rho^c$ occurs, we show in the next Lemma that we can bound this quantity independently of d.

Lemma 4.6. *If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , and (\mathbf{H}_3) . For all $d \in \mathbb{N}$, it holds*

$$\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d).$$

The function Υ is defined by

$$\Upsilon(d) = \begin{cases} \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa - 1)!}{\ln(\rho^{-1})^\kappa} \left(1 - \rho^{d+1} \sum_{i=0}^{\kappa-1} \frac{((d+1) \ln(\rho^{-1}))^i}{i!} \right) \right), \\ \text{if } d < \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa - 1)!}{\ln(\rho^{-1})^\kappa} \left(1 - e^{-(\kappa-1)} \sum_{i=0}^{\kappa-1} \frac{\left(\left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor \ln(\rho^{-1}) \right)^i}{i!} \right) \right) \\ + \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}, \\ \text{if } d = \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa - 1)!}{\ln(\rho^{-1})^\kappa} \left(1 - \rho^d \sum_{i=0}^{\kappa-1} \frac{(d \ln(\rho^{-1}))^i}{i!} \right) \right) + \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}, \\ \text{if } d > \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \end{cases}$$

With

$$\Upsilon(d) \leq v = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa \left(\frac{\kappa - 1}{\ln(\rho^{-1})e} \right)^{\kappa-1}.$$

We point out that the constant v is independent from d and $v \underset{\kappa \rightarrow \infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^\kappa}$ (using Stirling formula).

The proof of this lemma can be found in the appendix D.

4.2.4. McDiarmid’s inequality

We can now use McDiarmid’s inequality for $\tilde{S}_{\mathcal{I}}^{[d]}$ and prove Theorem 4.1.

Here, we distinguish two cases. Either, we have (\mathbf{H}_1^∞) and (\mathbf{H}_2^∞) , and we can use a classical version of McDiarmid’s inequality. The result is presented in Lemma 4.7. Either, we have only (\mathbf{H}_1^m) and (\mathbf{H}_2^m) for a finite value of m . In this case, the standard bounded difference hypothesis is not satisfied, thus we will use the notion of “ A -difference bound” and apply Lemma 4.3. The result is presented in Lemma 4.8.

Lemma 4.7. *If $(\mathbf{H}_1^\infty), (\mathbf{H}_2^\infty)$ and (\mathbf{H}_3) . Let $d \in \mathbb{N}$. It holds*

$$\begin{aligned} \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E} [\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) &\leq 2 \exp \left(-2 \frac{\varepsilon^2}{(nn_{\bar{B}}\rho^d \mathbb{V}_\infty)^2 + \sum_{i=1}^{N_2} (n_{\bar{B}}n_B \mathbb{V}_\infty \Upsilon(d))^2} \right) \\ &\leq 2 \exp \left(\frac{-2\varepsilon^2}{(n_{\bar{B}} \mathbb{V}_\infty)^2 (n^2 \rho^{2d} + N_2 (n_B \Upsilon(d))^2)} \right). \end{aligned}$$

Proof. Using Lemma 4.5 and 4.6 with hypotheses (\mathbf{H}_1^∞) and (\mathbf{H}_2^∞) , we get

- $|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d]'}| \leq c_i$ with $c_i = nn_{\bar{B}}\rho^d \mathbb{V}_\infty$ almost surely.
- $\forall i \in \bigcup_{t \in \mathcal{I}} \mathcal{V}(d\bar{\delta}, t), |\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d,i]}| \leq c_i$ with $c_i = n_{\bar{B}}n_B \mathbb{V}_\infty \Upsilon(d)$. almost surely.

Applying McDiarmid’s inequality yields the result. \square

Lemma 4.8. *Let $m \in \mathbb{N}$. If we assume $(\mathbf{H}_1^m), (\mathbf{H}_2^m), (\mathbf{H}_3)$ and (\mathbf{H}_4) . Let $d \in \mathbb{N}$, and $t_1, t_2 > 0$. There is $p \in (0, 1)$, such that, for $\varepsilon \geq 2npM + p\bar{c}$.*

$$\mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E} [\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) \leq 2 \left(p + \exp \left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2} \right) \right),$$

with

$$p \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}}n_B \mathbb{V}_m \Upsilon(d)}{t_2} \right)^m \text{ and } \bar{c} = t_1 + N_2 t_2.$$

Proof. First, we establish a A -difference bound (Definition 4.1). Let $t_1, t_2 > 0$, using Lemma 4.5, 4.6 and Markov's inequality, it holds

$$\forall t_1 > 0, \mathbb{P}(|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d]'}| \geq t_1) \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t_1} \right)^m.$$

Then, for all $i \in \bigcup_{t \in \mathcal{I}} \mathcal{V}(d\bar{\delta}, t)$, for all $t_2 > 0, \mathbb{P}(|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d,i]}| \geq t_2) \leq \left(\frac{n_{\bar{B}}n_{\bar{B}}\mathbb{V}_m\Upsilon(d)}{t_2} \right)^m$. We define the event $A = \bigcup_{i \in \bigcup_{t \in \mathcal{I}} \mathcal{V}(d\bar{\delta}, t)} (|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d,i]}| > t_2) \cup (|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d]'}| > t_1)$ and $p = 1 - \mathbb{P}(A)$.

By union bound, it holds $p \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}}n_{\bar{B}}\mathbb{V}_m\Upsilon(d)}{t_2} \right)^m$. We also define $\bar{c} = t_1 + N_2 t_2$. Therefore, applying Theorem 2.1 from [13], we get

$$\forall \varepsilon > 0, \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A] | \geq \varepsilon \right) \leq 2 \left(p + \exp \left(\frac{-2(\varepsilon - p\bar{c})^2}{t_1^2 + N_2 t_2^2} \right) \right).$$

Moreover

$$\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}] = \mathbb{P}(A)\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A] + \mathbb{P}(\bar{A})\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|\bar{A}].$$

Then

$$\begin{aligned} |\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}] - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A]| &= (1 - \mathbb{P}(A)) |\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|\bar{A}] - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A]| \\ &\leq p2nM \text{ with } M = \|\Phi\|_{\infty} \text{ (see Hypothesis } (\mathbf{H}_4)). \end{aligned}$$

Eventually, we get

$$\begin{aligned} \forall \varepsilon > 2npM, \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) \\ &= \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A] + \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A] - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon \right) \\ &\leq \mathbb{P} \left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}|A]| \geq \varepsilon - 2npM \right) \\ &\leq 2 \left(p + \exp \left(\frac{-2(\varepsilon - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2} \right) \right). \quad \square \end{aligned}$$

5. Concentration inequalities for $S_{\mathcal{I}}$

Now we are finally able to prove the result of section 3 and prove concentration inequalities for $S_{\mathcal{I}}$. To do this, we use the concentration inequalities from the previous section (with $\tilde{S}_{\mathcal{I}}^{[d]}$) and Lemma 4.1.

