2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

First Hitting Time Guarantees for Contractive Nonlinear Systems

Julien Walden Huang, Stephen Roberts and Jan-Peter Calliess

***Abstract*— We derive tight probabilistic bounds on the first hitting time of general classes of nonlinear autoregressive systems that can be linked to mean reverting stochastic processes. The obtained results are formulated such that they can be readily applied to models identified by machine learning techniques such as deep learning. As an application to finance, we show how our results can be utilised to inform statistical arbitrage trading strategies for which we provide probabilistic performance guarantees.**

I. Introduction

Over the past decade there has been a proliferation of machine learning based frameworks that have been researched and applied with the goal of enhancing time series modelling and system identification methods. These approaches facilitate a flexible function estimation that aims to capture non-linear relations in the underlying dynamics of the time series. Formally, one assumes that the time series data is obtained through the realisations of a stochastic data generating process (y_t) defined recursively by the equations:

$$y_{t+1} := \begin{cases} \psi(y_t, ..., y_{t-(d-2)}, y_{t-(d-1)}) + \epsilon_{t+1} \text{ for } t \geq 0 \\ a_{-t}, \text{ for } t \in \{-1, ..., -d\}. \end{cases} \tag{1}$$

where $\psi : \mathbb{R}^d \to \mathbb{R}$ is the transition function, $a \in \mathbb{R}^d$ represents the initial conditions and (ϵ_t) is a zero-mean stochastic process. The data set of past observations of (y_t) is utilised to construct an estimate $\hat{\psi} : \mathbb{R}^d \to \mathbb{R}$ of the dynamics ψ of (y_t). The flexibility of machine learning based system identification frameworks means that $\hat{\psi}$ can belong to a richer class of functions than classical autoregressive models such as AR, TAR or EXPAR models [1] which assume a more restrictive functional form. This modelling advantage has been shown to improve forecasting capabilities and prediction accuracy on benchmarks, case studies and empirical applications in a wide range of fields (e.g. [2], [3] or [4]). In spite of the their practical success, the theoretical understanding of the data generating process where ψ is identified by a general class of ML methods still pales in comparison to the wealth of theory available for classical autoregressive models.

In an effort to improve upon this state of affairs, we focus on developing widely applicable theoretical tools related to first hitting time guarantees and mean reversion of nonlinear autoregressive models compatible with representation class (such as neural networks) common in machine learning. These properties are of particular importance in a great many application domains, including in finance, control and ecology. In time series analysis, first hitting times and contractive dynamical systems have been extensively studied in a diverse range of contexts. For discrete time series, usual approaches involve either fitting an autoregressive AR(p) model or assuming underlying dynamics that are linear and stationary. The first hitting time probabilities and expectation are then computed numerically [5], [6] or can be lower bounded analytically in the case of the AR(1) model [7]. This approach has been explored in various domains: in statistical arbitrage and quantitative finance for optimal thresholds setting [8], [9], for predicting population extinction and time to extinction in ecology [10], signal detection and surveillance analysis [11] or structural health monitoring [12], [13]. For continuous time series, dynamics are usually assumed to follow the Ornstein-Uhlenbeck (OU) dynamics in which case the first hitting time probabilities can be obtained semi-analytically [14], [15] (and references therein) under some additional assumptions. Applications are numerous and involve, for example, hydrology [16], neuroscience [17] or quantitative finance [18], [19]. Note that, even though in aforementioned works specific forms of dynamic models were presupposed, the computation of the first hitting time probabilities had to rely on numerical approximation. Simplifying this computation is difficult even for simple dynamics and remains an open question for both Ornstein-Uhlenbeck models and AR(p) models.

What unifies these threads of works is that they provide an understanding of hitting times for time series whose dynamics conform to a specific (linear) structure. However, when those functions are identified by black-box machine learning algorithms, existing results are not applicable. Therefore, what is needed are theoretical bounds which can be computed for general classes of system dynamics that contain the ones arising in the context of machine learning-based black-box system identification.

In this work, we provide such bounds. In particular, we derive (contractive) Lipschitz conditions on the transition function sufficient to calculate our probabilistic hitting time bounds. As we explain, the conditions can be readily calculated for some of the most popular machine learning models. Our hitting time bounds are shown to be tight. While they involve a non-analytic definite integral, this can be computed numerically offline and its solutions could be stored in a look-up table. Moreover, we show how our results can be directly applied to inform trading decisions. Our hitting time

This work is supported by funding provided by the Oxford-Man Institute of Quantitative Finance.

Julien Walden Huang, Stephen Roberts Jan-Peter Calliess are with the Department of Engineering, University of Oxford and the Oxford-Man Institute of Quantitative Finance, Oxford, United Kingdom {julien.huang}@sjc.ox.ac.uk

bounds are shown to translate to probabilistic bounds on the returns of the ensuing trading strategy, provided the time series of the mean reverting synthetic asset satisfies the required contractive Lipschitz conditions.

II. Model assumptions

Let d $\in \mathbb{N}$, $\psi : \mathbb{R}^d \to \mathbb{R}$ and $a \in \mathbb{R}^d$. We assume that our time series is modelled by the stochastic nonlinear autoregressive (NAR) process $(y_t)_{t\in\mathbb{N}}$ starting at time 0 given by the dynamical system defined in equation (1) with transition function ψ and initial conditions $a \in \mathbb{R}^d$. The noise process $(\epsilon_t)_{t\in\mathbb{N}}$ is a zero mean stochastic process that is assumed to satisfy :

Assumption 1. *$(\epsilon_t)_{t\in\mathbb{N}}$ are independent and identically distributed random variables with bounded non-zero variance and a probability density function denoted by f_ϵ.*

To establish our bounds in subsequent sections, we need to make assumptions on the transition function ψ. We recall that for functions $f : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$, Lipschitz continuity is defined as:

