Mixed moving average field guided learning for spatio-temporal data

Imma Valentina Curato[1]**, Orkun Furat**[2]**,**
Lorenzo Proietti[3] **and Bennet Ströh**[4]

[1]*Faculty of Mathematics, TU Chemnitz,*
e-mail: imma-valentina.curato@math.tu-chemnitz.de

[2]*Department of Mathematics, Ulm University,*
e-mail: orkun.furat@uni-ulm.de

[3]*Faculty of Mathematics, TU Chemnitz,*
e-mail: lorenzo.proietti@math.tu-chemnitz.de

[4]*Department of Mathematics, Imperial College,*
e-mail: b.stroh@imperial.ac.uk

Abstract: Influenced mixed moving average fields are a versatile modeling class for spatio-temporal data. However, their predictive distribution is not generally known. Under this modeling assumption, we define a novel spatio-temporal embedding and a theory-guided machine learning approach that employs a generalized Bayesian algorithm to make ensemble forecasts. We use Lipschitz predictors to determine fixed-time and any-time PAC Bayesian bounds in the batch learning setting. Performing causal forecast is a highlight of our methodology as its potential application to data with temporal and spatial short and long-range dependence. We then test the performance of our learning methodology by using linear predictors and data sets simulated from a spatio-temporal Ornstein-Uhlenbeck process.

1. Introduction

Analyzing spatio-temporal data introduces various methodological challenges. These include determining models that can account for the serial correlation observed along their temporal and spatial dimensions and that, at the same time, can also enable forecasting tasks. Statistical models such as Gaussian processes [6], [30], [66], and [78]; spatio-temporal kriging [26], and [58]; space-time autoregressive moving average models [39]; point processes [40], and hierarchical models [26] are very versatile in modeling the spatio-temporal correlation observed in the data and can deliver forecasts once the variogram or the data distribution (up to a set of parameters) is carefully chosen in relation to the

studied phenomenon and practitioners' experience. In a nutshell, such choices allow access to the models' predictive distribution.

Suppose we want to avoid making any explicit choice regarding the data distribution. In this case, we can alternatively use deep learning methodologies to perform forecasting tasks in a spatio-temporal framework, see [5], [68], [76], [55] for a review, or a video frame prediction algorithm as in [57]. Deep learning techniques can successfully extract spatio-temporal features and learn the *inner law* of an observed spatio-temporal system. However, these models lack interpretability, i.e., it is not possible to infer the correlation and causal relationship between variables in different space-time points that the models consider, and typically, no proof of their generalization performances is available in a spatio-temporal framework for dependent data. On the other hand, the video prediction algorithm presented in [57] retains a *causal interpretation* of the relationship between different space-time points. However, as in the case of deep learning algorithms, there is no proof of their generalization performances.

This paper proposes a novel theory-guided machine learning methodology for spatio-temporal data that enables one-time ahead ensemble forecasts based on moment assumptions (no further assumptions on the data distribution are needed), an opportune spatio-temporal embedding, and a generalized Bayesian algorithm. A theory-guided machine learning methodology is a hybrid procedure that employs a stochastic model in synergy with a learning algorithm. Such methodologies have started to gain prominence in several scientific disciplines such as earth science, quantum chemistry, bio-medical science, climate science, and hydrology modeling as, for example, described in [18], [51], [64], [67], and [68]. It is important to emphasize that in these works the data are typically considered outputs of (deterministic) dynamical systems driven by partial differential equations. We define a theory-guided machine learning methodologies for data generated by a random field.

We call our methodology *mixed moving average field guided learning* or MMAF-guided learning. In particular, we analyze *raster data cubes* [62], which are described for dimension $d = 2$ in Section 3.1, and that are nowadays generated in environmental monitoring, from satellite observations, and climate and weather numerical models' outputs. Our methodology applies, in general, to raster data having spatial dimension $d \geq 1$, and we assume that such data are generated by an *influenced mixed moving average field* (MMAF, in short), see Definition 2.6. Such a class of random fields has been introduced in [28] and allows modeling the correlation and the causal relationship in different space-time points using ambit sets, see [8] [12] [59], and [60]. Such models have been so far employed to model data in environmental monitoring [45, 59], imaging analysis [50], and electricity networks [24]. They allow modeling Gaussian and non-Gaussian distributed data; they can be non-Markovian and have non-separable covariance functions. Moreover, they are stationary and θ-lex weakly dependent, as proven in [28, Section 3.3], and allow modeling temporal and spatial short and long-range dependence. A drawback of employing MMAF in forecasting tasks is that their predictive distribution is not explicitly known. To our knowledge, the only available results on the predictive distribution of an

MMAF in a spatio-temporal framework can be found in [59, Theorem 13] for a Gaussian spatio-temporal Ornstein-Uhlenbeck process.

We then select a generalized Bayesian algorithm, i.e., a so-called *randomized estimator*, which is a regular conditional probability on the class of the Lipschitz functions $\mathcal{H}$, see Definition 1.2. The latter is then employed to make ensemble forecasts. We call a function $h \in \mathcal{H}$ a *predictor*. Linear models, neural network architectures with feed-forward and convolutional modules [14, 74], and Lipschitz modifications of the transformer architecture [52, 63] are between the predictors belonging to $\mathcal{H}$. The selection of a *randomized estimator* is guided by the design of a spatio-temporal embedding of an observed *raster data cube*, which gives us the training data set on which the estimator can be learned. The precise construction of the spatio-temporal embedding is given in Section 3.1 and depends on several parameters that must be opportunely tuned. Therefore, given a *raster data cube*, there exist different ways to pre-process them into a training data set. Such a procedure ensures casual forecasts and good generalization performance, which, in the context of our paper, means obtaining non-vacuous PAC Bayesian bounds. In the paper, we present how to *guide* the design of a Dirac delta mass concentrated on the Empirical Risk Minimizer and a *randomized Gibbs estimator*.

1.1. Setting

Let $\boldsymbol{S} := ((\boldsymbol{X}_i, \boldsymbol{Y}_i)^\top)_{i \in \mathbb{Z}}$ be a random element defined on the canonical probability space $(\Omega, \mathcal{F}, \mathbb{P})$, see [16, Chapter 7] for more details on its definition. Here, each $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ is identically distributed and has values in $\mathcal{X} \times \mathcal{Y}$ (Euclidean Spaces). In particular, Ω is the space of all possible trajectories or realizations of the process $\boldsymbol{S}$. We further assume that $\boldsymbol{S}$ is a sample from an MMAF with finite second moments as carefully described in Section 3.1. $\boldsymbol{S_m}$ indicates a finite dimensional distribution of the process $\boldsymbol{S}$ of length m, and a *training data set* is one of its realization which, for a given $\omega \in \Omega$, we indicate with $S_m := ((X_i, Y_i)^\top)_{i=1}^m$ where $X_i = \boldsymbol{X}_i(\omega)$ and $Y_i = \boldsymbol{Y}_i(\omega)$. We also call $S := ((X_i, Y_i)^\top)_{i \in \mathbb{Z}}$ a realization from $\boldsymbol{S}$.

Let $\mathcal{H}$ be the set of all Lipschitz functions $h : \mathcal{X} \to \mathcal{Y}$, and $L : \mathcal{X} \times \mathcal{Y} \to [0, \infty]$ a loss function. We define the *generalization error (out-of-sample risk)* as

$$R(h) = \mathbb{E}[L(h(\boldsymbol{X}), \boldsymbol{Y})],$$

where $(\boldsymbol{X}, \boldsymbol{Y})$ indicates a general example belonging to $\boldsymbol{S}$, and the *empirical error (in-sample risk)*

$$r(h, \omega) = \frac{1}{m} \sum_{i=1}^m L(h(\boldsymbol{X_i}(\omega)), \boldsymbol{Y_i}(\omega)),$$

which is determined by a particular realization $\omega \in \Omega$ of the finite dimensional distribution $\boldsymbol{S}_m$. The function L is used to measure the discrepancy between a predicted output $h(\boldsymbol{X})$ and the true output $\boldsymbol{Y}$. Using the in-sample risk, we

measure the performance of a given predictor h just over an observed training data set S_m. In contrast, the out-of-sample risk gives us the performance of a predictor depending on the unknown distribution of the data $\mathbb{P}$. We then need a guarantee that a selected predictor will perform well when used on a set of out-of-sample observations, i.e., not belonging to S_m. We can also rephrase the problem as finding a predictor h for which the difference between the out-of-sample and in-sample risk $R(h) - r(h, \omega)$ is as small as possible. We call the latter *generalization gap*. The classical PAC framework aims to find a bound on the generalization gap that holds with high probability $\mathbb{P}$; see, for example, [75] and [81]. Such probability inequality is also called a *generalization bound*. The acronym PAC stands for Probably Approximately Correct and may be traced back to [80]. A PAC inequality states that with an arbitrarily high probability (hence "probably"), the performance (as provided by the generalization gap) of a learning algorithm is upper-bounded by a term decaying to an optimal value as more data is collected (hence "approximately correct"). Note that we drop the dependence on $\omega \in \Omega$ in $r(h)$ to ease the notations in the following.

In the paper, we use a PAC Bayesian approach, also known as *generalized Bayesian approach*. First, we select a *reference distribution* π on the space $(\mathcal{H}, \mathcal{T})$, where $\mathcal{T}$ indicates a σ-algebra on the space $\mathcal{H}$. The reference distribution gives a *structure* on the space $\mathcal{H}$, which we can interpret as our belief that certain predictors will perform better than others. The choice of π, therefore, is an indirect way to make the size of $\mathcal{H}$ come into play; see [21, Section 3] for a detailed discussion on the latter point. Therefore, π belongs to $\mathcal{M}_+^1(\mathcal{H})$, which denotes the set of all probability measures on the measurable set $(\mathcal{H}, \mathcal{T})$. We then aim to determine a *randomized estimator*. To introduce the latter, we need first the following definition.

Definition 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{H}, \mathcal{T})$ a measurable space, $\hat{\boldsymbol{h}} : \Omega \to \mathcal{H}$ a random element, and $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. The function*

$$\mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}}(\cdot|\mathcal{G})(\cdot) : \Omega \times \mathcal{T} \to [0,1]$$
$$(\omega, E) \to \mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}}(E|\mathcal{G})(\omega)$$

is a regular conditional distribution *of $\hat{\boldsymbol{h}}$ given $\mathcal{G}$ if:*

- *for any $E \in \mathcal{T}$, the function $\omega \to \mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}}(E|\mathcal{G})(\omega)$ is $\mathcal{G}$-measurable and a variant of the conditional probability $\mathbb{P}(\hat{\boldsymbol{h}} \in E|\mathcal{G})$, i.e.,*

$$\mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}}(E|\mathcal{G})(\cdot) = \mathbb{P}(\hat{\boldsymbol{h}} \in E|\mathcal{G})(\cdot) \ a.s.$$

- *for any $\omega \in \Omega$, $\mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}}(\cdot|\mathcal{G})(\omega)$ is a probability measure on $(\mathcal{H}, \mathcal{T})$.*

We assume throughout that the measurable space $(\mathcal{H}, \mathcal{T})$ is a Borel space, i.e. $\mathcal{T}$ is a countably generated σ-algebra. Then, the regular conditional distribution of $\hat{\boldsymbol{h}}$ given $\mathcal{G}$ exists, see [77, Theorem 5, Chapter II.7]. We can now give a formal definition of a *randomized estimator* that follows [22, Section 1].

Definition 1.2. *Let* $\boldsymbol{S} = ((\boldsymbol{X}_i, \boldsymbol{Y}_i)^\top)_{i \in \mathbb{Z}}$ *be a random vector on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, $(\mathcal{H}, \mathcal{T})$ *a Borel space, and* $\mathcal{G}_m = \sigma\{\boldsymbol{S}_m\}$. *We define a* random-ized estimator $\hat{\rho}$ *as the regular conditional distribution of* $\hat{\boldsymbol{h}}$ *given* $\mathcal{G}_m$, *i.e., for all* $\omega \in \Omega$, *and* $E \in \mathcal{T}$,

$$\hat{\rho}(E, \omega) := \mathbb{P}_{\hat{\boldsymbol{h}}|\mathcal{G}_m}(E|\mathcal{G}_m)(\omega).$$

From now on, we indicate with $\pi[\cdot]$ the expectation with respect to the reference distribution, and with $\hat{\rho}[\cdot]$ the conditional expectation with respect to $\mathcal{G}_m$. Moreover, we call $\hat{\rho}[R(h)]$ and $\hat{\rho}[r(h)]$ *the average generalization error and the average empirical error*, respectively. In particular, these notations indicate two random elements, i.e., for all $\omega \in \Omega$

$$\hat{\rho}[R(h)](\omega) := \int_{\mathcal{H}} R(h)\,\hat{\rho}(dh, \omega)$$

and

$$\hat{\rho}[r(h)](\omega) := \int_{\mathcal{H}} r(h)\,\hat{\rho}(dh, \omega).$$

We remark that the use of the letter h in the notation $\hat{\rho}[R(h)]$ and $\hat{\rho}[r(h)]$ plays the role of reminding the reader that we work on the hypothesis space $\mathcal{H}$. To evaluate the generalization performance of a randomized estimator $\hat{\rho}$, we determine a so-called *PAC Bayesian bound*, which is a bound on the *(average) generalization gap* defined as $\hat{\rho}[R(h)] - \hat{\rho}[r(h)]$ holding with high probability $\mathbb{P}$. Over the past two decades, PAC-Bayesian bounds have been successful in addressing various learning problems such as classification, sequential or batch learning, and deep learning [37, 42].

MMAF-guided learning applies to bounded loss functions, see Remark 3.20 for more details on this point. Throughout, for $\epsilon > 0$ called the *accuracy level*, we define the *truncated absolute loss* as

$$L^\epsilon(h(\boldsymbol{X}), \boldsymbol{Y})) = L(h(\boldsymbol{X}), \boldsymbol{Y})) \wedge \epsilon. \tag{1.1}$$

The generalization error is then indicated with $R^\epsilon(h) = \mathbb{E}[L^\epsilon(h(\boldsymbol{X}), \boldsymbol{Y})]$ and the empirical error with $r^\epsilon(h, \omega) = \frac{1}{m} \sum_{i=1}^m L^\epsilon(h(\boldsymbol{X}_{\boldsymbol{i}}(\omega)), \boldsymbol{Y}_{\boldsymbol{i}}(\omega))$ for $\omega \in \Omega$. We also drop in this case the dependence on $\omega \in \Omega$ in the empirical error's notation, and indicate the average generalization gap with $\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)]$. We then derive PAC Bayesian bound for the latter.

1.2. Contributions and outline

The PAC Bayesian bounds proven in the paper are the first results in the literature for θ-lex weakly dependent data, i.e., data generated by a stationary θ-lex weakly dependent random field $\boldsymbol{Z} = (\boldsymbol{Z}_t(x))_{(t,x) \in \mathbb{R} \times \mathbb{R}^d}$. The latter is a novel notion of dependence introduced in [28]. When considering a random field sampled on $\mathbb{Z} \times \mathbb{Z}^d$ such that $\mathbb{E}[|\boldsymbol{Z}_0(0)|^p] < \infty$ for $p > 1$, then θ-lex weak

dependence is a more general notion than $\alpha_{\infty,v}$ and $\phi_{\infty,v}$-mixing for random fields being $v \in \mathbb{N} \cup \{\infty\}$, and α and ϕ-mixing in the particular case of stochastic processes, see [28, Section 2.3]. In particular, an MMAF is a θ-lex weakly dependent random field, which can then be used to model very general frameworks; see Remark 3.22 and Appendix A.

We then analyze fixed-time and any-time PAC Bayesian bounds for data generated by an MMAF. In particular, our fixed-time bounds are explicitly stated in the function of *one single* θ-lex coefficient of the underlying field. In the literature, two comparable bounds exist for time series data, which are *explicitly* stated in the function of α-mixing and $\theta_{1,\infty}$-coefficient. The former is presented in [2, Section 3.2], and is a bound where it appears a series of α-mixing coefficients, which cannot be estimated from observed data and may diverge for power decaying coefficients. The latter is analyzed in [4, Section 3] for bounded $\theta_{1,\infty}$-coefficient, which also cannot be estimated from observed data. Instead, we can estimate the decay rate of the θ-lex-coefficients for specific MMAF models as the spatio-temporal Ornstein Uhlenbeck (STOU, in short) process and its mixed version called MSTOU process defined in [59, 60], respectively. In the paper, we also discuss the range of applicability of such estimation methodology to other types of MMAFs. The knowledge of the decay rate of the θ-lex coefficient of an MMAF is fundamental in our methodology and allows us to *guide* the choice of a *randomized estimator*, as detailed in Section 3.2, and to assess its generalization performance.

We start by proving a *fixed-time* PAC Bayesian bound that holds for all $\boldsymbol{S_m}$ with $m \geq 2$ and employs a novel exponential inequality for sums of weakly dependent processes. θ-lex weak dependence is a notion of *projective type* related to an L_1-norm, see [28] and Remark 2.15, and A.3. In regards to projective type dependence notions for L_p-norm and $p \in [1, \infty]$, there have been proven moment inequalities for partial sums of weakly dependent random fields in [17] for $p \in [2, \infty]$ and for stochastic processes and $p = 1$ in [36]. To the best of our knowledge, another exponential inequality for a projective type dependence notion was obtained just for $p = \infty$ in [4].

To give a complete overview of the range of applicability of MMAF-guided learning in the function of different choices of the spatio-temporal embedding, it is necessary to introduce a second fixed-time PAC Bayesian bound. The proof of such a result involves using an any-time PAC Bayesian bound. The latter is proven using Ville's maximal inequality for non-negative supermartingales [82], and it holds for all countable sequences of examples; this means *simultaneously* for each $\boldsymbol{S_m}$ and $m \geq 1$. Several any-time PAC Bayesian bounds exist in the literature for general dependent data frameworks as discussed in [43]. However, they do not hold in the so-called *batch learning case for dependent data*, which our proof covers for bounded losses.

We then combine the any-time bound with the results proven in [2] applied to a particular residual process. Hence, we obtain a fixed-time PAC Bayesian bound where a moment inequality is involved in its proof instead of an exponential one. In our examples and simulation results, the latter bound allows us to define *randomized estimators* using realistic spatio-temporal embeddings, which means

training data sets that can best preserve the serial correlation observed in a given *raster data cube*.

The paper is structured as follows. First, we review the MMAF framework in Section 2 and describe its causal interpretation. In this section, we also introduce the STOU and MSTOU processes. These are isotropic random fields for which we compute novel bounds for their θ-lex coefficients. The paper uses the latter to show feasible examples of MMAF-guided learning in the case of temporal and spatial short and long-range dependence. Section 3 introduces the spatio-temporal embedding applied to raster data cubes and the PAC Bayesian bounds. In Section 4, we give a step-by-step description on how to apply in the practice MMAF-guided learning and discuss the causal interpretation of the ensemble forecasts. We conclude by analyzing the performance of our methodology for a randomized Gibbs estimator on six simulated data sets from an STOU process with a Gaussian and a normal-inverse-Gaussian distributed Lévy seed. Appendix A contains further details on the dependence notions discussed in the paper and a review of the estimation methodologies for STOU and MSTOU processes. Appendix B contains detailed proofs of the theoretical results presented in the paper.

2. Mixed moving average fields

2.1. Notations

Throughout the paper, we indicate with $\mathbb{N}$ the set of positive integers, $\mathbb{N}_0$ the set of non-negative integers, and $\mathbb{R}^+$ the set of non-negative real numbers. As usual, we write $L^p(\Omega)$ for the space of (equivalence classes of) measurable functions $f : \Omega \to \mathbb{R}$ with finite L_p-norm $\|f\|_p$. When $\Omega = \mathbb{R}^n$ and $x \in \Omega$, $\|x\|_1$ and $\|x\|$ denote the L_1-norm and the Euclidean norm, respectively, and we define $\|x\|_\infty = \max_{j=1,\dots,n} |x^{(j)}|$, where $x^{(j)}$ represents the component j of the vector x.

To ease the notations in the following sections, we sometimes indicate the index set $\mathbb{R} \times \mathbb{R}^d$ by $\mathbb{R}^{1+d}$. $E \subset E'$ denotes a not necessarily proper subset E of a set E', $|E'|$ denotes the cardinality of E' and $dist(E, E') = \inf_{i \in E, j \in E'} \|i-j\|_\infty$ indicates the distance of two sets $E, E' \subset \mathbb{R}^{1+d}$. Let $n, k \geq 1$, and $F : \mathbb{R}^n \to \mathbb{R}^k$, we define $\|F\|_\infty = \sup_{t \in \mathbb{R}^n} \|F(t)\|$. We indicate with $\Gamma = \{i_1, \dots, i_u\} \in \mathbb{R}^{1+d}$ for $u \in \mathbb{N}$, a sequence of elements in $\mathbb{R}^{1+d}$. We then define the random vector $\boldsymbol{Z}_\Gamma = (\boldsymbol{Z}_{i_1}, \dots, \boldsymbol{Z}_{i_u})$. In general, we use bold notations when referring to random elements.

In the following, Lipschitz continuous is understood to mean globally Lipschitz. For $u \in \mathbb{N}$, $\mathcal{G}_u^*$ is the class of bounded functions from $\mathbb{R}^u$ to $\mathbb{R}$ and $\mathcal{G}_u$ is the class of bounded, Lipschitz continuous functions from $\mathbb{R}^u$ to $\mathbb{R}$ with respect to the distance $\|\cdot\|_1$ and define the Lipschitz constant as

$$Lip(h) = \sup_{x \neq y} \frac{|h(x) - h(y)|}{\|x - y\|_1}. \tag{2.1}$$

Hereafter, we often use the lexicographic order on $\mathbb{R}^{1+d}$. Let t and s be indicating a temporal and spatial coordinate. For distinct elements $y = (y_{1,t}, y_{1,s}, \ldots, y_{d,s})$ and $z = (z_{1,t}, z_{1,s}, \ldots, z_{d,s}) \in \mathbb{R}^{1+d}$ we say $y <_{lex} z$ if and only if $y_{1,t} < z_{1,t}$ or $y_{p,s} < z_{p,s}$ for some $p \in \{1, \ldots, d\}$ and $y_{1,t} = z_{1,t}$ and $y_{q,s} = z_{q,s}$ for $q = 1, \ldots, p-1$. Moreover, $y \leq_{lex} z$ if $y <_{lex} z$ or $y = z$ holds. Finally, let $z \in \mathbb{R}^{1+d}$, we define the set $V_z = \{y \in \mathbb{R}^{1+d} : y \leq_{lex} z\}$ and $V_z^r = V_z \cap \{y \in \mathbb{R}^{1+d} : \|z - y\|_\infty \geq r\}$ for $r > 0$. The definition of the set V_z^r is also used when referring to the lexicographic order on $\mathbb{Z}^{1+d}$.

2.2. *Definition and properties of MMAF*

Let $I = H \times \mathbb{R} \times \mathbb{R}^d$, where $H \subset \mathbb{R}^q$ for $q \geq 1$, and the Borel σ-algebra of I be denoted by $\mathcal{B}(I)$ and let $\mathcal{B}_b(I)$ contain all its Lebesgue bounded sets.

Definition 2.1. *A family of $\mathbb{R}$-valued random variables $\Lambda = \{\Lambda(B) : B \in \mathcal{B}_b(I)\}$ is called a Lévy basis on $(I, \mathcal{B}_b(I))$ if it is an independently scattered and infinitely divisible random measure. This means that:*

(i) For a sequence of pairwise disjoint elements of $\mathcal{B}_b(I)$, say $\{B_i, i \in \mathbb{N}\}$:

   – *$\Lambda(\bigcup_{i \in \mathbb{N}} B_i) = \sum_{i \in \mathbb{N}} \Lambda(B_i)$ almost surely when $\bigcup_{i \in \mathbb{N}} B_i \in \mathcal{B}_b(I)$*

   – *and $\Lambda(B_i)$ and $\Lambda(B_j)$ are independent for $i \neq j$.*

(ii) Let $B \in \mathcal{B}_b(I)$. Then, the random variable $\Lambda(B)$ is infinitely divisible, i.e., for any $i \in \mathbb{N}$, there exists a law μ_i such that the law $\mu_{\Lambda(B)}$ can be expressed as $\mu_{\Lambda(B)} = \mu_i^{*i}$, the i-fold convolution of μ_i with itself.

We refer the reader to [73] for more details on infinitely divisible distributions. In the following, we will restrict ourselves to Lévy bases which are homogeneous in space and time and factorizable, i.e., Lévy bases with characteristic function

$$\mathbb{E}\left[e^{iu\Lambda(B)}\right] = e^{\Phi(u)\Pi(B)} \tag{2.2}$$

for all $u \in \mathbb{R}$ and $B \in \mathcal{B}_b(I)$, where $\Pi = \pi \times \lambda_{1+d}$ is the product measure of the probability measure π on H and the Lebesgue measure λ_{1+d} on $\mathbb{R} \times \mathbb{R}^d$. Note that when using a Lévy basis defined on $I = \mathbb{R} \times \mathbb{R}^d$, $\Pi = \lambda_{1+d}$. Furthermore,

$$\Phi(u) = i\gamma u - \frac{1}{2}\sigma^2 u^2 + \int_{\mathbb{R}} \left(e^{iux} - 1 - iux\mathbf{1}_{[0,1]}(|x|)\right) \nu(dx) \tag{2.3}$$

is the cumulant transform of an infinitely divisible distribution with characteristic triplet (γ, σ^2, ν), where $\gamma \in \mathbb{R}$, $\sigma^2 \geq 0$ and ν is a Lévy-measure on $\mathbb{R}$, i.e.,

$$\nu(\{0\}) = 0 \quad \text{and} \quad \int_{\mathbb{R}} \left(1 \wedge x^2\right) \nu(dx) < \infty.$$

The quadruplet $(\gamma, \sigma^2, \nu, \pi)$ determines the distribution of the Lévy basis, and therefore, it is called its *characteristic quadruplet*. An important random

variable associated with the Lévy basis, is the so-called *Lévy seed*, which we define as the random variable Λ' having as cumulant transform (2.3), that is

$$\mathbb{E}\left[e^{iu\Lambda'}\right] = e^{\Phi(u)}. \tag{2.4}$$

By selecting different Lévy seeds, it is easy to compute the distribution of $\Lambda(B)$ for $B \in \mathcal{B}_b(I)$, for example, when $I = \mathbb{R} \times \mathbb{R}^d$. In the following two examples, we compute the Lévy bases used in generating the data sets in Section 4.2.

Example 2.2 (Gaussian Lévy basis)**.** *Let $\Lambda' \sim \mathcal{N}(\gamma, \sigma^2)$, then its characteristic function is equal to $\exp(i\gamma u - \frac{1}{2}\sigma^2 u^2)$. Because of (2.2), we have, in turn, that the characteristic function of $\Lambda(B)$ is equal to $\exp(i\gamma u \lambda_{1+d}(B) - \frac{1}{2}\sigma^2 \lambda_{1+d}(B)u^2)$. In conclusion, $\Lambda(B) \sim \mathcal{N}(\gamma\lambda_{1+d}(B), \sigma^2\lambda_{1+d}(B))$ for any $B \in \mathcal{B}_b(I)$.*

Example 2.3 (Normal Inverse Gaussian Lévy basis)**.** *Let K_1 denote the modified Bessel function of the third order and index 1. Then, for $x \in \mathbb{R}$, the NIG distribution is defined as*

$$f(x : \alpha, \beta, \mu, \delta)$$
$$= \alpha\delta(\pi^2(\delta^2 + (x-\mu)^2))^{-\frac{1}{2}} \exp(\delta\sqrt{\alpha^2-\beta^2} + \beta(x-\mu))K_1(\alpha\sqrt{\delta^2 + (x-\mu)^2}),$$

where α, β, μ and δ are parameters such that $\mu \in \mathbb{R}$, $\delta > 0$ and $0 \le |\beta| < \alpha$. Let $\Lambda' \sim NIG(\alpha, \beta, \mu, \delta)$, then by (2.2) we have that $\Lambda(B) \sim NIG(\alpha, \beta, \mu\lambda_{1+d}(B), \delta\lambda_{1+d}(B))$ for all $B \in \mathcal{B}_b(I)$.

We now follow [8] and [11] to formally define ambit sets.

Definition 2.4. *A family of ambit sets $(A_t(x))_{(t,x)\in\mathbb{R}\times\mathbb{R}^d}$ is a collection of subsets of $\mathbb{R} \times \mathbb{R}^d$, i.e. $A_t(x) \subset \mathbb{R} \times \mathbb{R}^d$ for every $(t,x) \in \mathbb{R} \times \mathbb{R}^d$, which satisfies the following properties:*

$$\begin{cases} A_t(x) = A_0(0) + (t,x), \ \textit{(Translation invariant)} \\ A_s(x) \subset A_t(x), \ \textit{for } s < t \\ A_t(x) \cap (t, \infty) \times \mathbb{R}^d = \emptyset \ \textit{(Non-anticipative)}. \end{cases} \tag{2.5}$$

We further assume that the random fields $\boldsymbol{Z} := (\boldsymbol{Z}_t(x))_{(t,x)\in\mathbb{R}\times\mathbb{R}^d}$ in the paper are *influenced*. By this name, we mean random fields defined on a given complete probability space $(\Omega, \mathcal{F}', P)$, equipped with the *filtration of influence* (in the sense of Definition 3.8 in [28]) $\mathbb{F} = (\mathcal{F}_{(t,x)})_{(t,x)\in\mathbb{R}\times\mathbb{R}^d}$ generated by Λ and the family of ambit sets $(A_t(x))_{(t,x)\in\mathbb{R}\times\mathbb{R}^d} \subset \mathbb{R} \times \mathbb{R}^d$, i.e., each $\mathcal{F}_{(t,x)}$ is the σ-algebra generated by the set of random variables $\{\Lambda(B) : B \in \mathcal{B}_b(H \times A_t(x))\}$, which are adapted to $\mathbb{F}$. We call our field *adapted* to the filtration of influence $\mathbb{F}$ if it is measurable with respect to the σ-algebra $\mathbb{F}$ for each $(t,x) \in \mathbb{R} \times \mathbb{R}^d$.

Moreover, we work with spatio-temporal stationary random fields in the following.

Definition 2.5 (Spatio-temporal stationarity)**.** *We say that $\boldsymbol{Z}$ is spatio-temporal stationary if for every $n \in \mathbb{N}$, $\tau \in \mathbb{R}$, $u \in \mathbb{R}^d$, $t_1, \ldots, t_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathbb{R}^d$, the joint distribution of $(\boldsymbol{Z}_{t_1}(x_1), \ldots, \boldsymbol{Z}_{t_n}(x_n))$ is the same as that of $(\boldsymbol{Z}_{t_1+\tau}(x_1 + u), \ldots, \boldsymbol{Z}_{t_n+\tau}(x_n + u))$.*

We simply use the term *stationary* throughout when referring to processes satisfying Definition 2.5. We can now formally define the stochastic model underlying our learning methodology.

Definition 2.6 (MMAF)**.** *Let $\Lambda = \{\Lambda(B), B \in \mathcal{B}_b(I)\}$ a Lévy basis, $f : H \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ a $\mathcal{B}(I)$-measurable function and $A_t(x)$ an ambit set. Then, the stochastic integral*

$$\boldsymbol{Z}_t(x) = \int_H \int_{A_t(x)} f(A, x - \xi, t - s) \; \Lambda(dA, d\xi, ds), \;\; (t, x) \in \mathbb{R} \times \mathbb{R}^d, \qquad (2.6)$$

is adapted to the filtration $\mathbb{F}$, stationary, and its distribution is infinitely divisible. We call the $\mathbb{R}$-valued random field $\boldsymbol{Z}$ an (influenced) mixed moving average field and f its kernel function.