5.1. General concentration inequalities

Lemma 5.1 (General concentration inequality for $S_{\mathcal{I}}$). *Let $d \in \mathbb{N}$. If (\mathbf{H}_1^{∞}) , (\mathbf{H}_2^{∞}) and (\mathbf{H}_3) are verified, the following relation holds for $\varepsilon > 2nn_{\bar{B}}\rho^d \mathbb{V}_{\infty}$.*

$$\mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq 2 \exp \left(\frac{-2(\varepsilon - 2nn_{\bar{B}}\rho^d \mathbb{V}_{\infty})^2}{(n_{\bar{B}}\mathbb{V}_{\infty})^2 (n^2 \rho^{2d} + N_2 (n_{\bar{B}}\Upsilon(d))^2)} \right). \quad (5.1)$$

Let $m \in \mathbb{N}$. If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) , are verified. For $t_1, t_2 > 0$, there is $p \in (0, 1)$ such that for $\varepsilon > 4npM + 2p\bar{c}$.

$$\begin{aligned} & \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \\ & \leq 2 \left(p + \exp \left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2} \right) \right) + \left(\frac{2nn_{\bar{B}}\rho^d \mathbb{V}_m}{\varepsilon} \right)^m, \end{aligned} \quad (5.2)$$

with

$$p \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}}n_B \mathbb{V}_m \Upsilon(d)}{t_2} \right)^m \quad \text{and } \bar{c} = t_1 + N_2 t_2.$$

Proof. If (\mathbf{H}_1^∞) and (\mathbf{H}_3) are verified Using Corollary 4.1, we have almost surely:

$$\begin{aligned} |S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| &= |S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]} + \tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}] + \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}] - \mathbb{E}[S_{\mathcal{I}}]| \\ &\leq |S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}| + |\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}]| + |\mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}] - \mathbb{E}[S_{\mathcal{I}}]| \\ &\leq |\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}]| + 2nn_{\bar{B}}\rho^d \mathbb{V}_\infty. \end{aligned}$$

Therefore, for $\varepsilon > 2nn_{\bar{B}}\rho^d \mathbb{V}_\infty$.

$$\begin{aligned} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) &\leq \mathbb{P}\left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[\tilde{S}_{\mathcal{I}}^{[d]}]| \geq \varepsilon - 2nn_{\bar{B}}\rho^d \mathbb{V}_\infty\right) \\ &\leq 2 \exp \left(\frac{-2(\varepsilon - 2nn_{\bar{B}}\rho^d \mathbb{V}_\infty)^2}{(n_{\bar{B}} \mathbb{V}_\infty)^2 (n^2 \rho^{2d} + N_2 (n_B \Upsilon(d))^2)} \right). \end{aligned}$$

If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) and (\mathbf{H}_3) are verified Using Corollary 4.1 and Markov inequality,

$$\forall t > 0, \mathbb{P}\left(|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}| \geq t\right) \leq \left(\frac{nn_{\bar{B}}\rho^d \mathbb{V}_m}{t} \right)^m.$$

Then

$$\begin{aligned} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) &= \mathbb{P}\left(|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]} + \tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}| + |\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(|S_{\mathcal{I}} - \tilde{S}_{\mathcal{I}}^{[d]}| \geq \frac{\varepsilon}{2}\right) + \mathbb{P}\left(|\tilde{S}_{\mathcal{I}}^{[d]} - \mathbb{E}[S_{\mathcal{I}}]| \geq \frac{\varepsilon}{2}\right) \\ &\leq 2 \left(p + \exp \left(\frac{-2(\frac{\varepsilon}{2} - (p\bar{c} + 2npM))^2}{t_1^2 + N_2 t_2^2} \right) \right) \\ &\quad + \left(\frac{2nn_{\bar{B}}\rho^d \mathbb{V}_m}{\varepsilon} \right)^m. \quad \square \end{aligned}$$

Remark 5.1. The choice of the parameter d is a trade-off between the quality of the approximation and the number of random variables involved in the McDiarmid's inequality. Indeed, when d is high, the approximation error decreases (see Lemma 4.5), conversely the number of random variables N_2 increases (see the bound for N_2 in the Lemma 4.2).

5.2. Concentration inequalities with optimized parameters

Equations (5.1) and (5.2) in Lemma 5.1 were established for each $d \in \mathbb{N}$ and each positive value of t_1, t_2 . Therefore, we can choose an appropriate value for each of these parameters to improve our bounds.

Moreover, the quantity N_2 is not easy to interpret and even less to estimate. Therefore, in the following theorems, we fix parameters d, t_1, t_2 and replace N_2 by an upper bound corresponding to the worst case.

Theorem 5.1 (Improved concentration inequality for $S_{\mathcal{I}}$, uniform contraction case). *If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) , for $\varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty$*

$$\mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq 2 \exp\left(\frac{-2(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty)^2}{(n_{\bar{B}}\mathbb{V}_\infty)^2(1 + n_{\bar{B}}n_B^3\Upsilon(d)^2nd^\kappa)}\right),$$

with $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$, $\tilde{d} = K_\rho(\infty) \ln(n)$ and $d = \lceil \tilde{d} \rceil = \lceil \frac{\ln(n)}{\ln(\rho^{-1})} \rceil$.

Proof. We use the previous Lemma 5.1 and set $\tilde{d} = \frac{\ln(n)}{\ln(\rho^{-1})}$ and $d = \lceil \tilde{d} \rceil$.

Therefore

$$d^\kappa = \left\lceil \frac{\ln(n)}{\ln(\rho^{-1})} \right\rceil^\kappa.$$

$$\rho^d \leq \rho^{\tilde{d}} = \exp(-\ln(\rho^{-1})d) = \frac{1}{n}.$$

From Lemma 4.2, we get $N_2 \leq N_1 n_d \leq n_{\bar{B}} n_B d^\kappa n$. Then, according to Lemma 5.1, it holds

$$\begin{aligned} & \forall \varepsilon > 2nn_{\bar{B}}\rho^d\mathbb{V}_\infty, \\ & \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq 2 \exp\left(\frac{-2(\varepsilon - 2nn_{\bar{B}}\rho^d\mathbb{V}_\infty)^2}{(n_{\bar{B}}\mathbb{V}_\infty)^2(n^2\rho^{2d} + N_2(n_B\Upsilon(d))^2)}\right) \\ & \Rightarrow \forall \varepsilon > 2n_{\bar{B}}\mathbb{V}_\infty, \\ & \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \leq 2 \exp\left(\frac{-2(\varepsilon - 2n_{\bar{B}}\mathbb{V}_\infty)^2}{(n_{\bar{B}}\mathbb{V}_\infty)^2(1 + n_{\bar{B}}n_B^3\Upsilon(d)^2nd^\kappa)}\right). \quad \square \end{aligned}$$