Definition 1 (Lipschitz continuity)**.** *For a domain $\mathcal{D} \subseteq \mathbb{R}^d$ constant, norm $\|\cdot\|$ and $\bar{L} \in \mathbb{R}_+$ we define the space of $\bar{L}$-*Lipschitz continuous functions *as*

$$\mathcal{L}_{\bar{L}}(\mathcal{D}, \|\cdot\|) := \left\{ f : \mathcal{D} \to \mathbb{R} | \forall x, x' \in \mathcal{D} : \left|f(x) - f(x')\right| \leq \bar{L} \left\|x - x'\right\| \right\}$$

where $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^d$. Constant $\bar{L}$ is called a Lipschitz constant *of any $f \in \mathcal{L}_{\bar{L}}(D, \|\cdot\|)$. Furthermore, the smallest $L^* > 0$ such that f is L^*- Lipschitz continuous is called the* best *Lipschitz constant of f.*

Then, a sufficient condition for our results is to assume ψ to be a contraction relative to the α^*-norm denoted by $\|\cdot\|_{\alpha^*}$. This norm is defined as: for $\alpha^* \in \mathbb{R}^d_{>0}$, $\|\cdot\|_{\alpha^*} : \mathbb{R}^d \to \mathbb{R}$ is the weighted l_1-norm

$$\forall x \in \mathbb{R}^d, \quad \|x\|_{\alpha^*} = \sum_{i=1}^{d} \alpha_i^* |x_i|.$$

The following definition states formally the Lipschitz-type smoothness condition that will be utilised to obtain the theoretical first hitting time guarantees for the stochastic process $(y_t)_{t\in\mathbb{N}}$.

Definition 2 (α^*-contracting process)**.** *Let $\mathcal{D} \subseteq \mathbb{R}^d$. An autoregressive process is called an α^*-*contracting process *on $\mathcal{D}$ if its transition function ψ is contained in $\mathcal{L}_1(\mathcal{D}, \|\cdot\|_{\alpha^*})$ and $\alpha^* \in \triangle_+ := \{x \in \mathbb{R}^d_{>0} | \sum_{i=1}^d x_i < 1\}$.*

Assumption 2. *Our time series $(y_t)_{t\in\mathbb{N}}$ is an α^* contracting process , i.e.*

$$\psi \in \mathcal{L}^{\alpha^*}(\mathcal{D}) := \mathcal{L}_1(\mathcal{D}, \|\cdot\|_{\alpha^*})$$

for some $\alpha^* \in \triangle_+$ and $\mathcal{D} = \mathbb{R}^d$.

One may wonder how this α^* contracting condition relates to a simpler Lipschitz assumption on ψ. Let $\mathcal{D} \subseteq \mathbb{R}^d$, $\alpha^* \in \triangle_+$ and $\bar{\alpha} := \sum_{i=1}^d \alpha_i$. We have the following relationship between Lipschitz function spaces: (1) $\mathcal{L}^{\alpha^*}(\mathcal{D}) \subseteq \mathcal{L}_{\bar{\alpha}}(\mathcal{D}, \|\cdot\|_\infty)$ and (2) for $\delta \in (0, 1)$, define $\alpha^* = (\frac{\delta}{d}, ..., \frac{\delta}{d})^\top$, then $\mathcal{L}_{\frac{\delta}{d}}(\mathcal{D}, \|\cdot\|_1) \subseteq \mathcal{L}^{\alpha^*}(\mathcal{D})$.
Therefore, although the α^* condition is notationally heavy, it is useful as it provides a weaker assumption than the alternative L_1 Lipschitz condition. Perhaps more importantly, the α^* condition provides additional flexibility that allows for the dependence on previous time lags to be greater than $\frac{1}{d}$ as long as the sum of the α^* coefficients is smaller than unity. This feature is useful in practice where models generally depend on recent time lags more. Finally, if the α^* coefficients are obtained by using a machine learning estimation of ψ then the input dimension of the estimation model can be greater than d with no explicit consequences on the α^* condition.

Notation 1 (Relevant matrices)**.** *For any $\alpha^* \in \triangle_+ := \{x \in \mathbb{R}^d_{\geq 0} | \sum_{i=1}^d x_i < 1\}$ and $T \in \mathbb{N}$, we define the associated matrices $A(T) \in \mathbb{R}^{T\times T}$ and $B \in \mathbb{R}^{d\times d}$. Here, $A(T)$ is a lower triangular banded matrix and B is a sparse matrix whose entries are given by:*

$$A(T)_{ij} := \begin{cases} 1, & \text{if } i-j = 0 \\ -\alpha^*_{(i-j)}, & \text{if } 0 < i-j \leq d \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

$$B_{ij} := \begin{cases} 1, & \text{if } i-j = 1 \\ \alpha^*_j, & \text{if } i = 1 \text{ and } 1 \leq j \leq d \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

for all $i, j \in \{1, ..., T\}$.

III. First hitting time guarantees

We will now state bounds on first hitting times of the time series generated by our dynamical system. We assume all definitions and assumptions introduced in Sec. II hold. Appealing to Banach's fixed point theorem one can show the existence of a unique fixed point for ψ: $y^* = \psi(y^*, \ldots, y^*)$. As we will see, the contractive properties of the time series result in a generalisation of mean-reverting behavior where the fixed point serves as the level to which the time series will tend to revert to in the long run after being exposed to a shock. More formally, we define the following.

Definition 3 (First hitting time)**.** *For $a \in \mathbb{R}^d$ with $a_1 > y^*$ and $\gamma \in [0, a_1 - y^*)$, we define the upper first hitting time of $(y_t)_{t\in\mathbb{N}}$:*

$$\tau^+_\gamma := \inf\{t \in \mathbb{N} | y_t - y^* < \gamma\} \ .$$

Similarly, for $a_1 < y^*$ and $\gamma \in [a_1 - y^*, 0)$, we define the lower first hitting time of $(y_t)_{t\in\mathbb{N}}$:

$$\tau^-_\gamma := \inf\{t \in \mathbb{N} | y_t - y^* > \gamma\} \ .$$

Initial value a_1 can be seen as having resulted from a "shock" in the time series and γ as a return barrier that indicates proximity to the long-run "mean" y^*. The first hitting times τ^+_γ and τ^-_γ are linked to the speed of mean reversion measured at various levels (γ). By conditioning on past hitting times and the last result of [20], one can show our following principal result:

Notation 2. *To alleviate notation, we denote the projection operator onto the $i-$th component by: $\pi_i : \mathbb{R}^d \to \mathbb{R}, (x_1, \dots, x_d)^\top \mapsto x_i$ for $i \in \{1, ..., d\}$. Furthermore, we denote the d-dimensional ones-vector: $\mathbf{1}_d = (1, ..., 1)^\top \in \mathbb{R}^d$.*

Theorem 1. *For $T \in \mathbb{N}$, define*

$$\mathfrak{I}^+_{(\alpha^*, y^*)}(T) := \int_{-b_1}^{\infty} ... \int_{-b_T}^{\infty} f_{\epsilon_{1:T}}(A(T)x)dx$$

where $A(T)$ is defined in (2), $f_{\epsilon_{1:T}}$ is the joint probability density function of any finite sequence of consecutive noise variables: $\epsilon_{1:T} := (\epsilon_1, ..., \epsilon_T)$ defined according to Assumption 1 and $b_i := \pi_1\big(B^i(a - y^*\mathbf{1}_d)\big) - \gamma$ for $i = 1, \dots, T$ where B is defined in (3). We have:

$$(i)\ \mathbb{P}\big(\tau^+_\gamma > T\big) \leq \mathfrak{I}^+_{(\alpha^*, y^*)}(T) < 1 \text{ and}$$

$$(ii)\ \mathbb{E}[\tau^+_\gamma] \leq 1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*, y^*)}(T).$$

Analogous bounds can be derived for $\mathbb{P}\big(\tau^-_\gamma > T\big)$ and $\mathbb{E}[\tau^-_\gamma]$.

Proof. Assumption 2 entails we have that the first hitting time τ^+_γ can be upper bounded with probability 1 by the first hitting time $\tau^z_\gamma := \inf\{t \in \mathbb{N} | z_t < \gamma\}$ of a linear AR(d) $(z_t)_{t\in\mathbb{N}}$ process with coefficients equal to the α^* vector, initial conditions $(a_1 - y^*, ..., a_d - y^*) \in \mathbb{R}^d$ and same noise process $(\epsilon_t)_{t\in\mathbb{N}}$ as $(y_t)_{t\in\mathbb{N}}$. Then, for an arbitrary $T \in \mathbb{N}$:

$$\mathbb{P}\big(\tau^z_\gamma > T\big) = \mathbb{P}\big(\min_{t\in\{1,...,T\}} z_t > \gamma\big) = \mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t > \gamma\}\big).$$

For every time step $t \in \{1, ..., T\}$, by iterating backwards from timestep t, we can re-express z_t as

$$z_t = \sum_{i=1}^{t} \beta(\alpha^*, t, i)\epsilon_i + \pi_1(B^t(a - y^*\mathbf{1}_d)). \tag{4}$$

where $\beta(\alpha^*, t, i)$ is a constant that depends on α^*, t and i. Define $\sigma^2 := var(\epsilon_1)$ which is finite and non-zero by Assumption 1. The independence and identical distributions of the noise variables imply $\forall s \leq t \in \mathbb{N}$,

$$\frac{cov(z_s, z_t)}{\sigma^2} = \sum_{i=1}^{s} \beta(\alpha^*, s, i)\beta(\alpha^*, t, i) = \langle M(T)_s, M(T)_t\rangle$$

where $M(T)_i$ denotes the i-th row of of a matrix $M \in \mathbb{R}^{T\times T}$. Therefore, the covariance matrix $V_T \in \mathbb{R}^{T\times T}$ of $(z_t)_{t\in\{1,...,T\}}$ is given by $V_T = \sigma^2 M(T)M(T)^\top$. From (35) in [20], we have that the sample covariance of $(z_t)_{t\in\{1,...,T\}}$ is known and given by $V_T^{-1} = \frac{1}{\sigma^2} A_{\alpha^*}(T)A_{\alpha^*}(T)^\top$ with $A_{\alpha^*}(T)$ is defined in (2). By uniqueness of the square root of a matrix and of the inverse matrix we obtain an explicit expression for $M(T)$: $M(T)^{-1} = A_{\alpha^*}(T)^\top$.

Using the above relation, equation (4) and $\det(A_{\alpha^*}(T)) = 1$, we obtain the bound:

$$\mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t > \gamma\}\big) \leq \mathbb{P}\big(M(T)\epsilon_{1:T} > -b\big)$$

$$= \int_{-b_1}^{\infty} ... \int_{-b_T}^{\infty} \frac{f_{\epsilon_{1:T}}(A_{\alpha^*}(T)\begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix})}{|\det(A_{\alpha^*}(T)^{-1})|} dx_1 ... dx_T$$
$$= \mathfrak{I}^+_{(\alpha^*, y^*)}(T).$$

The second statement of 1 follows almost immediately. For $N \in \mathbb{N}$, define E_N as the partial sum $E_N := \sum_{T=1}^{N} T\mathbb{P}\big(\tau^+_\gamma = T\big)$ then

$$E_N = \sum_{T=1}^{N} T\mathbb{P}\big(\tau^+_\gamma = T\big) = \sum_{T=1}^{N}\sum_{t=1}^{T} \mathbb{P}\big(\tau^+_\gamma = T\big)$$
$$= \sum_{t=1}^{N}\sum_{T=t}^{N} \mathbb{P}\big(\tau^+_\gamma = T\big) = \sum_{T=1}^{N} \mathbb{P}\big(\tau^+_\gamma \geq T\big)$$
$$= 1 + \sum_{T=1}^{N} \mathbb{P}\big(\tau^+_\gamma > T\big) \leq 1 + \sum_{T=1}^{N} \mathfrak{I}^+_{(\alpha^*, y^*)}(T).$$

Taking limits on both sides of the inequality yields

$$\mathbb{E}[\tau^+_\gamma] = \sum_{T=1}^{\infty} T\mathbb{P}\big(\tau^+_\gamma = T\big) \leq 1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*, y^*)}(T).$$

□

Theorem 1 provides a lower bound on the cumulative density function of the first hitting times of $(y_t)_{t\in\mathbb{N}}$ as it returns to the fixed point of its autoregressive model. By varying the choice of barrier γ, the bound given in Theorem 1 can be used as a theoretical guarantee on the speed of mean reversion of $(y_t)_{t\in\mathbb{N}}$.