Remark 2.7. *On a technical level, we assume all stochastic integrals in this paper to be well defined in the sense of Rajput and Rosinski [65]. For more details, including sufficient conditions on the existence of the integral as well as the explicit representation of the characteristic triplet of the MMAF's infinitely divisible distribution (which can be directly determined from the characteristic quadruplet of Λ), we refer to [28, Section 3.1]. In the latter, there can also be found a multivariate definition of a Lévy basis and an MMAF.*

Important examples of MMAFs are the spatio-temporal Ornstein-Uhlenbeck field (STOU) and the mixed spatio-temporal Ornstein-Uhlenbeck field (MSTOU), whose properties have been thoroughly analyzed in [59] and [60]. There are also interesting time series models in the MMAF framework, which we present in section A.5.

Example 2.8 (STOU process)**.** *Let $\Lambda = \{\Lambda(B), B \in \mathcal{B}_b(I)\}$ be a Lévy basis, $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ a $\mathcal{B}(I)$-measurable function defined as $f(s, \xi) = \exp(-As)$ for $A > 0$, and $A_t(x)$ be defined as in (2.9). Then, the STOU is defined as*

$$\boldsymbol{Z}_t(x) := \int_{A_t(x)} \exp(-A(t - s))\Lambda(ds, d\xi). \qquad (2.7)$$

The STOU is a stationary and Markovian random field. Moreover, an STOU exhibits exponential temporal autocorrelation (just like the temporal Ornstein-Uhlenbeck process) and has a spatial autocorrelation structure determined by the shape of the ambit set. In addition, this class of fields admits non-separable autocovariances, which are desirable in practice, see Example A.7.

Example 2.9 (MSTOU process)**.** *An MSTOU process is defined by mixing the parameter A in the definition of an STOU process; that is, we assume that π has support in $H = (0, \infty)$. This modification allows the determination of*

random fields with power-decaying autocovariance functions, see example A.9. Let $\Lambda = \{\Lambda(B), B \in \mathcal{B}_b(I)\}$ be a Lévy basis, $f : (0, \infty) \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ a $\mathcal{B}(I)$-measurable function defined as $f(A, s, \xi) = \exp(-As)$, and $A_t(x)$ be defined as in (2.9). Moreover, let $l(A)$ be the density of π with respect to the Lebesgue measure such that

$$\int_0^\infty \frac{1}{A^{d+1}} l(A) dA < \infty.$$

Then, the MSTOU is defined as

$$\boldsymbol{Z}_t(x) := \int_0^\infty \int_{A_t(x)} \exp(-A(t-s)) \Lambda(dA, ds, d\xi). \tag{2.8}$$

The MMAF framework has a *causal interpretation* under the following assumption.

Assumption 2.10. *For a $c > 0$, we consider*

$$A_t(x) := \left\{ (s, \xi) \in \mathbb{R} \times \mathbb{R}^d : s \leq t \text{ and } \|x - \xi\| \leq c|t - s| \right\}. \tag{2.9}$$

To explain why using cone-shaped ambit sets allows to have such an interpretation, we borrow the concept of *lightcone* from special relativity.

A lightcone describes the possible paths that the light can make in space-time leading to a space-time point (t, x) and the ones that lie in its future. In the context of our paper, we use their geometry to identify the space-time points having a causal relationship. For a point (t, x), $c > 0$ and by using the Euclidean norm to assess the distance between different space-time points, we define a *lightcone* as the set

$$\mathcal{A}_t^{light}(x) = \left\{ (s, \xi) \in \mathbb{R} \times \mathbb{R}^d : \|x - \xi\| \leq c|t - s| \right\}.$$

The set $\mathcal{A}_t^{light}(x)$ can be split into two disjoint sets, namely, $A_t(x)$ and $A_t(x)^+$. The set $A_t(x)$ is called *past lightcone*, and its definition corresponds to the one of a cone-shaped ambit set (2.9).

The set

$$A_t(x)^+ = \{(s, \xi) \in \mathbb{R} \times \mathbb{R}^d : s > t \text{ and } \|x - \xi\| \leq c|t - s|\}, \tag{2.10}$$

is called instead the *future lightcone*. By using an influenced MMAF on a cone-shaped ambit set as the underlying model, we implicitly assume that the following sets

$$l^-(t,x) = \{\boldsymbol{Z}_s(\xi) : (s, \xi) \in A_t(x) \setminus (t, x)\} \text{ and } l^+(t,x) = \{\boldsymbol{Z}_s(\xi) : (s, \xi) \in A_t(x)^+\} \tag{2.11}$$

are respectively describing the values of the field that have a direct influence on the determination of $\boldsymbol{Z}_t(x)$ and the future field values influenced by $\boldsymbol{Z}_t(x)$. We can then uncover the causal relationship between space-time points described above by estimating the constant c from observed data, which we call *the speed*

of information propagation in the physical system under analysis. A similar approach to modeling causal relationships can be found in several machine learning frameworks, as in [57] and [72]. The sets (2.11) are employed in the former to perform video frame prediction and in the latter to discover coherent structures, as defined in [48], in spatio-temporal physical systems. Hence, the concept of *lightcones* enters into play in the definition of their algorithm. In statistical modeling, we typically have two equivalent approaches towards causality: structural causal models, which rely on the use of directed acyclical graphs (DAG) [61], and Rubin causal models, which rely upon the potential outcomes framework [71]. The concept of causality employed in this paper can be inscribed into the latter. In fact, by using MMAFs on cone-shaped ambit sets, the set $l^+(t, x)$ describes the possible future outcomes that can be observed starting from the spatial position (t, x). Finally, we consider the following definitions of temporal and spatial short and long-range dependence in the paper.

Definition 2.11 (Short and long range dependence)**.** *A random field* $(\mathbf{Z}_t(x))_{(t,x)\in\mathbb{R}\times\mathbb{R}^d}$
is said to have temporal short-range dependence *if*

$$\int_0^\infty Cov(\mathbf{Z}_t(x), \mathbf{Z}_{t+\tau}(x)) \, d\tau < \infty,$$

and temporal long-range dependence *if the integral above is infinite. Similarly, an isotropic random field, see Definition A.10, has* spatial short-range dependence *if*

$$\int_0^\infty C(r) \, dr < \infty,$$

where $Cov(\mathbf{Z}_t(x), \mathbf{Z}_t(x+u)) = C(|u|)$ *and* $r = |u|$. *It is said to have* spatial long-range dependence *if the integral is infinite.*

Under Assumption 2.10, an STOU process admits temporal and spatial short-range dependence, whereas an MSTOU process can admit temporal and spatial short and long-range dependence by carefully modeling the parameter A.

Example 2.12. *Let* $\mathbf{Z}$ *be an MSTOU process as defined in Example 2.9 for* $d = 1$, *Assumption 2.10 hold, and* $l(A) = \frac{\beta^\alpha}{\Gamma(\alpha)} A^{\alpha-1} \exp(-\beta A)$ *be the Gamma density with shape and rate parameters* $\alpha > d+1$ *and* $\beta > 0$. *From the calculations in Example A.9 and by setting* $u = 0$, *then* $\mathbf{Z}$ *has temporal short-range dependence because for* $\alpha > 3$, *the integral*

$$\int_0^\infty Cov(\mathbf{Z}_t(x), \mathbf{Z}_{t+\tau}(x)) \, d\tau = \frac{c\beta^\alpha Var(\Lambda')}{2(\alpha-2)(\alpha-1)} \int_0^\infty (\beta+\tau)^{-(\alpha-2)} \, d\tau$$
$$= \frac{c\beta^3 Var(\Lambda')}{2(\alpha-1)(\alpha-2)(\alpha-3)}$$

is finite. This integral is infinite for $2 < \alpha \le 3$, *and we then say that the MSTOU process has temporal long-range dependence. We obtain spatial short*

or long-range dependence for the same choice of parameters. In fact, for $r = |u|$ and $\tau = 0$, and $\alpha > 3$, the integral

$$\int_0^\infty C(r)\,dr = \frac{c\beta^\alpha Var(\Lambda')}{2(\alpha-2)(\alpha-1)} \int_0^\infty (\beta + r/c)^{-(\alpha-2)}\,dr$$
$$= \frac{c\beta^3 Var(\Lambda')}{2(\alpha-1)(\alpha-2)(\alpha-3)}$$

converges, whereas it diverges for $2 < \alpha \leq 3$.

2.3. Weak dependence coefficients in MMAF-guided learning

We start by giving the definitions of the dependence notions involved in our learning methodology.

Definition 2.13. *Let $\mathbf{Z}$ be an $\mathbb{R}$-valued random field. Then, $\mathbf{Z}$ is called θ-lex-weakly dependent if*

$$\theta_{lex}(r) = \sup_{u,v \in \mathbb{N}} \theta_{u,v}(r) \xrightarrow[r \to \infty]{} 0,$$

where

$$\theta_{u,v}(r) = \sup \left\{ \frac{|Cov(F(\mathbf{Z}_\Gamma), G(\mathbf{Z}_{\Gamma'}))|}{\|F\|_\infty v Lip(G)},\ F \in \mathcal{G}_u^*,\ G \in \mathcal{G}_v,\ \Gamma,\ \Gamma',\ |\Gamma| = u,\ |\Gamma'| = v \right\}$$

for $\Gamma = \{t_{i_1}, \ldots, t_{i_u}\} \in \mathbb{R}^{1+d}$, $\Gamma' = \{t_{j_1}, \ldots, t_{j_v}\} \in \mathbb{R}^{1+d}$ such that $\Gamma \in V_{\Gamma'}^r = \bigcap_{l=1}^v V_{t_{j_l}}^r$ for $t_{j_l} \in \Gamma'$. We call $(\theta_{lex}(r))_{r \in \mathbb{R}^+}$ the θ-lex-coefficients.

MMAFs are θ-lex weakly dependent random fields, as proven in Proposition B.1. Moreover, Definition 2.13 is an extension to the random field case of a dependence notion developed for *causal processes* called θ-weak dependence.

Definition 2.14. *Let $\mathbf{Z}$ be an $\mathbb{R}$-valued stochastic process. Then, $\mathbf{Z}$ is called θ-weakly dependent if*

$$\theta(k) = \sup_{u \in \mathbb{N}} \theta_u(k) \xrightarrow[k \to \infty]{} 0,$$

where

$$\theta_u(k) = \sup \left\{ \frac{|Cov(F(\mathbf{Z}_\Gamma), G(\mathbf{Z}_{j_1}))|}{\|F\|_\infty Lip(G)},\ F \in \mathcal{G}_u^*,\ G \in \mathcal{G}_1,\ \Gamma, |\Gamma| = u \right\}.$$

for $\Gamma = \{t_{i_1}, \ldots, t_{i_u}\} \in \mathbb{R}$ such that $i_1 \leq i_2 \leq \ldots \leq i_u \leq i_u + k \leq j_1$. We call $(\theta(k))_{k \in \mathbb{R}^+}$ the θ-coefficients.

We use the following result extensively in Section 3 proofs.

Remark 2.15 (Projective-type representation of θ-weak dependence)**.**
Let $(X_t)_{t\in\mathbb{Z}}$ be a real-valued θ-lex weakly dependent process, $\mathcal{L}_1 = \{g : \mathbb{R} \to \mathbb{R}, g \in \mathcal{G}_1, Lip(g) \leq 1\}$, $s \in \mathbb{N}$, $j_1 \in \mathbb{Z}$, and $\mathcal{M} = \sigma\{\boldsymbol{X}_t : t \leq j_1 \text{ and } |t-j_1| \geq s\}$, then it is showed in [33, Proposition 2.3] that

$$\theta(s) = \sup_{j_1 \in \mathbb{Z}} \sup_{g \in \mathcal{L}_1} \|\mathbb{E}[g(\boldsymbol{X}_{j_1})|\mathcal{M}] - \mathbb{E}[g(\boldsymbol{X}_{j_1})]\|_1. \tag{2.12}$$

An alternative proof of this result can also be found in [28, Lemma 5.1].

Remark 2.16 (About the parameter k and r in Definition 2.14 and 2.13)**.**
Let us start by assuming that $\boldsymbol{Z} = (\boldsymbol{Z}_i)_{i\in\mathbb{R}}$ is a sequence of independent and identically distributed (in short, i.i.d.) random variables, then for any $F \in \mathcal{G}_u^*$ with $u \in \mathbb{N}$, $G \in \mathcal{G}_1$, and selecting a set $\Gamma = \{i_1, \ldots, i_u\} \in \mathbb{R}$ such that $i_1 \leq \ldots \leq i_u \leq i_u + k \leq j$, we have that

$$Cov(F(\boldsymbol{Z}_\Gamma), G(\boldsymbol{Z}_j)) = 0,$$

and $\theta(k) = 0$ for all $k \in \mathbb{R}^+$. The process $\boldsymbol{Z}$ is θ-weakly dependent and the parameter k is encoding the distance between the marginals

$$\boldsymbol{P} := (\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_u}), \text{ and } \boldsymbol{F} := (\boldsymbol{Z}_j).$$

In terms of the process $\boldsymbol{Z}$, the σ-algebras generated by $\boldsymbol{P}$ and $\boldsymbol{F}$ represent past *and* future *events. Obviously, in the case of a sequence of independent random variables, the past plays no role in the unfolding of the future. However, if we consider a θ-weakly dependent process $\boldsymbol{Z}$, then it satisfies the inequality*

$$Cov(F(\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_u}), G(\boldsymbol{Z}_j)) \leq 2\|F\|_\infty Lip(G)\theta(k).$$

The result above has been proven, for example, in [27] and gives a criterion to measure explicitly *the dependence between past and future. Here, the past is progressively forgotten for $k \to \infty$. The parameter k expresses the distance at which we are evaluating the influence of the past on how the future unfolds, and the coefficients $\theta(k)$ is a measure of how fast the past is forgotten.*

Similar considerations can be made in the case in which we consider $\boldsymbol{Z} = (\boldsymbol{Z}_i)_{i\in\mathbb{R}^{1+d}}$ a θ-lex weakly dependent random field for the constant $r \in \mathbb{R}^+$. However, $\boldsymbol{P}$ and $\boldsymbol{F}$ represent the marginals of lexicographically ordered elements in this case. The lexicographic order in $\mathbb{R}^{1+d}$ substitutes the natural temporal order for stochastic processes defined on $\mathbb{R}$.

In the MMAF modeling framework, we can show general formulas computing upper bounds of the θ-lex coefficients. The latter is given as a function of the characteristic quadruplet of the driving Lévy basis Λ and the kernel function f in (2.6), see Proposition B.1.

For $d = 1, 2$, when the MMAF has a kernel function with no spatial component, we can compute a bound for the θ-lex coefficients expressed in terms of the covariances of the field $\boldsymbol{Z}$, which can be computed as shown in Section A.2. These bounds have an expression that allows us to use standard statistical inference tools to infer the decay rate parameter of the coefficients; see, for example,

the estimators (2.14) and (2.15). In the general framework described in Proposition B.1, however, similar estimation methodologies are not yet available and remain an interesting open problem.

Proposition 2.17. *Let Λ be an $\mathbb{R}$-valued Lévy basis with characteristic quadruplet $(\gamma, \sigma^2, \nu, \pi)$ and $f : H \times \mathbb{R} \to \mathbb{R}$ a $\mathcal{B}(H \times \mathbb{R})$-measurable function not depending on the spatial dimension, i.e.,*

$$\boldsymbol{Z}_t(x) = \int_H \int_{A_t(x)} f(A, t - s)\Lambda(dA, ds, d\xi), \quad (t, x) \in \mathbb{R}^{1+d}. \tag{2.13}$$

(i) For $d = 1$, if $\int_{|x|>1} x^2 \nu(dx) < \infty$ and $\gamma + \int_{|x|>1} x\nu(dx) = 0$, then $\boldsymbol{Z}$ is θ-lex weakly dependent and

$$\theta_{lex}(r) \leq 2 \left(2Var(\Lambda') \int_0^\infty \int_{A_0(0) \cap A_0(r \min(2,c))} f(A, -s)^2 ds d\xi \pi(dA) \right)^{\frac{1}{2}}$$
$$= 2\sqrt{2Cov(\boldsymbol{Z}_0(0), \boldsymbol{Z}_0(r \min(2, c)))},$$

where $Var(\Lambda') = \sigma^2 + \int_{\mathbb{R}} x^2 \nu(dx)$.

(ii) For $d = 2$, if $\int_{|x|>1} x^2 \nu(dx) < \infty$ and $\gamma + \int_{|x|>1} x\nu(dx) = 0$, then $\boldsymbol{Z}$ is θ-lex weakly dependent and

$$\theta_{lex}(r) \leq 2 \left(2Cov \left(\boldsymbol{Z}_0(0, 0), \boldsymbol{Z}_0 \left(r \min \left(1, \frac{c}{\sqrt{2}} \right), r \min \left(1, \frac{c}{\sqrt{2}} \right) \right) \right) \right.$$
$$\left. + 2Cov \left(\boldsymbol{Z}_0(0, 0), \boldsymbol{Z}_0 \left(r \min \left(1, \frac{c}{\sqrt{2}} \right), -r \min \left(1, \frac{c}{\sqrt{2}} \right) \right) \right) \right)^{\frac{1}{2}}.$$

The proof of the results above is given in Appendix B.1.

Notation 2.18. *In general, we indicate the bounds of the θ-lex coefficients determined in Proposition B.1, Corollary B.2 or Proposition 2.17 using the sequence $(\tilde{\theta}_{lex}(r))_{r \in \mathbb{R}^+}$ where*

$$\theta_{lex}(r) \leq 2\tilde{\theta}_{lex}(r).$$

Definition 2.19. *For $r \in \mathbb{R}^+$, $\bar{\alpha} > 0$, if $\tilde{\theta}_{lex}(r) \leq \bar{\alpha} \exp(-\lambda r)$ we say that $\boldsymbol{Z}$ admits exponentially decaying θ-lex coefficients, whereas if $\tilde{\theta}_{lex}(r) \leq \bar{\alpha} r^{-\lambda}$, we say that $\boldsymbol{Z}$ admits power decaying θ-lex coefficients.*

We give below examples of MMAF with exponential and power-decaying θ-lex coefficients.

Example 2.20. *Let $d = 1$ and $\boldsymbol{Z}$ be an STOU as in Def. 2.8. If $\int_{|x|>1} x^2 \nu(dx) < \infty$, $\gamma + \int_{|x|>1} x\, \nu(dx) = 0$, then $\boldsymbol{Z}$ is θ-lex weakly dependent with*

$$\tilde{\theta}_{lex}(r) = \left(Var(\Lambda') \int_{A_0(0) \cap (A_0(\psi) \cup A_0(-\psi))} \exp(2As)\, ds\, d\xi \right)^{\frac{1}{2}}$$

$$\leq \Big(2 Var(\Lambda') \int_{A_0(0) \cap A_0(\psi)} \exp(2As)\, ds\, d\xi \Big)^{\frac{1}{2}}$$

$$= \Big(2 Var(\Lambda') \int_{-\infty}^{-\frac{\psi}{2c}} \int_{\psi+cs}^{-cs} \exp(2As)\, ds\, d\xi \Big)^{\frac{1}{2}} = \Big(\frac{c}{A^2} Var(\Lambda') \exp\Big(\frac{-A\psi}{c} \Big) \Big)^{\frac{1}{2}}$$

$$= \Big(\frac{c}{A^2} Var(\Lambda') \exp\Big(- \underbrace{\frac{A\min(2,c)}{c}}_{2\lambda} r \Big) \Big)^{\frac{1}{2}}$$

$$= \sqrt{2 Cov(\boldsymbol{Z}_0(0), \boldsymbol{Z}_0(r\min(2,c)))} := \bar{\alpha} \exp(-\lambda r),$$

where $\lambda > 0$ and $\bar{\alpha} > 0$. Because the temporal and spatial autocovariance functions of an STOU are exponential, see (A.4), the model admits temporal and spatial short-range dependence. By estimating the parameter vector $\theta_0 = \{A, c, Var(\Lambda')\}$ using the methodologies revised in Appendix A.3, we can estimate the parameter λ using the following plug-in estimator

$$\boldsymbol{\lambda^*} = \frac{\min(2, \boldsymbol{c^*})}{2\boldsymbol{c^*}}, \tag{2.14}$$

where the estimators $\boldsymbol{A^*}$ and $\boldsymbol{c^*}$ are defined in (A.19). This estimator is consistent because of [59, Theorem 12] and the continuous mapping theorem. Furthermore, by using an estimator of the parameter $Var(\Lambda')$, we can also obtain a consistent estimator for the parameter $\bar{\alpha}$.

Example 2.21. *Let $d = 1$ and $\boldsymbol{Z}$ be an MSTOU as defined in Example 2.12. If $\int_{|x|>1} x^2\, \nu(dx) < \infty$, $\gamma + \int_{|x|>1} x\, \nu(dx) = 0$, then $\boldsymbol{Z}$ is θ-lex weakly dependent with*

$$\tilde{\theta}_{lex}(r) \leq \Big(\frac{c}{A^2} Var(\Lambda') \int_0^{\infty} \exp\Big(\frac{-A\psi}{c} \Big) \pi(dA) \Big)^{\frac{1}{2}}$$

$$= \Big(\frac{Var(\Lambda')c\beta^{\alpha}}{(\beta + \psi/c)^{\alpha-2}(\alpha - 2)(\alpha - 1)} \Big)^{\frac{1}{2}}$$

$$= \Big(\frac{Var(\Lambda')c\beta^{\alpha}}{(\alpha - 2)(\alpha - 1)} \Big(\beta + \frac{r\min(2,c)}{c} \Big)^{-(\alpha-2)} \Big)^{\frac{1}{2}}$$

$$= \sqrt{2 Cov(\boldsymbol{Z}_0(0), \boldsymbol{Z}_0(r\min(2,c)))} := \bar{\alpha} r^{-\lambda},$$

where $\lambda = \frac{\alpha-2}{2}$ and $\bar{\alpha} > 0$. As already addressed in Example 2.12, for $2 < \alpha \leq 3$, that is $0 < \lambda \leq \frac{1}{2}$, the model admits temporal and spatial long-range dependence. Instead, for $\alpha > 3$, that is $\lambda > \frac{1}{2}$, the model admits temporal and spatial short range dependence.

For the model used in Example 2.21, an estimator for

$$\boldsymbol{\lambda^*} = \frac{\boldsymbol{\alpha^*} - 2}{2} \tag{2.15}$$

where the vector of parameters $\theta_1 = \{\alpha, \beta, c, Var(\Lambda')\}$ is estimated using a GMM estimator $\boldsymbol{\theta_1^*} = \{\boldsymbol{\alpha^*}, \boldsymbol{\beta^*}, \boldsymbol{c^*}, \boldsymbol{Var(\Lambda')^*}\}$.
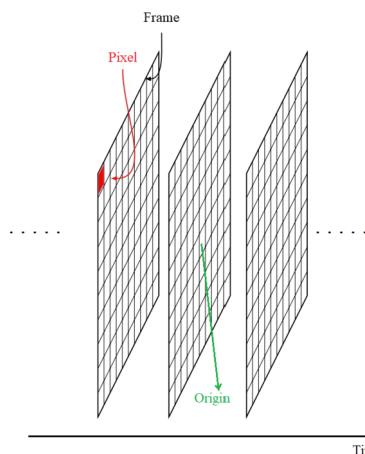
FIG 1. *Raster data cube's spatio-temporal index set with origin in* (t_0, x_0).

The estimator of λ depends on the chosen data generating process and the spatial dimension of the data. For further details on the parametric estimation methodologies nowadays available for other fields belonging to the MMAF class, we refer the reader to the Appendices A.3, A.4 and A.5.

3. Mixed moving average field guided learning

3.1. Pre-processing N frames

In this section, we describe MMAF-guided learning for a spatial dimension $d = 2$. Let $(\tilde{Z}_t(x))_{(t,x) \in \mathbb{T} \times \mathbb{L}}$ be an observed data set on a regular lattice $\mathbb{L} \subset \mathbb{R}^2$ across times $\mathbb{T} = \{t_0 + h_t, \ldots, t_0 + h_t N\}$ for $h_t \in \mathbb{R}$, such that

$$\tilde{Z}_t(x) = \mu_t(x) + Z_t(x) \tag{3.1}$$

holds, and no measurement errors are present in the observations. Here, $\mu_t(x)$ is a deterministic function, and $Z_t(x)$ are considered realizations from a zero mean stationary (influenced) MMAF.

We represent graphically the regular spatial lattice $\mathbb{L}$ as a *frame made of a finite amount of pixels*, i.e., squared-cells representing each of them a unique spatial position $x \in \mathbb{L}$, see Figure 1. In several applications, such as satellite imagery, a pixel refers to a spatial cell of several square meters. In the paper, we assume that a pixel represents the spatial point $x \in \mathbb{R}^2$ corresponding to the center of the pictured squared cell. We then use the name *pixel* and spatial position throughout interchangeably. In total, N frames represent the spatio-temporal index set of the observed data set. This terminology is often used to describe *raster data cubes* [62]. We define in Section 3.1.1 a spatio-temporal embedding for data with such structure. For dimension $d = 1$, we consider that

the pixel collapses in the point $x \in \mathbb{R}$ that describes, see Figure 3. MMAF-guided learning also applies to spatial index sets of dimension $d > 2$. However, we do not represent the spatial positions using pixels in such cases.

We call (t_0, x_0) the origin of the space-time grid, see Figure 1, and h_t and h_s the time and space discretization step in the observed data set, i.e. the distance between two pixels along the temporal and spatial dimensions.

MMAF-guided learning has the target to determine one-time ahead ensemble forecasts of the field $\boldsymbol{Z}$ in a pixel x^*. We do not consider further the problem of estimating the deterministic function $\mu_t(x)$ when performing forecasting tasks, i.e., we assume our data set to be generated by a zero mean MMAF from now on. We refer the reader to [26] for a review of how to estimate the function $\mu_t(x)$.

In the following sections, we need to define a *training data set* to employ in our learning methodology. Therefore, we pre-process the set of indices represented by the N frames to select a set of different examples, i.e., input-output pairs, to include in the training data set S_m.

3.1.1. Spatio-temporal embedding

Let us consider a stationary random field $(\boldsymbol{Z}_t(x))_{(t,x) \in \mathbb{Z} \times \mathbb{L}}$, and select a pixel position x^* in $\mathbb{L}$. We define the input-output vectors

$$\boldsymbol{X}_i = \boldsymbol{L}_{\boldsymbol{p}}^-(t_0 + ia, x^*), \quad \text{and} \quad \boldsymbol{Y}_i = \boldsymbol{Z}_{t_0+ia}(x^*), \quad \text{for } i \in \mathbb{Z}, \qquad (3.2)$$

where

$$\boldsymbol{L}_{\boldsymbol{p}}^-(t, x^*) = (\boldsymbol{Z}_{i_1}(\xi_1), \ldots, \boldsymbol{Z}_{i_{a(p,c)}}(\xi_{a(p,c)}))^\top, \qquad (3.3)$$

with indices selected in the set

$$\mathcal{I}(t, x^*) := \{(i_s, \xi_s) : \|x^* - \xi_s\| \le c\,(t - i_s) \text{ for } 0 \lessdot t - i_s \leqq p, \text{ and } (i_s, \xi_s) <_{lex} (i_{s+1}, \xi_{s+1})\}, \qquad (3.4)$$

for $t = t_0 + ia$ with $i \in \mathbb{Z}$, and $c > 0$. We call $a(p, c) := |\mathcal{I}(t, x^*)|$ and assume that is constant for all $t = t_0 + ia$ and $i \in \mathbb{Z}$. The parameters $a > 0$ and $p > 0$ are multiples of h_t such that $a = a_t h_t$, $p = p_t h_t$, $a_t, p_t \in \mathbb{N}$ and $a_t \ge p_t + 1$. We note that each element of $\boldsymbol{L}_{\boldsymbol{p}}^-(t, x^*) \subset l^-(t, x^*)$ for $t = t_0 + ia$ and $i \in \mathbb{Z}$, where $l^-(t, x^*)$ is defined in (2.11) and identifies the set of all points in $\mathbb{R} \times \mathbb{R}^d$ that could possibly influence the realization $\boldsymbol{Z}_t(x^*)$. The sampling leading to (3.2) can be performed starting by a pixel x^* for which the index set $\mathcal{I}(t_0 + ia, x^*) \subseteq \mathbb{Z} \times \mathbb{L}$ for all i. Further details on why the examples need to be structured following the index sets $\mathcal{I}(t, x^*)$ are given in the next remark.

Remark 3.1 (Geometry and lexicographic order of the examples)**.** *The sets $\mathcal{I}(t, x^*)$ are chosen with a geometry that is inherited by the definition of the cone-shaped ambit set in (2.9). Such a geometry allows us to give a causal interpretation of the one-time ahead ensemble forecast, as shown in Section 4. Moreover, we store in the input vectors $\boldsymbol{X}_i$ values of the fields $\boldsymbol{Z}$ with indices in*
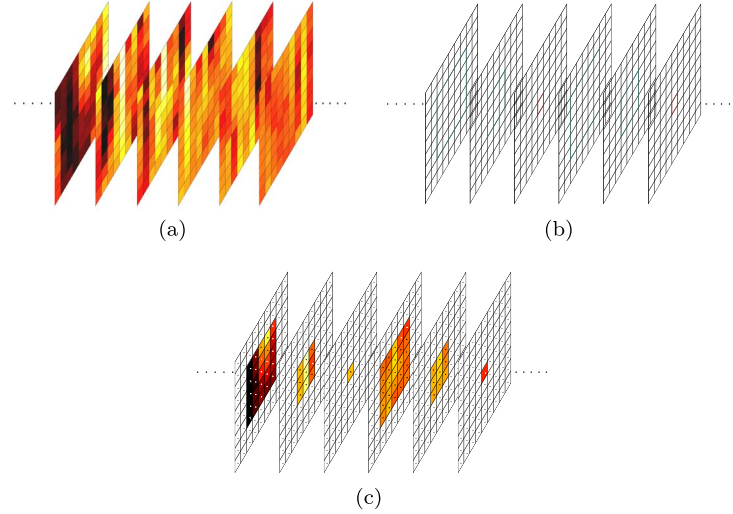
FIG 2. *(a) Observed raster data cube. (b) Spatio-temporal embedding: the red pixel identifies the spatial point x^*, and for $i \in \mathbb{Z}$ the set $\mathcal{I}(t_0 + ia, x^*)$ defined in (3.4) is represented by the pixels in the green boxes. The parameters in use in this sampling are $c = \sqrt{2}$, $p_t = 2$, $a_t = 3$, $h_t = h_s = 1$. (c) A realization S from the cone-shaped sampling process.*

lexicographic order. This choice implies that $((\boldsymbol{X}_{i_1}, \boldsymbol{Y}_{i_1}), \ldots, (\boldsymbol{X}_{i_u}, \boldsymbol{Y}_{i_u}))$ and $(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ for $u \in \mathbb{N}$ and $j \in \mathbb{Z}$ are lexicographically ordered marginals of the field $\boldsymbol{Z}$. This allows, in turn, to precisely assess how the θ-lex weakly dependence of the data generating process $\boldsymbol{Z}$ (defined w.r.t. lexicographically ordered marginals, see Definition 2.13) is inherited by the process $\boldsymbol{S}$ and to understand how the θ-lex weakly dependence plays a role in the definition of the randomized estimators in Section 3.2. Modifications of this representation may be needed when using convolutional or transformer architectures as predictors. This latter issue is outside the scope of the present paper but constitutes an important future research direction of MMAF-guided learning.