Remark 5.2. *The notation $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$ has been introduced in anticipation to Corollary 5.2 and the notation $K_\rho(m) = \frac{1 - \frac{1}{m}}{\ln(\rho^{-1})}$.*

Theorem 5.2 (Improved concentration inequality for $S_{\mathcal{I}}$, weak contraction case). *If we assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) . It holds, for $\varepsilon \geq 4n_{\bar{B}}n_Bd^{2\kappa}\rho^{-1}n^{\frac{2}{m}}L(n)$*

$$\begin{aligned} & \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) \\ & \leq 2 \exp\left(\frac{-2\rho^2\left(\frac{\varepsilon}{2} - 2n_{\bar{B}}n_Bd^{2\kappa}\rho^{-1}n^{\frac{2}{m}}L(n)\right)^2}{(n_{\bar{B}}\mathbb{V}_m n^{\frac{2}{m}})^2(1 + nn_{\bar{B}}n_B^3\Upsilon(d)^2d^\kappa)}\right) \end{aligned}$$

$$+ \frac{\rho^m}{n} \left(2n_{\bar{B}}n_B d^\kappa + \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)} \right)^m \right).$$

With $K_\rho(m) = \frac{1-\frac{1}{\rho}}{\ln(\rho^{-1})}$, $\tilde{d} = K_\rho(m) \ln(n)$, $d = \lceil \tilde{d} \rceil = \left\lceil \frac{(1-\frac{1}{\rho}) \ln(n)}{\ln(\rho^{-1})} \right\rceil$.

$$\text{And } L(n) = \left(\frac{n_{\bar{B}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^\kappa (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}} n} + n_{\bar{B}} n_B^2 \Upsilon(d) \right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}}.$$

The proof can be found in the Appendix E.

Remark 5.3. In Theorems 5.1 and 5.2, quantities $n_{\bar{B}}, n_B, \mathbb{V}_m, \mathbb{V}_\infty, \kappa$ and ρ are constants and the function $\Upsilon(d)$ can always be bound by a constant v independent of d (see Lemma 4.6).

Remark 5.4. In Theorem 5.2, an extra additive term appears. This term decreases rapidly with n and is therefore often not harmful in practice.

Moreover, adding appropriate hypotheses about m can lead to a fully exponential concentration inequality.

5.3. Expected deviation bounds

Expected deviation bounds may be obtained from previous theorems. In this case, we want to bound $\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|]$.

Corollary 5.1. We assume (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) , it holds

$$\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \leq n_{\bar{B}} \mathbb{V}_\infty \left(2 + \sqrt{\frac{\pi}{2} (1 + n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa n)} \right),$$

with $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$ and $d = \lceil K_\rho(\infty) \ln(n) \rceil$.

Proof.

$$\begin{aligned} \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] &= \int_0^\infty \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &= \int_0^{2n_{\bar{B}} \mathbb{V}_\infty} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &\quad + \int_{2n_{\bar{B}} \mathbb{V}_\infty}^\infty \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &\leq 2n_{\bar{B}} \mathbb{V}_\infty + \int_{2n_{\bar{B}} \mathbb{V}_\infty}^\infty 2 \exp\left(\frac{-2(t - 2n_{\bar{B}} \mathbb{V}_\infty)^2}{(n_{\bar{B}} \mathbb{V}_\infty)^2 (1 + n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa n)} \right) dt \\ &\leq 2n_{\bar{B}} \mathbb{V}_\infty + n_{\bar{B}} \mathbb{V}_\infty \sqrt{\frac{\pi}{2} (1 + n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa n)}. \quad \square \end{aligned}$$

Corollary 5.2. We assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) , it holds

$$\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \leq 2 \frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \sqrt{\frac{\pi}{2} (1 + n n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa)} + 4n_{\bar{B}} n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)$$

$$+ 2\rho^m M \left(2n_{\bar{B}}n_B d^\kappa + \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)} \right)^m \right),$$

with $K_\rho(m) = \frac{1-\frac{1}{m}}{\ln(\rho^{-1})}$, $d = \lceil K_\rho(m) \ln(n) \rceil$ and $L(n) = \left(\frac{n_{\bar{B}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^\kappa (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}} n} + n_{\bar{B}} n_B^2 \Upsilon(d) \right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}}$.

Proof.

$$\begin{aligned} & \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \\ &= \int_0^\infty \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &= \int_0^{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &+ \int_{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^\infty \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &\leq 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) + \int_{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^{2nM} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq t) dt \\ &\text{(because } S_{\mathcal{I}} \leq nM \text{ (see hypothesis } (\mathbf{H}_4)\text{))} \\ &\leq 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) + \int_{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^{2nM} \\ &\quad \times 2 \exp \left(\frac{-2\rho^2 \left(\frac{t}{2} - 2n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) \right)^2}{\left(n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}} \right)^2 (1 + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa)} \right) dt \\ &+ 2M\rho^m \left(2n_{\bar{B}}n_B d^\kappa + \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)} \right)^m \right). \end{aligned}$$

Moreover,

$$\begin{aligned} & \int_{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}^{2nM} 2 \exp \left(\frac{-2\rho^2 \left(\frac{t}{2} - 2n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) \right)^2}{\left(n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}} \right)^2 (1 + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa)} \right) dt \\ &= 4 \int_0^{2nM} \exp \left(\frac{-2\rho^2 t^2}{\left(n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}} \right)^2 (1 + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa)} \right) dt \\ &\leq 2 \frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \sqrt{\frac{\pi}{2} (1 + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa)}. \end{aligned}$$

Therefore,

$$\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] \leq 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n) + 2 \frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \sqrt{\frac{\pi}{2} (1 + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa)}$$

$$+ 2\rho^m M \left(2n_{\bar{B}}n_B d^\kappa + \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)} \right)^m \right). \quad \square$$

Using these moment inequalities, we can also analyze the limit behavior of $\mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|]$.

Corollary 5.3.

- We assume (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) . It holds

$$\begin{aligned} \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] &\leq G_1(\kappa, \rho, \mathbb{V}_\infty, n_{\bar{B}}, n_B, n) \\ &\underset{n \rightarrow \infty}{\sim} n_B n_{\bar{B}} \mathbb{V}_\infty v \sqrt{\frac{\pi}{2} n_B n_{\bar{B}} (K_\rho(\infty))^\kappa n \ln(n)^\kappa}, \\ &=_{n \rightarrow \infty} \mathcal{O} \left(\sqrt{n \ln(n)^\kappa} \right), \end{aligned}$$

with $K_\rho(\infty) = \frac{1}{\ln(\rho^{-1})}$.