Remark 1. *Comments on the behaviour of $\mathfrak{I}^+_{(\alpha^*, y^*)}(T)$:*

1) *$\forall T \in \mathbb{N}$, $\mathfrak{I}^+_{(\alpha^*, y^*)}(T)$ is decreasing in γ.*
2) *$\forall T \in \mathbb{N}_{>d}$, if $\forall\, i$, $\alpha^*_i \leq \beta^*_i$ and $\exists\, j$ s.t. $\alpha^*_j < \beta^*_j$ then $\mathfrak{I}^+_{(\alpha^*, y^*)}(T) < \mathfrak{I}^+_{(\beta^*, y^*)}(T)$.*
3) *$\forall T \in \mathbb{N} : \lim_{\|\alpha^*\|_1 \to 0} \mathfrak{I}^+_{(\alpha^*, y^*)}(T) = \frac{1}{2^T}$.*

Proof. In this proof, we modify previous notation to emphasize dependence on the parameters studied in Remark 1 and we define the following notation: for every $T \in \mathbb{N}$, we denote by $J_T \subseteq \mathbb{R}^T$ the set $J_T := \prod_{i=1}^{T}[-b_i, \infty)$ where $b_i := \big(B^i_{\alpha^*}|a - y^*\mathbf{1}_d|)\big)_1 - \gamma$. (i): Consider $\gamma_1, \gamma_2 \in [0, a_1 - y^*)$ with $\gamma_1 \leq \gamma_2$. We have $\forall i = 1, ..., T$, $b_i(\gamma_1) \geq b_i(\gamma_2)$ which implies that $J_T(\gamma_1) \supseteq J_T(\gamma_2)$ and subsequently

$$\mathfrak{I}^+_{(\alpha^*, y^*)}(T, \gamma_1) = \int_{J_T(\gamma_1)} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx$$
$$\geq \int_{J_T(\gamma_2)} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx = \mathfrak{I}^+_{(\alpha^*, y^*)}(T, \gamma_2).$$

(ii): Consider $\alpha^*, \beta^* \in \triangle_+$ with $\forall\, i$, $\alpha^*_i \leq \beta^*_i$ and such that $\exists\, j$ with $\alpha^*_j < \beta^*_j$. It follows that $\forall t \in \mathbb{N}_{>d}$, $z_t(\alpha^*) < z_t(\beta^*))$ (where $z_t(\alpha^*)$ ($z(\beta^*)$ denotes a linear AR(d) process defined with coefficients equal to α^* (β^*)) and noise process $(\epsilon_t)_{t\in\mathbb{N}}$. Then, $\forall T \in \mathbb{N}_{>d}$,

$$\mathfrak{I}^+_{(\alpha^*, y^*)}(T) = \mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t(\alpha^*) > \gamma\}\big) \leq \mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t(\beta^*) > \gamma\}\big)$$
$$= \mathfrak{I}^+_{(\beta^*, y^*)}(T).$$

(iii): Let $J_T^{(-1)} := \{x \in \mathbb{R}^T | A_{\alpha^*}(T)^{-1}x \in J_T\}$ where

$A_{\alpha^*}(T)^{-1}$ is well-defined as $A_{\alpha^*}(T)$ is a lower triangular matrix with ones on the diagonal. Then $\mathfrak{I}^+_{(\alpha^*,y^*)}(T) = \mathbb{P}\big(M(T)\epsilon(T) > -\mathbf{b}\big) = \int_{J_T} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx$

$= \int_{J_T^{(-1)}} f_{\epsilon_{1:T}}(x)dx = \int_{\mathbb{R}^T} f_{\epsilon_{1:T}}(x)\mathbf{1}_{J_T^{(-1)}}(x)dx.$

where $\mathbf{1}_A(x)$ denotes the indicator of a subset A. Define $g(x,\alpha^*) = f_{\epsilon_{1:T}}(x)\mathbf{1}_{J_T^{(-1)}}$. Since g verifies all the conditions of Theorem 5.6 of [21] for $\alpha_0^* := 0$, we have $\int_{\mathbb{R}^T} g(x,\alpha^*)dx$ is continuous in α_0^*. This implies that $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ is continuous in $\alpha^* = 0$ and therefore

$$\lim_{\|\alpha^*\|_1 \to 0} \mathfrak{I}^+_{(\alpha^*,y^*)}(T) = \mathfrak{I}^+_{(0,y^*)}(T) = \frac{1}{2^T}$$

. □

The multi-dimensional integral expression stated in $\mathfrak{I}^+$ corresponds to the computation of the orthant probabilities of a T-dimensional random vector. This computation can be done using quadrature, sparse grids or Monte-Carlo methods and dedicated software libraries exist [22]. Furthermore, this computation can be done offline and a look-up table can be created. In the case where $(\epsilon_t)_{t\in\mathbb{N}}$ is i.i.d. Gaussian, the T-dimensional random vector has distribution: $\mathcal{N}(b, V^{-1})$. Extensive research has been done to optimise the numerical evaluation of this type of expression and fast quasi Monte Carlo methods can be used for accurate computation for $T < 100$ [23].

The following result gives a condition under which $\mathbb{E}[\tau_\gamma^+)$ is guaranteed to be finite.

Proposition 1. *If, instead of Assumption 1, $(\epsilon_t)_{t\in\mathbb{N}}$ is assumed to be a Gaussian white noise process then $\mathbb{E}[\tau_\gamma^+] < \infty$.*

Proof. Follows from proof of Theorem 1 and [5]. □

Proposition 1 can be used to characterise the data generating process defined in (1) as mean reverting in the sense that the expected crossing time is finite for any choice of initial conditions $a \in \mathbb{R}^d$ and barrier $\gamma \in [0, a_1 - y^*)$. In particular, Proposition 1 holds when $\gamma = 0$ and the barrier is equal to the mean of the process. It is important to note however that this does not provide any information on long-run convergence of the data generating process. One may also wonder how tight the bounds given in Theorem 1 are and whether they can be improved for the assumptions used in this paper. The following result shows that the bounds stated in Theorem 1 cannot be improved for $\alpha^* \in \triangle_+$.