Last but not least, it is important to notice that the considerations above rule out the choice of overlapping *examples, i.e., the possible choice of examples that maintain the cone-shaped geometry but have indices not following the lexicographic order. If we were to make this choice, we would then work with a training data set that does not inherit the dependence structure of the data generating process $\boldsymbol{Z}$; see Section 3.1.2 for more details.*

The sequence $((\boldsymbol{X}_i, \boldsymbol{Y}_i)^\top)_{i \in \mathbb{Z}}$ is composed of identically distributed random vectors for all $i \in \mathbb{Z}$. We call $\boldsymbol{S} := ((\boldsymbol{X}_i, \boldsymbol{Y}_i)^\top)_{i \in \mathbb{Z}}$ *a cone-shaped sampling process*. An example of a realization S of the sampling scheme, together with its related spatio-temporal embedding's description can be found in Figure 2. The distribution of $\boldsymbol{S}$ is indicated throughout by $\mathbb{P}$.

Next, let us assume to observe a data set $(Z_t(x))_{(t,x) \in \mathbb{T} \times \mathbb{L}}$, and that we want to determine a one-time ahead ensemble forecast in the pixel x^*. We define

$m := \left\lfloor \frac{N}{a_t} \right\rfloor$ and a training data set $S_m = ((X_i, Y_i)^\top))_{i=1}^m$ as a realization from the *cone-shaped sampling process* $\boldsymbol{S}$ of fixed length m. In particular,

$$X_i = L_p^-(t_0 + ia, x^*), \text{ and } Y_i = Z_{t_0+ia}(x^*) \text{ for } i = 1, \dots, m, \qquad (3.5)$$

where

$$L_p^-(t, x^*) = (Z_{i_1}(\xi_1), \dots, Z_{i_{a(p,c)}}(\xi_{a(p,c)}))^\top, \text{ and } (i_s, \xi_s) \in \mathcal{I}(t, x^*) \qquad (3.6)$$

for $s = 1, \dots, a(p, c)$ and $t = t_0 + ia$ with $i = 1, \dots, m$. We assume that the parameters a and p follow the constraints in Table 1. The index set used to define the training data set (3.5) is a spatio-temporal embedding in the set $\mathbb{R} \times \mathbb{R}^2$. A similar interpretation can be given for the index set defining the cone-shaped sampling process (3.2).

TABLE 1
Parameters involved in pre-processing N observed frames.

Parameters	Constraints	Interpretation
$a := a_t h_t$	$p_t + 1 \leq a_t \leq \left\lfloor \frac{N}{2} \right\rfloor$	*translation vector*
$p := p_t h_t$	$1 \leq p_t < \left\lfloor \frac{N}{2} \right\rfloor - 1$	*past time horizon*
$m := \left\lfloor \frac{N}{a_t} \right\rfloor$	$2 \leq m < N$	*number of examples in S_m*

For an observed raster data cube, we know the value of the constants N, h_t, and h_s, and it remains to select the parameters a_t and p_t. We discuss the selection of the parameter a_t in Section 3.2 and of p_t in Section 4.2, respectively.

3.1.2. Study of the dependence structure of $\boldsymbol{L}$ and $\boldsymbol{L}^\epsilon$

We analyze in this section the dependence structure of the processes $\boldsymbol{L} := (L(h(\boldsymbol{X}_i), \boldsymbol{Y}_i))_{i \in \mathbb{Z}}$ and $\boldsymbol{L}^\epsilon := (L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i))_{i \in \mathbb{Z}}$, where L and L^ϵ are the loss functions defined in Section 1.1, and h is a Lipschitz predictor.

Proposition 3.2. *Let $\boldsymbol{S}$ be the cone-shaped sampling process defined by (3.2), then $\boldsymbol{L}$ is a θ-weakly dependent process for all $h \in \mathcal{H}$. Moreover, for $a, p > 0$, $k \in \mathbb{N}$, and $r = ka - p > 0$, it has coefficients*

$$\theta(k) \leq \tilde{d}(Lip(h)a(p,c) + 1)\left(\frac{2}{\tilde{d}}\mathbb{E}[|\boldsymbol{Z}_t(x) - \boldsymbol{Z}_t^{(r)}(x)|] + \theta_{lex}(r)\right), \qquad (3.7)$$

where $\tilde{d} > 0$ is a constant independent of r, and $\boldsymbol{Z}_t^{(r)}(x) := \boldsymbol{Z}_t(x) \wedge r$.

Remark 3.3 (Locally Lipschitz predictor)**.** *Let the predictor h be a locally Lipschitz function such that $h(0) = 0$ and*

$$|h(x) - h(y)| \leq \tilde{c}\|x - y\|_1(1 + \|x\|_1 + \|y\|_1) \text{ for } x, y \in \mathbb{R}^{a(p,c)},$$

for $\tilde{c} > 0$. Moreover, let $\boldsymbol{Z}$ be a stationary and θ-lex weakly dependent random field such that $|\boldsymbol{Z}| \leq C$ almost surely. An easy generalization of Proposition 3.2,

leads to show that $\boldsymbol{L}$ is θ-weakly dependent with coefficients

$$\theta(k) \leq \tilde{d}(\tilde{c}(1+2C)a(p,c)+1)\Big(\frac{2}{\tilde{d}}\mathbb{E}[|\boldsymbol{Z}_t(x)-\boldsymbol{Z}_t^{(r)}(x)|]+\theta_{lex}(r)\Big), \qquad (3.8)$$

where $r = ka - p > 0$ for $a, p > 0$ and $k \in \mathbb{N}$, $\tilde{d} > 0$ is a constant independent of r, and $\boldsymbol{Z}_t^{(r)}(x) := \boldsymbol{Z}_t(x) \wedge r$.

Remark 3.4. *The spatio-temporal embedding discussed in Section 3.1.1 also apply to θ-weakly dependent time series models $\boldsymbol{Z}$. In such case, the parameter $c = 0$, $\boldsymbol{S}$ is a flat cone-shaped sampling process, and (straightforwardly) a θ-weakly dependent process with coefficients satisfying the bound (3.7).*

In the case of MMAFs, we have obtained explicit bounds for the θ-lex coefficients in Propositions B.1 and 2.17. We now prove that a more refined bound than (3.7) for the θ-coefficients of the process $\boldsymbol{L}$ can be given in this setting.

We consider the following assumption.

Assumption 3.5. *Let $\boldsymbol{Z}$ be an MMAF under the Assumption 2.10 and such that $\mathbb{E}[|\boldsymbol{Z}_t(x)|^2] < \infty$. Let $N \in \mathbb{N}$, $\mathbb{T} = \{t_0 + h_t, \ldots, t_0 + h_t N\}$ and $\mathbb{L} \subset \mathbb{R}^2$. S_m and S are realization from $(\boldsymbol{Z}_t(x))_{(t,x) \in \mathbb{Z} \times \mathbb{L}}$ following the spatio-temporal embedding defined in Section 3.1.1.*

Proposition 3.6. *Let Assumption 3.5 hold. Then $\boldsymbol{L}$ is a θ-weakly dependent process for all $h \in \mathcal{H}$ with coefficients*

$$\theta(k) \leq 2(Lip(h)a(p,c)+1)\widetilde{\theta}_{lex}(r), \qquad (3.9)$$

where $r = ka - p > 0$ for $a, p > 0$ and $k \in \mathbb{N}$. In particular, for linear predictors, i.e., $h_\beta(X) = \beta_0 + \beta_1^T X$, for $\beta := (\beta_0, \beta_1)^\top \in B$ and $B = \mathbb{R}^{a(p,c)+1}$, we have that $\boldsymbol{L}$ is a θ-weakly dependent process for all $\beta \in B$ with coefficients

$$\theta(k) \leq 2(\|\beta_1\|_1 + 1)\widetilde{\theta}_{lex}(r). \qquad (3.10)$$

Lemma A.2 straightforwardly implies that the process $\boldsymbol{L}^\epsilon$ is θ-weakly dependent with the same θ-coefficients as the process $\boldsymbol{L}$ under the assumptions of Proposition 3.2, Remark 3.3 and Proposition 3.6.

3.2. PAC Bayesian bounds for MMAF generated data

Three essential results, namely, the change of measure theorem of Donsker and Varadhan [34], the Markov's inequality, and an exponential inequality, are typically employed to prove a *fixed-time PAC Bayesian bound*, which holds for a given choice of m (which in our framework is related to a given number N of frames). We find in [13] a *scheme of proof* for PAC Bayesian bounds that summarizes the above. There is, however, another scheme of proof described in [23], which allows us to obtain any-time PAC Bayesian bounds, which hold simultaneously for all m. Such methodology avoids using the Markov's and the exponential inequalities by substituting them with the Ville's inequality for

non-negative supermartingale [82]. To the best of our knowledge, in the dependent case, the existing proofs of such bounds are holding in the so-called *online framework* and make use of martingale properties as detailed in [23] and [43].

In this section, we prove a novel exponential inequality for θ-weakly dependent processes (the dependence property of $\boldsymbol{L}^\epsilon$) such to be capable of applying the scheme of proof of [13] and obtaining a fixed-time bound. Moreover, we also prove an any-time bound for θ-weakly dependent processes in the *batch learning framework*. We then define a novel *scheme of proof* for fixed time bounds that combines an any-time bound with the Markov's inequality and the projective-type property of the θ-weak dependence (2.12). The last part of this *scheme of proof* is inspired by the analysis made in [2]. Our examples and simulation studies show that the latter bound allows us to *guide* a randomized estimator employing realistic spatio-temporal embeddings, which means training data sets that can best preserve the serial correlation observed in a given *raster data cube*. From the definition of the spatio-temporal embedding in Section 3.1.1, because the MMAFs are θ-lex weakly dependent fields, the more the parameter a_t becomes bigger, the more the examples in S_m are less correlated. This phenomenon is a consequence of the dependence structure of the data generating process and significantly impacts the generalization performance of the employed estimators.

For a measurable space $(\mathcal{H}, \mathcal{T})$ and for any $(\rho, \pi) \in \mathcal{M}_+^1(\mathcal{H})^2$, where $\rho \ll \pi$ means that ρ is absolutely continuous respect to π with Radon-Nikodym derivative $\frac{d\rho}{d\pi}$, the PAC Bayesian bounds introduced in this section employ the *Kullback-Leibler divergence*

$$KL(\rho, \pi) = \left\{ \begin{array}{ll} \rho\left[\log \frac{d\rho}{d\pi}\right] & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise} \end{array} \right. ,$$

and the *f-divergences* defined as

$$D_f(\rho, \pi) = \left\{ \begin{array}{ll} \pi\left[f\left(\frac{d\rho}{d\pi}\right)\right] & \text{if } \rho \ll \pi \\ +\infty & \text{otherwise} \end{array} \right. .$$

For $\phi_p(x) = x^p$, we have that the chi-square divergence corresponds to $D_{\phi_2-1}(\rho, \pi)$.

We prove next an exponential inequality for θ-weakly dependent processes using the notations introduced in Section 3.1.

Theorem 3.7. *Let $\boldsymbol{Z}$ be an $\mathbb{R}$-valued stationary θ-weakly dependent process, and $f : \mathbb{R} \to [a, b]$, for $a, b \in \mathbb{R}$ such that $(f(\boldsymbol{Z}_i))_{i \in \mathbb{Z}}$ is itself θ-weakly dependent. Let $l = \lfloor \frac{m}{k} \rfloor$, for $m, k \in \mathbb{N}$, such that $l \geq 2$ and $0 < s < \frac{3l}{|b-a|}$, then*

$$\mathbb{E}\left[\exp\left(\frac{s}{m}\sum_{i=1}^m (f(\boldsymbol{Z}_i) - \mathbb{E}[f(\boldsymbol{Z}_i)])\right)\right] \leq \exp\left(\frac{s^2 Var(f(\boldsymbol{Z}_1))}{2l\left(1 - \frac{s|b-a|}{3l}\right)}\right) + \exp(s|b-a|)\theta(k)s,$$

$$(3.11)$$

$$\mathbb{E}\left[\exp\left(\frac{s}{m}\sum_{i=1}^m (\mathbb{E}[f(\boldsymbol{Z}_i)] - f(\boldsymbol{Z}_i))\right)\right] \leq \exp\left(\frac{s^2 Var(f(\boldsymbol{Z}_1))}{2l\left(1 - \frac{s|b-a|}{3l}\right)}\right) + \exp(s|b-a|)\theta(k)s.$$

$$(3.12)$$

Note that increasing the parameter k makes the bounds in (3.11) and (3.12) become tighter because $\theta(k)$ goes to zero when k goes to infinity.

We have shown in the last section that the process $\boldsymbol{L}^\epsilon$ is θ-weakly dependent. We then apply Theorem 3.7 in the scheme of proof described in [13] and determine a *fixed-time PAC Bayesian bound*.

Theorem 3.8 (Fixed-time PAC Bayesian bound- *Type I*)**.** *Let* $0 < \epsilon < 3$, $l = \left\lfloor \frac{m}{k} \right\rfloor$ *such that* $l \geq 2$, *and Assumption 3.5 holds. If* $\pi \in \mathcal{M}_+^1(\mathcal{H})$ *such that* $\pi[\theta(k)] < \infty$, $\hat{\rho}$ *is a randomized estimator in the sense of Definition (1.2)*, $\hat{\rho}(\cdot, \omega) \ll \pi(\cdot)$ *for all* $\omega \in \Omega$, *and* $\delta \in (0, 1)$, *then*

$$
\mathbb{P}\Big\{ \forall \hat{\rho} : |\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)]| \leq \Big(KL(\hat{\rho}, \pi) + \log\Big(\frac{1}{\delta}\Big) \Big) \frac{1}{\sqrt{l}}
$$
$$
+ \frac{1}{\sqrt{l}} \log \Big(\exp\Big(\frac{3\epsilon^2}{2(3 - \epsilon)}\Big) + \pi\Big[3\sqrt{l}\exp(3\sqrt{l})\theta(k)\Big] \Big) \Big\} \geq 1 - 2\delta,
$$
$$
(3.13)
$$

where $\theta(k)$ *is a* θ-*coefficient of the process* $\boldsymbol{L}^\epsilon$.

For any possible choice of the parameters defining the spatio-temporal embedding leading to $\boldsymbol{S}_m$, Theorem 3.8 gives us a PAC Bayesian bound for a randomized estimator $\hat{\rho}$. The fastest convergence rate that can be obtained in this framework is $\mathcal{O}(m^{-1/2})$ when choosing the parameter $k = 1$.

Given an accuracy level ϵ, if the right-hand side of the PAC Bayesian bound is less than ϵ with a high probability (i.e., the bound is not vacuous), we say that a randomized estimator $\hat{\rho}$ has *good generalization performance*. The smaller the right-hand side becomes, the better the performance of a randomized estimator in forecasting is. We then aim to select a spatio-temporal embedding such as the right-hand side of the inequality (3.13) is small as possible. For the reader's convenience, we have summarized all the parameters appearing in MMAF-guided learning in Table 2.

TABLE 2
Overview of parameters appearing in MMAF-guided learning.

Parameters	Type	Interpretation
ϵ	Hyperparameter	Accuracy level
N	Given Parameter	Number of frames
h_t	Given Parameter	time step
h_s	Given Parameter	space step
λ	Unknown Parameter	Decay rate of the θ-lex coefficients
c	Unknown Parameter	Speed of information propagation
p_t	Hyperparameter	Length of the past included in each input X_i
a_t	User Choice Parameter	Translation vector
k	User Choice Parameter	k-th θ-coefficient of the process $\boldsymbol{L}^\epsilon$

In the following, we show in our examples and theorems how to *guide* the design of two types of randomized estimators, see Definition 1.2:

- *a Dirac delta mass concentrated on the empirical risk minimizer*, defined for all $E \in \mathcal{T}$ and $\omega \in \Omega$ as

$$\hat{\rho}(E, \omega) := \delta_{\hat{\beta}(\omega)}, \qquad (3.14)$$

where $\hat{\boldsymbol{\beta}} : \Omega \to \mathcal{H}$ defines the *empirical risk minimizer*

$$\hat{\boldsymbol{\beta}} := arg \inf_{\beta} r^\epsilon(\beta); \qquad (3.15)$$

- and *the randomized Gibbs estimator* $\bar{\rho}$ defined for all $E \in \mathcal{T}$ and $\omega \in \Omega$ as

$$\bar{\rho}(E, \omega) := \frac{\int_E \exp(-\sqrt{m} r^\epsilon(h)) \pi(dh)}{\int_{\mathcal{H}} \exp(-\sqrt{m} r^\epsilon(h)) \pi(dh)}, \qquad (3.16)$$

where $r^\epsilon(h) := r(h) \wedge \epsilon$ and $r(h)$ is the *empirical error*.

Note that the definition of a randomized estimator depends on $\boldsymbol{S}_m$. Therefore, different values of the parameters $\epsilon, a_t, p_t, \lambda$ and c give us different distributions to be used in forecasting tasks.

Remark 3.9 (Possible selection rule for $\sqrt{l}$ and the parameter a_t in relation to the bound (3.13))**.** *Differently from the classical PAC Bayesian for independently and identically distributed data, see reviews [1] and [42], the parameter l is tuned in the bounds (3.13) through the choice of the spatio-temporal embedding. We now give several examples of the latter that can be chosen to tighten the right-hand side of the bound (3.13). In Proposition 3.6, we have determined for a Lipschitz predictor a bound from above of the θ-coefficient $\theta(k)$ related to the process $\boldsymbol{L}^\epsilon$. If we employ this result, we get a PAC Bayesian bound where we are capable of dividing in the last term the contribution of a predictor h from the dependence structure of the underlying MMAF field $\boldsymbol{Z}$ expressed through the θ-lex coefficients. By applying Proposition 3.6, the $Lip(h)$ appears in the bound.*

Therefore for $l \geq 2$, and $r = ka - p$, we obtain the bound

$$\mathbb{P}\Big\{ \forall \hat{\rho} : |\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)]| \leq \Big(KL(\hat{\rho}, \pi) + \log\Big(\frac{1}{\delta}\Big) \Big) \frac{1}{\sqrt{l}}$$

$$+ \frac{1}{\sqrt{l}} \log \Big(\exp\Big(\frac{3\epsilon^2}{2(3 - \epsilon)}\Big)$$

$$+ \pi \Big[3\sqrt{l} \exp(3\sqrt{l}) 2(Lip(h) a(p, c) + 1) \bar{\alpha} \, \tilde{\theta}_{lex}(r) \Big] \Big) \Big\} \geq 1 - 2\delta. \qquad (3.17)$$

We now consider that the parameters $\epsilon, p_t, \lambda, c,$ and $\bar{\alpha}$ (which determines how tight $\tilde{\theta}(r)$ is as a bound of $\theta(k)$) have been selected and we concentrate on the selection of the parameter a_t. We refer the reader to Section 4 for a discussion on the selection of the other parameters.

For $k = 1$, we obtain a convergence rate of $\mathcal{O}(m^{-\frac{1}{2}})$, and we can pre-process the data choosing the smallest value of a_t satisfying the following inequalities

$$\begin{cases} -\lambda h_t(a_t - p_t) + 3\sqrt{\frac{N}{a_t}} < 0 & \text{for exp. decaying } \theta\text{-lex coef., see Def. 2.19,} \\ -\lambda \log(h_t(a_t - p_t)) + 3\sqrt{\frac{N}{a_t}} < 0 & \text{for power decaying } \theta\text{-lex coef., see Def, 2.19.} \end{cases}$$

Under these choices, we obtain that $\exp(3\sqrt{l})\tilde{\theta}_{lex}(r) \leq 1$, and the bound (3.17) tightens. For $k > 1$, we obtain a convergence rate of $\mathcal{O}((m/k)^{-\frac{1}{2}})$, and we can pre-process the data choosing the smallest value of a_t satisfying the following inequalities

$$\begin{cases} -\lambda h_t(ka_t - p_t) + 3\sqrt{\frac{N}{a_t}} < 0 & \text{for exp. decaying } \theta\text{-lex coef.,} \\ -\lambda \log(h_t(ka_t - p_t)) + 3\sqrt{\frac{N}{a_t}} < 0 & \text{for power decaying } \theta\text{-lex coef..} \end{cases}$$

By choosing $k > 1$, we could, however, select a lower value of a_t.

The choice of the parameter a_t discussed in this section is inversely proportional to the parameter k and proportional to the parameter p_t. This means we obtain more examples (i.e., longer training data sets S_m) when k increases and p_t decreases. Hence, a careful choice of the parameter a_t must be done even to obtain training data sets with $m \geq 1$, especially when the data generating process $\boldsymbol{Z}$ admits power decaying θ-lex coefficients.

Example 3.10 (*Guided* randomized estimators of type (3.14))**.** *Let us assume to observe two data sets with $N = 20000$ from an STOU and an MSTOU process as defined in Example 2.20 and 2.21, respectively. The parameters $\lambda = \frac{1}{2}, p_t = 1, c = 1, h_s = 1$ and $h_s = 1$ are associated to the STOU data set and $\lambda = \frac{1}{2}, \alpha = 3, p_t = 1, c = 1, h_t = 1$, and $h_s = 1$ to the MSTOU data set. The former is a serially correlated data set having temporal and spatial short-range dependence and exponentially decaying θ-lex coefficients, whereas the latter possesses temporal and spatial long-range dependence and power decaying θ-lex coefficients. Let us choose a_t following the rule in Remark 3.9 in the case of linear predictors. We call $\boldsymbol{S^1_m}$ the sampling from the STOU, where $a_t = 92$ and $m = 217$, and $\boldsymbol{S^2_m}$ the one from the MSTOU, where $a_t = 8742$ and $m = 2$. We have that the dimension of the input space of our examples is $a(p, c) = 3$ and we consider the parametric space $B \subset \{\beta = (\beta_0, \beta_1) : \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^3 \text{ and } \|\beta\|_1 \leq 1\}$ such that $\text{card}(B) = M < \infty$ for $M \in \mathbb{N}$. We look at the bound (3.17) for $k = 1$. We consider as reference distribution a uniform distribution π on B. For a realization S^1_m or S^2_m, we have that*

$$KL(\delta_{\hat{\beta}} \| \pi) = \sum_{\beta \in B} \log \Big(\frac{\delta_{\hat{\beta}}\{\beta\}}{\pi\{\beta\}} \Big) \delta_{\hat{\beta}}\{\beta\} = \log \frac{1}{\pi\{\hat{\beta}\}} = \log(M). \tag{3.18}$$

It is crucial to notice that the bigger the cardinality of the space B is, the more the term $\log(M)$ and the bound increase.

Finally, for $\delta \in (0, 1)$, we obtain for $\boldsymbol{S^1_m}$ and $\boldsymbol{S^2_m}$ (obviously w.r.t. different distribution $\mathbb{P}$ depending on their data generation process) the bound

$$\mathbb{P}\Big\{ |R^\epsilon(\hat{\boldsymbol{\beta}}) - r^\epsilon(\hat{\boldsymbol{\beta}})| \leq \log \Big(\frac{M}{\delta} \Big) \frac{1}{\sqrt{m}} + \frac{1}{\sqrt{m}} \log \Big(\exp \Big(\frac{3\epsilon^2}{2(3 - \epsilon)} \Big) $$
$$+ 3\sqrt{m}\pi[2(\|\beta_1\|_1 + 1)\bar{\alpha}] \Big) \Big\} \geq 1 - 2\delta. \tag{3.19}$$

Let us assume to work with an accuracy level $\epsilon = 1$, $M = 100$, $\bar{\alpha} = 1$, $\delta = 0.025$. Then, the generalization gap with respect to $\boldsymbol{S}_m^1$ is less than 0.98 with at least 95% probability, whereas for $\boldsymbol{S}_m^2$, we can just prove that the generalization gap is less than 9.40 with a probability of at least 95%. The latter is obviously a vacuous bound.

The randomized Gibbs estimator is the *minimizer* of the right hand-side of the bound (3.13) and gives the best generalization performance in the class $\mathcal{M}_+^1(\mathcal{H})$ and has the best possible rate of convergence for $k = 1$. The result below is also called an *oracle inequality* in the literature.

Theorem 3.11 (PAC Bayesian bound for the randomized Gibbs estimator-Type I). *Let $0 < \epsilon < 3$, $m \geq 2$, and Assumption 3.5 holds. If π is a distribution on $\mathcal{M}_+^1(\mathcal{H})$ such that $\pi[\theta(1)] < \infty$, $\hat{\rho}$ is a randomized estimator in the sense of Definition (1.2), $\hat{\rho}(\cdot, \omega) \ll \pi(\cdot)$ for all $\omega \in \Omega$, $\bar{\rho}$ is the randomized Gibbs estimator defined in (3.16), and $\delta \in (0, 1)$*

$$
\begin{aligned}
\mathbb{P}\Big\{ \bar{\rho}[R^\epsilon(h)] \leq \inf_{\hat{\rho}} &\Big(\hat{\rho}[R^\epsilon(h)] + \Big(KL(\hat{\rho}, \pi) + \log\Big(\frac{1}{\delta}\Big)\frac{2}{\sqrt{m}}\Big) \\
&+ \frac{2}{\sqrt{m}} \log\Big(\exp\Big(\frac{3\epsilon^2}{2(3-\epsilon)}\Big) + \pi\Big[3\sqrt{m}\exp(3\sqrt{m})\theta(1)\Big]\Big)\Big) \Big\} \geq 1 - 2\delta.
\end{aligned}
$$
(3.20)

We now focus on determining the any-time PAC Bayesian bound in our framework. Let us define the filtration $\mathbb{F} = (\mathcal{F}_m)_{m \in \mathbb{N}_0}$ where $\mathcal{F}_m = \sigma(\boldsymbol{S}_m)$ for all $m \geq 1$, i.e., the filtration generated by the cone-shaped sampling process, and $\mathcal{F}_0$ is equal to the trivial sigma-algebra. For $\epsilon > 0$, we then define the process $(f_i(\boldsymbol{S}, h))_{i \in \mathbb{N}_0}$ as

$$
\begin{cases}
f_0(\boldsymbol{S}, h) = & 0 \\
f_m(\boldsymbol{S}, h) = & \eta \sum_{i=1}^m (L^\epsilon(h(X_i), Y_i) - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]) \\
& -\eta \sum_{i=1}^m (\mathbb{E}[L^\epsilon(h(X_i), Y_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]) \\
& -\frac{\eta^2}{2} m\epsilon^2, \text{ for } m \in \mathbb{N}.
\end{cases}
$$

We can then prove an any-time PAC-Bayesian bound, which allows a choice of the accuracy level $\epsilon > 3$ and, therefore, it complements the results given in Theorem 3.8 for a bounded loss.

Theorem 3.12. *Let $\epsilon > 0$ and Assumption 3.5 holds. If $\pi \in \mathcal{M}_+^1(\mathcal{H})$ such that $\pi[\theta(1)] < \infty$, $\hat{\rho}$ is a randomized estimator in the sense of Definition (1.2), $\hat{\rho}(\cdot, \omega) \ll \pi(\cdot)$ for all $\omega \in \Omega$, and $\delta \in (0, 1)$, then*

$$
\begin{aligned}
\mathbb{P}\Big\{ \forall \hat{\rho}, \forall m \geq 1 : \hat{\rho}[r^\epsilon(h)] - \hat{\rho}[R^\epsilon(h)] &\leq \frac{KL(\hat{\rho}, \pi) + \log\Big(\frac{1}{\delta}\Big)}{m\eta} + \frac{\eta}{2}\epsilon^2 \\
+ \hat{\rho}\Big[\frac{1}{m} \sum_{i=1}^m \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)|\mathcal{F}_{i-1}] &- \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]\Big]\Big\} \geq 1 - \delta.
\end{aligned}
$$
(3.21)

We call the process $\frac{1}{m}\sum_{i=1}^{m}(\mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)])_{i\in\mathbb{N}}$ the *residual process*.

Following the same line of proof, it can also be proven that an any-time PAC Bayesian bound holds for the average generalization gap $\hat{\rho}[R^{\epsilon}(h)] - \hat{\rho}[r^{\epsilon}(h)]$, by using the following definition of $f_m(\boldsymbol{S},h)$, namely,

$$
\begin{cases}
f_0(\boldsymbol{S},h) = & 0 \\
f_m(\boldsymbol{S},h) = & \eta\sum_{i=1}^{m}(\mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)] - L^{\epsilon}(h(X_i),Y_i)) \\
& -\eta\sum_{i=1}^{m}(\mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)] - \mathbb{E}[L^{\epsilon}(h(X_i),Y_i)|\mathcal{F}_{i-1}]) \\
& -\frac{\eta^2}{2}m\epsilon^2, \text{ for } m\in\mathbb{N}.
\end{cases}
$$

The following result follows straightforwardly by applying a union bound.

Corollary 3.13. *Let $\epsilon > 0$ and Assumption 3.5 holds. If $\pi\in\mathcal{M}_+^1(\mathcal{H})$ such that $\pi[\theta(1)] < \infty$, $\hat{\rho}$ is a randomized estimator in the sense of Definition (1.2), $\hat{\rho}(\cdot,\omega)\ll\pi(\cdot)$ for all $\omega\in\Omega$, and $\delta\in(0,1)$, then*

$$
\mathbb{P}\Big\{\forall\hat{\rho},\forall m\geq 1: |\hat{\rho}[R^{\epsilon}(h)] - \hat{\rho}[r^{\epsilon}(h)]| \leq \frac{KL(\hat{\rho},\pi) + \log\left(\frac{1}{\delta}\right)}{\eta m} + \frac{\eta\epsilon^2}{2}
$$
$$
+ \hat{\rho}\Big[\Big|\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^{\epsilon}(h(\boldsymbol{X}_i),\boldsymbol{Y}_i)]\Big|\Big]\Big\} \geq 1-2\delta.
$$
(3.22)

We can now give the proof of a fixed-time PAC Bayesian bound obtained by using the bound (3.22), a Markov's inequality, and the projective-type property (2.12). In this new scheme of proof, it is determined a fixed-time PAC Bayesian bound of the *residual process* following [2, Theorem 1].

Theorem 3.14 (Fixed-time PAC Bayesian bound- *Type II*)**.** *Let $\epsilon > 0$ and Assumption 3.5 holds. If $\pi\in\mathcal{M}_+^1(\mathcal{H})$ such that $\pi[\theta(1)] < \infty$, $\hat{\rho}$ is a randomized estimator in the sense of Definition (1.2), $\hat{\rho}(\cdot,\omega)\pi(\cdot)$ for all $\omega\in\Omega$, and $\delta\in(0,1)$, then*

$$
\mathbb{P}\Big\{\forall\hat{\rho}: |\hat{\rho}[r^{\epsilon}(h)] - \hat{\rho}[R^{\epsilon}(h)]| \leq \frac{KL(\hat{\rho},\pi) + \log\left(\frac{1}{\delta}\right)}{\eta m}
$$
$$
+ \frac{\eta\epsilon^2}{2} + \Big(\epsilon\pi\Big[\frac{\theta(1)}{\delta}\Big](D_{\phi_2-1}(\hat{\rho},\pi)+1)\Big)^{\frac{1}{2}}\Big\} \geq 1-3\delta,
$$
(3.23)

where $\theta(1)$ is a θ-coefficient of the process $\boldsymbol{L}^{\epsilon}$.