- We assume (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_2) and (\mathbf{H}_3) . It holds

$$\begin{aligned} \mathbb{E}[|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|] &\leq G_2(\kappa, \rho, \mathbb{V}_m, n_{\bar{B}}, n_B, n) \\ &\underset{n \rightarrow \infty}{\sim} \frac{n_{\bar{B}} n_B \mathbb{V}_m v n^{\frac{2}{m}}}{\rho} \sqrt{2\pi n_{\bar{B}} n_B K_\rho(m)^\kappa \ln(n)^\kappa n} \\ &=_{n \rightarrow \infty} \mathcal{O} \left(n^{\frac{2}{m}} \sqrt{n \ln(n)^\kappa} \right), \end{aligned}$$

with $K_\rho(m) = \frac{1 - \frac{1}{m}}{\ln(\rho^{-1})}$.

Proof. We use Corollaries 5.1 and 5.2 and the bound $\forall d \in \mathbb{N}, \Upsilon(d) \leq v$ (see Lemma 4.6). □

Remark 5.5. Theorem 5.2, Corollaries 5.2 and 5.3 with hypotheses (\mathbf{H}_1^m) and (\mathbf{H}_2^m) are useful only if $m > 4$. If this is not the case $\lim_{n \rightarrow +\infty} n^{\frac{2}{m} - \frac{1}{2}} \neq 0$, hence $\lim_{n \rightarrow +\infty} \mathbb{E} \left[\frac{|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]|}{n} \right] \neq 0$. In the context of learning theory, this means that the empirical risk may not converge to theoretical risk (see Section 3.4).

Appendix A: Proof of Corollary 3.3

Proof. If (\mathbf{H}_1^∞) , (\mathbf{H}_2^∞) and (\mathbf{H}_3) are verified

$$\begin{aligned} \mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)] &= \int_{t=0}^\infty \mathbb{P} \left(\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|) \geq t \right) dt \\ &= \int_{t=0}^{\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_\infty}{n}\right)} \mathbb{P} \left(\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|) \geq t \right) dt \\ &+ \int_{t=\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_\infty}{n}\right)}^\infty \mathbb{P} \left(\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|) \geq t \right) dt. \end{aligned}$$

From one side,

$$\int_{t=0}^{\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)} \mathbb{P}\left(\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|) \geq t\right) dt \leq \exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right).$$

From the other side,

$$\begin{aligned} & \int_{t=\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)}^{\infty} \mathbb{P}\left(\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|) \geq t\right) dt \\ &= \int_{t=\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)}^{\infty} \mathbb{P}\left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \geq \frac{\ln(t)}{s}\right) dt \\ &= \int_{t=\exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)}^{\infty} \exp\left(\frac{-2n^2\left(\frac{\ln(t)}{s} - \frac{2n_{\bar{B}}\mathbb{V}_{\infty}}{n}\right)^2}{(n_{\bar{B}}\mathbb{V}_{\infty})^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}\right) dt \\ &\leq \exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n}\right) \int_{t=1}^{\infty} 2 \exp\left(\frac{-2n^2 \ln(t)^2}{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}\right) dt. \end{aligned}$$

Then,

$$\begin{aligned} & \int_{t=1}^{\infty} 2 \exp\left(\frac{-2n^2 \ln(t)^2}{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}\right) dt \\ &= 2 \int_{t=0}^{\infty} \exp\left(\frac{-2n^2 t^2}{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)} + t\right) dt \\ &\leq \frac{n_{\bar{B}}\mathbb{V}_{\infty}s}{n} \sqrt{2\pi(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)} \\ &\quad \times \exp\left(\frac{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}{8n^2}\right). \end{aligned}$$

It yields,

$$\begin{aligned} & \mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)] \\ &\leq \exp\left(\frac{2sn_{\bar{B}}\mathbb{V}_{\infty}}{n} + \frac{(n_{\bar{B}}\mathbb{V}_{\infty}s)^2(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}{8n^2}\right) \\ &\quad \times \left(1 + \frac{n_{\bar{B}}\mathbb{V}_{\infty}s}{n} \sqrt{2\pi(1 + An_{\bar{B}}n_{\bar{B}}^3\kappa!^2 \lceil \ln(n) \rceil^{\kappa} n)}\right). \end{aligned}$$

If (\mathbf{H}_1^m) , (\mathbf{H}_2^m) , (\mathbf{H}_3) and (\mathbf{H}_4) are verified The demonstration is very similar to the previous point. Indeed, we have,

$$\begin{aligned} & \mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)] \\ &\leq \exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right) + \int_{t=\exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(Ms)} \mathbb{P}\left(|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})| \geq \frac{\ln(t)}{s}\right) dt \end{aligned}$$

$$\begin{aligned} &\leq \exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right) + \int_{t=\exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(Ms)} \\ &\times 2 \exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\ln(t)}{2s} - \frac{L_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{(Hn_{\bar{B}})^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}\right) + \frac{\rho^m L_2(n)}{n} dt. \end{aligned}$$

And,

$$\begin{aligned} &\int_{t=\exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right)}^{\exp(Ms)} 2 \exp\left(\frac{-2n^{2-\frac{4}{m}}\left(\frac{\ln(t)}{2s} - \frac{L_1(n)}{n^{1-\frac{2}{m}}}\right)^2}{(Hn_{\bar{B}})^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}\right) dt \\ &= \exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right) \int_{t=1}^{\infty} 2 \exp\left(\frac{-n^{2-\frac{4}{m}} \ln(t)^2}{2(Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}\right) dt \\ &= 2 \exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right) \int_{t=0}^{\infty} \exp\left(\frac{-n^{2-\frac{4}{m}} t^2}{2(Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)} + t\right) dt \\ &= 2 \exp\left(\frac{2sL_1(n)}{n^{1-\frac{2}{m}}}\right) \frac{Hn_{\bar{B}}s}{n^{1-\frac{2}{m}}} \sqrt{2\pi(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)} \\ &\times \exp\left(\frac{(Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}{2n^{2-\frac{4}{m}}}\right) \\ &\leq 2 \exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4sL_1(n) + (Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)\right)\right) \\ &\times \frac{Hn_{\bar{B}}s}{n^{1-\frac{2}{m}}} \sqrt{2\pi(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}. \end{aligned}$$