Proposition 2 (Tightness)**.** *The upper bounds in Theorem 1 are tight for all $\alpha^* \in \triangle_+$.*

Proof. Proposition 2 follows from the proof of Theorem 1. □

The proof of Theorem 1 shows that the bounds are tight when the dynamics of $(y_t)_{t\in\mathbb{N}}$ can be represented by a linear autoregressive model (AR(p), $p \in \mathbb{N}$). In particular, for $\alpha^* \in \triangle_+$, this implies that any non-linear model that is Lipschitz continuous with respect to $\|\cdot\|_{\alpha^*}$ and has a fixed point y^*, will have its first hitting time probabilities and expectation upper bounded by a linear AR process with coefficients α^* and intercept c (specified such that the mean of the AR process is y^*).

An improvement on the tightness of the upper bounds given in Theorem 1 can be obtained by considering local α^*-contraction conditions. This is given in the following corollary.

Corollary 1. *Assume that the assumptions of this section hold and consider $a \in \mathbb{R}^d$ and $\gamma \in [0, a_1 - y^*)$.*
Define $\mathcal{D}^* = \prod_{i=1}^d \mathbb{R}_{\geq b_i}$ where $b_i = \min\{\gamma, \min_{j\leq i} a_j\}$, then $\forall T \in \mathbb{N}$, the α^* coefficients used to compute $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ in Theorem 1 can be replaced by $\alpha^*_{D^*} \in \mathbb{R}^d$ where $\alpha^*_{D^*}$ satisfies $\psi|_{\mathcal{D}^*} \in \mathcal{L}^{\alpha^*_{D^*}}(D^*)$.

Proof. The proof of Corollary 1 follows from the proof of Theorem 1 and Remark 1.2. □

IV. Estimation of α^* from machine learning models

The following subsection explains how the α^* coefficients utilised in this paper can be obtained for learning based system identification frameworks that are sufficiently regular. Assume that the following time series data: $\mathcal{S}_n = \{y_t\}_{t\in\{1,\dots,n\}}$ can be observed and that an autoregressive machine learning forecasting model $\hat{\psi}$ has been fitted to the data. Then, using $\hat{\psi}$ as a replacement for the transition function ψ one aims to estimate the α^* coefficients. To do this, a main advantage of the theoretical results obtained so far in this paper is the intuitive formulation of the Lipschitz type conditions that were used. In particular, if $\hat{\psi}$ is differentiable, we can utilise the partial derivatives of $\hat{\psi}$. This can be done by using the following result:

Proposition 3. *If the domain $\mathcal{D} \subseteq \mathbb{R}^d$ of ψ is convex and $\psi \in C^1(\mathcal{D})$ then $\psi \in \mathcal{L}^{\alpha^*}(\mathcal{D})$ with $\alpha_i^* = \sup_{x\in\mathcal{D}} |\frac{\partial\psi}{\partial x_i}(x)|$.*

Proof. Follows directly from an application of the multivariate version of the mean value theorem and the convexity of $\mathcal{D}$. □

From Proposition 3, we have that if there exists $\{\alpha_i^*\}_{i\in\{1,\dots,d\}}$ such that $\max_{x\in\mathcal{D}} |\frac{\partial\hat{\psi}}{\partial x_i}(x)| \leq \alpha_i^*$ for all $i \in \{1,\dots,d\}$ and $\sum_{i=1}^d \alpha_i^* < 1$ then Theorem 1 and Proposition 1 can be applied. While the computation of Lipschitz constants of machine learning models is difficult (with the exception of some non-parametric frameworks [24]), estimating the gradients of a learned model is generally more straightforward. In particular, for nonlinear autoregressive models that rely on neural networks, backpropagation can be used to compute the partial derivatives and existing deep learning libraries (e.g. Pytorch or Tensorflow) can be leveraged (see torch.autograd/tf.GradientTape).

This type of input-output partial derivative computation has been extensively used in computer vision and explainable AI for input sensitivity analysis [25], [26] and has started

Method	$L_1(20)$	$L_1(40)$	$\mathbb{E}[\tau\|\tau \leq 40]$
AR(1): $x_{t+1} = 0.9x_t + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 8.96$			
Non-linear Estim.	0.12 ± 0.067	0.06 ± 0.033	9.05 ± 1.23
AR(1) Estim.	0.087 ± 0.046	0.052 ± 0.027	7.95 ± 0.048
AR(3): $x_{t+1} = 0.7x_t + 0.15x_{t-1} + 0.05x_{t-2} + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 7.66$			
Non-linear Estim.	0.102 ± 0.032	0.1 ± 0.08	11.86 ± 1.13
AR(1) Estim.	0.074 ± 0.042	0.038 ± 0.02	7.43 ± 0.58
ESTAR(1): $x_{t+1} = 0.4x_t + 0.3x_t(1 - e^{-\frac{x_t^2}{2}}) + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 3.4$			
Non-linear Estim.	0.093 ± 0.021	0.04 ± 0.009	4.77 ± 0.61
AR(1) Estim.	0.085 ± 0.0095	0.038 ± 0.004	3.57 ± 0.04
Neur. Net.: $x_{t+1} = NN(x_t, x_{t-1}, x_{t-2}) + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 7.36$			
Non-linear Estim.	0.17 ± 0.061	0.06 ± 0.03	10 ± 1.33
AR(1) Estim.	0.085 ± 0.037	0.08 ± 0.03	3.8 ± 1.2

TABLE I: Performance of the first hitting time bounds in practice.

to expand to other application areas. Existing heuristics and techniques from these approaches can therefore be leveraged in order to compute the α^* coefficients. Furthermore, for several nonparametric machine learning model choices that utilise neural networks, it is possible to incorporate gradient learning directly into the model fitting process which would offer a more direct way of estimating the $\{\alpha_i^*\}_{i\in\{1,\dots,d\}}$ coefficients [27].