Remark 3.15 (Selection rule for the parameter a_t in relation to the bound (3.23))**.** *We now discuss an exemplary spatio-temporal embedding which can be chosen such to tighten the right-hand side of the bound (3.23). We consider that the parameters ϵ, p_t, λ and c have been selected (the parameter k does not enter in the determination of the spatio-temporal embedding in this case), and we*

concentrate on the selection of the parameter a_t. We refer the reader to Section 4 for a discussion on the selection of the other parameters. Let us choose $\eta = \frac{1}{\sqrt{m}}$, and select a_t as the smallest constant such that $\tilde{\theta}_{lex}(a-p) \leq \frac{1}{2m}$, which implies

$$\begin{cases} -\lambda h_t(a_t - p_t) - \log\left(\frac{a_t}{2N}\right) \leq 0 & \textit{for exp. decaying } \theta\textit{-lex coeff., see Def. 2.19,} \\ -\lambda \log(h_t(a_t - p_t)) - \log\left(\frac{a_t}{2N}\right) \leq 0 & \textit{for power decaying } \theta\textit{-lex coeff., see Def. 2.19} \end{cases}$$

So doing we have that the third addend in the right-hand side of the bound has the same order of magnitude as the other terms, which gives us the convergence rate of $\mathcal{O}(m^{-\frac{1}{2}})$. We can notice that with respect to the selection rule in Remark 3.9, we can guide the design of a randomized estimator using a lower value of the parameter a_t. This ultimately means that we are working with a training data set that admits a stronger serial correlation along the temporal and spatial dimension than the ones obtained using the selection rule in Remark 3.9.

The bound (3.23) can also be used in different ways to guide the selection of a randomized estimator. For example, when working with the estimator (3.14), we can choose a value of η that minimizes the right-hand side of the bound and then select a spatio-temporal embedding that makes the bound not vacuous. Examples of such a spatio-temporal embedding are used in Example 3.17.

Remark 3.16 (Hostile Framework)**.** *Theorem 3.14 employs a well-known result in the PAC Bayesian literature for dependent and stationary data, analyzed in [2]. The authors determine a general fixed-time bound holding for unbounded losses; see Theorem 1. In particular, let $\delta \in (0,1)$, $p > 1$, and $q = \frac{p}{p-1}$, with probability at least $1 - \delta$, they prove that for any $\hat{\rho}$*

$$\hat{\rho}[R(h)] \leq \hat{\rho}[r(h)] + \left(\frac{\mathcal{M}_{\phi_q,n}}{\delta}\right)^{\frac{1}{q}} (D_{\phi_p-1}(\hat{\rho},\pi) + 1)^{\frac{1}{p}}.$$

This bound obviously applies to our set-up, but it cannot be used in its general shape to guide *the selection of a randomized estimator, which is the primal target of our paper. The reason for this is that we cannot write the term $\mathcal{M}_{\phi_q,n}$ as a function of the θ-coefficients of the process $\boldsymbol{L}^{\boldsymbol{\epsilon}}$, similarly, as done in [2] in the case of α-mixing processes. When working with such a dependence notion, the term $\mathcal{M}_{\phi_q,n}$ can be controlled using Lemma 3 (for bounded losses) and Theorem 3 (for unbounded losses) in [35]. To the best of our knowledge, no proofs in the literature extend such results for θ-weakly dependent processes. Moreover, even if we could determine such a proof, following the methodology described in [2], we will end up estimating the term $\mathcal{M}_{\phi_q,n}$ from above using $\sum_{j \in \mathcal{Z}} \theta(j)$. This estimate diverges for specific power decaying θ-coefficient sequences, such as those discussed in Example 2.21.*

Therefore, we developed novel results where the PAC Bayesian bounds depend just on one single *θ-coefficient and are also not equal to infinity for a given data set generated by a temporal and spatial long-range dependent MMAF.*

Example 3.17 (*Guided* randomized estimators of type (3.14) and comparisons between the bounds (3.17) and (3.23))**.** *We use the data described in Example 3.10. We also maintain the same assumptions on the set B and the reference*

distribution π. Moreover, we select parameters $\epsilon = 1, M = 100, \delta = \frac{0.05}{3}$, and $c = p_t = h_t = 1$. From [2, Proposition 1], we have that $(D_{\phi_2-1}(\hat{\rho}, \pi) + 1) = M$. We compute first the bound (3.23) for the estimator (3.14) and the data set observed from the STOU process. We use the bound (3.10) for the θ-coefficients of the process $\boldsymbol{L}^\epsilon$, and assume that $\bar{\alpha} = 1$. Moreover, we select the parameter as $\eta = \sqrt{\frac{2\log(M/\delta)}{\epsilon^2 m}}$. Such choice minimizes the right-hand side of the bound (3.23) in this framework, we then obtain for $\delta \in (0,1)$ that

$$\mathbb{P}\Big\{|R^\epsilon(\hat{\boldsymbol{\beta}}) - r^\epsilon(\hat{\boldsymbol{\beta}})| \leq \sqrt{\frac{2\log(M/\delta)}{m}} + \Big(\epsilon\frac{M2(\pi[\|\beta_1\|_1]+1)\bar{\alpha}\tilde{\theta}_{lex}(a-p)}{\delta}\Big)^{\frac{1}{2}}\Big\} \geq 1 - 3\delta. \tag{3.24}$$

We then select the smallest parameter a_t such that $\tilde{\theta}_{lex}(a-p) \leq \frac{\delta}{4Mm}$. We obtain an $\boldsymbol{S_m^1}$, where $a_t = 34$ and $m = 588$, and a generalization gap less than or equal to 0.21 with at least 95% probability. With the same choice of spatio-temporal embedding and computing directly the bound (3.19) for $\delta = 0.025$, we obtain that the generalization gap is less than or equal to 2.99 with at least 95% probability. Using instead the spatio-temporal embedding defined in Example 3.10 for the STOU data set, we obtain that the generalization gap computed with (3.24) is less than 0.28 versus a 0.98 obtained from (3.19).

For the data set observed from the MSTOU process, we now compute the bound (3.24) by employing the selection rule in Remark 3.15. We then obtain a spatio-temporal embedding $\boldsymbol{S_m^2}$ for $a_t = 1170$ and $m = 17$ and a generalization error less or equal to 27.51 with probability at least 95%. With the same choice of spatio-temporal embedding and computing directly the bound (3.19) for $\delta = 0.025$, we obtain a generalization gap less than or equal to 5.66 with at least 95% probability instead. Using the spatio-temporal embedding defined in Example 3.10 for the MSTOU data set, we obtain that the generalization gap computed with (3.23) is equal to 18.97 versus a 9.40 obtained from (3.19).

From this simple example, we can notice how, for the temporal and spatial short-range data set, the bound obtained in (3.24) is tighter than the one presented in (3.19). In the temporal and spatial long-range case, it seems, however, that with all spatio-temporal embeddings so far considered, we can just obtain a vacuous bound. However, there is an aspect of the bound (3.24) that we have not truly used so far in our evaluations. All the results above hold under the choice of the accuracy level $\epsilon = 1$. If we choose $\epsilon = 1000$ (which goes outside the range of validity of the bound (3.19)) and compute again the bound in (3.24) for the choice of $a_t = 1170$, we obtain a not-vacuous bound. In fact, under this choice of the accuracy level, we obtain a generalization error less or equal to 838.84 with a probability at least of 95%. Therefore, for different reasons that in the case of temporal and spatial short-range data, the bound obtained in 3.24 is the one to employ when analyzing long-range data.

From the description of possible selection rules for the parameter a_t and the Examples 3.10 and 3.17, different questions may arise. The first regards the vacuousness of the bounds (3.13) and (3.23) observed in some of our examples when employing temporal and spatial long-range data. Secondly, it is important

to question the dependence of the bound from the constant $Lip(h)$ and the accuracy level parameter ϵ. The three remarks below tackle these issues and present interesting future research directions for our work.

Remark 3.18 (Spatio-temporal embeddings for temporal and spatial long-range dependent data sets)**.** *In general, the bounds (3.13) and (3.23) depend non-linearly on the set of parameters listed in Table 2, and assessing their magnitude in the temporal and spatial long-range dependence framework for MMAF is still an open problem.*

It is important to highlight that the selection of the spatio-temporal embedding depends on the discretization step in time h_t. Such constant is related, in practical applications, to the frequency of the observed data and has a significant impact on the tightness of the bound. From the selection rules in Remarks 3.9 and 3.15, we can see how the parameter value a_t decreases when $h_t > 1$.

Moreover, the bound (3.23) holds for $\epsilon > 0$ and this gives us better generalization performance as long as we consider an $\epsilon \geq 3$ (which goes outside the range of validity of the bound (3.13)), see Example 3.17.

Remark 3.19 (Are the fixed-time bounds (3.13) and (3.23) depending on the value of the $Lip(h)$?)**.** *There is no explicit dependence on the $Lip(h)$ in the bounds (3.13) and (3.23). Such a coefficient appears if we employ Proposition 3.6 to estimate the θ-coefficients in the bound. However, we do not have a formal proof of the tightness of this estimation.*

When working with deep neural network predictors *computing their Lipschitz constant is a complex numerical task; see [38] and [74]. In this framework, we could use the bound (3.9) for the θ-coefficients of the process $\boldsymbol{L}^\epsilon$ and then work with a numerical approximation for the $Lip(h)$. However, future research should focus on determining tight estimates for the θ-coefficients and assessing their dependence on the $Lip(h)$ in detail. Such analysis could greatly help the empirical computation of PAC Bayesian bounds for (random) deep learning architectures and extend the range of applicability of MMAF-guided learning in practical applications.*

Remark 3.20 (The importance of the accuracy level ϵ)**.** *For the time being, the proofs of Theorems 3.8 and 3.14 work just if $\boldsymbol{L}^\epsilon$ is bounded. Such an assumption allows us to apply the projective-type representation of the θ-coefficients discussed in Remark 2.15. This, in turn, allows us to be capable of* guiding *a randomized estimator, i.e., choosing the parameters defining the spatio-temporal embedding by controlling the magnitude of the $\theta(1)$-coefficient of the process $\boldsymbol{L}^\epsilon$. It is important to highlight that, as observed at the end of Example 3.17 for a temporal and spatial long-range data set, or in the sensitivity analysis conducted for a temporal and spatial short-range data set in Section 4.2, the performances of the employed randomized estimators improve when $\epsilon \geq 3$.*

In the following theorem, we give an alternative fixed-time bound for the randomized Gibbs estimator, obtained using Theorem (3.14). In this case, the estimator no longer minimizes the right-hand side of the inequality.

Theorem 3.21 (PAC Bayesian bound for the randomized Gibbs estimator-*Type II*)**.** *Let $\epsilon > 0$ and Assumption 3.5 hold. If $m \geq 1$, $\eta = \frac{1}{\sqrt{m}}$, π is a distribution on $\mathcal{H}$ such that $\pi[\theta(1)] < \infty$, $\bar{\rho}$ is the randomized Gibbs estimator defined in (3.16), and $\delta \in (0, 1)$*

$$
\mathbb{P}\Big\{\bar{\rho}[R^\epsilon(h)] \leq \inf_{\hat{\rho}} \Big(\hat{\rho}[R^\epsilon(h)] + \Big(KL(\hat{\rho}, \pi) + \log\Big(\frac{1}{\delta}\Big)\Big)\frac{2}{\sqrt{m}}\Big) + \frac{\epsilon^2}{\sqrt{m}}
$$
$$
+ 2\Big(\epsilon\pi\Big[\frac{\theta(1)}{\delta}\Big](D_{\phi_2-1}(\bar{\rho}, \pi) + 1)\Big)^{\frac{1}{2}}\Big\} \geq 1 - 4\delta,
$$
$$(3.25)$$

where $\theta(1)$ is a θ-coefficient of the process $\boldsymbol{L}^\epsilon$.

Several results on fixed-time PAC Bayesian bounds in a dependent framework can be found in the batch setting, but only for time series models. We review them in the remark below.

Remark 3.22 (Are there PAC Bayesian bounds for dependent data with faster convergence rates?)**.** *In [2], the authors determine an oracle inequality with a rate of $\mathcal{O}(m^{-\frac{1}{2}})$ under the assumption that $((\boldsymbol{X}_i, \boldsymbol{Y}_i)^\top)_{i\in\mathbb{Z}}$ is generated by a stationary and α-mixing process, see [19] for a detailed explanation of the properties of this dependence notion, with coefficients $(\alpha_j)_{j\in\mathbb{Z}}$ such that $\sum_{j\in\mathbb{Z}} \alpha_j < \infty$. Such a bound employs the chi-squared divergence and holds for unbounded losses. It is important to highlight that the randomized estimator obtained by minimization of the PAC Bayesian bound is not a Gibbs estimator in this framework. An explicit bound for linear predictors can be found in their [2, Corollary 2]. This result holds under the assumption that $\pi[\|\beta\|^6] < \infty$.*

In [4], the authors prove oracle inequalities for a Gibbs estimator and data generated by a stationary and bounded $\theta_{1,\infty}$-weakly dependent process– such dependence notion extends the concept of ϕ-mixing discussed in [69]– or a causal Bernoulli shift process. Models with bounded $\theta_{1,\infty}$-weak coefficients are causal Bernoulli shifts with bounded innovations, uniform ϕ-mixing sequences, and dynamical systems; see [4] for more details. The oracle inequality is here obtained for an absolute loss function and has a rate of $\mathcal{O}(m^{-\frac{1}{2}})$. An extension of this work for Lipschitz loss functions under ϕ-mixing [49] can be found in [3]. Here, the authors show an oracle inequality for a Gibbs estimator with the optimal rate $\mathcal{O}(m^{-1})$. This rate is considered optimal in the i.i.d literature, and for a squared loss function, [20].

Interesting results in the literature of PAC Bayesian bounds for heavy-tailed data (albeit identically distributed) can also be found in [41] and [47].

MMAF-guided learning has the potential to be extended to general θ-lex weakly dependent models because of the results in Proposition 3.2. Another possible extension of the methodology is related to using bounded and local Lipschitz losses; see Remark 3.3. Considering unbounded losses is beyond the scope of the present paper and the content of future research treated in [29].
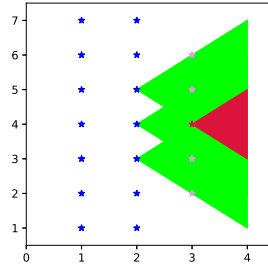
FIG 3. *The x- and y-axes represent the time and spatial dimension, respectively. We picture the last* 3 *frames of a data set with spatial dimension* $d = 1$ *where the blue stars represent the pixels used in the definition of the training data set, and the violet stars represent the space-time points where it is possible to provide forecasts with MMAF-guided learning for* $p_t = c = h_t = 1$. *Note that the forecast in the time-spatial position* $(4, 3)$ *lies in the intersection (red area) of the future lightcones* $A_2(5)^+$, $A_2(4)^+$ *and* $A_2(3)^+$ *as defined in (2.10) and represented with green cones.*

4. Ensemble forecasts using MMAF-guided learning

4.1. Practical implementation and casual forecast

The knowledge of all the parameters in Table 2 allows us to have a precise definition of the spatio-temporal embedding defined in Section 2.1. So far, we have only discussed the selection of the parameter a_t. The parameter p_t is an *hyperparameter* in our learning methodology. By establishing a finite grid of values $I \subset \mathbb{N}$ for the parameter p_t, we could introduce in (3.13) and (3.23) an explicit dependence on its possible values. So doing, we could approach the general question of minimizing the right-hand-side of the bound (also known as, *exact minimization*) in the function of every possible randomized estimator $\hat{\rho}$ and the grid of values I, see Section 2.1 in [1] for exemplary calculations. Similarly, we could approach the selection of the parameter ϵ and obtain a methodology for the *selection of all the hyperparameters* involved in MMAF-guided learning. Such an issue is outside the scope of the present paper. However, we provide in our numerical experiments in Section 4.2 a sensitivity analysis about the hyperparameters ϵ and p_t.

If we assume that our data are generated by an STOU or an MSTOU process, i.e., we are assuming that the data admits exponential or power-decaying θ-lex coefficients, as discussed in Section A.3 and A.4, there are several methodologies available for the estimation of the parameters c and λ. There are also other feasible model set-ups for time series models, such as when $c = 0$. In this framework, several estimation methodologies can be employed to estimate the model's parameters of an MMAF, see A.5.

Remark 4.1. *When using the spatio-temporal embeddings described in Remark 3.9 and 3.15, similarly to the kriging literature, we need an inference step before being capable of delivering one-time ahead ensemble forecasts. In this literature, it is often assumed that the estimated parameters used in the*

calculation of kriging weights and kriging variances are the true one*, see [25, Chapter 3] and [26, Chapter 6] for a discussion on the range of applicability of such estimates. We implicitly make the same assumptions if, for example, we use estimated values for the parameters λ and c.*

It remains an interesting open problem to understand the interplay of the estimates' bias of the parameters involved in the computation of the PAC Bayesian bounds (3.13) and (3.23). One of the biggest problems of this analysis relies upon disentangling the effect of the bias of the constant c introduced in the preprocessing step, which changes the length of the input-features vector X_i.

We detail now how to *guide* the design of a randomized Gibbs estimation $\bar{\rho}$ and make one-time ahead ensemble forecasts. Differently from the notations so far employed, we indicate the training data set by $S_m^{x^*}$ to remark the dependence of the training data set on the pixel position x^* where we perform our forecasts. In general, the learning methodology applies to any pixel x^* for which $\mathcal{I}(t_0 + ia, x^*) \subseteq \mathbb{T} \times \mathbb{L}$ for all $i = 1, \ldots, N$.

Application Steps:

(i) We observe a raster data cube whose spatio-temporal index set is described by N frames and we estimate the parameters λ and c (and $\bar{\alpha}$ if necessary).

(ii) We fix a pixel position x^* and choose a value for the accuracy level ϵ and the hyperparameter p_t. We then select the parameter a_t as suggested in Remark 3.9 or 3.15 and determine the spatio-temporal embedding.

(iii) We determine the training data set $S_m^{x^*}$ using all available N frames, which correspond to a specific realization $\omega \in \Omega$, i.e., $\boldsymbol{S}_m^{x^*}(\omega) = S_m^{x^*}$.

(iv) We then draw β from the distribution $\bar{\rho}(\cdot, \omega)$, following the definition of the randomized Gibbs estimator in (3.16).

(v) We perform a so-called *ensemble forecast* by repeating point (iv) several times.

A one-time ahead forecast corresponds to the space-time point $((t_0 + Na) + h_t, x^*)$ and it is given by $\langle L_p^-((t_0 + Na) + h_t), \beta \rangle$, where β is a draw from the randomized Gibbs distribution. Therefore, we can make a forecast in a future time point $t = (t_0 + Na) + h_t$ as long as the set $\mathcal{I}((t_0 + Na) + h_t, x^*)$, as defined in (3.4), has cardinality $a(p, c)$.

For each β, the forecast of the field we obtain in the space-time point $((t_0 + Na) + h_t, x^*)$ lies in the intersections of the future light cones of the space-time points belonging to $L_p^-((t_0 + Na) + h_t, x^*)$. In Figure 3, we give an example of a one-time ahead forecast performed for an MMAF for $d = 1$. As we can see, MMAF-guided learning enables us to make plausible forecasts in space-time points (under the causality concept induced by the ambit sets, see Section 2.2) starting from the set of inputs we observe.

4.2. Linear predictors: an example with simulated data

We work in the hypothesis space $\mathcal{H}' = \{h_\beta(X) = \beta_0 + \beta_1^T X, \text{ for } \beta := (\beta_0, \beta_1)^\top \in B\}$, where $B \subset \mathbb{R}^{(a(p,c)+1)}$. We use simulated observations from an STOU, i.e.,

a temporal and spatial short-range dependent field with exponentially decaying θ-lex weakly dependent coefficients. We simulate four data sets $(Z_t(x))_{(t,x)\in\mathbb{T}\times\mathbb{L}}$ from a zero mean STOU process by employing the diamond grid algorithm introduced in [59] for $d = 1$. The time and spatial discretization steps are chosen as $h_t = h_s = 0.05$ on the spatio-temporal interval $[0, 100] \times [0, 10]$. Therefore $\mathbb{T} = \{h_t, \ldots, 2,000 h_t\}$ and $\mathbb{L} = \{0, \ldots, 200\}$. We use the unidimensional frames related to the time indices $\mathbb{T}^{train} = \{0, h_t, \ldots, 1,999 h_t\}$ to determine the training data sets and the one corresponding to $\mathbb{T}^{test} = \{2000 h_t\}$ as a test set. We choose as distribution for the Lévy seed Λ' a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.5$, and an $NIG(\alpha, \beta, \mu, \delta)$ distribution with $\alpha = 5, \beta = 0, \delta = 0.2$ and $\mu = 0$; see Exercises 2.2 and 2.3. We use the latter distribution to test the behavior of MMAF-guided learning for different sets of heavy-tailed data. We generate data with different seeds for the Lévy basis realizations. Moreover, the speed of information propagation c is equal to one for all generated data sets. Finally, we choose different mean reverting parameters, namely $A = 1$ or 4. Therefore, λ is equal to $\frac{1}{2}$ if $A = 1$ or 2 if $A = 4$. We call these data sets GAU10, GAU1A4, NIG1A4, and NIG10. For the NIG Lévy seed described above, we generate two further spatio-temporal data sets called $(Z_t^i(x))_{(t,x)\in\mathbb{T}_i\times\mathbb{L}}$ for $i = 1, 2$ on the spatio-temporal interval $[0, 1000] \times [0, 10]$. We choose $h_t = h_s = 0.05$, $\mathbb{L} = \{0, \ldots, 200\}$ and time indices $\mathbb{T}_1 = \{18,000 h_t, \ldots, 20,000 h_t\}$ and $\mathbb{T}_2 = \{0, h_t, \ldots, 20,000 h_t\}$. Such data sets are called NIG1 in the following. A summary scheme of the data's characteristics is given in Table 3. For the NIG1 data sets, we use the unidimensional frames corresponding to $\mathbb{T}_1^{train} = \{18,000 h_t, \ldots, 19,999 h_t\}$ and $\mathbb{T}_2^{train} = \{0, h_t, \ldots, 19,999 h_t\}$ to determine the training data sets, respectively, and $\mathbb{T}_1^{test} = \{20,000 h_t\}$ and $\mathbb{T}_2^{test} = \{20,000 h_t\}$ as test sets. For all the generated data sets, we perform a one-time ahead ensemble forecast for the pixels corresponding to $\mathbb{L}' = \{1, \ldots, 199\}$.

TABLE 3
Overview on simulated data sets with $c = 1$ and spatial dimension $d = 1$.

Data Set	Mean Reverting Parameter	Lévy seed	Random generator seed
GAU1A4	$A = 4$	Gaussian	1
GAU10	$A = 1$	Gaussian	10
NIG1	$A = 1$	NIG	1
NIG1A4	$A = 4$	NIG	1
NIG10	$A = 1$	NIG	10

We follow the steps detailed in Section 4.1 to use MMAF-guided learning in practice. We start by estimating the parameters c and λ. We use the estimators ($A.19$) presented in Section A.3 and the plug-in estimator (2.14). Table 4 gives the results for each data set. We then use such estimates and the frames corresponding to $\mathbb{L}' \times \mathbb{T}^{train}$, $\mathbb{L}' \times \mathbb{T}_1^{train}$, or $\mathbb{L}' \times \mathbb{T}_2^{train}$ in the selection of the parameter a_t by following the rules in Remark 3.9 for $k = 1$, and Remark 3.15. We then analyze the performance of the different randomized Gibbs estimators obtained from a particular choice of the parameter a_t when the reference distribution is assumed to be multivariate standard Gaussian. We do not give an

empirical evaluation of the right-hand side of the bounds (3.20) and (3.25) and base our assessment of the performance of the different estimators on how narrow their inter-quartile range on a 50 member ensemble forecast is. Note that each forecast we make has a casual interpretation as described in Figure 3.

We start by conducting two different experiments to showcase the performance of our methodology for $\epsilon = 2.99$ and $p_t = 1$. We use as a baseline model a linear model where the estimation of the parameter vector β is performed using the empirical risk minimizer defined in (3.15).

In the first experiment, we use the data sets GAU1A4, GAU10, NIG1A4, NIG10 and work with the training data sets $S_m^{x^*}$ described by the Tables 5 and 6.

TABLE 4
Estimations of parameters A, c and λ.

Data Set	A^*	c^*	λ^*	Frames Used
GAU1A4	3.9684	0.9958	1.9715	$\mathbb{T} \times \mathbb{L}$
GAU10	0.8429	0.9978	0.4196	$\mathbb{T} \times \mathbb{L}$
NIG1	1.0186	1.0018	0.5111	$\mathbb{T}_1 \times \mathbb{L}$
NIG1	1.0186	1.0019	0.5112	$\mathbb{T}_2 \times \mathbb{L}$
NIG1A4	4.0308	1.0076	2.0461	$\mathbb{T} \times \mathbb{L}$
NIG10	0.9728	1.0021	0.4884	$\mathbb{T} \times \mathbb{L}$

TABLE 5
Parameters defining the training data sets $S_m^{x^*}$ used in the first experiment for each pixel.

Selection of parameters as in Remark 3.9							
GAU1A4		GAU10		NIG1A4		NIG10	
a_t	124	a_t	346	a_t	121	a_t	313
m	16	m	6	m	17	m	7
k	1	k	1	k	1	k	1

TABLE 6
Parameters defining the training data sets $S_m^{x^*}$ used in the first experiment for each pixel.

Selection of parameters as in Remark 3.15							
GAU1A4		GAU10		NIG1A4		NIG10	
a_t	47	a_t	156	a_t	45	a_t	139
m	41	m	13	m	43	m	15
k	1	k	1	k	1	k	1

An acceptance-rejection algorithm with a Gaussian proposal determines a draw β from the randomized Gibbs estimator. We show in Figure 4 the min-max range of the two ensemble forecasts (dark grey for the parameters in Table 5, and light grey for the one in Table 6) as well as the inter-quartile ranges of a 50-member ensemble forecast for each pixel in $\mathbb{L}'$ compared with the test set. As the plots clearly show, using a randomized Gibbs estimator, we obtain an inter-quartile range that contains the test set for each spatial position $x^* \in \mathbb{L}'$. Moreover, the ranges seem to have a similar behavior independently of the test set and the Lévy seeds. The randomized Gibbs estimators guided by the choices
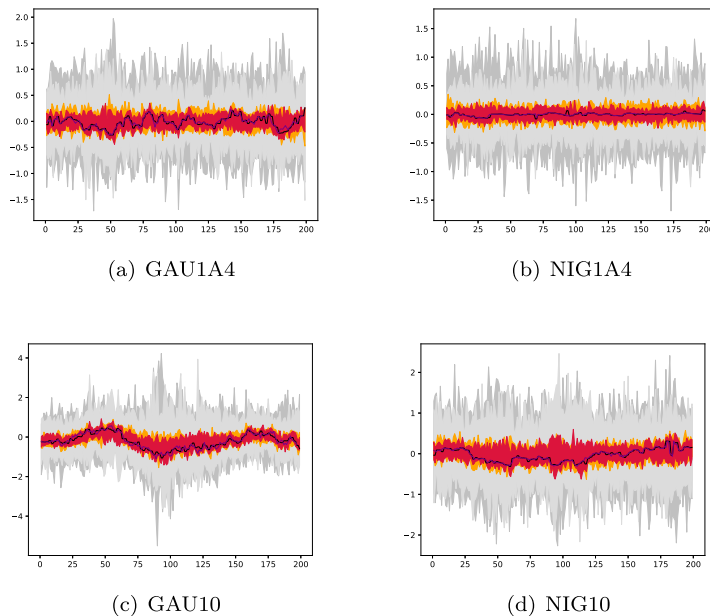
(a) GAU1A4   (b) NIG1A4

(c) GAU10   (d) NIG10

FIG 4. *Min-max and inter-quartile range of a* 50-*member ensemble forecast for the training data sets* $S_m^{x^*}$ *described in Tables 5 and 6. Dark grey and orange colors represent the ranges related to the use of Table 5, whereas the light grey and red colors represent the ranges related to the use of Table 6. The test set is depicted with a thick black line, while the forecasts obtained using the baseline estimator (3.15) are represented with a violet thick line. The training data set* $S_m^{x^*}$ *described in Table 6 are used for computing the baseline estimates with respect to the GAU1A4, NIG1A4, GAU10, and NIG10 data sets in (a), (b), (c) and (d), respectively. The x-axis represents the pixels in* $\mathbb{L}'$, *whereas the forecast values are along the y-axis.*

of parameters in Table 6 have narrower inter-quartile ranges (in red) with respect to the ones guided by the choices of parameters in Table 5 (in orange).

Let us define the average Relative Mean Absolute Error (averRMAE) to compare our forecasts with the baseline model. Let $P = |\mathbb{L}'|$, we define the average Relative Mean Absolute Error as

$$averRMAE = \frac{1}{P} \sum_{i=1}^{P} \frac{|Z_t(x_i) - \hat{Z}_t(x_i)|}{|Z_t(x_i)|},$$

where $\hat{Z}_t(x_i)$ is the one-time ahead forecast obtained with the linear model for all i. The observations in our simulated data sets have an order of magnitude (on average) of 10^{-3}. Tables 7 and 8 show that the empirical risk minimizer cannot capture any significant digit, as also seen in Figure 4. Our ensemble forecasts give, at least, an interval where the one-time ahead forecasts can lie.

In the second experiment, we analyze the performance of randomized Gibbs estimators for the data sets NIG1. For each pixel, we work with the data sets

TABLE 7

averRMAE for the baseline estimator following the selection rule in Remark 3.9. The NIG1 data set's averRMAE has been computed with respect to S_m^{2,x^*} described in Table 9.

	GAU1A4	GAU10	NIG1A4	NIG10	NIG1
linear	0.0304	0.0176	0.0168	0.0127	0.0142

TABLE 8

averRMAE for the baseline estimator following the selection rule in Remark 3.15. The NIG1 data set's averRMAE has been computed with respect to S_m^{2,x^*} described in Table 10.

	GAU1A4	GAU10	NIG1A4	NIG10	NIG1
linear	0.0303	0.0176	0.0156	0.0128	0.0138

S_m^{1,x^*} and S_m^{2,x^*} described in Tables 9 and 10, obtaining the ensemble forecasts in Figure 5.

(a) NIG1's ensemble forecast based on Table 9

(b) NIG1's ensembe forecast based on Table 10
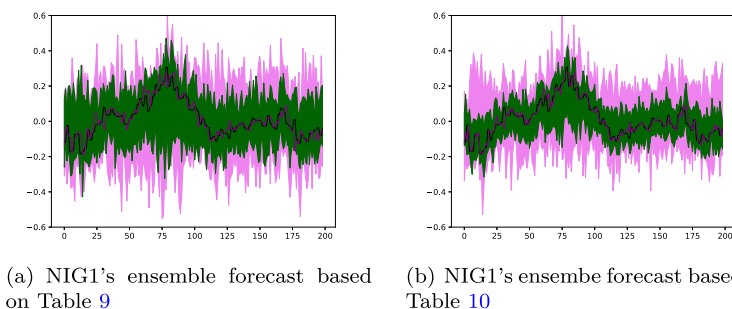
FIG 5. *Inter-quartile range of a 50-member ensemble forecast using S_m^{1,x^*} (in violet) and S_m^{2,x^*} (in green) as defined in Table 9 (a) and Table 10 (b). The test set and the baseline forecasts are depicted with thick black and violet lines, respectively. The training data sets S_m^{2,x^*} in Table 9 and 10 are respectively used as baseline estimates in (a) and (b). The x-axis represents the pixels in $\mathbb{L}'$, whereas the forecast values are along the y-axis.*

TABLE 9

Parameters describing the training data sets used in the second experiment at each pixel x^*.