Consequently,

$$\begin{aligned} &\mathbb{E}[\exp(s|R(\hat{f}) - \mathcal{R}_{emp}(\hat{f})|)] \\ &\leq \exp\left(\frac{1}{2n^{1-\frac{4}{m}}}\left(4sL(n) + (Hn_{\bar{B}}s)^2(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)\right)\right) \\ &\times \left(1 + \frac{\exp(Ms)\rho^m L_2(n)}{n} + \frac{2Hn_{\bar{B}}s}{n^{1-\frac{2}{m}}} \sqrt{2\pi(1 + En_{\bar{B}}n_B^3(\kappa!)^2 \lceil \ln(n) \rceil^\kappa n)}\right). \square \end{aligned}$$

Appendix B: Proof of Lemma 4.1

Proof. Let's prove by induction that for each i in $[0, d]$,

$$\forall t \in \mathbb{Z}^\kappa, \forall d \in \mathbb{N}, \|X_t - \tilde{X}_t^{[i]}\|_m \leq \rho^i \mathbb{V}_m. \tag{B.1}$$

For $i = 0$, $\tilde{X}_t^{[0]} = \bar{X}$ and \bar{X} is drawn from the law μ_X . Therefore, using hypothesis **(H₂^m)**, we get

$$\forall t \in \mathbb{Z}^\kappa, \|X_t - \tilde{X}_t^{[0]}\|_m \leq \mathbb{V}_m.$$

Thus Equation (B.1) is verified for $i = 0$.

Moreover, if we suppose that (B.1) is verified for i . For each $t \in \mathbb{Z}^\kappa$, it holds

$$\begin{aligned} \|X_t - \tilde{X}_t^{[i+1]}\|_m &\leq \|F((X_{t+s})_{s \in \mathcal{B}}, \varepsilon_t) - F((\tilde{X}_{t+s}^{[i]})_{s \in \mathcal{B}}, \varepsilon_t)\|_m \\ &\leq \sum_{s \in \mathcal{B}} \lambda_s \|X_{t+s} - \tilde{X}_{t+s}^{[i]}\|_m \leq \sum_{s \in \mathcal{B}} \lambda_s \rho^i \mathbb{V}_m \leq \rho^{i+1} \mathbb{V}_m. \end{aligned}$$

Consequently, by induction, Equation (B.1) is verified for each i in $[0, d]$. Finally, setting $i = d$ yields Lemma 4.1. \square

Appendix C: Proof of Lemma 4.5

Proof. We first show that $\|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d]'}\|_m \leq nn_{\bar{\mathcal{B}}}\rho^d \mathbb{V}_m$

$$\begin{aligned} \|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d]'}\|_m &= \left\| \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}}) - \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]'})_{s \in \bar{\mathcal{B}}}) \right\|_m \\ &\leq \sum_{t \in \mathcal{I}} \|\Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}}) - \Phi((\tilde{X}_{t+s}^{[d]'})_{s \in \bar{\mathcal{B}}})\|_m \\ &\leq \sum_{t \in \mathcal{I}} \sum_{s \in \bar{\mathcal{B}}} \|\tilde{X}_{t+s}^{[d]} - \tilde{X}_{t+s}^{[d]'}\|_m \\ &\leq nn_{\bar{\mathcal{B}}}\rho^d \mathbb{V}_m \text{ using Lemma 4.4.} \end{aligned}$$

We then bound $\|\tilde{S}_{\mathcal{I}}^{[d]} - \tilde{S}_{\mathcal{I}}^{[d,i]}\|_m$. For all t and $i \in \mathbb{Z}^\kappa$ we have the following properties

$$\tilde{X}_t^{[d]} \neq \tilde{X}_t^{[d,i]} \iff i \in \mathcal{V}(d\delta, t) \iff t \in \mathcal{V}(d\delta, i).$$

$$\begin{aligned} \|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m &= \left\| \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d]})_{s \in \bar{\mathcal{B}}}) - \sum_{t \in \mathcal{I}} \Phi((\tilde{X}_{t+s}^{[d,i]})_{s \in \bar{\mathcal{B}}}) \right\|_m \\ &\leq \sum_{t \in \mathcal{I}} \sum_{s \in \bar{\mathcal{B}}} \|\tilde{X}_{t+s}^{[d]} - \tilde{X}_{t+s}^{[d,i]}\|_m \\ &\leq \sum_{t \in \mathbb{Z}^\kappa} \sum_{s \in \bar{\mathcal{B}}+t} \|\tilde{X}_s^{[d]} - \tilde{X}_s^{[d,i]}\|_m. \end{aligned}$$

Only random variables $\tilde{X}_t^{[d,i]}$ with $t \in \mathcal{V}(d\delta, i)$ are impacted by the replacement of ε_i by ε'_i . Thus, at worst, each $t \in \mathcal{V}(d\delta, i)$ appears $n_{\bar{\mathcal{B}}}$ times in the sum. Therefore, we get

$$\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq \sum_{t \in \mathcal{V}(d\delta, i)} n_{\bar{\mathcal{B}}} \|\tilde{X}_t^{[d,i]} - \tilde{X}_t^{[d]}\|_m.$$

Moreover, for all t in \mathbb{Z}^κ

$$\mathcal{V}(d\delta, t) = \bigcup_{c=0}^d \mathcal{V}(c\delta, t) \setminus \mathcal{V}((c-1)\delta, t) \quad \text{with} \quad \mathcal{V}(0, t) = t \quad \text{and} \quad \mathcal{V}(-1, t) = \emptyset.$$

And by definition of $\mathcal{V}(d\delta, t)$ (see Section 2.1), it holds

$$\forall r_1 \neq r_2, (\mathcal{V}(r_1\delta, t) \setminus \mathcal{V}((r_1 - 1)\delta, t)) \cap (\mathcal{V}(r_2\delta, t) \setminus \mathcal{V}((r_2 - 1)\delta, ts)) = \emptyset.$$

Therefore, $\bigcup_{c=0}^d \mathcal{V}(c\delta, t) \setminus \mathcal{V}(c-1, t)$ is a partition of the set $\mathcal{V}(d\delta, t)$ and we can rewrite the previous inequality as

$$\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq n_{\bar{B}} \sum_{c=0}^d \left(\sum_{s \in \mathcal{V}(c\delta, i) \setminus \mathcal{V}((c-1)\delta, i)} \|\tilde{X}_s^{[d,i]} - \tilde{X}_s^{[d]}\|_m \right).$$

Using Lemma 4.4. We get

$$\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq n_{\bar{B}} \sum_{c=0}^d \text{Card}(\mathcal{V}(c\delta, i) \setminus \mathcal{V}((c-1)\delta, i)) \rho^c \mathbb{V}_m.$$

$\forall c \in \mathbb{N}$, $\mathcal{V}(c\delta, i)$ is a κ -orthotope and

$$\text{Card}(\mathcal{V}(c\delta, i)) = \prod_{j=1}^{\kappa} (2c\delta_j + 1).$$