In Table I, we illustrate how the upper bounds given in Theorem 1 can perform in a in practtical example. Using a standard autoregressive model, we generate a time series of fixed length: 1000 timesteps and model the noise with a Gaussian distribution. Then, utilising a 2-layer Neural Network with sigmoid activation we estimate the α^* constants of the underlying function and compute $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ for $1 \leq T \leq 40$ (We stop at 40 as the first hitting time probabilities are approximately equal to 0 for $T \geq 40$). These values are compared against the "true" first hitting probabilities of the stochastic process defined by the selected autoregression function. We compute the (averaged) $L_1(20)$-error: $\frac{1}{20}\sum_{T=1}^{20} |\mathfrak{I}^+_{(\alpha^*,y^*)}(T) - \mathbb{P}(\tau_\gamma^+ > T)|$, (averaged) $L_1(40)$-error: $\frac{1}{40}\sum_{T=1}^{40} |\mathfrak{I}^+_{(\alpha^*,y^*)}(T) - \mathbb{P}(\tau_\gamma^+ > T)|$ and the value of the estimation of the conditional expectation $\mathbb{E}[\tau|\tau \leq 40]$. These values are averaged over 10 simulations and the standard deviation of the obtained results is also stated. As a benchmark, we estimate the first hitting time probabilities of the time series using an AR(1) model as is most commonly done in practice (see discussion in introduction). As the AR(1) estimation approach aims to directly estimate the first hitting time of the stochastic process as opposed to ensuring a bound, one expects it to be more precise than the nonlinear estimation approach in terms of $L_1(20)$ and $L_1(40)$ metrics. While this can be observed for some values in Table I, we have that in the majority of computed loss metrics our proposed approach is competitive with the results of the AR(1) first hitting time estimation method. The estimated values of $\mathbb{E}[\tau|\tau \leq 40]$ then illustrates the fact that the nonlinear estimation method aims to ensure a lower bound on the first hitting times of the time series.

One caveat to the discussion of this section is that the robustness of the estimation of the α^* coefficients can be difficult to obtain as it depends strongly on the precision of the system idenfication method. Some research on robust estimation of the gradient/partial derivatives for neural network based approaches can be found (e.g. see [28] [29]): however the impact of the estimation error on the partial derivatives estimates and thereby on the α^* estimates remains an open question that we will explore in future work.

V. Application to Statistical Arbitrage

Statistical Arbitrage. In this section, we utilise our theoretical results in the context of statistical arbitrage (*statarb*). In general terms, statarb can be defined as any trading framework that utilises interdependencies between the price time series of financial assets to construct a mean reverting synthetic asset which can be studied to obtain trading signals. A simple example of statarb is given below.

Example 1. *(Simple pairs trading) One trades a synthetic asset $Y = X - X'$ which is the difference of two underlying assets X and X'. Here a* long *position (buying) can be assumed by buying X and (short) selling X'. And a* short *(selling) position by (short) selling X and buying X'. A pairs trading strategy then aims to profit by leveraging the mean reverting behaviour of the synthetic asset which, in a loose sense, means that it tends to oscillate around or converge to a fixed mean value (y^*). It enters a long trade whenever the price of the synthetic asset reaches a threshold level U_1 that is far below the mean. It closes the long trade whenever the asset price has reverted back to a level L_1 close to the mean by selling it. Conversely, the strategy assumes a short position if it reaches a level U_2 that is far above the mean closes out the position upon reaching a level L_2 near the mean. This is illustrated in Figure 1[a]. Here we traded a simulated synthetic asset employing our strategy and show the U_1, L_1 thresholds.*

The notion of mean reversion and the choice of design parameters U[1] and L vary throughout the literature, but typically depend on specific, often quite restrictive assumptions on the functional form of the dynamics of the synthetic asset.

For a review of the extensive statarb literature, the reader is referred to [8]. As we focus here on the optimisation of the trading strategy given the synthetic asset, relevant approaches to our research can be found in [18], [30]. In particular, the former assumes that an underlying Ornstein-Uhlenbeck (OU) model holds and fits an AR(1) process in order to derive optimal trading thresholds and policies that optimise standard trading measures. Extensions for OU modelling with jump processes [31], stop-loss rules [32] and regime-switching [33] have also been developed. Apart from a few exceptions (e.g. [34], [35]) which do not focus directly on threshold setting, most of the relevant research has ignored the use of more general nonlinear processes applicable in settings where the synthetic asset has been identified with a (nonlinear) machine learning method.

[1] w.l.o.g. the index notation is omitted as the short and long positions can be treated in similar ways.

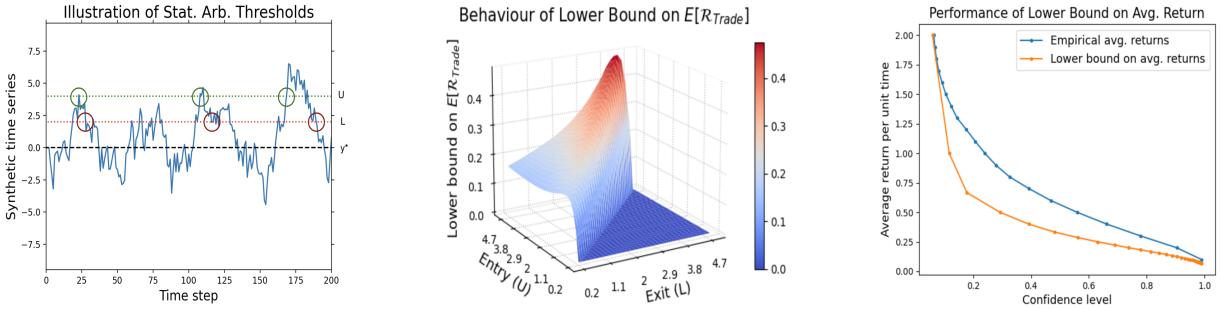

Fig. 1: **[a]**: *Statistical arbitrage thresholds for the short position. Positions are opened (green circles) when the time series hits U (green line) and closed (red circles) when it subsequently hits L (red line). The dashed black line represents the "fixed point" y^*.* **[b]**: *Dependence of the expected return lower bound guarantee (Eq. 6) on the entry threshold (U) and exit threshold (L). Here, $\alpha^* = (0.7, 0.15, 0.05)$ and U, L are given in units of noise standard deviation.* **[c]**: *Setting thresholds $U = 4.4, L = 2.2$, the bound on $\mathcal{R}_{Trade}(U, L)$ given in Eq. 5 is illustrated empirically for various choices of confidence levels (p) by computing the empirical distribution of the returns for positions opened and closed at thresholds (U, L).*

.