Selection of parameters as in Remark 3.9			
S_m^{1,x^*}		S_m^{2,x^*}	
a_t	303	a_t	652
m	7	m	32
k	1	k	1

By comparing the inter-quartile range of the ensemble forecasts, we see that in both cases, the green range representing the forecasts related to S_m^{2,x^*} is contained in the purple range, which represents the inter-quartile range of the forecasts made using the data set S_m^{1,x^*}. Both of them include the test set. The amplitude of the inter-quartile range reduces when the number of observations in the data set increases. The forecasts of the baseline model are performed

TABLE 10

Parameters describing the training data sets used in the second experiment at each pixel x^*.

Selection of parameters as in Remark 3.15			
S_m^{1,x^*}		S_m^{2,x^*}	
a_t	134	a_t	207
m	15	m	96
k	1	k	1

using the data set S_m^{2,x^*} and have a high averRMAE as reported in Tables 7 and 8.

We want now to analyze our methodology's sensitivity to the hyperparameter p_t when a_t is selected following the rules in Remarks 3.9 and 3.15, respectively. The selection of the parameter a_t is proportional to the values of p_t. Therefore, the smaller we choose this parameter, the more examples we obtain in $S_m^{x^*}$. We work in this experiment with the data sets GAU1A4 and NIG1A4. However, we obtained the same conclusions for all the other data sets employed in our study. We generate ensemble forecasts (and their respective inter-quartile ranges) for $p_t = 1, 8, 15$. All inter-quartile ranges in Figure 6 contain the test set, and a significant increase of p_t has a negative impact on the inter-quartile range amplitude. For this reason, we have chosen the parameter $p_t = 1$ in our previous experiments.

TABLE 11

Parameters describing the training data sets $S_m^{x^*}$ used in the third experiment for each pixel.

Selection of parameters as in Remark 3.9											
GAU1A4						NIG1A4					
p_t	1	p_t	8	p_t	15	p_t	1	p_t	8	p_t	15
a_t	124	a_t	129	a_t	134	a_t	121	a_t	127	a_t	132
m	16	m	15	m	14	m	17	m	15	m	14
k	1	k	1	k	1	k	1	k	1	k	1

TABLE 12

Parameters describing the training data sets $S_m^{x^*}$ used in the third experiment for each pixel.

Selection of parameters as in Remark 3.15											
GAU1A4						NIG1A4					
p_t	1	p_t	8	p_t	15	p_t	1	p_t	8	p_t	15
a_t	47	a_t	53	a_t	58	a_t	45	a_t	52	a_t	58
m	41	m	37	m	33	m	43	m	39	m	35
k	1	k	1	k	1	k	1	k	1	k	1

Finally, we analyze the sensitivity of our methodology to the choice of the accuracy level ϵ. We work in this experiment with the data set GAU10 and the related $S_m^{x^*}$ described in Tables 5 and 6. However, we obtained the exact same conclusions for all the other data sets employed in our study. We choose $\epsilon = 1, 2, 2.99$ for the $S_m^{x^*}$ in Table 5 and $\epsilon = 1, 2.99, 5$ for the training data set in Table 6. We remind that the bounds (3.20) and (3.25) work for $0 < \epsilon < 3$ and $\epsilon > 0$, respectively. Also, in this experiment, we plot the inter-quartile ranges

(a) GAU1A4's ensemble forecast based on Table 11

(b) GAU1A4's ensemble forecast based on Table 12

(c) NIG1A4's ensemble forecast based on Table 11

(d) NIG1A4's ensemble forecast based on Table 12
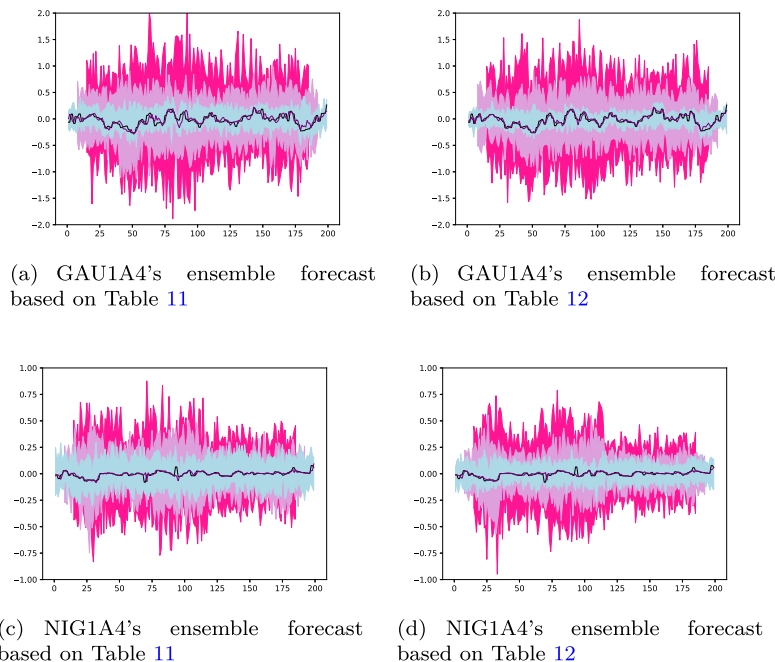
FIG 6. *Inter-quantile range of a* 50-*member ensemble forecast for* $p_t = 1$ *(blue),* $p_t = 8$ *(lilac),* $p_t = 15$ *(magenta) using the training data sets described in Table 11 and Table 12. The test set and the baseline forecasts are depicted in black and violet thick lines, respectively. The training data sets* $S_m^{x^*}$ *in Table 12 are respectively used as baseline estimates in (a), (b), (c) and (d). The x-axis represents the pixels in* $\mathbb{L}'$*, whereas the forecast values are along the y-axis.*

(a) GAU10's ensemble forecast based on Table 5

(b) GAU10's ensemble forecast based on Table 6
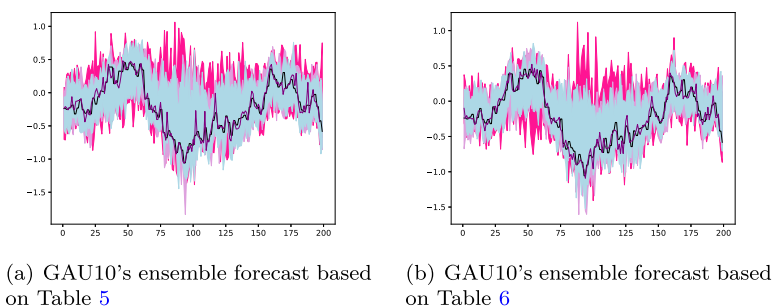
FIG 7. *(a) Inter-quantile range of a* 50-*member ensemble forecast for* $\epsilon = 1$ *(magenta),* $\epsilon = 2$ *(lilac),* $\epsilon = 2.99$ *(blue) using the training data set described in Table 5 for the GAU10. The test set and the baseline forecasts are depicted in thick black and violet lines, respectively. The training data sets* $S_m^{x^*}$ *in Table 5 is used as baseline estimates for* $\epsilon = 2.99$*. (b) Inter-quantile range of a* 50-*member ensemble forecast for* $\epsilon = 1$ *(magenta),* $\epsilon = 2.99$ *(lilac),* $\epsilon = 5$ *(blue) using the training data sets described in Table 6 for the GAU10. The test set and the baseline forecasts are depicted in thick black and violet lines, respectively. The training data sets* $S_m^{x^*}$ *in Table 6 is used as baseline estimates for* $\epsilon = 5$*. The x-axis represents the pixels in* $\mathbb{L}'$*, whereas the forecast values are along the y-axis.*

obtained for the different randomized Gibbs estimators in Figure 7 and observe that the bigger the parameter ϵ, the narrower the amplitude of the ranges. We choose in our experiment $\epsilon = 2.99$ because it is the bigger ϵ for which both the bound (3.20) and (3.25) are defined.

5. Conclusions

We define a novel theory-guided machine learning methodology for spatio-temporal data called MMAF-guided learning, which works in the class of the Lipschitz functions, e.g., linear functions and several types of neural network modules. Our methodology applies to raster data cubes, and it works under the assumption that such data are generated by an influenced mixed moving average field (MMAF, in short) defined on a cone-shaped ambit set. Such random fields are strictly stationary, θ-lex weakly dependent, and can be employed to model Gaussian and non-Gaussian distributed data. Moreover, they can be non-Markovian and admit non-separable covariance functions.

We show fixed-time and any-time PAC Bayesian bounds in this framework. All our bounds are determined for a bounded loss and depend on *one single* θ-lex coefficients of the underlying MMAF. In particular, our bounds hold for temporal and spatial short and long-range dependent data.

To enable one-time ahead ensemble forecasts, we need an estimate of the decay rate of the θ-lex coefficients of the underlying MMAF. Such an estimation is feasible, for example, in the case of STOU and MSTOU processes. We can then define spatio-temporal embeddings such that they make the right-hand side of the fixed-time PAC Bayesian bounds, proven in the paper, not vacuous. The analyzed embeddings give us possible training data sets for learning *randomized estimators* that have good generalization performance with a high probability. We can then determine one-time ahead ensemble forecasts.

In conclusion, we test the learning procedure for a *guided* randomized Gibbs estimator and a Gaussian reference distribution on the class of linear models. Our estimator has a convergence rate of $\mathcal{O}(m^{-\frac{1}{2}})$. We simulate a set of six data sets from an STOU process with Gaussian and NIG Lévy seed and determine (50 members) ensemble forecasts. We obtain that the inter-quartile ranges of our forecasts always contain the test set and are narrower when the number of observations in the training data set increases. Moreover, our forecasts have a causal interpretation induced by the ambit sets of the data-generating process known as Rubin's potential outcomes framework.

Appendix A: Appendix

A.1. *Weak dependence notions for causal processes and (influenced) MMAF*

In this section, we discuss in more detail the dependence notions called θ-weak dependence and θ-lex weak dependence. The latter notion has been introduced

in [28, Definition 2.1] as an extension to the random field case of the notion of θ-weak dependence satisfied by causal stochastic processes [32]. This notion of dependence is presented in Definition 2.14. However, the notion of θ-lex weak dependence given in Definition 2.13 slightly differs from the one given in [28, Definition 2.1] and represents an extension to the random field case of the θ-weak dependence notion defined in [33, Remark 2.1]. Note that the definitions of θ-weak dependence in [32] and [33, Remark 2.1] differ because of the cardinality of the marginal distributions on which the function G is computed, namely, $G \in \mathcal{G}_1$ in the former and $G \in \mathcal{G}_\nu$ for $\nu \in \mathbb{N}$ in the latter.

Let us now analyze the relationship between θ-weak dependence, α-mixing, and ϕ-mixing. Most of the PAC Bayesian literature for stationary and heavy-tailed data employs the following two mixing conditions, see Remark 3.22, namely α-mixing and ϕ-mixing. The results in the Lemma below give us a proof that the θ-weak dependence is more general than α-mixing and ϕ-mixing and therefore describes the dependence structure of a bigger class of models.

Let $\mathcal{M}$ and $\mathcal{V}$ be two sub-sigma algebras of $\mathcal{F}$. First of all, the strong mixing coefficient [70] is defined as

$$\alpha(\mathcal{M}, \mathcal{V}) = \sup\{|P(M)P(V) - P(M \cap V)|, M \in \mathcal{M}, V \in \mathcal{V}\}.$$

A stochastic process $\boldsymbol{X}$ is said to be α-mixing if

$$\alpha(r) = \alpha(\sigma\{\boldsymbol{X}_s, s \leq 0\}, \sigma\{\boldsymbol{X}_s, s \geq r\})$$

converges to zero as $r \to \infty$. The ϕ-mixing coefficient has been introduced in [49] and defined as

$$\phi(\mathcal{M}, \mathcal{V}) = \sup\{|P(V|M) - P(V)|, M \in \mathcal{M}, V \in \mathcal{V}, P(M) > 0\}.$$

A stochastic process $\boldsymbol{X}$ is said to be ϕ-mixing if

$$\phi(r) = \phi(\sigma\{\boldsymbol{X}_s, s \leq 0\}, \sigma\{\boldsymbol{X}_s, s \geq r\})$$

converges to zero as $r \to \infty$.

Lemma A.1. *Let $(\boldsymbol{X}_t)_{t \in \mathbb{Z}}$ be a stationary real-valued stochastic process such that $\mathbb{E}[|\boldsymbol{X}_0|^q] < \infty$ for some $q > 1$. Then,*

(a) *$\theta(r) \leq 2^{\frac{2q-1}{q}} \alpha(r)^{\frac{q-1}{q}} \|\boldsymbol{X}_0\|_q \leq 2^{\frac{q-1}{q}} \phi(r)^{\frac{q-1}{q}} \|\boldsymbol{X}_0\|_q$, and*

(b) *θ-weak dependence is a more general dependence notion than α-mixing and ϕ-mixing.*

Proof. The proof of the first inequality at point (a) is proven in [28, Proposition 2.5] using the representation of the θ-coefficients (2.12). The proof of the second inequality follows from a classical result in [19, Proposition 3.11]. In [28, Proposition 2.7], a stochastic process is defined, which is θ-weak dependent but neither α-mixing or ϕ-mixing. □

As seen in Definition 2.13 by using the lexicographic order in $\mathbb{R}^{1+d}$, an opportune extension of θ-weak dependence valid for random fields can be defined.

The definition of θ-lex coefficients for $G \in \mathcal{G}_1$ is given in [28, Definition 2.1]. The latter can be represented as $\theta_{lex}^v(r) := \sup_{u \in \mathbb{N}}\{\theta_{u,v}(r)\}$ for $v = 1$. Therefore, an alternative way to define the θ-lex coefficients in Definition 2.13 is obviously

$$\theta_{lex}(r) = \sup_{v \in \mathbb{N}} \theta_{lex}^v(r), v \in \mathbb{N} \text{ for all } r \in \mathbb{R}^+. \tag{A.1}$$

The following Lemma has important applications in the following sections.

Lemma A.2. *Let $\boldsymbol{Z}$ be a θ-lex weakly dependent random field and $M > 0$, then $\boldsymbol{Z}_i^M = \boldsymbol{Z}_i \vee (-M) \wedge M$ is θ-lex weakly dependent.*

Proof. Let $u, v \in \mathbb{N}$, $M > 0$, $F \in \mathcal{G}_u^*$, $G \in \mathcal{G}_v$, and $i_1, i_2, \ldots, i_u \in V_{\Gamma'}^r$ where $\Gamma' = \{j_1, \ldots, j_v\}$. Let $F^M(\boldsymbol{Z}_{i_1}, \ldots, \boldsymbol{Z}_{i_u}) = F(\boldsymbol{Z}_{i_1}^M, \ldots, \boldsymbol{Z}_{i_u}^M)$, and $G^M(\boldsymbol{Z}_{j_1}, \ldots, \boldsymbol{Z}_{j_v}) = G(\boldsymbol{Z}_{j_1}^M, \ldots, \boldsymbol{Z}_{j_v}^M)$. We have that F^M is a bounded function on $(\mathbb{R}^n)^u$ and G^M is a bounded and Lipschitz function on $(\mathbb{R}^n)^v$ (with the same Lipschitz coefficients as the function G). Let $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_v)$ and $(\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_v) \in (\mathbb{R}^n)^v$, then

$$|G^M(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_v) - G^M(\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_v)| \leq Lip(G) \sum_{i=1}^v |\boldsymbol{Z}_i^M - \tilde{\boldsymbol{Z}}_i^M|$$

$$\leq Lip(G) \sum_{i=1}^v |\boldsymbol{Z}_i - \tilde{\boldsymbol{Z}}_i|.$$

Hence, it holds that

$$|Cov(F(\boldsymbol{Z}_{i_1}^M, \ldots, \boldsymbol{Z}_{i_u}^M), G(\boldsymbol{Z}_{j_1}^M, \ldots, \boldsymbol{Z}_{j_v}^M))| \leq \|F\|_\infty v Lip(G) \theta_{lex}(r),$$

where $\theta_{lex}(r)$ are the θ-coefficients of the field $\boldsymbol{Z}$. So, the field $\boldsymbol{Z}_t^M(x)$ is θ-lex weakly dependent. $\qquad\square$

Note that the above result also holds for $\boldsymbol{X}$ a θ-weakly dependent process. Therefore, the truncated $\boldsymbol{X}_t^M = \boldsymbol{X}_t \vee (-M) \wedge M$ is a θ-weakly dependent process.

The notion of θ-lex weak dependence also admits a *projective-type representation.*

Remark A.3. *Let $(\boldsymbol{Z}_t)_{t \in \mathbb{Z}^{1+d}}$ be a real-valued and θ-lex weakly dependent random field, $\mathcal{L}_1 = \{g : \mathbb{R} \to \mathbb{R}, g \in \mathcal{G}_v, Lip(g) \leq 1\}$ and $\Gamma' = \{j_1, \ldots, j_v\} \in \mathbb{Z}^{1+d}$ such that $|\Gamma'| = v$. Let $r \in \mathbb{N}$, and $\mathcal{M} = \sigma\{\boldsymbol{Z}_t : t \in V_{\Gamma'}^r \subset \mathbb{Z}^{1+d}\}$, then it holds that*

$$\theta_{lex}(r) = \sup_{v \in \mathbb{N}} \sup_{\Gamma'} \sup_{g \in \mathcal{L}_1} \|\mathbb{E}[g(\boldsymbol{Z}_{j_1}, \ldots, \boldsymbol{Z}_{j_v})|\mathcal{M}] - \mathbb{E}[g(\boldsymbol{Z}_{j_1}, \ldots, \boldsymbol{Z}_{j_v})]\|_1. \tag{A.2}$$

The result above follows by readily applying [28, Lemma 5.1].

We now use the representation of the θ-lex coefficients (A.2) to understand its relationships to $\alpha_{\infty,v}$-mixing and $\phi_{\infty,v}$-mixing for $v \in \mathbb{N} \cup \{\infty\}$. These notions are defined in [31], and they are strong mixing notions used in the study of stationary random fields.

In general, for $u, v \in \mathbb{N} \cup \{\infty\}$, given coefficients

$$\alpha_{u,v}(r) = \sup\{\alpha(\sigma(\boldsymbol{Z}_\Gamma), \sigma(\boldsymbol{Z}_{\Gamma'})), \Gamma, \Gamma' \in \mathbb{R}^{1+d}, |\Gamma| \leq u, |\Gamma'| \leq v, dist(\Gamma, \Gamma') \geq r\},$$

and

$$\phi_{u,v}(r) = \sup\{\phi(\sigma(\boldsymbol{Z}_\Gamma), \sigma(\boldsymbol{Z}_{\Gamma'})), \Gamma, \Gamma' \in \mathbb{R}^{1+d}, |\Gamma| \leq u, |\Gamma'| \leq v, dist(\Gamma, \Gamma') \geq r\}.$$

a random field $\boldsymbol{Z}$ is said to be $\alpha_{u,v}$-mixing or $\phi_{u,v}$-mixing if the coefficients $(\alpha_{u,v}(r))_{r \in \mathbb{R}^+}$ or $(\phi_{u,v}(r))_{r \in \mathbb{R}^+}$ converge to zero as $r \to \infty$. We then have the following result.

Lemma A.4. *Let* $(\boldsymbol{Z}_t)_{t \in \mathbb{Z}^{1+d}}$ *be a stationary real-valued random field such that* $\mathbb{E}[|\boldsymbol{Z}_0|^q] < \infty$ *for some* $q > 1$*. Then, for* $v \in \mathbb{N} \cup \{\infty\}$*,*

(a) $\theta_{lex}(r) \leq 2^{\frac{2q-1}{q}} \alpha_{\infty,v}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q \leq 2^{\frac{q-1}{q}} \phi_{\infty,v}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q$*, and*

(b) *it holds that* θ*-lex weak dependence is more general than* $\alpha_{\infty,v}$*-mixing and* α*-mixing in the special case of stochastic processes. Moreover,* θ*-lex weak dependence is more general than* $\phi_{\infty,v}$*-mixing.*

Proof. From the proof of [28, Proposition 2.5], we have that

$$\theta_{lex}^1(r) \leq 2^{\frac{2q-1}{q}} \alpha_{\infty,1}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q.$$

Because of (A.1) and [19, Proposition 3.11], we have that

$$\theta_{lex}(r) \leq 2^{\frac{2q-1}{q}} \alpha_{\infty,1}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q \leq 2^{\frac{q-1}{q}} \phi_{\infty,1}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q.$$

Equally,

$$\theta_{lex}(r) \leq 2^{\frac{2q-1}{q}} \alpha_{\infty,v}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q \leq 2^{\frac{q-1}{q}} \phi_{\infty,v}(r)^{\frac{q-1}{q}} \|\boldsymbol{Z}_0\|_q.$$

The proof of the point (b) follows directly by [28, Proposition 2.7]. In fact $\theta_{lex}(r) = \theta_{1,\infty}(r)$ following the notations of [33, Definition 2.3] and the process used in the proof of the Proposition is θ-lex weakly dependent but neither $\alpha_{\infty,v}$, α or $\phi_{\infty,v}$-mixing. □

A.2. Autocovariance structure of MMAF and isotropy

Moment conditions for MMAFs are typically expressed in function of the characteristic quadruplet of its driving Lévy basis and the kernel function f.

Proposition A.5. *Let* $\boldsymbol{Z}$ *be an* $\mathbb{R}$*-valued MMAF driven by a Lévy basis with characteristic quadruplet* $(\gamma, \sigma^2, \nu, \pi)$ *with kernel function* $f : H \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ *and defined on an ambit set* $A_t(x) \subset \mathbb{R} \times \mathbb{R}^d$*.*

(i) *If* $\int_{|x|>1} |x|\nu(dx) < \infty$ *and* $f \in L^1(H \times \mathbb{R} \times \mathbb{R}^d) \cap L^2(H \times \mathbb{R} \times \mathbb{R}^d)$ *the first moment of* $\boldsymbol{Z}$ *is given by*

$$\mathbb{E}[\boldsymbol{Z}_t] = \mathbb{E}(\Lambda') \int_H \int_{A_t(x)} f(A, -s, -\xi) ds \, d\xi \, \pi(dA),$$

where $\mathbb{E}(\Lambda') = \gamma + \int_{|x| \geq 1} x \, \nu(dx)$*.*

(ii) If $\int_{\mathbb{R}} x^2 \, \nu(dx) < \infty$ and $f \in L^2(H \times \mathbb{R} \times \mathbb{R}^d)$, then $\boldsymbol{Z} \in L^2(\Omega)$ and

$$Var(\boldsymbol{Z}_t(x)) = Var(\Lambda') \int_H \int_{\mathbb{R} \times \mathbb{R}^d} f(A, -s, -\xi)^2 ds \, d\xi \, \pi(dA),$$

$$Cov(\boldsymbol{Z}_0(0), \boldsymbol{Z}_t(x))$$
$$= Var(\Lambda') \int_H \int_{A_0(0) \cap A_t(x)} f(A, -s, -\xi) f(A, t - s, x - \xi) \, ds \, d\xi \, \pi(dA),$$

and

$$Corr(\boldsymbol{Z}_0(0), \boldsymbol{Z}_t(x))$$
$$= \frac{\int_H \int_{A_0(0) \cap A_t(x)} f(A, -s, -\xi) f(A, t-s, x-\xi) \, ds \, d\xi \, \pi(dA)}{\int_H \int_{\mathbb{R} \times \mathbb{R}^d} f(A, -s, -\xi)^2 ds \, d\xi \, \pi(dA)},$$

where $Var(\Lambda') = \sigma^2 + \int_{\mathbb{R}^d} xx' \nu(dx)$.

(iii) If $\sigma^2 = 0$, $\int_{|x| \in \mathbb{R}} |x| \, \nu(dx) < \infty$, and $f \in L^1(H \times \mathbb{R} \times \mathbb{R}^d)$, then the first moment of $\boldsymbol{Z}$ is given by

$$E[\boldsymbol{Z}_t(x)] = \int_H \int_{A_t(x)} f(A, -s, -\xi) \Big(\gamma_0 + \int_{\mathbb{R}} x\nu(dx) \Big) ds \, d\xi \, \pi(dA),$$

where

$$\gamma_0 := \gamma - \int_{|x| \le 1} x \, \nu(dx). \tag{A.3}$$

Proof. Immediate from [73, Section 25] and [28, Theorem 3.3]. $\qquad\square$

From Proposition A.5, we can evince that the autocovariance function of an MMAF depends on the variance of the Lévy seed Λ', the kernel function f and the distribution π of the random parameter A.

We give below the explicit expression of the autocovariance functions for an STOU and MSTOU process.

Example A.6. *Let $\boldsymbol{Z}$ ad defined in Example 2.8, $u \in \mathbb{R}^d, \tau \in \mathbb{R}$, and $\mathbb{E}[\boldsymbol{Z}_t(x)^2] < \infty$. Then,*

$$Cov(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x + u)) = Var(\Lambda') \exp(-A\tau) \int_{A_t(x) \cap A_{t+\tau}(x+u)} \exp(-2A(t - s)) ds \, d\xi, \tag{A.4}$$

and

$$Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x + u)) = \frac{\exp(-A\tau) \int_{A_t(x) \cap A_{t+\tau}(x+u)} \exp(-2A(t - s)) \, ds \, d\xi}{\int_{A_t(x)} \exp(-2A(t - s)) \, ds \, d\xi}. \tag{A.5}$$

Example A.7. *Let $\boldsymbol{Z}$ ad defined in Example 2.8 and $d = 1$, then*

$$\rho^T(\tau) := Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x)) = \exp(-A|\tau|), \tag{A.6}$$

$$\rho^S(u) := Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_t(x+u)) = \exp\left(-A\frac{|u|}{c}\right), \tag{A.7}$$

$$\rho^{ST}(\tau, u) := Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x+u)) = \min\left(\exp(-A|\tau|), \exp\left(-\frac{A|u|}{c}\right)\right). \tag{A.8}$$

Example A.8. *Let $\boldsymbol{Z}$ be defined as in Example 2.9, $u \in \mathbb{R}^d$ and $\tau \in \mathbb{R}$, and $\mathbb{E}[\boldsymbol{Z}_t(x)^2] < \infty$, then*

$$Cov(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x+u)) =$$
$$= Var(\Lambda') \exp(-A\tau) \int_0^\infty \int_{A_t(x) \cap A_{t+\tau}(x+u)} \exp(-2A(t-s)) ds \, d\xi \, l(A) dA, \tag{A.9}$$

$$Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x+u)) =$$
$$\frac{\exp(-A\tau) \int_0^\infty \int_{A_t(x) \cap A_{t+\tau}(x+u)} \exp(-2A(t-s)) \, ds \, d\xi \, l(A) dA}{\int_0^\infty \int_{A_t(x)} \exp(-2A(t-s)) \, ds \, d\xi \, l(A) dA}. \tag{A.10}$$

Example A.9. *Let $\boldsymbol{Z}$ be defined as in Example 2.9 for $d = 1$, Assumption 2.10 hold, and $l(A) = \frac{\beta^\alpha}{\Gamma(\alpha)} A^{\alpha-1} \exp(-\beta A)$ be the Gamma density with shape and rate parameters $\alpha > d + 1$ and $\beta > 0$. For $d = 1$, $u \in \mathbb{R}$ and $\tau \in \mathbb{R}$*

$$Var(\boldsymbol{Z}_t(x)) = \frac{Var(\Lambda')c\beta^2}{2(\alpha-2)(\alpha-1)} \tag{A.11}$$

$$Cov(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x+u)) = \frac{Var(\Lambda')c\beta^\alpha}{2(\beta + \max\{|\tau|, |u|/c\})^{\alpha-2}(\alpha-2)(\alpha-1)}, \tag{A.12}$$

$$\rho^{ST}(\tau, u) := Corr(\boldsymbol{Z}_t(x), \boldsymbol{Z}_{t+\tau}(x+u)) = \left(\frac{\beta}{\beta + \max\{|\tau|, |u|/c\}}\right)^{\alpha-2}. \tag{A.13}$$

It follows the definition of an *isotropic* spatio-temporal random field.

Definition A.10 (Isotropy). *Let $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$. A spatio-temporal random field $(\boldsymbol{Z}_t(x))_{(t,x) \in \mathbb{R} \times \mathbb{R}^d}$ is called* isotropic *if its spatial covariance:*

$$Cov(\boldsymbol{Z}_t(x), \boldsymbol{Z}_t(x+u)) = C(|u|), \ \forall u \in \mathbb{R}^d$$

for some positive definite function C.

STOU and MSTOU processes defined on cone-shaped ambit sets are isotropic random fields.

A.3. Inference on STOU processes

Let us start by explaining the available estimation methodologies for the parameter vector $\theta_0 = \{A, c, Var(\Lambda')\}$ under the STOU modeling assumption when

the spatial dimension $d = 1$. Throughout, we refer to the notations used in Example 2.20.

We have two ways of estimating the parameter vector θ_0 in such a scenario. The first one is presented in [59]. Here, the parameters A and c are first estimated using *normalized spatial and temporal variograms* defined as

$$\gamma^S(u) := \frac{\mathbb{E}((\boldsymbol{Z}_t(x) - \boldsymbol{Z}_t(x - u))^2)}{Var(\boldsymbol{Z}_t(x))} = 2(1 - \rho^S(u)) = 2\Big(1 - \exp\Big(-\frac{Au}{c}\Big)\Big),$$
(A.14)

and

$$\gamma^T(\tau) := \frac{\mathbb{E}((\boldsymbol{Z}_t(x) - \boldsymbol{Z}_{t-\tau}(x))^2)}{Var(\boldsymbol{Z}_t(x))} = 2(1 - \rho^T(\tau)) = 2(1 - \exp(-A\tau)), \quad \text{(A.15)}$$

where ρ^S and ρ^T are defined in Example A.7. Note that normalized variograms are used to separate the estimation of the parameters A and c from the parameter $Var(\Lambda')$. Let $N(u)$ be the set containing all the pairs of indices at mutual spatial distance u for $u > 0$ and the same observation time. Let $N(\tau)$ be the set containing all the pairs of indices where the observation times are at a distance $\tau > 0$ and have the same spatial position. $|N(u)|$ and $|N(\tau)|$ give the number of the obtained pairs, respectively. Moreover, let $\hat{\boldsymbol{k}}_2$ be the empirical variance which is defined as

$$\hat{\boldsymbol{k}}_2 = \frac{1}{(D-1)} \sum_{i=1}^{D} \boldsymbol{Z^2}_{t_i}(x_i),$$
(A.16)

where D denotes the sample size. The empirical normalized spatial and temporal variograms are then defined as follows:

$$\hat{\boldsymbol{\gamma}}^S(u) = \frac{1}{|N(u)|} \sum_{i,j \in N(u)} \frac{(\boldsymbol{Z}_{t_i}(x_i) - \boldsymbol{Z}_{t_j}(x_j))^2}{\hat{\boldsymbol{k}_2}}$$
(A.17)

$$\hat{\boldsymbol{\gamma}}^T(\tau) = \frac{1}{|N(\tau)|} \sum_{i,j \in N(\tau)} \frac{(\boldsymbol{Z}_{t_i}(x_i) - \boldsymbol{Z}_{t_j}(x_j))^2}{\hat{\boldsymbol{k}_2}}.$$
(A.18)

By matching the empirical and the theoretical forms of the normalized variograms, we can estimate A and c by employing the estimators

$$\boldsymbol{A^*} = -\tau^{-1} \log\Big(1 - \frac{\hat{\boldsymbol{\gamma}}^T(\tau)}{2}\Big), \text{ and } \boldsymbol{c^*} = -\frac{\boldsymbol{A^*}u}{\log\Big(1 - \frac{\hat{\boldsymbol{\gamma}}^S(u)}{2}\Big)}.$$
(A.19)

Alternatively, we can use a least square methodology to estimate the parameters A and c, i.e. (A.17) and (A.18) are computed at several lags, and a least-squares estimation is used to fit the computed values to the theoretical curves. The authors in [59] use the methodology discussed in [54] to achieve the last target. We refer the reader also to [25, Chapter 2] for further discussions and examples of possible variogram model fitting. The parameter $Var(\Lambda')$ can be estimated

by matching the second-order cumulant of the STOU with its empirical counterpart. The consistency of this estimation procedure is proven in [59, Theorem 12].