We recall that $n_{\mathcal{B}} = \prod_{j=1}^{\kappa} (2\delta_j + 1) = \text{Card}(\mathcal{B}) + 1$. Then, for all $c > 1$,

$$\begin{aligned} \text{Card}(\mathcal{V}(c\delta, i) \setminus \mathcal{V}((c-1)\delta, i)) &= \text{Card}(\mathcal{V}(c\delta, i)) - \text{Card}(\mathcal{V}((c-1)\delta, i)) \\ &= \prod_{j=1}^{\kappa} (2c\delta_j + 1) - \prod_{j=1}^{\kappa} (2(c-1)\delta_j + 1) \\ &\leq c^{\kappa} \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c} \right) - (c-1)^{\kappa} \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c-1} \right) \\ &\leq \prod_{j=1}^{\kappa} \left(2\delta_j + \frac{1}{c-1} \right) (c^{\kappa} - (c-1)^{\kappa}) \\ &\leq \prod_{j=1}^{\kappa} (2\delta_j + 1) (c^{\kappa} - (c-1)^{\kappa}) \\ &\leq n_{\mathcal{B}} (c^{\kappa} - (c-1)^{\kappa}) \leq n_{\mathcal{B}} \kappa c^{\kappa-1}. \end{aligned}$$

Eventually, we get $\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq n_{\bar{B}} \mathbb{V}_m \left(1 + \sum_{c=1}^d n_{\mathcal{B}} \kappa c^{\kappa-1} \rho^c \right)$. \square

Appendix D: Proof of Lemma 4.6

Proof. Let $p \in \mathbb{N}$ and $(a, b) \in \mathbb{R}^2$.

We first compute $I_p = \int_a^b t^p \rho^t dt$.

$$I_p = \int_a^b t^p \rho^t dt = \left[t^p \frac{\rho^t}{\ln(\rho)} \right]_a^b + \frac{p}{\ln(\rho^{-1})} \int_a^b t^{p-1} \rho^t dt = \frac{a^p \rho^a - b^p \rho^b}{\ln(\rho^{-1})} + \frac{p}{\ln(\rho^{-1})} I_{p-1}.$$

And

$$I_0 = \int_a^b \rho^t dt = \frac{\rho^a - \rho^b}{\ln(\rho^{-1})}.$$

By induction, we get

$$I_p = \sum_{i=0}^p \frac{a^i \rho^a - b^i \rho^b}{\ln(\rho^{-1})^{p-i+1}} \times \frac{p!}{i!} = \frac{p!}{\ln(\rho^{-1})^{p+1}} \sum_{i=0}^p \frac{(a^i \rho^a - b^i \rho^b) \ln(\rho^{-1})^i}{i!}. \quad (\text{D.1})$$

Let $f(t) = t^p \rho^t$; we have $f'(t) = t^{p-1} \rho^t (p - t \ln(\rho^{-1}))$. Then $f'(t) = 0 \iff t = \frac{p}{\ln(\rho^{-1})}$. Thus f is increasing on $[0, \frac{p}{\ln(\rho^{-1})}]$ and decreasing on $[\frac{p}{\ln(\rho^{-1})}, +\infty]$.

Applying this to our case,

- if $d < \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$. Then: $\sum_{c=1}^d c^{\kappa-1} \rho^c \leq \int_0^{d+1} t^{\kappa-1} \rho^t dt$.
- if $d = \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$, Then $\sum_{c=1}^d c^{\kappa-1} \rho^c \leq \int_0^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor} t^{\kappa-1} \rho^t dt + \left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1}$.
- if $d > \lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor$. Then $\sum_{c=1}^d c^{\kappa-1} \rho^c \leq \int_0^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor} t^{\kappa-1} \rho^t dt + \left(\frac{\kappa-1}{\ln(\rho^{-1})e}\right)^{\kappa-1} + \int_{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor}^d t^{\kappa-1} \rho^t dt$.

Using Equation (D.1), we get

- $\int_0^{d+1} t^{\kappa-1} \rho^t dt \leq \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} \left(1 - \rho^{d+1} \sum_{i=0}^{\kappa-1} \frac{((d+1) \ln(\rho^{-1}))^i}{i!}\right)$.
- $\int_{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor}^d t^{\kappa-1} \rho^t dt \leq \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} \left(\rho^{\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor} \sum_{i=0}^{\kappa-1} \frac{(\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \rfloor \ln(\rho^{-1}))^i}{i!} - \rho^d \sum_{i=0}^{\kappa-1} \frac{(d \ln(\rho^{-1}))^i}{i!}\right)$.

We recall that,

$$\begin{aligned} \|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m &\leq n_{\bar{B}} \mathbb{V}_m \left(1 + n_B \kappa \sum_{c=1}^d c^{\kappa-1} \rho^c\right) \\ &\leq n_{\bar{B}} n_B \mathbb{V}_m \kappa \left(\frac{1}{n_B \kappa} + \sum_{c=1}^d c^{\kappa-1} \rho^c\right). \end{aligned}$$

We now define the function Υ .

$$\Upsilon : \mathbb{N} \mapsto \mathbb{R}$$

$$\Upsilon(d) = \begin{cases} \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} \left(1 - \rho^{d+1} \sum_{i=0}^{\kappa-1} \frac{((d+1) \ln(\rho^{-1}))^i}{i!} \right) \right), \\ \text{if } d < \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} \left(1 - e^{-(\kappa-1)} \sum_{i=0}^{\kappa-1} \frac{\left(\left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor \ln(\rho^{-1}) \right)^i}{i!} \right) \right) \\ + \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}, \\ \text{if } d = \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \\ \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} \left(1 - \rho^d \sum_{i=0}^{\kappa-1} \frac{(d \ln(\rho^{-1}))^i}{i!} \right) \right) + \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}, \\ \text{if } d > \left\lfloor \frac{\kappa-1}{\ln(\rho^{-1})} \right\rfloor. \end{cases}$$

Consequently, we proved that,

$$\|\tilde{S}_{\mathcal{I}}^{[d,i]} - \tilde{S}_{\mathcal{I}}^{[d]}\|_m \leq n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d).$$

It is important to note that we can easily verify that the function Υ is increasing and bounded by its limit when $d \rightarrow \infty$. It holds

$$\begin{aligned} \forall d \in \mathbb{N}, \Upsilon(d) &\leq \lim_{d \rightarrow \infty} \Upsilon(d) = \kappa \left(\frac{1}{n_B \kappa} + \frac{(\kappa-1)!}{\ln(\rho^{-1})^\kappa} + \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1} \right) \\ &\leq \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}. \end{aligned}$$

Thereafter, we will denote $v = \frac{1}{n_B} + \frac{\kappa!}{\ln(\rho^{-1})^\kappa} + \kappa \left(\frac{\kappa-1}{\ln(\rho^{-1})e} \right)^{\kappa-1}$. Moreover, we emphasize that using Stirling formula, we can prove that $v \underset{\kappa \rightarrow \infty}{\sim} \frac{\kappa!}{\ln(\rho^{-1})^\kappa}$. \square

Appendix E: Proof of Theorem 5.2

Proof. From Lemma 4.2, we get $N_2 \leq N_1 n_d \leq n_{\bar{B}} n_B d^\kappa n$.