Our approach. By contrast, we merely need to make the more general assumption that our synthetic asset Y is governed by a general nonlinear AR process as per Eq. 1. In our framework, we consider Y to be mean reverting if it is contracting, in which case we choose the "mean" to coincide with the fixed point, i.e. $m = y^*$.

Using our theoretical results, we show how, having obtained estimates for the α^*-coefficients and the first hitting time bounds from Theorem 1, we can inform the selection of the entry and exit trading thresholds U, L such that we get a probabilistic guarantee on the holding time of the trade and expected return. An illustrative example with a simulated synthetic asset is given in Figure 1[b, c]. To tune U, L and understand the profitability properties of the trades of the strategy, we are interested in bounds involving the following variables:

Definition 4 (Informal definition of trading variables)**.**

- $r(U, L, c)$*: return of a single trade at thresholds (U, L) and transaction cost c.*
- $S(U, L)$*: time taken to close positions once they have been opened (with threshold (U,L)).*
- $\mathcal{R}_{Trade}(U, L, c) := \frac{r(U,L,c)}{S(U,L)}$*: average return of a single trade per unit of time with thresholds (U, L).*

Under common noise assumptions (e.g. Gaussian with a finite variance of $\sigma := var(\epsilon_2)$), we can utilise Theorem 1 to obtain an upper bound $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)$ on $S(U, L)$ that holds with high probability $p \in [0, 1)$: $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p) := \min\{T \in \mathbb{N} | \, 1 - \mathfrak{I}^+_{(\alpha^*,y^*)}(T) \geq p\} \in \mathbb{N} \cup \{\infty\}$ where $\mathfrak{I}^+_{(\alpha^*,y^*)}$ depends on the choice of U, L and σ. This upper bound can then be used to set a probabilistic guarantee on the average return per unit of time:

Corollary 2. *Let $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)$ be as defined above,*

$$\mathbb{P}\Big(\mathcal{R}_{Trade}(U, L) \geq \frac{r(U, L, c)}{\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)}\Big) \geq p \qquad (5)$$

where $p \in [0, 1)$ is a chosen confidence level. Furthermore, we have

$$\mathbb{E}[\mathcal{R}_{Trade}(U, L)] \geq \frac{r(U, L, c)}{1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*,y^*)}(T)}. \qquad (6)$$

Proof. This result follows directly from Theorem 1 and Jensen's inequality. □

As a lower bound for $r(U, L, c)$ is generally easily obtainable by considering the difference in value of the underlying positions at U and L, (5) and (6) can be used to determine trading thresholds that guarantee in expectation or with high probability a sufficiently high average return per unit time. The final optimisation of the trading thresholds will then also depend on the number of times the position entry threshold U is hit (i.e. the number of times a position in the underlying securities can be opened), the desired duration of the trade and the average return per unit of time of other trading opportunities in the portfolio. Figure 1[b] provide an illustration of the behaviour of the lower bound guarantees on $\mathbb{E}[\mathcal{R}_{Trade}(U, L)]$ stated in (6) for various values of U and L. These lower bounds were computed in the context of a simple case of a single mean reverting asset (implies $r(U, L, c) \geq U - L$) when the dynamics of the synthetic asset were assumed to be α^*-Lipschitz contracting with $\alpha^* = (0.7, 0.15, 0.05)$. For a specific choice of U, L, Figure 1[c] illustrates the lower bound stated in (5). To obtain the bound, the relation $r(U, L, c) \geq U - L$ was utilised. The experiments were run 5000 times by simulating from a neural network (4-layers, Relu activation, trained on daily equity data) with $\alpha^* = (0.7, 0.15, 0.05)$ in order to obtain the illustrated empirical distribution. As expected, for each confidence level p the curve representing the lower bound given in (5) lies beneath the curve representing the empirically estimated $(1 - p)$-th quantile of the average return per unit of time.

VI. Conclusions

In this work, we derived novel first hitting time bounds for general classes of nonlinear systems. In contrast to existing work, we did not need to impose strong requirements on the functional form of the transition function. Instead,

our bounds rested on contraction conditions relative to a weighted norm. Such conditions can be readily verified for a great many machine learning models such as neural networks (e.g. via partial gradients automatically derived by popular packages such as tensorflow.) We also provided a synthetic example of a trading application of where our hitting time bounds can be leveraged to inform a strategy's position changes such that probabilistic and expected lower bounds on the return can be guaranteed. Forthcoming work will extend these results to consider locally contractive dynamical systems (see Section IV) and provide an empirical evidence of the potential profitability of this approach can be when applied to learning-based trading of financial assets. Of course the generality of our results might also suggest they could be employed in a wide range of other disciplines where hitting times are of interest, such as in econometrics, ecology and control.

REFERENCES

[1] J. G. De Gooijer *et al.*, *Elements of nonlinear time series analysis and forecasting*. Springer, 2017, vol. 37.

[2] N. Kohzadi, M. S. Boyd, B. Kermanshahi, and I. Kaastra, "A comparison of artificial neural network and time series models for forecasting commodity prices," *Neurocomputing*, vol. 10, no. 2, pp. 169–181, 1996.

[3] M. Ghiassi, H. Saidane, and D. Zimbra, "A dynamic artificial neural network model for forecasting time series events," *International Journal of Forecasting*, vol. 21, no. 2, pp. 341–362, 2005.