A second possible methodology for estimating the vector θ_0 employs a generalized method of moment estimator (GMM), as in [60]. It is essential to notice that by using such an estimator, we cannot separate the parameter $Var(\Lambda')$ from the estimation of the parameters A and c. Instead, all moment conditions must be combined into one optimization criterion, and all the estimations must be found simultaneously. Consistency and asymptotic normality of the GMM estimator are discussed in [60] and [28], respectively.

For $d \geq 2$, a least square methodology is still applicable for estimating the variogram's parameters. The estimator used in [59] is a normalized version of the least-square estimator for spatial variogram's parameters discussed in [54], which also applies for $d > 1$. This method, paired with a method of moments (matching the second-order cumulant of the field $\boldsymbol{Z}$ with its empirical counterparts), allows estimating the parameter $Var(\Lambda')$. The GMM methodology discussed in [60] also continues to apply for $d \geq 2$. However, when the spatial dimension is increasing, the shape of the *normalized variograms* and the *field's moments* become more complex, and higher computational effort is required to navigate through the high dimensional surface of the optimization criterion behind least-squares or GMM estimators.

A.4. Inference for MSTOU processes

When estimating the parameter vector $\theta_1 = \{\alpha, \beta, c, Var(\Lambda')\}$ under an MSTOU modeling assumption– see, for example, solely the shape of the coefficients in Example 2.21– it is evident that the shape of the autocorrelation function, and therefore of the normalized temporal and spatial variograms, become more complex for increasing d. As already addressed in the previous sections, when estimating the parameters (α, β, c) alone, we can use the least-squares type estimator discussed in [54]. Moreover, by pairing the latter with a method of moments or using a GMM estimator, we can estimate the complete vector θ_1.

A.5. Time series models

In the MMAF framework, we can also find time series models. The latter are θ-weakly dependent.

Example A.11 (Time series case)**.** *The supOU process studied in [7] and [12] is an example of a* causal *mixed moving average process. Let the kernel function* $f(A, s) = e^{-As}1_{[0,\infty)}(s)$, $A \in \mathbb{R}^+$, $s \in \mathbb{R}$ *and* Λ *a Lévy basis on* $\mathbb{R}^+ \times \mathbb{R}$ *with generating quadruple* $(\gamma, \sigma^2, \nu, \pi)$ *such that*

$$\int_{|x|>1} \log(|x|)\,\nu(dx) < \infty, \ and \ \int_{\mathbb{R}^+} \frac{1}{A}\pi(dA) < \infty, \tag{A.20}$$

then the process

$$\boldsymbol{Z}_t = \int_{\mathbb{R}^+} \int_{-\infty}^{t} e^{-A(t-s)} \, \Lambda(dA, ds) \tag{A.21}$$

is well defined for each $t \in \mathbb{R}$ and strictly stationary and called a supOU process where A represents a random mean reversion parameter.

If $\mathbb{E}(\Lambda') = 0$ and $\int_{|x|>1} |x|^2 \nu(dx) < \infty$, the supOU process is θ-weakly dependent with coefficients

$$\theta_Z(r) \le \Big(\int_{\mathbb{R}^+} \int_{-\infty}^{r} e^{-2As} \sigma^2 \, ds \, \pi(dA) \Big)^{\frac{1}{2}} = \Big[Var(\Lambda') \int_{\mathbb{R}^+} \frac{e^{-2Ar}}{2A} \, \pi(dA) \Big]^{\frac{1}{2}} \tag{A.22}$$

$$= Cov(\boldsymbol{Z}_0, \boldsymbol{Z}_{2r})^{\frac{1}{2}},$$

where $Var(\Lambda') = \sigma^2 + \int_{\mathbb{R}} x^2 \nu(dx)$, by using Theorem 3.11 in [12].

If $\mathbb{E}(\Lambda') = \mu$ and $\int_{|x|>1} |x|^2 \nu(dx) < \infty$, the supOU process is θ-weakly dependent with coefficients

$$\theta_Z(r) \le \Big(Cov(\boldsymbol{Z}_0, \boldsymbol{Z}_{2r}) + \frac{4\mu^2}{Var(\Lambda')^2} Cov(\boldsymbol{Z}_0, \boldsymbol{Z}_r)^2 \Big)^{\frac{1}{2}}. \tag{A.23}$$

If $\int_{\mathbb{R}} |x| \nu(dx) < \infty$, $\sigma^2 = 0$, $\gamma_0 = \gamma - \int_{|x| \le 1} x \, \nu(dx) > 0$ and $\nu(\mathbb{R}^-) = 0$, where $\mathbb{R}^-$ identifies the set of the negative real numbers, then the supOU process admits θ-coefficients

$$\theta_Z(r) \le \mu \int_{\mathbb{R}^+} \frac{e^{-Ar}}{A} \, \pi(dA), \tag{A.24}$$

and when in addition $\int_{|x|>1} |x|^2 \nu(dx) < \infty$

$$\theta_Z(r) \le \frac{2\mu}{Var(\Lambda')} Cov(\boldsymbol{Z}_0, \boldsymbol{Z}_r). \tag{A.25}$$

Note that the necessary and sufficient condition $\int_{\mathbb{R}^+} \frac{1}{A} \pi(dA)$ for the supOU process to exist is satisfied by many continuous and discrete distributions π, see [79, Section 2.4] for more details. For example, a probability measure π being absolutely continuous with density $\pi' = x^h l(x)$ and regularly varying at zero from the right with $h > 0$, i.e. l is slowly varying at zero and satisfies the above condition. If moreover, $l(x)$ is continuous in $(0. + \infty)$ and $\lim_{x \to 0^+} l(x) > 0$ exists, it holds that

$$Cov(\boldsymbol{Z}_0, \boldsymbol{Z}_r) \sim \frac{C}{r^h}, \text{ with a constant } C > 0 \text{ and } r \in \mathbb{R}^+$$

where for $h \in (0, 1)$ the supOU process exhibits long memory and for $h > 1$ short memory. In this set-up, concrete examples where the covariances are calculated explicitly can be found in [9] and [27].

Another interesting example of MMAFs is given by the class of *trawl processes*. A distinctive feature of these processes is that one can model the correlation structure independently from the marginal distribution, see [10] for further details on their definition. In the case of trawl processes, we also have available in the literature likelihood-based methods for estimating their parameters; see [15] for further details.

In general, the generalized method of moments is employed to estimate the parameters of an MMAF, see [27, 59, 60].

Appendix B: Appendix

B.1. Bounds for the θ-lex coefficients of MMAF

In [28, Proposition 3.11], it is given a general methodology to show that an MMAF $\boldsymbol{Z}$ is θ-lex weakly dependent. Given that the definition of θ-lex-weak dependence used in the paper slightly differs from the one given in [28], the proof of Proposition B.1 differs from the one of [28, Proposition 3.11]. Proposition 2.17 is a novel computation of a θ-lex coefficients' bound of an MMAF when the kernel does not depend on the spatial component.

Before giving a detailed account of these proofs, let us state first some notations. Let $r > 0$, $\{(t_{j_1}, x_{j_1}), \ldots (t_{j_v}, x_{j_v})\} = \Gamma' \in \mathbb{R}^{1+d}$ and $\{(t_{i_1}, x_{i_1}), \ldots, (t_{i_u}, x_{i_u})\} = \Gamma \in V_{\Gamma'}^r$ such that $|\Gamma| = u$ and $|\Gamma'| = v$ for $(u, v) \in \mathbb{N} \times \mathbb{N}$. We call the *truncated* (influenced) MMAF the vector

$$\boldsymbol{Z}_{\Gamma'}^{(\psi)} = \left(\boldsymbol{Z}_{t_{j_1}}^{(\psi)}(x_{j_1}), \ldots, \boldsymbol{Z}_{t_{j_v}}^{(\psi)}(x_{j_v}) \right)^\top, \tag{B.1}$$

where $\psi := \psi(r)$ for $r > 0$. In particular, for all $a \in \{1, \ldots, u\}$ and a $b \in \{1, \ldots, v\}$, ψ has to be chosen such that it exists a set $B_{t_{j_b}}^{\psi}(x_{j_b})$ with the following properties.

- $|B_{t_{j_b}}^{\psi}(x_{j_b})| \to \infty$ as $r \to \infty$ for all b, and
- $I_{i_a} = H \times A_{t_{i_a}}(x_{i_a})$ and $I_{j_b} = H \times B_{t_{j_b}}^{\psi}(x_{j_b})$ are disjoint sets or intersect on a set $H \times O$, where $O \subset \mathbb{R}^{1+d}$ and $dim(O) < d + 1$, for all a and b

Let us now assume that it is possible to construct the sets $B_{t_{j_b}}^{\psi}(x_{j_b})$. Then, since $\pi \times \lambda_{1+d}(H \times O) = 0$ and by the definition of a Lévy basis, it follows that

$$\boldsymbol{Z}_{t_{i_a}}(x_{i_a}) = \int_H \int_{A_{t_{i_a}}(x_{i_a})} f(A, t_{i_a} - s, x_{i_a} - \xi) \Lambda(dA, ds, d\xi) \text{ and}$$

$$\boldsymbol{Z}_{t_{j_b}}^{(\psi)}(x_{j_b}) = \int_H \int_{B_{t_{j_b}}^{\psi}(x_{j_b})} f(A, t_{j_b} - s, x_{j_b} - \xi) \Lambda(dA, ds, d\xi),$$

and $\boldsymbol{Z}_\Gamma$ and $\boldsymbol{Z}_{\Gamma'}^{(\psi)}$ are independent. Hence, for $F \in \mathcal{G}_u^*$ and $G \in \mathcal{G}_v$, $F(\boldsymbol{Z}_\Gamma)$ and

$G(\boldsymbol{Z}_{\Gamma'}^{(\psi)})$ are also independent. Now

$$
\begin{aligned}
&|Cov(F(\boldsymbol{Z}_\Gamma), G(\boldsymbol{Z}_{\Gamma'}))| \\
&\leq |Cov(F(\boldsymbol{Z}_\Gamma), G(\boldsymbol{Z}_{\Gamma'}^{(\psi)}))| + |Cov(F(\boldsymbol{Z}_\Gamma), G(\boldsymbol{Z}_{\Gamma'}) - G(\boldsymbol{Z}_{\Gamma'}^{(\psi)}))| \\
&= |E[(G(\boldsymbol{Z}_{\Gamma'}) - G(\boldsymbol{Z}_{\Gamma'}^{(\psi)}))F(\boldsymbol{Z}_\Gamma)] - E[G(\boldsymbol{Z}_{\Gamma'}) - G(\boldsymbol{Z}_{\Gamma'}^{(\psi)})]E[F(\boldsymbol{Z}_\Gamma)]| \\
&\leq 2\|F\|_\infty E[|G(\boldsymbol{Z}_{\Gamma'}) - G(\boldsymbol{Z}_{\Gamma'}^{(\psi)})|] \qquad\qquad\qquad\qquad\qquad\text{(B.2)} \\
&\leq 2\mathrm{Lip}(G)\|F\|_\infty \sum_{l=1}^{v} E[|\boldsymbol{Z}_{t_{j_l}}(x_{j_l}) - \boldsymbol{Z}_{t_{j_l}}^{(\psi)}(x_{j_l})|] = \\
&= 2\mathrm{Lip}(G)\|F\|_\infty v E[|\boldsymbol{Z}_{t_{j_1}}(x_{j_1}) - \boldsymbol{Z}_{t_{j_1}}^{(\psi)}(x_{j_1})|],
\end{aligned}
$$

because an (influenced) MMAF is a stationary random field. To show that a field satisfy Definition 2.13, is then enough to prove that $E[|\boldsymbol{Z}_{t_{j_1}}(x_{j_1}) - \boldsymbol{Z}_{t_{j_1}}^{(\psi)}(x_{j_1})|]$ in the above inequality converges to zero as $r \to \infty$. The proofs of Proposition B.1 and 2.17 below differ in the definition of the sequence ψ and the sets $B_{t_{j_b}}^{\psi}(x_{j_b})$.

Proposition B.1. *Let Λ be an $\mathbb{R}$-valued Lévy basis with characteristic quadruplet $(\gamma, \sigma^2, \nu, \pi)$, $f : H \times \mathbb{R}^{1+d} \to \mathbb{R}$ a $\mathcal{B}(H \times \mathbb{R}^{1+d})$-measurable function and $\boldsymbol{Z}_t(x)$ be defined as in (2.6).*

(i) If $\int_{|x|>1} x^2\nu(dx) < \infty$, $\gamma + \int_{|x|>1} x\nu(dx) = 0$ and $f \in L^2(H \times \mathbb{R}^{1+d})$, then $\boldsymbol{Z}$ is θ-lex-weakly dependent and

$$
\theta_{lex}(r) \leq 2\Big(\int_H \int_{-\infty}^{\rho(r)} Var(\Lambda') \int_{\|\xi\|\leq cs} f(A, -s, -\xi)^2 ds d\xi \pi(dA)\Big)^{\frac{1}{2}}.
$$

(ii) If $\int_{|x|>1} x^2\nu(dx) < \infty$ and $f \in L^2(H \times \mathbb{R}^{1+d}) \cap L^1(H \times \mathbb{R}^{1+d})$, then $\boldsymbol{Z}$ is θ-lex-weakly dependent and

$$
\begin{aligned}
\theta_{lex}(r) \leq 2\Bigg(&\int_H \int_{-\infty}^{\rho(r)} Var(\Lambda') \int_{\|\xi\|\leq cs} f(A, -s)^2 \, ds d\xi \pi(dA) \\
&+ \Big| \int_H \int_{-\infty}^{\rho(r)} \mathbb{E}(\Lambda') \int_{\|\xi\|\leq cs} f(A, -s) \, ds d\xi \pi(dA) \Big|^2 \Bigg)^{\frac{1}{2}}.
\end{aligned}
$$

(iii) If $\int_{\mathbb{R}} |x|\,\nu(dx) < \infty$, $\sigma^2 = 0$ and $f \in L^1(H \times \mathbb{R}^{1+d})$ with γ_0 defined in (A.3), then $\boldsymbol{Z}$ is θ-lex-weakly dependent and

$$
\begin{aligned}
\theta_{lex}(r) \leq 2\Bigg(&\int_H \int_{-\infty}^{\rho(r)} \int_{\|\xi\|\leq cs} |f(A, -s)\gamma_0| \, ds \, d\xi \pi(dA) \\
&+ \int_H \int_{-\infty}^{\rho(r)} \int_{\|\xi\|\leq cs} \int_{\mathbb{R}} |f(A, -s)x| \, \nu(dx) \, ds \pi(dA) \Bigg).
\end{aligned}
$$

The results above hold for all $r > 0$ with

$$
\rho(r) = \frac{-r\min(1/c, 1)}{\sqrt{(d+1)(c^2+1)}}, \qquad\qquad\qquad\qquad\text{(B.3)}
$$

$Var(\Lambda') = \sigma^2 + \int_{\mathbb{R}} x^2\,\nu(dx)$ *and* $\mathbb{E}(\Lambda') = \gamma + \int_{|x|\geq 1} x\nu(dx)$.

Proof of Proposition B.1. In this proof, we assume that $B^{\psi}_{t_{j_b}}(x_{j_b}) = A_{t_{j_b}}(x_{j_b})\backslash V^{\psi}_{(t_{j_b},x_{j_b})}$, where $\psi = \psi(r) := \frac{r}{\sqrt{(d+1)(c^2+1)}}$.

(i) Using the translation invariance of $A_t(x)$ and $V^{(\psi)}_{(t,x)}$ we obtain

$$E[|\boldsymbol{Z}_{t_{j_1}}(x_{j_1}) - \boldsymbol{Z}^{(\psi)}_{t_{j_1}}(x_{j_1})|]$$

$$\leq \left(\int_H \int_{A_0(0)\cap V^{\psi}_0} Var(\Lambda')f(A,-s,-\xi)^2 d\xi ds\pi(dA) \right)^{\frac{1}{2}}$$

$$= \left(\int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} Var(\Lambda') \int_{\|\xi\|\leq cs} f(A,-s,-\xi)^2\, d\xi ds\pi(dA) \right)^{\frac{1}{2}},$$

where we have used Proposition A.5-(ii) to bound the L_1-distance from above. Overall, we obtain

$$\theta_{lex}(r) \leq 2 \left(\int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} Var(\Lambda') \int_{\|\xi\|\leq cs} f(A,-s,-\xi)^2 d\xi ds\pi(dA) \right)^{\frac{1}{2}},$$

which converges to zero as r tends to infinity by applying the dominated convergence theorem.

(ii) By applying Proposition A.5-(i) and (ii), we obtain

$$E[|\boldsymbol{Z}_{t_{j_1}}(x_{j_1}) - \boldsymbol{Z}^{(\psi)}_{t_{j_1}}(x_{j_1})|]$$

$$\leq \left(\int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} Var(\Lambda') \int_{\|\xi\|\leq cs} f(A,-s,-\xi)^2 d\xi ds\pi(dA) \right.$$

$$\left. + \left(\int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} \mathbb{E}(\Lambda') \int_{\|\xi\|\leq cs} f(A,-s,-\xi) d\xi ds\pi(dA) \right)^2 \right)^{\frac{1}{2}}.$$

Finally, we proceed similarly to proof (i) and obtain the desired bound.

(iii) We apply now Proposition A.5-(iii). Then,

$$E[|\boldsymbol{Z}_{t_{j_1}}(x_{j_1}) - \boldsymbol{Z}^{(\psi)}_{t_{j_1}}(x_{j_1})|]$$

$$\leq \left(\int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} \int_{\|\xi\|\leq cs} |f(A,-s,-\xi)\gamma_0| d\xi ds\pi(dA) \right.$$

$$\left. + \int_H \int_{-\infty}^{\frac{-r\min(1/c,1)}{\sqrt{(d+1)(c^2+1)}}} \int_{\|\xi\|\leq cs} \int_{\mathbb{R}} |f(A,-s,-\xi)y|\nu(dy) d\xi ds\pi(dA) \right).$$

The bound for the θ-lex-coefficients is obtained following the proof line in (i).

$\square$

Proposition B.1 gives general bounds for the θ-lex coefficients of MMAF. For example, it can also be used to compute upper bounds for the θ-lex-coefficients of an MSTOU process for $d > 2$ which Proposition 2.17 does not cover.

Corollary B.2. *Let $\mathbf{Z}$ be an MSTOU process as in Definition 2.9 and $(\gamma, \sigma^2, \nu, \pi)$ be the characteristic quadruplet of its driving Lévy basis. Moreover, let the mean reversion parameter A be $Gamma(\alpha, \beta)$ distributed with density $l(A) = \frac{\beta^\alpha}{\Gamma(\alpha)} A^{\alpha-1} \exp(-\beta A)$ where $\alpha > d + 1$ and $\beta > 0$.*

(i) *If $\int_{|x|>1} x^2 \, \nu(dx) < \infty$ and $\gamma + \int_{|x|>1} x\nu(dx) = 0$, then $\mathbf{Z}$ is θ-lex-weakly dependent. Let $c \in [0,1]$, then for*

$d=1,$

$$\theta_{lex}(r) \leq 2 \left(\frac{cVar(\Lambda')\beta^\alpha}{2\Gamma(\alpha)} \left(\frac{\Gamma(\alpha-2)}{(2\psi+\beta)^{\alpha-2}} + \frac{2\psi\Gamma(\alpha-1)}{(2\psi+\beta)^{\alpha-1}} \right) \right)^{\frac{1}{2}},$$

and for

$d \geq 2,$

$$\theta_{lex}(r) \leq 2 \left(V_d(c) \frac{d!Var(\Lambda')\beta^\alpha}{2^{d+1}} \sum_{k=0}^{d} \frac{(2\psi)^k}{k!(2\psi+\beta)^{\alpha-d-1+k}} \frac{\Gamma(\alpha-d-1+k)}{\Gamma(\alpha)} \right)^{\frac{1}{2}}.$$

Let $c > 1$, then for

$d \in \mathbb{N},$

$$\theta_{lex}(r) \leq 2 \left(V_d(c) \frac{d!Var(\Lambda')\beta^\alpha}{2^{d+1}} \sum_{k=0}^{d} \frac{\left(\frac{2\psi}{c}\right)^k}{k! \left(\frac{2\psi}{c}+\beta\right)^{\alpha-d-1+k}} \frac{\Gamma(\alpha-d-1+k)}{\Gamma(\alpha)} \right)^{\frac{1}{2}}.$$

The above implies that, in general, $\theta_{lex}(r) = \mathcal{O}(r^{\frac{(d+1)-\alpha}{2}})$.

(ii) *If $\int_{\mathbb{R}} |x| \, \nu(dx) < \infty$, $\Sigma = 0$ and γ_0 as defined in (A.3), then $\mathbf{Z}$ is θ-lex-weakly dependent. Let $c \in (0,1]$, then for*

$$d \in \mathbb{N}, \quad \theta_{lex}(r) \leq 2V_d(c)d!\beta^\alpha \gamma_{abs} \sum_{k=0}^{d} \frac{\psi^k}{k!(\psi+\beta)^{\alpha-d-1+k}} \frac{\Gamma(\alpha-d-1+k)}{\Gamma(\alpha)},$$

whereas for $c > 1$ and

$$d \in \mathbb{N}, \quad \theta_{lex}(r) \leq 2V_d(c)d!\beta^\alpha \gamma_{abs} \sum_{k=0}^{d} \frac{\left(\frac{\psi}{c}\right)^k}{k! \left(\frac{\psi}{c}+\beta\right)^{\alpha-d-1+k}} \frac{\Gamma(\alpha-d-1+k)}{\Gamma(\alpha)},$$

where $\gamma_{abs} = |\gamma_0| + \int_{\mathbb{R}} |x|\nu(dx)$, $V_d(c)$ denotes the volume of the d-dimensional ball with radius c, and $\psi := \psi(r) = \frac{1}{\sqrt{c^2+1}} \frac{r}{d+1}$.

Proof. Proof of this corollary can be obtained by modifying the proof of [28, Section 3.7] in line with the calculations performed in Proposition B.1. □

The results of the Corollary above imply that, in general, $\theta_{lex}(r) = \mathcal{O}(r^{(d+1)-\alpha})$. We give now the proof of Proposition 2.17.

Proof of Proposition 2.17. (i) Without loss of generality, let us determine the truncated set when $(t_{j_b}, x_{j_b}) = (0, 0)$. We use, to this end, two auxiliary ambit sets translated by a value $\psi > 0$ along the spatial axis, namely, the cones $A_0(\psi)$ and $A_0(-\psi)$, as illustrated in Figure 8-(a), (c) for $c \leq 1$ and in Figure 8-(b), (d) for $c > 1$. Then, we set the truncated integration set to $B_0^\psi(0) = A_0(0) \backslash (A_0(\psi) \cup A_0(-\psi))$. Since $(t_{i_a}, x_{i_a}) \in V_{(0,0)}^r$, it is sufficient to choose ψ such that the integration set of $\boldsymbol{Z}_0^{(\psi)}(0)$ is a subset of $(V_{(0,0)}^r)^c$. To this end, the three intersecting points $(\frac{-\psi}{2c}, \frac{-\psi}{2})$, $(\frac{-\psi}{c}, 0)$ and $(\frac{-\psi}{2c}, \frac{\psi}{2})$ have to be inside the set $(V_{(0,0)}^r)^c$, as illustrated in Figure 8-(e) for $c \leq 1$ and in Figure 8-(f) for $c > 1$. Clearly, this leads to the conditions $\psi \leq rc$, $\psi \leq 2r$ and $\psi \leq 2rc$, which are satisfied for $\psi = r\min(2, c)$. Hence, by using Proposition A.5-(ii), we have that

$$\theta_{lex}(r) \leq 2Var(\Lambda')^{\frac{1}{2}} \left(\int_0^\infty \int_{A_0(0) \cap (A_0(\psi) \cup A_0(-\psi))} f(A, -s)^2 ds d\xi \pi(d\lambda) \right)^{\frac{1}{2}}$$

$$= 2Var(\Lambda')^{\frac{1}{2}} \left(\int_0^\infty \int_{A_0(0) \cap A_0(\psi)} f(A, -s)^2 ds d\xi \pi(d\lambda) \right.$$

$$+ \int_0^\infty \int_{A_0(0) \cap A_0(-\psi)} f(A, -s)^2 ds d\xi \pi(d\lambda)$$

$$\left. - \int_0^\infty \int_{A_0(0) \cap A_0(\psi) \cap A_0(-\psi)} f(A, -s)^2 ds d\xi \pi(d\lambda) \right)^{\frac{1}{2}}$$

$$\leq 2\sqrt{2Cov(\boldsymbol{Z}_0(0), \boldsymbol{Z}_0(r\min(2, c)))}.$$

which converges to zero as $r \to \infty$ for the dominated convergence theorem.

(ii) In this proof, we indicate the spatial components and write $x_{i_a} = (y_{i_a}, z_{i_a}) \in \mathbb{R} \times \mathbb{R}$ and $x_{j_b} = (y_{j_b}, z_{j_b}) \in \mathbb{R} \times \mathbb{R}$ for $a \in \{1, \ldots, u\}$ and $b \in \{1, \ldots, v\}$. Without loss of generality, let us then determine the truncated set when $(t_{j_b}, y_{j_b}, z_{j_b}) = (0, 0, 0)$. To this end, we use four additional ambit sets that are translated by a value $\psi > 0$ along both spatial axis, namely, the cones $A_0(\psi, \psi)$, $A_0(\psi, -\psi)$, $A_0(-\psi, \psi)$ and $A_0(-\psi, -\psi)$, as illustrated in Figure 10-(c) for $c \leq 1$ and in Figure 10-(d) for $c > 1$). Then, we set the truncated integration set to $B_0^\psi(0) = A_0(0, 0) \backslash (A_0(\psi, \psi) \cup A_0(\psi, -\psi) \cup A_0(-\psi, \psi) \cup A_0(-\psi, -\psi))$. Since $(t_{i_a}, y_{i_a}, z_{i_a}) \in V_{(0,0,0)}^r$, it is sufficient to
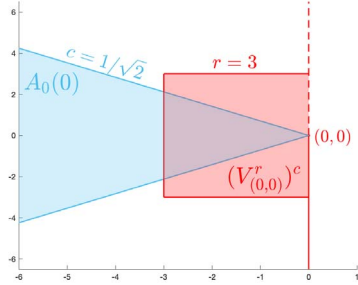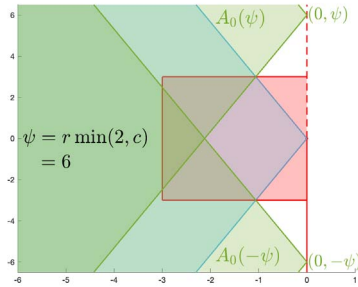
(a) Integration set $A_0(0)$ for $c = 1/\sqrt{2}$ together with the complement of $V_{(0,0)}^r$ for $r = 3$.
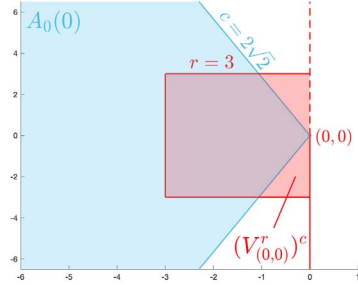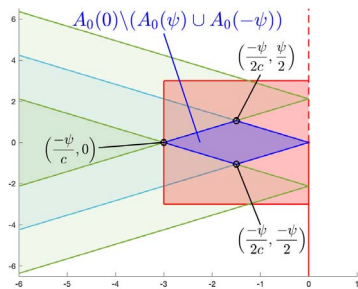
(b) Integration set $A_0(0)$ for $c = 2\sqrt{2}$ together with the complement of $V_{(0,0)}^r$ for $r = 3$.
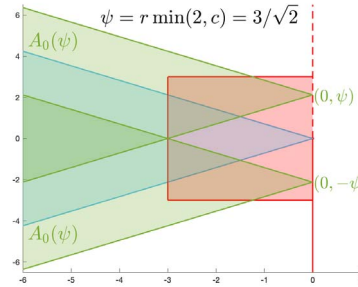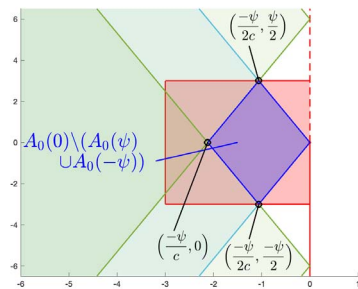
(c) Integration set $A_0(0)$ together with $A_0(\psi)$ and $A_0(-\psi)$ for $\psi = r\min(2,c)$.

(d) Integration set $A_0(0)$ together with $A_0(\psi)$ and $A_0(-\psi)$ for $\psi = r\min(2,c)$.

(e) Integration set of $\boldsymbol{Z}_0^{(\psi)}(0)$ together with the complement of $V_{(0,0)}^r$.

(f) Integration set of $\boldsymbol{Z}_0^{(\psi)}(0)$ together with the complement of $V_{(0,0)}^r$.

Fig 8. *Exemplary integration set and truncated integration set of an MMAF $\boldsymbol{Z}_t(x)$ for $(t,x) = (0,0)$.*

choose ψ such that the integration set of $\boldsymbol{Z}_0^{(\psi)}(0,0)$ is a subset of $(V_{(0,0,0)}^r)^c$, i.e.

$$\sup_{b \in B_0^\psi(0)} \|b\|_\infty \leq r. \tag{B.4}$$

In the following, we prove that the choice $\psi = r\min(1, c/\sqrt{2})$ is sufficient for (B.4) to hold. We investigate cross-sections of the truncated integration set $B_0^\psi(0)$ along the time axis. For a fixed time point t, we call this cross-section B^t and, similarly, we denote the cross-section of an ambit set by A^t. Note that the cross sections of our ambit sets along the time axis are circles with radius $|ct|$ (see also Figure 9).