We set $\tilde{d} = \frac{(1-\frac{1}{m}) \ln(n)}{\ln(\rho^{-1})}$, $d = \lceil \tilde{d} \rceil$, $t_1 = \frac{n^{1+\frac{1}{m}} \rho^{\tilde{d}} n_{\bar{B}} \mathbb{V}_m}{\rho(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}}$ and $t_2 = \frac{n^{\frac{2}{m}} n_B n_{\bar{B}} \Upsilon(d) \mathbb{V}_m}{\rho}$.

Then

$$\begin{aligned} p &\leq \left(\frac{nn_{\bar{B}} \rho^d \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2} \right)^m \\ &\leq \left(\frac{nn_{\bar{B}} \rho^{\tilde{d}} \mathbb{V}_m}{t_1} \right)^m + N_2 \left(\frac{n_{\bar{B}} n_B \mathbb{V}_m \Upsilon(d)}{t_2} \right)^m \end{aligned}$$

$$\begin{aligned} &\leq \frac{\rho^m n_{\bar{B}} n_B d^\kappa}{n} + n_{\bar{B}} n_B d^\kappa n \frac{\rho^m}{n^2} \\ &\leq \frac{2n_{\bar{B}} n_B d^\kappa \rho^m}{n}. \end{aligned}$$

We also note that $\rho^{\bar{d}} = \rho^{\bar{d}} = \exp(-\ln(\rho^{-1})K_\rho(m)\ln(n)) = \exp(-(1 - \frac{1}{m})\ln(n)) = n^{-(1-\frac{1}{m})}$.

On the other hand,

$$\begin{aligned} t_1^2 + N_2 t_2^2 &\leq \left(\frac{n^{1+\frac{1}{m}} \rho^{\bar{d}} n_{\bar{B}} \mathbb{V}_m}{\rho(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}} \right)^2 + n_{\bar{B}} n_B d^\kappa n \left(\frac{n^{\frac{2}{m}} n_B n_{\bar{B}} \Upsilon(d) \mathbb{V}_m}{\rho} \right)^2 \\ &= \left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right)^2 \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{2}{m}}} + n n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa \right). \end{aligned}$$

And

$$\begin{aligned} \bar{c} = t_1 + N_2 t_2 &\leq \frac{n^{1+\frac{1}{m}} \rho^{\bar{d}} n_{\bar{B}} \mathbb{V}_m}{\rho(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}} + n_{\bar{B}} n_B d^\kappa n \left(\frac{n^{\frac{2}{m}} n_B n_{\bar{B}} \Upsilon(d) \mathbb{V}_m}{\rho} \right) \\ &\leq \left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}} + n_{\bar{B}} n_B^2 \Upsilon(d) d^\kappa n \right). \end{aligned}$$

Therefore,

$$\begin{aligned} &2npM + p\bar{c} \\ &\leq \frac{2n_{\bar{B}} n_B d^\kappa \rho^m}{n} \left(\left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}}} + n_{\bar{B}} n_B^2 \Upsilon(d) d^\kappa n \right) + 2nM \right) \\ &\leq 2n_{\bar{B}} n_B d^\kappa \rho^m \left(\left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho} \right) \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}} n} + n_{\bar{B}} n_B^2 \Upsilon(d) d^\kappa \right) + 2M \right) \\ &\leq 2n_{\bar{B}} n_B d^{2\kappa} \rho^m n^{\frac{2}{m}} \left(\left(\frac{n_{\bar{B}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^\kappa (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}} n} + n_{\bar{B}} n_B^2 \Upsilon(d) \right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}} \right) \\ &\leq 2n_{\bar{B}} n_B d^{2\kappa} \rho^m n^{\frac{2}{m}} L(n). \end{aligned}$$

With $L(n) = \left(\frac{n_{\bar{B}} \mathbb{V}_m}{\rho} \right) \left(\frac{1}{d^\kappa (n_{\bar{B}} n_B d^\kappa)^{\frac{1}{m}} n} + n_{\bar{B}} n_B^2 \Upsilon(d) \right) + \frac{2M}{d^\kappa n^{\frac{2}{m}}}$.

Consequently using $p \leq \frac{2n_{\bar{B}} n_B d^{2\kappa} \rho^m}{n}$. It holds $\forall \varepsilon > 4n_{\bar{B}} n_B d^{2\kappa} \rho^m n^{\frac{2}{m}} L(n)$,

$$\begin{aligned} &\exp\left(\frac{-2\left(\frac{\varepsilon}{2} - (p\bar{c} + 2npM)\right)^2}{t_1^2 + N_2 t_2^2}\right) \\ &\leq \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{B}} n_B d^{2\kappa} \rho^m n^{\frac{2}{m}} L(n)\right)^2}{\left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho}\right)^2 \left(\frac{1}{(n_{\bar{B}} n_B d^\kappa)^{\frac{2}{m}}} + n n_{\bar{B}} n_B^3 \Upsilon(d)^2 d^\kappa\right)}\right). \end{aligned}$$

In particular; $\forall \varepsilon > 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)$,

$$\begin{aligned} & \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - (p\bar{c} + 2npM)\right)^2}{t_1^2 + N_2 t_2^2}\right) \\ & \leq \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)\right)^2}{\left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho}\right)^2 \left(\frac{1}{(n_{\bar{B}}n_B d^\kappa)^{\frac{2}{m}}} + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa\right)}\right). \end{aligned}$$

Finally,

$$\begin{aligned} \forall \varepsilon > 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n), \left(\frac{2nn_{\bar{B}}\rho^d \mathbb{V}_m}{\varepsilon}\right)^m & \leq \left(\frac{2nn_{\bar{B}}\rho^d \mathbb{V}_m}{4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)}\right)^m \\ & \leq \frac{\rho^m}{n} \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)}\right)^m. \end{aligned}$$