[4] M. Valipour, M. E. Banihabib, and S. M. R. Behbahani, "Comparison of the arma, arima, and the autoregressive artificial neural network models in forecasting the monthly inflow of dez dam reservoir," *Journal of hydrology*, vol. 476, pp. 433–441, 2013.

[5] G. K. Basak and K.-W. R. Ho, "Level-crossing probabilities and first-passage times for linear processes," *Advances in applied probability*, pp. 643–666, 2004.

[6] E. Di Nardo, "On the first passage time for autoregressive processes," 2008.

[7] A. Novikov, R. Melchers, E. Shinjikashvili, and N. Kordzakhia, "First passage time of filtered poisson process with exponential shape function," *Probabilistic engineering mechanics*, vol. 20, no. 1, pp. 57–65, 2005.

[8] C. Krauss, "Statistical arbitrage pairs trading strategies: Review and outlook," *Journal of Economic Surveys*, vol. 31, no. 2, pp. 513–545, 2017.

[9] H. Puspaningrum, Y.-X. Lin, and C. M. Gulati, "Finding the optimal pre-set boundaries for pairs trading strategy based on cointegration technique," *Journal of Statistical Theory and Practice*, vol. 4, no. 3, pp. 391–419, 2010.

[10] J. M. Ferguson and J. M. Ponciano, "Predicting the process of extinction in experimental microcosms and accounting for interspecific interactions in single-species time series," *Ecology letters*, vol. 17, no. 2, pp. 251–259, 2014.

[11] M. Frisén and C. Sonesson, "Optimal surveillance based on exponentially weighted moving averages," *Sequential Analysis*, vol. 25, no. 4, pp. 379–403, 2006.

[12] M. Mollineaux and R. Rajagopal, "Structural health monitoring of progressive damage," *Earthquake Engineering & Structural Dynamics*, vol. 44, no. 4, pp. 583–600, 2015.

[13] H. Y. Noh, K. K. Nair, A. S. Kiremidjian, and C. Loh, "Application of time series based damage detection algorithms to the benchmark experiment at the national center for research on earthquake engineering (ncree) in taipei, taiwan," *Smart Structures and Systems*, vol. 5, no. 1, pp. 95–117, 2009.

[14] A. Lipton and V. Kaushansky, "On the first hitting time density of an ornstein-uhlenbeck process," *arXiv preprint arXiv:1810.02390*, 2018.

[15] R. Martin, M. Kearney, and R. Craster, "Long-and short-time asymptotics of the first-passage time of the ornstein–uhlenbeck and other mean-reverting processes," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 13, p. 134001, 2019.

[16] A. J. Fisher, D. A. Green, A. V. Metcalfe, and K. Akande, "First-passage time criteria for the operation of reservoirs," *Journal of hydrology*, vol. 519, pp. 1836–1847, 2014.

[17] P. Lánskỳ and C. E. Smith, "The effect of a random initial value in neural first-passage-time models," *Mathematical biosciences*, vol. 93, no. 2, pp. 191–215, 1989.

[18] W. K. Bertram, "Analytic solutions for optimal statistical arbitrage trading," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 11, pp. 2234–2243, 2010.

[19] Z. Zeng and C.-G. Lee, "Pairs trading: optimal thresholds and profitability," *Quantitative Finance*, vol. 14, no. 11, pp. 1881–1893, 2014.

[20] J. Wise, "The autocorrelation function and the spectral density function," *Biometrika*, vol. 42, no. 1/2, pp. 151–159, 1955.

[21] J. Elstrodt, *Maß-und Integrationstheorie*. Springer, 1996, vol. 7.

[22] T. Hahn, "Cuba—a library for multidimensional numerical integration," *Computer Physics Communications*, vol. 168, no. 2, pp. 78–95, 2005.

[23] A. Genz and F. Bretz, *Computation of multivariate normal and t probabilities*. Springer Science & Business Media, 2009, vol. 195.

[24] J.-P. Calliess, S. J. Roberts, C. E. Rasmussen, and J. Maciejowski, "Lazily adapted constant kinky inference for nonparametric regression and model-reference adaptive control," *Automatica*, vol. 122, p. 109216, 2020.

[25] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *The Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[26] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[27] A. A. Ismail, H. Corrada Bravo, and S. Feizi, "Improving deep learning interpretability by saliency guided training," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[28] P. Cardaliaguet and G. Euvrard, "Approximation of a function and its derivative with a neural network," *Neural Networks*, vol. 5, no. 2, pp. 207–220, 1992.

[29] W. Wang, P. Yu, L. Lin, and T. Tong, "Robust estimation of derivatives using locally weighted least absolute deviation regression," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2157–2205, 2019.

[30] R. J. Elliott, J. Van Der Hoek*, and W. P. Malcolm, "Pairs trading," *Quantitative Finance*, vol. 5, no. 3, pp. 271–276, 2005.

[31] J. Stübinger and S. Endres, "Pairs trading with a mean-reverting jump–diffusion model on high-frequency data," *Quantitative Finance*, vol. 18, no. 10, pp. 1735–1751, 2018.

[32] T. Leung and X. Li, "Optimal mean reversion trading with transaction costs and stop-loss exit," *International Journal of Theoretical and Applied Finance*, vol. 18, no. 03, p. 1550020, 2015.

[33] Y. Bai and L. Wu, "Analytic value function for optimal regime-switching pairs trading rules," *Quantitative Finance*, vol. 18, no. 4, pp. 637–654, 2018.

[34] C. L. Dunis, J. Laws, and B. Evans, "Trading futures spread portfolios: applications of higher order and recurrent networks," *The European Journal of Finance*, vol. 14, no. 6, pp. 503–521, 2008.

[35] C. L. Dunis, J. Laws, P. W. Middleton, and A. Karathanasopoulos, "Trading and hedging the corn/ethanol crush spread using time-varying leverage and nonlinear models," *The European Journal of Finance*, vol. 21, no. 4, pp. 352–375, 2015.