$t \in \left(\frac{-\psi}{\sqrt{2}c}, 0\right]$: As the distance between the center of the circle $A_0^t(0,0)$ and the centers of the circles $A_0^t(\psi,\psi)$, $A_0^t(-\psi,\psi)$, $A_0^t(\psi,-\psi)$, $A_0^t(-\psi,-\psi)$ is $\sqrt{2}\psi$, respectively, the set $A_0^t(0,0)$ (which is a circle with radius $|ct|$) is disjoint from every of the additional ambit sets' cross-sections at t and hence $B^t = A_0^t(0,0)$ (see Figure 9-(a)). Clearly, we obtain

$$\sup_{t \in \left(-\psi/(\sqrt{2}c), 0\right]} \sup_{b \in B^t} \|b\| = \sup_{t \in \left(-\psi/(\sqrt{2}c), 0\right]} \max(c|t|, |t|)$$

$$= \max\left(\frac{\psi}{\sqrt{2}}, \frac{\psi}{\sqrt{2}c}\right). \tag{B.5}$$

$t \in \left(-\frac{\psi}{c}, -\frac{\psi}{\sqrt{2}c}\right]$: For such t the set $A_0^t(0,0)$ intersects with every additional ambit sets' cross-section (see Figure 9-(b)). However, as the additional ambit sets' cross-sections do not intersect with each other, the point $p_1(t) = (t, c|t|, 0) \in B^t$ on the boundary of $A_0^t(0,0)$ (see the red point in Figure 9-(b)) is not excluded from B^t by any additional ambit set. Note that symmetry makes it sufficient to look at $p_1(t)$. Hence, we obtain

$$\sup_{t \in \left(-\psi/c, -\psi/(\sqrt{2}c)\right]} \sup_{b \in B^t} \|b\| = \sup_{t \in \left(-\psi/c, -\psi/(\sqrt{2}c)\right]} \|p_1(t)\|$$

$$= \max\left(\psi, \frac{\psi}{c}\right). \tag{B.6}$$

$t \in \left(-\frac{\sqrt{2}\psi}{c}, -\frac{\psi}{c}\right]$: For such t the set $A_0^t(0,0)$ intersects with every additional ambit sets' cross-section. Such intersections additionally restrict $A_0^t(0,0)$ (see Figure 9-(c)). Straightforward calculations show that the point where the boundaries of $A_0^t(-\psi,\psi)$ and $A_0^t(-\psi,\psi)$ as well as the set $A_0^t(0,0)$ intersect, say $p_2(t)$, is given by $(t, 0, \psi - \sqrt{(ct)^2 - \psi^2})$ (see red point in Figure 9-(c)). Note that symmetry makes it sufficient to look at $p_2(t)$. We obtain

$$\sup_{t \in \left(-\sqrt{2}\psi/c, -\psi/c\right]} \sup_{b \in B^t} \|b\| = \sup_{t \in \left(-\sqrt{2}\psi/c, -\psi/c\right]} \|p_2(t)\|$$

$$= \max\left(\psi, \frac{\sqrt{2}\psi}{c}\right). \tag{B.7}$$

$t \le -\frac{\sqrt{2}\psi}{c}$: In the following, we show that for such t, the set $A^t(0,0)$ is entirely included in the union of the additional ambit sets' cross sections. Clearly, this is true if the upper point where the boundaries of $A_0^t(-\psi,\psi)$ and $A_0^t(-\psi,\psi)$ intersect, say $p_3(t)$, is outside of $A_0^t(0,0)$ (see the red point in Figure 9-(d)). Note that symmetry makes it sufficient to look at $p_3(t)$. As straightforward calculations show that $p_3(t) = (t,0,\psi+\sqrt{(ct)^2-\psi^2})$, this is true if $\psi + \sqrt{(ct)^2-\psi^2} \ge c|t|$, or equivalently $(\psi+\sqrt{(ct)^2-\psi^2})^2 \ge (ct)^2$. Moreover, we have

$$\psi^2 + 2\psi\sqrt{(ct)^2-\psi^2} + (ct)^2 - \psi^2 \ge (ct)^2 \iff \psi \ge 0. \tag{B.8}$$

In view of condition (B.4) we combine (B.5), (B.6) and (B.7) and set $\psi = r\min(1,c/\sqrt{2})$, which also satisfies (B.8).

In addition to the cross-sectional views from Figure 9, we give a full three-dimensional view of the set $B_0^{\psi}(0)$ for $c \le 1$ in Figure 10-(e) and for $c > 1$ in Figure 10-(f) that highlight the points on the boundary of $B_0^{\psi}(0)$ with maximal ∞-norm for $\psi = r\min(1,c/\sqrt{2})$.

Therefore, because of Proposition A.5-(ii), we can conclude that

$\theta_{lex}(r)$

$$\le 2Var(\Lambda')^{\frac{1}{2}}\left(\int_0^\infty\!\!\!\int_{A_0(0,0)\cap(A_0(\psi,\psi)\cup A_0(-\psi,\psi)\cup A_0(\psi,-\psi)\cup A_0(-\psi,-\psi))} f(A,-s)^2 ds d\xi\pi(d\lambda)\right)^{\frac{1}{2}}$$

$$= 2Var(\Lambda')^{\frac{1}{2}}\left(\int_0^\infty\int_{A_0(0,0)\cap A_0(\psi,\psi)} f(A,-s)^2 ds d\xi\pi(d\lambda)\right.$$

$$+\int_0^\infty\int_{A_0(0,0)\cap A_0(-\psi,\psi)} f(A,-s)^2 ds d\xi\pi(d\lambda)$$

$$+\int_0^\infty\int_{A_0(0,0)\cap A_0(\psi,-\psi)} f(A,-s)^2 ds d\xi\pi(d\lambda)$$

$$+\int_0^\infty\int_{A_0(0,0)\cap A_0(-\psi,-\psi)} f(A,-s)^2 ds d\xi\pi(d\lambda)$$

$$-\int_0^\infty\int_{A_0(0)\cap A_0(\psi,-\psi)\cap A_0(-\psi,-\psi)} f(A,-s)^2 ds d\xi\pi(d\lambda)$$

$$-\int_0^\infty\int_{A_0(0)\cap A_0(-\psi,\psi)\cap(A_0(\psi,-\psi)\cup A_0(-\psi,-\psi))} f(A,-s)^2 ds d\xi\pi(d\lambda)$$

$$\left.-\int_0^\infty\int_{A_0(0)\cap A_0(\psi,\psi)\cap(A_0(-\psi,\psi)\cup A_0(\psi,-\psi)\cup A_0(-\psi,-\psi))} f(A,-s)^2 ds d\xi\pi(d\lambda)\right)^{\frac{1}{2}}$$

(a) Cross sections of the auxiliary ambit sets and $A_0^t(0,0)$ for $t = -1$ and $c = 2$.

(b) Cross sections of the auxiliary ambit sets and $A_0^t(0,0)$ for $t = -5/4$ and $c = 2$.

(c) Cross sections of the auxiliary ambit sets and $A_0^t(0,0)$ for $t = -7/4$ and $c = 2$.

(d) Cross sections of the auxiliary ambit sets and $A_0^t(0,0)$ for $t = -2.2$ and $c = 2$.
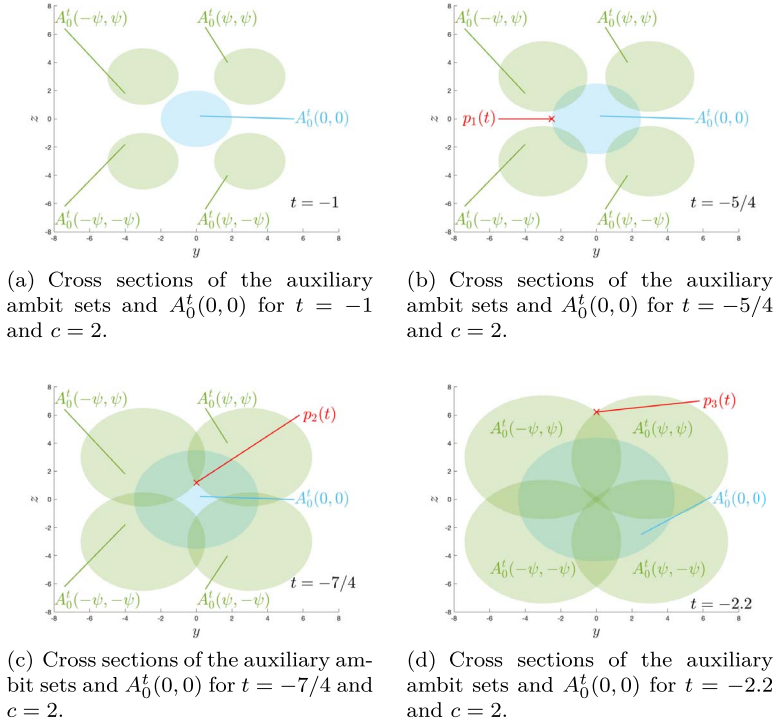
FIG 9. *Cross sections of the auxiliary ambit sets and $A_0^t(0,0)$ at different time points for $\psi = 3$.*

$$\leq 2\sqrt{2Cov(\boldsymbol{Z}_0(0,0), \boldsymbol{Z}_0(\psi,\psi)) + 2Cov(\boldsymbol{Z}_0(0,0), \boldsymbol{Z}_0(\psi,-\psi))}.$$

which converges to zero as $r \to \infty$ for the dominated convergence theorem. $\qquad\square$

B.2. Proofs of Section 3

Proof of Proposition 3.2. We drop the bold notations indicating random fields and stochastic processes in the following. Let $h \in \mathcal{H}$, we call $L_i = L(h(X_i), Y_i)$ for $i \in \mathbb{Z}$, $Z_t^{(M)}(x) := Z_t(x) \vee (-M) \wedge M$ for $M > 1$, and $L_i^{(M)} = L(h(X_i^{(M)}), Y_i^{(M)})$ where

$$X_i^{(M)} = L_p^{-(M)}(t_0 + ia, x^*), \quad \text{and} \quad Y_i^{(M)} = Z_{t_0+ia}^{(M)}(x^*), \quad \text{for } i \in \mathbb{Z},$$

and

$$L_p^{-(M)}(t, x^*) = \{Z_s^{(M)}(\xi) : (s, \xi) \in \mathbb{Z} \times \mathbb{L}, \ \|x^* - \xi\| \leq c\,(t - s) \text{ and } t - s \leq p\}.$$
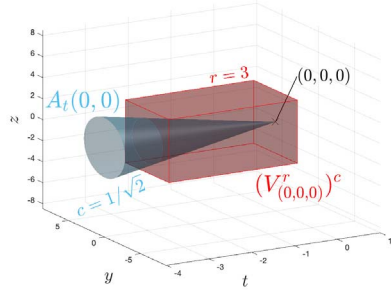
(a) Integration set $A_0(0,0)$ for $c = 1/\sqrt{2}$ together with the complement of $V^r_{(0,0,0)}$ for $r = 3$.
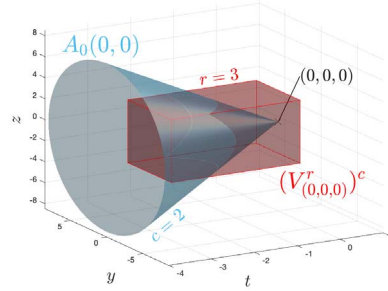
(b) Integration set $A_0(0,0)$ for $c = 2$ together with the complement of $V^r_{(0,0,0)}$ for $r = 3$.
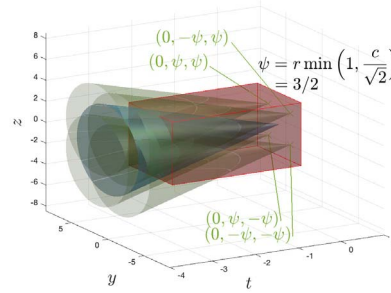
(c) Integration set $A_0(0,0)$ together with $A_0(\psi,\psi)$, $A_0(-\psi,\psi)$, $A_0(\psi,-\psi)$ and $A_0(-\psi,-\psi)$ for $\psi = r\min(1, c/\sqrt{2})$.
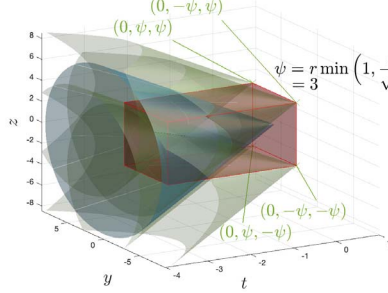
(d) Integration set $A_0(0,0)$ together with $A_0(\psi,\psi)$, $A_0(-\psi,\psi)$, $A_0(\psi,-\psi)$ and $A_0(-\psi,-\psi)$ for $\psi = r\min(1, c/\sqrt{2})$.
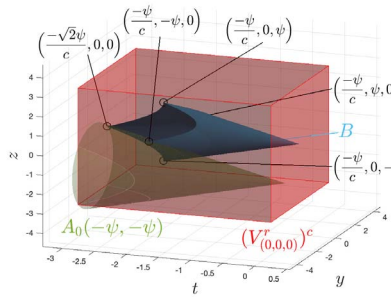
(e) B is the integration set of $\boldsymbol{Z}_0^{(\psi)}(0,0)$. In addition, we illustrate $A_0(-\psi,-\psi)$ for $c = 1/\sqrt{2}$ together with the complement of $V^r_{(0,0,0)}$ for $r = 3$.
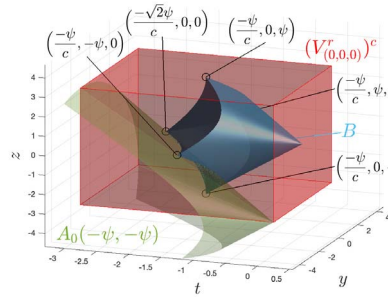
(f) B is the integration set of $\boldsymbol{Z}_0^{(\psi)}(0,0)$. In addition, we illustrate $A_0(-\psi,-\psi)$ for $c = 2$ together with the complement of $V^r_{(0,0,0)}$ for $r = 3$.

Fig 10. *Integration set and truncated integration set of an MMAF $\boldsymbol{Z}_t(y,z)$ for $d = 2$.*

for $t = t_0 + ia$ with $i \in \mathbb{Z}$. For $u \in \mathbb{N}$, $i_1 \leq i_2 \leq \ldots \leq i_u < i_u + k = j$ with $k \in \mathbb{N}$, let us consider the marginal of the field

$$\Big((X_{i_1}, Y_{i_1}), \ldots, (X_{i_u}, Y_{i_u}), (X_j, Y_j)\Big), \tag{B.9}$$

and let us define

$$\Gamma = \{(t_i, x_i) \in \mathbb{Z}^{1+d} : Z_{t_i}(x_i) \in L_p^-(t_0 + i_s a, x^*) \text{ or } (t_i, x_i) = (t_0 + i_s a, x^*)$$
$$\text{for } s = 1, \ldots, u\},$$

and

$$\Gamma' = \{(t_i, x_i) \in \mathbb{Z}^{1+d} : Z_{t_i}(x_i) \in L_p^-(t_0 + ja, x^*) \text{ or } (t_i, x_i) = (t_0 + ja, x^*)\}.$$

Then $r = dist(\Gamma, \Gamma')$. In particular $\Gamma \in V_{\Gamma'}^r$, and $r = (j - i_u)a - p$. For $F \in \mathcal{G}_u^*$ and $G \in \mathcal{G}_1$, then

$$|Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j))| \tag{B.10}$$
$$\leq |Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j) - G(L_j^{(M)}))| \tag{B.11}$$
$$+ |Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j^{(M)}))|. \tag{B.12}$$

The summand (B.11) is less than or equal to

$$2\|F\|_\infty Lip(G)\mathbb{E}[|L_j - L_j^{(M)}|]$$
$$\leq 2\|F\|_\infty Lip(G)(\mathbb{E}[|Y_j - Y_j^{(M)}|] + \mathbb{E}[|h(X_j) - h(X_j^{(M)})|])$$
$$\leq 2\|F\|_\infty Lip(G)(Lip(h)a(p,c) + 1)\mathbb{E}[|Z_{t_1}(x_1) - Z_{t_1}^{(M)}(x_1)|]$$

by stationarity of the field Z, and because L and h are Lipschitz functions. Moreover, the function $G(L_j^{(M)})$ belongs to $\mathcal{G}_{a(p,c)+1}$. Let $(X, Y), (X', Y') \in \mathbb{R}^{a(p,c)+1}$, then

$$|G(L(h(X^{(M)}), Y^{(M)})) - G(L(h(X'^{(M)}), Y'^{(M)}))|$$
$$\leq Lip(G)|L(h(X^{(M)}), Y^{(M)}) - L(h(X'^{(M)}), Y'^{(M)})|$$
$$\leq Lip(G)(|h(X^{(M)})) - h(X'^{(M)})| + |Y^{(M)} - Y'^{(M)}|$$
$$\leq Lip(G)(Lip(h) + 1)(\|X - X'\|_1 + |Y - Y'|),$$

and $Lip(G(L_j^{(M)})) \leq Lip(G)(Lip(h) + 1)$.

Because Z is a θ-lex weakly dependent random field, (B.12) is less than or equal to

$$\tilde{d}\|F\|_\infty Lip(G)(Lip(h) + 1)\theta_{lex}(r).$$

We choose now $M = r$ and obtain that (B.10) is less than or equal than

$$\|F\|_\infty Lip(G)\tilde{d}(Lip(h)a(p,c) + 1)\Big(\frac{2}{\tilde{d}}\mathbb{E}[|Z_{t_1}(x_1) - Z_{t_1}^{(r)}(x_1)|] + \theta_{lex}(r)\Big).$$

The quantity above converges to zero as $r \to \infty$. Therefore, $(L_i)_{i \in \mathbb{Z}}$ is a θ-weakly dependent process. □

Remark B.3. *Note that when working with a general θ-weakly dependent random field, as in Proposition 3.2, we do not employ the truncated random field defined in (B.1). This happens because a θ-lex weakly dependent field is not generally defined as an integral driven by a Lévy basis on an ambit set. The field (B.1) can just be employed in the MMAF framework. In a more general framework, without further information on the structure of the field, we can just employ the truncated random field $\boldsymbol{Z}_t^{(M)}(x) := \boldsymbol{Z}_t(x) \vee (-M) \wedge M$.*

Proof of Proposition 3.6. We drop the bold notations indicating random fields and stochastic processes in the following. We call $L_i = L(h(X_i), Y_i)$ for $i \in \mathbb{Z}$, as defined in Proposition B.1. Moreover, we employ the truncated field $Z_t^{(\psi)}(x)$ and define $L^{(\psi)} = (L(h(X_i^{(\psi)}), Y_i^{(\psi)}))_{i \in \mathbb{Z}}$ where

$$X_i^{(\psi)} = L_p^{-(\psi)}(t_0 + ia, x^*), \quad \text{and} \quad Y_i^{(\psi)} = Z_{t_0+ia}^{(\psi)}(x^*), \quad \text{for } i \in \mathbb{Z},$$

and

$$L_p^{-(\psi)}(t, x^*) = \{Z_s^{(\psi)}(\xi) : (s, \xi) \in \mathbb{Z} \times \mathbb{L}, \ \|x^* - \xi\| \le c\,(t-s) \text{ and } t - s \le p\}.$$

for $t = t_0 + ia$ with $i \in \mathbb{Z}$. For $u \in \mathbb{N}$, $i_1 \le i_2 \le \ldots \le i_u < i_u + k = j$ with $k \in \mathbb{N}$, let us consider the marginal of the field

$$\Big((X_{i_1}, Y_{i_1}), \ldots, (X_{i_u}, Y_{i_u}), (X_j, Y_j)\Big), \tag{B.13}$$

and let us define

$$\Gamma = \{(t_i, x_i) \in \mathbb{Z}^{1+d} \colon Z_{t_i}(x_i) \in L_p^-(t_0 + i_s a, x^*) \text{ or } (t_i, x_i) = (t_0 + i_s a, x^*)$$
$$\text{for } s = 1, \ldots, u\},$$

and

$$\Gamma' = \{(t_i, x_i) \in \mathbb{Z}^{1+d} \colon Z_{t_i}(x_i) \in L_p^-(t_0 + ja, x^*) \text{ or } (t_i, x_i) = (t_0 + ja, x^*)\}.$$

Then $r = dist(\Gamma, \Gamma')$. In particular $\Gamma \in V_{\Gamma'}^r$, and $r = (j - i_u)a - p$. For $F \in \mathcal{G}_u^*$ and $G \in \mathcal{G}_1$, then

$$|Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j))|$$
$$\le |Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j) - G(L_j^{(\psi)}))| \tag{B.14}$$
$$+ |Cov(F(L_{i_1}, \ldots, L_{i_u}), G(L_j^{(\psi)}))|. \tag{B.15}$$

The summand (B.15) is equal to zero because $\Gamma \in V_{\Gamma'}^r$, see proof of Proposition 2.17 for more details about this part of the proof. We can then bound (B.14) from above by

$$2\|F\|_\infty Lip(G)\mathbb{E}[|L_j - L_j^{(\psi)}|]$$
$$\le 2\|F\|_\infty Lip(G)(\mathbb{E}[|Y_j - Y_j^{(\psi)}|] + \mathbb{E}[|h(X_j) - h(X_j^{(\psi)})|]) \tag{B.16}$$

$$\leq 2\|F\|_\infty Lip(G)(Lip(h)a(p,c)+1)\mathbb{E}[|Z_{t_1}(x_1)-Z_{t_1}^{(\psi)}(x_1)|] \tag{B.17}$$

where (B.16) holds because L is a function with Lipschitz constant equal to one, and (B.17) holds given that h is Lipschitz.

When we work with linear functions, we consider h parameterized with respect to the set $B \in \mathbb{R}^{a(p,c)}$

$$\mathbb{E}[|h_\beta(X_j) - h_\beta(X_j^{(\psi)})|\|] = \mathbb{E}\Big[\Big|\sum_{l=1}^{a(p,c)} \beta_{1,l}(Z_{t_l}(x_l) - Z_{t_l}^{(\psi)}(x_l))\Big|\Big]. \tag{B.18}$$

By stationarity of the field Z, we have that (B.18) is smaller or equal than $\|\beta_1\|_1\mathbb{E}[|Z_{t_1}(x_1) - Z_{t_1}^{(\psi)}(x_1)|]$. Overall, we have that (B.14) is smaller or equal than

$$2\|F\|_\infty Lip(G)(Lip(h)a(p,c)+1)\mathbb{E}[|Z_{t_1}(x_1)-Z_{t_1}^{(\psi)}(x_1)|] \quad \text{for } h \text{ a Lipschitz function,}$$

or it is smaller or equal than

$$2\|F\|_\infty Lip(G)(\|\beta_1\|_1+1)\mathbb{E}[|Z_{t_1}(x_1) - Z_{t_1}^{(\psi)}(x_1)|] \text{ for } h_\beta \text{ a linear function.}$$

Because of the properties of the truncated field $Z_{t_1}^{(\psi)}(x_1)$, we have that the above bounds converge to zero as $r \to \infty$. Therefore, L is a θ-weakly dependent process. $\qquad\square$

The proof of Theorem 3.7 uses a blocks technique introduced in the papers [44] and [56]. Such results are based on the use of several lemmas. To ease the complete understanding of the proof of Theorem 3.7, we prove these Lemmas below, given that they undergo several modifications in our framework. Let us start by partitioning a set $\{1, 2, \ldots, m\}$ into k blocks. Each block will contain $l = \lfloor \frac{m}{k} \rfloor$ terms. Let $h = m - kl < k$ denote the remainder when we divide m by k. We now construct k blocks such that the number of elements in the jth-block is defined by

$$\bar{l}_j = \begin{cases} l+1 & \text{if } j = 1, 2, \ldots, h \\ l & \text{if } j = h+1, \ldots, k \end{cases}.$$

Let $(\boldsymbol{U}_i)_{i\in\mathbb{Z}}$ a stationary process, and $\boldsymbol{V}_m = \sum_{i=1}^m \boldsymbol{U}_i$, for $j = 1, \ldots, k$ we define the j-th block as

$$\boldsymbol{V}_{j,m} = \boldsymbol{U}_j + \boldsymbol{U}_{j+k} + \ldots \boldsymbol{U}_{j+(\bar{l}_j-1)k} = \sum_{i=1}^{\bar{l}_j} \boldsymbol{U}_{j+(i-1)k}$$

such that

$$\boldsymbol{V}_m = \sum_{j=1}^k \boldsymbol{V}_{j,m} = \sum_{j=1}^k \sum_{i=1}^{\bar{l}_j} \boldsymbol{U}_{j+(i-1)k}.$$

For $j = 1, 2, \ldots, k$, let us define $p_j = \frac{\bar{l}_j}{m}$. It follows that $\sum_{j=1}^k p_j = \frac{1}{m} \sum_{j=1}^k \bar{l}_j = 1$.

Lemma B.4. *For all $s \in \mathbb{R}$*

$$\mathbb{E}\Big[\exp\Big(s\frac{\boldsymbol{V}_m}{m}\Big)\Big] \leq \sum_{j=1}^{k} p_j \mathbb{E}\Big[\exp\Big(s\frac{\boldsymbol{V}_{j,m}}{\bar{l}_j}\Big)\Big]$$

The proof of the result above is due to Hoeffding [46].

Lemma B.5. *Let the assumptions of Theorem 3.7 hold and define the process $(\boldsymbol{U}_i)_{i\in\mathbb{Z}}$ such that $\boldsymbol{U}_i := f(\boldsymbol{Z}_i) - \mathbb{E}[f(\boldsymbol{Z}_i)]$. For all $j = 1, 2, \ldots, k$, $l \geq 2$ and $0 < s < \frac{3l}{|b-a|}$*

$$M_{\bar{l}_j}(s) = \Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j}\exp\Big(\frac{s\,\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big] - \prod_{i=1}^{\bar{l}_j}\mathbb{E}\Big[\exp\Big(\frac{s\,\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big]\Big|$$

$$\leq \exp(s\,|b-a|)\theta(k)s \tag{B.19}$$

The same result holds when defining the process $(\boldsymbol{U}_i)_{i\in\mathbb{Z}}$ for $\boldsymbol{U}_i = \mathbb{E}[f(\boldsymbol{Z}_i)] - f(\boldsymbol{Z}_i)$.

Proof. Let us first discuss the case when $\boldsymbol{U}_i = f(\boldsymbol{Z}_i) - \mathbb{E}[f(\boldsymbol{Z}_i)]$, we have that the process $\boldsymbol{U}$ has mean zero and $|\boldsymbol{U}_i| \leq |b-a|$. Let us define $\mathcal{F}_j = \sigma(\boldsymbol{U}_i, i \leq j)$.

$$M_{\bar{l}_j} := \Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j}\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big] - \prod_{i=1}^{\bar{l}_j}\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big]\Big|$$

$$= \Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j-1}\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big|\mathcal{F}_{j+(\bar{l}_j-2)k}\Big]\Big]$$

$$- \prod_{i=1}^{\bar{l}_j}\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big]\Big|$$

$$\leq \Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j-1}\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big(\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big|\mathcal{F}_{j+(\bar{l}_j-2)k}\Big]$$

$$- \mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big]\Big)\Big]\Big|$$

$$+ \Big|\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big]\Big|\Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j-1}\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big] - \prod_{i=1}^{\bar{l}_j-1}\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)k}}{\bar{l}_j}\Big)\Big]\Big|$$

$$\leq \Big\|\prod_{i=1}^{\bar{l}_j-1}\exp\Big(\frac{s|\boldsymbol{U}_{j+(i-1)k}|}{\bar{l}_j}\Big)\Big\|_{\infty}\mathbb{E}\Big[\Big|\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big|\mathcal{F}_{j+(\bar{l}_j-2)k}\Big]$$

$$- \mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(\bar{l}_j-1)k}}{\bar{l}_j}\Big)\Big]\Big|\Big]$$

$$+ \Big\|\exp\Big(\frac{s|\boldsymbol{U}_{j+(\bar{l}_j-1)k}|}{\bar{l}_j}\Big)\Big\|_{\infty}M_{\bar{l}_j-1}$$

The above is then less than or equal to

$$\exp\left(\frac{s(\bar{l}_j-1)|b-a|}{\bar{l}_j}\right)\exp\left(\frac{-sa}{\bar{l}_j}\right)\mathbb{E}\left[\left|\mathbb{E}\left[\exp\left(\frac{sf(\boldsymbol{Z}_{j+(\bar{l}_j-1)k})}{\bar{l}_j}\right)\middle|\mathcal{F}_{j+(\bar{l}_j-2)k}\right]\right.\right.$$
(B.20)

$$\left.\left.-\mathbb{E}\left[\exp\left(\frac{sf(\boldsymbol{Z}_{j+(\bar{l}_j-1)k})}{\bar{l}_j}\right)\right]\right|\right]+\exp\left(\frac{s|b-a|}{\bar{l}_j}\right)M_{\bar{l}_j-1}.$$

Note that the function $g(x)=\dfrac{\exp\left(\frac{sx}{\bar{l}_j}\right)}{\exp\left(\frac{sb}{\bar{l}_j}\right)\frac{s}{\bar{l}_j}}1_{[a,b]}(x)$ is in $\mathcal{L}_1$ for each s. We then use the projective-type representation of the θ-coefficients of $f(\boldsymbol{Z})$, see Remark 2.15, and obtain that

$$M_{\bar{l}_j}\leq\exp\left(\frac{s\bar{l}_j|b-a|}{\bar{l}_j}\right)\theta(k)\frac{s}{\bar{l}_j}+\exp\left(\frac{s|b-a|}{\bar{l}_j}\right)M_{\bar{l}_j-1}. \qquad \text{(B.21)}$$

Let now, $u=\exp\left(\frac{s|b-a|}{\bar{l}_j}\right)$, we have that

$$M_{\bar{l}_j}\leq\theta(k)u^{\bar{l}_j}\frac{s}{\bar{l}_j}+uM_{\bar{l}_j-1}$$

$$\leq(\bar{l}_j-2)\theta(k)u^{\bar{l}_j}\frac{s}{\bar{l}_j}+u^{\bar{l}_j-2}\left|\mathbb{E}\left[\exp\left(\frac{s\boldsymbol{U}_j}{\bar{l}_j}\right)\exp\left(\frac{s\boldsymbol{U}_{j+k}}{\bar{l}_j}\right)\right]-\mathbb{E}\left[\exp\left(\frac{s\boldsymbol{U}_j}{\bar{l}_j}\right)\right]\right.$$

$$\left.\mathbb{E}\left[\exp\left(\frac{s\boldsymbol{U}_{j+k}}{\bar{l}_j}\right)\right]\right|$$

$$\leq(\bar{l}_j-2)\theta(k)u^{\bar{l}_j}\frac{s}{\bar{l}_j}+u^{\bar{l}_j-1}\left|\mathbb{E}\left[\exp\left(\frac{s\boldsymbol{U}_{j+k}}{\bar{l}_j}\right)|\mathcal{F}_j\right]-\mathbb{E}\left[\exp\left(\frac{s\boldsymbol{U}_{j+k}}{\bar{l}_j}\right)\right]\right|$$

$$\leq(\bar{l}_j-1)\theta(k)u^{\bar{l}_j}\frac{s}{\bar{l}_j}=(\bar{l}_j-1)\exp\left(s|b-a|\right)\theta(k)\frac{s}{\bar{l}_j}.$$

In conclusion, for all $j=1,\ldots,k$ (and remembering that $\bar{l}_j=l$ or $\bar{l}_j=l+1$)

$$M_{\bar{l}_j}\leq\exp(s|b-a|)\theta(k)s.$$

For $\boldsymbol{U}_i=\mathbb{E}[f(\boldsymbol{Z}_i)]-f(\boldsymbol{Z}_i)$, the proof follows the same line as above provided that the point (B.20) is changed as follows:

$$\exp\left(\frac{s(\bar{l}_j-1)|b-a|}{\bar{l}_j}\right)\exp\left(\frac{sb}{\bar{l}_j}\right)\mathbb{E}\left[\left|\mathbb{E}\left[\exp\left(\frac{-sf(\boldsymbol{Z}_{j+(\bar{l}_j-1)k})}{\bar{l}_j}\right)\middle|\mathcal{F}_{j+(\bar{l}_j-2)k}\right]\right.\right.$$

$$\left.\left.-\mathbb{E}\left[\exp\left(\frac{-sf(\boldsymbol{Z}_{j+(\bar{l}_j-1)k})}{\bar{l}_j}\right)\right]\right|\right]+\exp\left(\frac{s|b-a|}{\bar{l}_j}\right)M_{\bar{l}_j-1}.$$

Note that the function $g(x)=\dfrac{\exp\left(\frac{sx}{\bar{l}_j}\right)}{\exp\left(\frac{-sa}{\bar{l}_j}\right)\frac{s}{\bar{l}_j}}1_{[-b,-a]}(x)$ is in $\mathcal{L}_1$ for each s. We then use the projective-type representation of the θ-coefficients of $f(\boldsymbol{Z})$ and obtain again that the inequality (B.21) holds. $\qquad\square$

Lemma B.6. *Let the Assumptions of Theorem 3.7 hold and define the process* $(\boldsymbol{U}_i)_{i \in \mathbb{Z}}$ *such that* $\boldsymbol{U}_i = f(\boldsymbol{Z}_i) - \mathbb{E}[f(\boldsymbol{Z}_i)]$. *For all* $j = 1, 2, \ldots, k$, $l \geq 2$ *and* $0 < s < \frac{3l}{b-a}$

$$\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{V}_{j,m}}{\bar{l}_j}\Big)\Big] \leq \exp\Big(\frac{s^2 \mathbb{E}[\boldsymbol{U}_1^2]}{2l\Big(1 - \frac{s|b-a|}{3l}\Big)}\Big) + \exp(s\,|b-a|)\theta(k)s$$

The same result holds when defining the process $(\boldsymbol{U}_i)_{i \in \mathbb{Z}}$ *for* $\boldsymbol{U}_i = \mathbb{E}[f(\boldsymbol{Z}_i)] - f(\boldsymbol{Z}_i)$.