Using Equation (5.2), for $\varepsilon > 4n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)$, we get

$$\begin{aligned} \mathbb{P}(|S_{\mathcal{I}} - \mathbb{E}[S_{\mathcal{I}}]| \geq \varepsilon) & \leq 2 \exp\left(\frac{-2\left(\frac{\varepsilon}{2} - 2n_{\bar{B}}n_B d^{2\kappa} \rho^{-1} n^{\frac{2}{m}} L(n)\right)^2}{\left(\frac{n_{\bar{B}} \mathbb{V}_m n^{\frac{2}{m}}}{\rho}\right)^2 \left(\frac{1}{(n_{\bar{B}}n_B d^\kappa)^{\frac{2}{m}}} + nn_{\bar{B}}n_B^3 \Upsilon(d)^2 d^\kappa\right)}\right) \\ & \quad + \frac{2n_{\bar{B}}n_B d^\kappa \rho^m}{n} + \frac{\rho^m}{n} \left(\frac{\mathbb{V}_m}{2n_B d^{2\kappa} L(n)}\right)^m. \quad \square \end{aligned}$$

References

- [1] ALQUIER, P., BERTIN, K., DOUKHAN, P. and GARNIER, R. (2020). High-dimensional VAR with low-rank transition. *Statistics and Computing* 1–15. [MR4108695](#)
- [2] ALQUIER, P., DOUKHAN, P. and FAN, X. (2019). Exponential inequalities for nonstationary Markov chains. *Dependence Modeling* 7 150–168. [MR3962650](#)
- [3] ALQUIER, P., WINTENBERGER, O. et al. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli* 18 883–913. [MR2948906](#)
- [4] BALLESTER, C., BERTALMIO, M., CASELLES, V., SAPIRO, G. and VERDERA, J. (2001). Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing* 10 1200–1211. [MR1851781](#)
- [5] BARTLETT, P. L., BOUCHERON, S. and LUGOSI, G. (2000). Model Selection and Error Estimation. *SSRN Electronic Journal*.
- [6] BERTAIL, P. and CIOLEK, G. (2019). New Bernstein and Hoeffding type inequalities for regenerative Markov chains. *Latin American Journal of Probability and Mathematical Statistics* 16 259. [MR3914351](#)

- [7] BICKEL, P. J., RITOV, Y., TSYBAKOV, A. B. et al. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics* **37** 1705–1732. [MR2533469](#)
- [8] BLUMER, A., EHRENFUCHT, A., HAUSSLER, D. and WARMUTH, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)* **36** 929–965. [MR1072253](#)
- [9] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press. [MR3185193](#)
- [10] CHAZOTTES, J.-R., COLLET, P., KÜLSKE, C. and REDIG, F. (2007). Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields* **137** 201–225. [MR2278456](#)
- [11] CHEN, L. and WU, W. B. (2016). Stability and asymptotics for autoregressive processes. *Electronic Journal of Statistics* **10**. [MR3579674](#)
- [12] CHEN, L. and WU, W. B. (2018). Concentration inequalities for empirical processes of linear time series. *Journal of Machine Learning Research* **18** 1-46. [MR3845530](#)
- [13] COMBES, R. (2015). An extension of McDiarmid’s inequality. *arXiv:1511.05240 [cs, math, stat]*. arXiv: [1511.05240](#).
- [14] DEDECKER, J., DOUKHAN, P. and FAN, X. (2019). Deviation inequalities for separately Lipschitz functionals of composition of random functions. *Journal of Mathematical Analysis and Applications* **479** 1549–1568. [MR3987922](#)
- [15] DEDECKER, J., DOUKHAN, P., LANG, G., RAFAEL, L. R. J., LOUHICHI, S. and PRIEUR, C. (2007). Weak dependence. In *Weak dependence: With examples and applications* 9–20. Springer. [MR2338725](#)
- [16] DEDECKER, J. and FAN, X. (2015). Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications* **125** 60–90. [MR3274692](#)
- [17] DOUKHAN, P., POMMERET, D., RYNKIEWICZ, J. and SALHI, Y. (2017). A class of random field memory models for mortality forecasting. *Insurance: Mathematics and Economics* **77** 97–110. [MR3724939](#)
- [18] DOUKHAN, P. and TRUQUET, L. (2007). A fixed point approach to model random fields. *Alea* **3** 111–132. [MR2349804](#)
- [19] FREUND, Y., MANSOUR, Y., SCHAPIRE, R. E. et al. (2004). Generalization bounds for averaged classifiers. *The annals of statistics* **32** 1698–1722. [MR2089139](#)
- [20] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58** 13–30. [MR0144363](#)
- [21] KENTON, J. D. M.-W. C. and TOUTANOVA, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT* 4171–4186.
- [22] KONTOROVICH, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *Proceedings of the 31st International Conference on Machine Learning* (E. P. XING and T. JEBARA, eds.). *Proceedings of*

- Machine Learning Research* **32** 28–36. PMLR, Beijing, China.
- [23] KONTOROVICH, L. A. and RAMANAN, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability* **36** 2126–2158. [MR2478678](#)
- [24] KUTIN, S. (2002). Extensions to McDiarmid’s inequality when differences are bounded with high probability. *Department Computer Science, University of Chicago, Chicago, IL. Technical report TR-2002-04*.
- [25] KUZNETSOV, V. and MOHRI, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In *Advances in neural information processing systems* 541–549.
- [26] LINDVALL, T. (2002). *Lectures on the coupling method*. Courier Corporation. [MR1924231](#)
- [27] LUGOSI, G. (2002). *Pattern Classification and Learning Theory In Principles of Nonparametric Learning* 1–56. Springer Vienna, Vienna. [MR1987656](#)
- [28] MASSART, P. (2007). *Concentration inequalities and model selection* **6**. Springer. [MR2319879](#)
- [29] MCCOY, B. M. and WU, T. T. (2014). *The two-dimensional Ising model*. Courier Corporation. [MR3838431](#)
- [30] MCDIARMID, C. (1989). *On the method of bounded differences*. In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference. London Mathematical Society Lecture Note Series* 148–188. Cambridge University Press. [MR1036755](#)
- [31] MOHRI, M. and ROSTAMIZADEH, A. (2010). Stability Bounds for Stationary ϕ -mixing and β -mixing Processes. *Journal of Machine Learning Research* **11** 789–814. [MR2600630](#)
- [32] PAULIN, D. et al. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability* **20**. [MR3383563](#)
- [33] SANCHEZ-PEREZ, A. (2015). Time series prediction via aggregation: an oracle bound including numerical cost. In *Modeling and Stochastic Learning for Forecasting in High Dimensions* 243–265. Springer. [MR3588112](#)
- [34] SCHUSTER, M. and PALIWAL, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **45** 2673–2681.
- [35] SURGAILIS, D. (2008). A quadratic ARCH(∞) model with long memory and Lévy stable behavior of squares. *Advances in Applied Probability* **40** 1198–1222. [MR2488538](#)
- [36] VIRMAUX, A. and SCAMAN, K. (2018). Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) 3835–3844. Curran Associates, Inc.
- [37] WARNKE, L. (2016). On the Method of Typical Bounded Differences. *Combinatorics, Probability and Computing* **25** 269–299. [MR3455677](#)