Proof.

$$\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{V}_{j,m}}{\bar{l}_j}\Big)\Big] = \mathbb{E}\Big[\exp\Big(\sum_{i=1}^{\bar{l}_j} \frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big] \leq \prod_{i=1}^{\bar{l}_j} \mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big]$$

$$+ \Big|\mathbb{E}\Big[\prod_{i=1}^{\bar{l}_j} \exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big] - \prod_{i=1}^{\bar{l}_j} \mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big]\Big|$$

$$= \mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big]^{\bar{l}_j} + M_{\bar{l}_j} \tag{B.22}$$

We have that $\mathbb{E}[\boldsymbol{U}_{j+(i-1)\,k}] = 0$ by definition of the process $\boldsymbol{U}$, and $\frac{\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}$ satisfies the Bernstein moment condition (Remark A1 [56]) with $K_1 = \frac{|b-a|}{3\bar{l}_j}$. Hence, for $\bar{l}_j \geq 2$ and $0 < s < \frac{3\bar{l}_j}{|b-a|}$

$$\mathbb{E}\Big[\exp\Big(\frac{s\boldsymbol{U}_{j+(i-1)\,k}}{\bar{l}_j}\Big)\Big] \leq \exp\Big(\frac{s^2 \mathbb{E}[(\boldsymbol{U}_{j+(i-1)k}/\bar{l}_j)^2]}{2\Big(1 - \frac{s|b-a|}{3\bar{l}_j}\Big)}\Big). \tag{B.23}$$

Because $\bar{l}_j \geq l$, we can conclude that the inequality above holds for $l \geq 2$. Moreover, by stationarity of the process $\boldsymbol{U}$ and since for all $j = 1, 2, \ldots, k$, we can bound (B.22) uniformly with respect to the index j by using Lemma B.5, and noticing that

$$0 < s < \frac{3l}{|b-a|} \leq \frac{3\bar{l}_j}{|b-a|},$$
and then
$$\Big(1 - \frac{s|b-a|}{3\bar{l}_j}\Big) \geq \Big(1 - \frac{s|b-a|}{3l}\Big).$$

The same proof applies when defining $\boldsymbol{U}_i = \mathbb{E}[f(\boldsymbol{Z}_i)] - f(\boldsymbol{Z}_i)$. $\qquad\square$

Proof of Theorem 3.7. By combining Lemmas B.4, B.5, B.6, for $0 < s < \frac{3l}{|b-a|}$:

$$\mathbb{E}\Big[\exp\Big(s\frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{Z}_i) - \mathbb{E}[f(\boldsymbol{Z}_i)]\Big)\Big] = \mathbb{E}\Big[\exp\Big(s\frac{1}{m}\sum_{i=1}^{m} \boldsymbol{U}_i\Big)\Big]$$

$$= \mathbb{E}\Big[\exp\Big(\frac{s}{m} \sum_{j=1}^{k} \boldsymbol{V}_{j,m} \Big) \Big]$$

$$= \mathbb{E}\Big[\exp\Big(\frac{s}{m} \sum_{j=1}^{k} \sum_{i=1}^{\bar{l}_j} \boldsymbol{U}_{j+(i-1)\,k} \Big) \Big]$$

$$\leq \sum_{j=1}^{k} p_j \mathbb{E}\Big[\exp\Big(\frac{s \boldsymbol{V}_{j,m}}{\bar{l}_j} \Big) \Big] \qquad \text{(B.24)}$$

$$\leq \exp\left(\frac{s^2 Var(f(\boldsymbol{Z}_1))}{2l\Big(1 - \frac{s|b-a|}{3l}\Big)} \right) + \exp(s\,|b-a|)\theta(k)s, \tag{B.25}$$

where $\boldsymbol{V}_{j,m} = \sum_{i=1}^{\bar{l}_j} \boldsymbol{U}_{j+(i-1)\,k}$. The inequalities (B.24) and (B.25) hold because of Lemma B.4 and Lemma B.6, respectively. We have then proved the inequality (3.11). The same proof applies to showing the bound (3.12) by defining $\boldsymbol{U}_i = \mathbb{E}[f(\boldsymbol{Z}_i)] - f(\boldsymbol{Z}_i)$. $\qquad \square$

We remind the reader that the proof of Theorem 3.8 and 3.14 use the below Lemma that we recall for completeness.

Lemma B.7 (Legendre transform of the Kullback-Leibler divergence function)**.** *Let $(\mathcal{H}, \mathcal{T})$ be a measurable space (note that in the case of hypotheses of type $\{f_\theta : \theta \in \Theta\}$ then $\mathcal{H} = \Theta$). For any $\pi \in \mathcal{M}_+^1(\mathcal{H})$, for any measurable function $f : \mathcal{H} \to \mathbb{R}$ such that $\pi[\exp(f)] \leq \infty$, we have that*

$$\pi[\exp(f)] = \exp\Big(\sup_{\hat{\rho} \in \mathcal{M}_+^1(\mathcal{H})} \hat{\rho}[f] - KL(\hat{\rho}, \pi) \Big),$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as f is upper bounded on the support of π, the supremum with respect to $\hat{\rho}$ in the right-hand side is reached for the Gibbs distribution with Radon-Nikodym derivative w.r.t. π equal to $\frac{\exp(f)}{\pi[\exp(h)]}$.

The proof of Lemma (B.7) has been known since the work of Kullback [53] in the case of a finite space B, whereas Donsker and Varadhan have proved the general case [34]. Given this result, we are ready to prove our PAC Bayesian bounds.

Proof of Theorem 3.8. Let us choose $f(s) = \frac{s}{\epsilon}$ for $0 < s < \epsilon$, which satisfies the assumptions of Theorem 3.7 and has support in $[0,1]$. We have that $R^\epsilon(h) - r^\epsilon(h) = \frac{\epsilon}{m}(\sum_{i=1}^{m} \mathbb{E}[f(\boldsymbol{L}_i^\epsilon)] - f(\boldsymbol{L}_i^\epsilon))$. We have, in this case, that $\boldsymbol{U}_i = \mathbb{E}[f(\boldsymbol{L}_i)] - f(\boldsymbol{L}_i^\epsilon)$ such that $\mathbb{E}[\boldsymbol{U}_i] = 0$ and $|\boldsymbol{U}_i| \leq 1$. Note that the process $f(\boldsymbol{L}_i^\epsilon)_{i \in \mathbb{Z}}$ has the same θ-weak coefficients of the process $(\boldsymbol{L}_i^\epsilon)_{i \in \mathbb{Z}}$ because f is a 1-Lipschitz function. By Theorem 3.7 applied for $0 < \epsilon \sqrt{l} < 3\sqrt{l}$, and the bound (3.9),

$$\mathbb{E}\Big[\exp\Big(\sqrt{l}\,(R^\epsilon(h) - r^\epsilon(h)) \Big) \Big] = \mathbb{E}\Big[\exp\Big(\epsilon\sqrt{l}\frac{1}{m} \sum_{i=1}^{m} \boldsymbol{U}_i \Big) \Big] \leq \exp\Big(\frac{3\epsilon^2}{2(3-\epsilon)} \Big)$$

$$+ 3\sqrt{l}\exp(3\sqrt{l})\theta(k) \tag{B.26}$$

where the last equality holds because of the particular shape of the chosen function f.

We follow the scheme of proof developed by [13].

$$\sqrt{l}\,\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)] = \hat{\rho}[\sqrt{l}\,(R^\epsilon(h) - r^\epsilon(h))]$$
$$\leq KL(\hat{\rho}||\pi) + \log(\pi[\exp(\sqrt{l}\,(R^\epsilon(h) - r^\epsilon(h)))]) \quad (\mathbb{P}\text{-a.s. by Lemma B.7}). \tag{B.27}$$

We have that $\pi[\exp(\sqrt{l}\,(R^h(\beta) - r^\epsilon(h)))] := \boldsymbol{A_m}$ is a random variable on $\boldsymbol{S_m}$. By Markov's inequality, for $\delta \in (0,1)$

$$\mathbb{P}\Big(\boldsymbol{A_m} \leq \frac{\mathbb{E}[\boldsymbol{A_m}]}{\delta}\Big) \geq 1 - \delta.$$

This, in turn, implies that with probability at least $1 - \delta$ over $\boldsymbol{S_m}$

$$\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)]$$
$$\leq \frac{KL(\hat{\rho}||\pi) + \log\frac{1}{\delta}}{\sqrt{l}} + \frac{1}{\sqrt{l}}\log\Big(\pi\Big[\mathbb{E}\Big[\exp\Big(\sqrt{l}(R^\epsilon(h) - r^\epsilon(h))\Big)\Big]\Big]\Big) \tag{B.28}$$
$$\leq \frac{KL(\hat{\rho}||\pi) + \log\frac{1}{\delta}}{\sqrt{l}} + \frac{1}{\sqrt{l}}\log\Big(\pi\Big[\exp\Big(\frac{3\epsilon^2}{2(3-\epsilon)}\Big) + 3\sqrt{l}\exp(3\sqrt{l})\theta(k)\Big]\Big), \tag{B.29}$$

where (B.28) holds by swapping the expectation over $\boldsymbol{S_m}$ and over π using Fubini's Theorem, and (B.29) is obtained by using (B.26). Similarly, it can be proven that the bound (B.29) holds for $\hat{\rho}[R^\epsilon(h)] - \hat{\rho}[r^\epsilon(h)]$. We then conclude by using an union bound. □

Proof of Theorem 3.11. Let $\frac{d\bar{\rho}}{d\pi} = \frac{\exp(-\sqrt{m}r^\epsilon(h))}{\pi[\exp(-\sqrt{m}r^\epsilon(h))]}$.

$$\bar{\rho} = arg\inf_{\hat{\rho}}\Big(KL(\hat{\rho}||\pi) - \hat{\rho}[-\sqrt{m}r^\epsilon(h)]\Big) = arg\inf_{\hat{\rho}}\Big(\frac{KL(\hat{\rho}||\pi)}{\sqrt{m}} + \hat{\rho}[r^\epsilon(h)]\Big),$$

and by using one side of the bound (3.13) for $k = 1$, for all $\delta \in (0,1)$

$$\mathbb{P}\Big\{\bar{\rho}[R^\epsilon(h)] \leq \inf_{\hat{\rho}}\Big(\hat{\rho}[r^\epsilon(h)] + \Big(KL(\hat{\rho},\pi) + \log\Big(\frac{1}{\delta}\Big)\Big)\frac{1}{\sqrt{m}}\Big)$$
$$+ \frac{1}{\sqrt{m}}\log\Big(\pi\Big[\exp\Big(\frac{3\epsilon^2}{2(3-\epsilon)}\Big) + 3\sqrt{m}\exp(3\sqrt{m})\theta(1)\Big]\Big)\Big\} \geq 1 - \delta$$

We now substitute to $\hat{\rho}[r^\epsilon(\beta)]$ the other side of the bound (3.13), we obtain that

$$\mathbb{P}\Big\{\bar{\rho}[R^\epsilon(h)] \leq \inf_{\hat{\rho}}\Big(\hat{\rho}[R^\epsilon(h)] + \Big(KL(\hat{\rho},\pi) + \log\Big(\frac{1}{\delta}\Big)\Big)\frac{2}{\sqrt{m}}\Big) \tag{B.30}$$

$$+ \frac{2}{\sqrt{m}} \log\left(\pi\left[\exp\left(\frac{3\epsilon^2}{2(3-\epsilon)}\right) + 3\sqrt{m}\exp(3\sqrt{m})\theta(1)\right]\right)\right\} \geq 1 - 2\delta$$

(B.31)

by using an union bound. □

Proof of Theorem 3.12. First of all, we show that the process $(\exp(f_i(\boldsymbol{S}, h))_{i\in\mathbb{N}_0}$ is a super-martingale. Note that

$$\begin{cases} f_0(\boldsymbol{S}, h) & = 0 \\ f_m(\boldsymbol{S}, h) & = \eta \sum_{i=1}^{m}(L^\epsilon(h(X_i), Y_i) - \mathbb{E}[L^\epsilon(h(X_i), Y_i)|\mathcal{F}_{i-1}]) - \frac{\eta^2}{2}m\epsilon^2, \text{ for } m\in\mathbb{N}. \end{cases}$$

In fact, the process $\eta \sum_{i=1}^{m}(L^\epsilon(h(X_i), Y_i) - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)])$ minus the residual process is equal to a sum of martingale differences with respect to the filtration $(\mathcal{F}_m)_{m\in\mathbb{N}_0}$, namely

$$\eta \sum_{i=1}^{m}(L^\epsilon(h(X_i), Y_i) - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)|\mathcal{F}_{i-1}]).$$

We have that

$$\mathbb{E}_{m-1}[\exp(f_m(\boldsymbol{S}, h)] = \exp(f_{m-1}(\boldsymbol{S}, h))\mathbb{E}_{m-1}[\exp(\eta(L^\epsilon(h(\boldsymbol{X}_m), \boldsymbol{Y}_m)$$
$$- \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_m), \boldsymbol{Y}_m)|\mathcal{F}_{m-1}]) - \frac{\eta^2}{2}\epsilon^2)] \leq 1,$$

by applying the conditional Hoeffding's Lemma. By direct application of Theorem 4 in [23], we obtain that for all $m \geq 1$ and $\delta \in (0, 1)$, we have that

$$\hat{\rho}\left[\eta \sum_{i=1}^{m}(L^\epsilon(h(X_i), Y_i) - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)])\right] \leq KL(\hat{\rho}, \pi) + \log\left(\frac{1}{\delta}\right) + \frac{\eta^2}{2}m\epsilon^2$$
$$+ \hat{\rho}\left[\eta \sum_{i=1}^{m}\mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]\right]$$

with probability at least $1 - \delta$. Dividing both sides of the inequality by m, we obtain that for all $m \geq 1$ and $\delta \in (0, 1)$

$$\hat{\rho}[r^\epsilon(h) - R^\epsilon(h)] \leq \frac{KL(\hat{\rho}, \pi) + \log\left(\frac{1}{\delta}\right)}{\eta m} + \frac{\eta}{2}\epsilon^2$$
$$+ \hat{\rho}\left[\frac{1}{m} \sum_{i=1}^{m}\mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]\right].$$

(B.32)

with probability at least $1 - \delta$.

□

Proof of Theorem 3.14. Let us apply the scheme of proof used in [2, Theorem 1] to the residual process $\Delta_m(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)|\mathcal{F}_{i-1}] - \mathbb{E}[L^\epsilon(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)]$.

We have that

$$
\begin{aligned}
\hat{\rho}[|\Delta_m(h)|] &= \int_{\mathcal{H}} |\Delta_m(h)| \frac{d\hat{\rho}}{d\pi} d\pi \\
&\leq \Big(\int_{\mathcal{H}} |\Delta_m(h)|^2 \, d\pi \Big)^{\frac{1}{2}} \Big(\int_{\mathcal{H}} \Big(\frac{d\hat{\rho}}{d\pi} \Big)^2 d\pi \Big)^{\frac{1}{2}} \text{(Cauchy-Schwarz's Inequality)} \\
&\leq \Big(\epsilon \int_{\mathcal{H}} |\Delta_m(h)| \, d\pi \Big)^{\frac{1}{2}} \Big(D_{\phi_2-1}(\hat{\rho}, \pi) + 1 \Big)^{\frac{1}{2}}.
\end{aligned}
$$

We then apply the Markov's inequality to $\int_{\mathcal{H}} |\Delta_m(h)| \, d\pi$, and for $\delta \in (0,1)$ obtain that with probability at least $1 - \delta$

$$
\begin{aligned}
\hat{\rho}[|\Delta_m(h)|] &\leq \Big(\epsilon \frac{\mathbb{E}[\int_{\mathcal{H}} |\Delta_m(h)| \, d\pi]}{\delta} \Big)^{\frac{1}{2}} \Big(D_{\phi_2-1}(\hat{\rho}, \pi) + 1 \Big)^{\frac{1}{2}} \\
&\leq \Big(\epsilon \frac{\pi[\mathbb{E}[|\Delta_m(h)|]]}{\delta} \Big)^{\frac{1}{2}} \Big(D_{\phi_2-1}(\hat{\rho}, \pi) + 1 \Big)^{\frac{1}{2}} \quad \text{(Fubini's Theorem)} \\
&\leq \Big(\epsilon \frac{\pi[\theta(1)]}{\delta} \Big)^{\frac{1}{2}} \Big(D_{\phi_2-1}(\hat{\rho}, \pi) + 1 \Big)^{\frac{1}{2}} \quad \text{(Definition 2.12)}
\end{aligned}
$$

The last inequality holds because $\boldsymbol{L}$ is a θ-weakly dependent process as proven in Proposition 3.6, and the truncation of this process through the accuracy level ϵ means applying a projection function to the process $\boldsymbol{L}$ which has $Lip(h) = 1$, see Erratum in the arXiv version of [27] for a detailed explanation. We can then apply the projective-type representation of the θ-coefficients in Remark 2.15. We then use an union bound to combine the result above with the any-time bound in Corollary 3.13. □

Acknowledgments

Funding

References

[1] Alquier, P. (2024). User-friendly introduction to PAC-Bayes bounds. *Found. Trends Mach. Learn.*, **17** 17–303. https://ieeexplore.ieee.org/document/10413338

[2] Alquier, P. and Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Mach. Learn.*, **107** 887–902. MR3780170

[3] Alquier, P. and Li, X and Wintenberger, O. (2013). Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 65–93. doi: 10.2478/demo-2013-0004

[4] Alquier, P. and Wintenberger, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli*, **18** 883–913. MR2948906

[5] Amato, F. and Guignard, F. and Robert, S. and Kanevski, M. (2020). A novel framework for spatio-temporal prediction of environmental data using deep learning. *Sci. Rep.*, **10**. doi: 10.1038/s41598-020-79148-7

[6] Bai, Y. and Song, P. X.-K. and Raghunathan, T. E. (2012). Joint composite estimating functions in spatiotemporal models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **74** 799–824. MR2988907

[7] Barndorff-Nielsen, O. E. (2001). Superposition of Ornstein-Uhlenbeck type processes. *Theory Probab. Appl.*, **45** 175–194. MR1967758

[8] Barndorff-Nielsen, O. E. and Benth, F. E. and Veraart, A. E. D. (2018). *Ambit Stochastics*. Springer, Cham. MR3839270

[9] Barndorff-Nielsen, O. E. and Leonenko, N. N. (2005). Spectral properties of superpositions of Ornstein-Uhlenbeck type processes. *Methodol. Comput. Appl. Probab.*, **7** 335–352. MR2210585

[10] Barndorff-Nielsen, O. E. and Lunde, A. and Shepard, N. and Veraart, A. E. D. (2014). Integer-valued trawl processes: a class of stationary infinitely divisible processes. *Scand. J. Stat.*, **41** 693–724. MR3249424

[11] Barndorff-Nielsen, O. E. and Schmiegel, J. (2004). Spatio-temporal modeling based based on Lévy processes, and its applications to turbulence. *Russian Math. Surveys*, **59** 65–90. MR2068843

[12] Barndorff-Nielsen, O. E. and Stelzer, R. (2011). Multivariate supOU processes. *Ann. Appl. Probab.*, **21** 140–182. MR2759198

[13] Bégin, L. and Germain, P. and Laviolette, F. and Roy, J.-F. PAC-Bayesian bounds based on the Rényi divergence. In *Proceedings of the* 19*th annual International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR, 2018.

[14] Behrmann, J. and Grathwohl, W. and Chen, R. T. Q. and Duvenaud, D. and Jacobsen, J.-H. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

[15] Bennedsen, M. and Lunde, A. and Shephard, N. and Veraart, A. E. D. (2023) Inference and forecasting for continuous-time integer-valued trawl processes. *J. Econometrics*, **236**. MR4614574

[16] Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, Inc., New York. MR1324786

[17] BLANCHARD, G. and CARPENTIER, A. and ZADOROZHNYI, O. (2024). Moment inequalities for sums of weakly dependent random fields. *Bernoulli*, **30** 2501–2520. MR4746617

[18] BOUALLÈGUE, Z. B. and WEYN, J. A. and CLARE, M. C. A. and DRAM-SCH, J. and DUEBEN, P. and CHANTRY, M. (2024). Improving Medium-Range Ensemble Weather Forecasts with Hierarchical Ensemble Transform-ers. *Artif. Intell. Earth Syst.*, **3** e230027. doi: 10.1175/AIES-D-23-0027.1.

[19] BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions, Volume 1.* Kendrick Press, Heber City, UT. MR2325294

[20] BUNEA, F. and TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggre-gation for Gaussian regression. *Ann. Statist.*, **35** 1674–1697. MR2351101

[21] CATONI, O. (2004). *Statistical learning theory and stochastic optimization. Lecture Notes in Math., 1851.* Springer-Verlag, Berlin. MR2163920

[22] CATONI, O. (2007). *PAC-Bayesian supervised classification: the thermody-namics of statistical learning. IMS Lecture Notes Monogr. Ser., 56.* Insti-tute of Mathematical Statistics, Beachwood, OH. MR2483528

[23] CHUGG, B. and WANG, H. and RAMDAS, A. (2023). A unified recipe for deriving (time-uniform) PAC-Bayes bounds. *J. Mach. Learn. Res.*, **24** 1–61. MR4690321

[24] COURGEAU, V. and VERAART, A. E. D. (2022). High-frequency estima-tion of the Lévy-driven graph Ornstein-Uhlenbeck process. *Electron. J. Stat.*, **16** 4863–4925. MR4489241

[25] CRESSIE, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons, Inc., New York. MR1239641

[26] CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-Temporal data.* John Wiley & Sons, Inc., Hoboken, NJ. MR2848400

[27] CURATO, I. V. and STELZER, R. (2019). Weak dependence and GMM estimation for supOU and mixed moving average processes. *Electron. J. Stat.*, **13** 310–360. MR3910486

[28] CURATO, I. V. and STELZER, R. and STRÖH. B. (2022). Central limit theorems for stationary random fields under weak dependence with appli-cation to ambit and mixed moving average fields. *Ann. Appl. Probab.*, **32** 1814–1861. MR4430002

[29] CURATO, I. V. and STERNKOPF, J. (2025). Mixed moving average field guided learning with unbounded losses. *Work in Progress.*

[30] DAVIS, R. A. and KLÜPPELBERG, C. and STEINKOHL, C. (2013). Statis-tical inference for max-stable processes in space and time. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **75** 791–819. MR3124792

[31] DEDECKER, J. (1998). A central limit theorem for stationary random fields. *Probab. Theory Related Fields*, **110** 397–426. MR1616496

[32] DEDECKER, J. and DOUKHAN, P. (2003). A new covariance inequality and applications. *Stochastic Process. Appl.*, **106**, 63–80. MR1983043

[33] DEDECKER, J. and DOUKHAN, P. and LANG, G. and RAFAEL, L. R. J and LOUHICHI, S. and PRIEUR, C. (2007). *Weak Dependence: With Examples and Applications. Lect. Notes Stat., 190* Springer, New York. MR2338725

[34] DONSKER, M. D. and VARADHAN, S. R. S. (1976). Asymptotic evaluation

of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math.*, **29** 389–461. MR0428471

[35] DOUKHAN, P. (1994). *Mixing. Properties and Examples. Lect. Notes Stat., 85.* Springer-Verlag, New York. MR1312160

[36] DOUKHAN, P. and NEUMANN, M. H. (2007). Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Process. Appl.*, **117** 878–903. MR2330724

[37] DZIUGAITE, G. K. and ROY, M. D. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

[38] FAZLYAB, M. and ROBEY, A. and HASSANI, H. and MORARI, M. and G. J. PAPPAS, G. J. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[39] GLASBEY, C. A. and ALLCROFT, D. J. (2008). A spatiotemporal auto-regressive moving average model for solar radiation. *J. R. Stat. Soc. Ser. C*, **57** 343–355. MR2440013

[40] GONZÁLEZ, J. A. and RODRÍGUEZ-CORTÉS, F. J. and CRONIE, O. and MATEU, J. (2016). Spatio-temporal point process statistics: a review. *Spat. Stat.*, **18** 505–544. MR3575505

[41] GRÜNWALD P. D. and MEHTA, N. A. (2020). Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *J. Mach. Learn. Res.*, **21** 1–80. MR4095335

[42] GUEDJ, B. (2019). A primer on PAC-Bayesian learning. *SMF 2018: Congrès de la Société Mathématique de France*, 391–413. MR4508923

[43] HADDOUCHE, M. and GUEDJ, B. (2022). PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales. *Trans. Mach. Learn. Res.*. https://openreview.net/forum?id=qxrwt6F3sf

[44] HANG, H. and STEINWART, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.*, **45** 708–743. MR3650398

[45] HIGDON, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, 37–56. MR2059819

[46] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58** 13–30. MR0144363

[47] HOLLAND, M. PAC-Bayes under potentially heavy tails. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2715–2724, 2019.

[48] HOLMES, P. and LUMLEY, J. L. and BERKOOZ, G. and ROWLEY, C. W. (2012). *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge University Press, Cambridge, UK. MR2920624

[49] IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Teor. Verojatnost. i Primenen.*, **7** 361–392. MR0148125

[50] JÓNSDÓTTIR, K. Ý. and RØNN-NIELSEN, A. and MOURIDSEN, K. and

Vedel Jensen, E. B. (2013). Lévy-based modelling in brain imaging. *Scand. J. Stat.*, **40** 511–529. MR3091695

[51] Karpatne, A. and Atluri, G. and Faghmous, J. H. and Steinbach, M. and Banerjee, A. and Ganguly, A. and Shekhar, S and Samatova, N. and Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.*, **29** 2318–2331. doi: 10.1109/TKDE.2017.2720168

[52] Kim, H. and Papamakarios, G. and Mnih, A. (2021). The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, **139** 5562–5571. PMLR, 2021.

[53] Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London. MR0103557

[54] Lahiri, S. N. and Lee, Y. and Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *J. Stat. Plan. Inference.*, **103** 62–85. MR1896984

[55] Liu, B. and Wang, M. and Li, Y. and Chen, H. and Li, J. (2021). Deep learning for spatio-temporal sequence forecasting: a survey. *J. Beijing Univ.Tech.*, **47** 925–941. doi: 10.11936/bjutxb2020120037

[56] Modha, D. S. and Masry, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory*, **42** 2133–2145. MR1447519

[57] Montañez, G. D. and Shalizi, C. R. The LICORS cabinet: nonparametric light cone methods for spatio-temporal modeling. In *2017 International Joint Conference on Neural Networks*, pages 2811–2819, 2017.

[58] Montero, J. M. and Fernández-Avilés, G. and Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. John Wiley & Sons, Ltd. doi: 10.1002/9781118762387

[59] Nguyen, M. and Veraart, A. E. D. (2017). Spatio-temporal Ornstein–Uhlenbeck processes: theory, simulation and statistical inference. *Scand. J. Stat.*, **44** 46–80. MR3619694

[60] Nguyen, M. and Veraart, A. E. D. (2018). Bridging between short-range and long-range dependence with mixed spatio-temporal Ornstein-Uhlenbeck processes. *Stochastics*, **90** 1023–1052. MR3854526

[61] Pearl, J. and Glymour, M. and Jewell, N. P. (2016). *Causal inference in statistics a primer*. John Wiley & Sons, Ltd., Chichester. MR3497861

[62] Pebesma, E. J. and Bivand, R. S. (2023). *Spatial Data Science: with applications in R*. Chapman and Hall/CRC. doi: 10.1201/9780429459016

[63] Qi, X. and Wang, J. and Chen, Y. and Shi, Y. and Zhang, L. (2023). LipsFormer: Introducing lipschitz continuity to vision transformers. https://openreview.net/forum?id=cHf1DcCwcH3

[64] Raissi, M. and Perdikaris, P. and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, **378** 686–707. MR3881695

[65] Rajput, B. S. and Rosinski, J. (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Related Fields*, **82** 451–487.

MR1001524

[66] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2005). *Gaussian processes for Machine Learning*. The MIT Press Cambridge, MA. MR2514435

[67] REGAZZONI, F. and PAGANI, S, and SALVADOR, M. DEDE', L. and QUAR-TERONI, A. (2024). Learning the intrinsic dynamics of spatio-temporal processes through Latent Dynamics Networks. *Nat. Commun.*, **15**. doi: 10.1038/s41467-024-45323-x

[68] REICHSTEIN, M. and CAMPS-VALLS, G. and STEVENS, B. and JUNG, M. and DENZLER, J. and CARVALHAIS, N. and PRABHAT (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, **556** 195–204. doi: 10.1038/s41586-019-0912-1

[69] RIO., E. (1996.) Sur le théorème de Berry–Esseen pour les suites faiblement dépendantes. *Probab. Theory Related Fields*, **104** 255–282. MR1373378

[70] ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. U.S.A.*, **42** 43–47. MR0074711

[71] RUBIN, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *J. Amer. Statist. Assoc.*, **100** 322–331. MR2166071

[72] RUPE, A. and KUMAR, N. and EPIFANOV, V. and KASHINATH, K. and PAVLZK, O. and SCHIMBACH, F. and PATWARY, M. MAIDANOV, S. and LEE, V. and PRABHAT, and CRUTCHFIELD, J. P. DisCo: physics-based unsupervised discovery of coherent structures in spatiotemporal systems. In *IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments*, pages 75–87, 2019.

[73] SATO, K.-I. (2013). *Lévy processes and infinitely divisible distributions.* Cambridge University Press, Cambridge, 2013. MR3185174

[74] SCAMAN, K. and VIRMAUX, A. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Syste*, pages 3839–3848, 2019.

[75] SHAWE-TAYLOR, J. and BARTLETT, P. L. and WILLIAMSON, R. C. and ANTHONY, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Trans. Inform. Theory*, **44** 1926–1940. MR1664055

[76] SHI, X. and GAO, Z. and LAUSEN, L. and WANG, H. and YEUNG, D.-Y. and WANG, W.-K. and WOO, W.-C. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems 30*, pages 5617–5627, 2017.

[77] SHIRYAEV, A. N. (2016). *Probability 1*, 3rd Edition. Springer, New York. MR3467826

[78] STEIN, M. L. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.*, **100** 310–321. MR2156840

[79] STELZER, R. and TOSSTDORF, T. and WITTLINGER, M. (2015). Moment based estimation of supOU processes and a related stochastic volatility model. *Stat. Risk Model.*, **32** 1–24. MR3331779

[80] VALIANT, L. G. (1984). A theory of the learnable. *Commun. ACM*, **27** 1134–1142. doi: 10.1145/1968.1972

[81] VAPNIK, V. N. (2000). *The nature of Statistical Learning Theory.* Springer-

Verlag, New York. MR1719582

[82] VILLE, J. (1939). Étude critique de la notion de collectif. *The Mathematical Gazette*, **23** 490–491. MR3533075