First Hitting Time Guarantees for Nonlinear Time Series Models

Julien Huang [1] **Jan-Peter Calliess** [1]

Abstract

We derive tight probabilistic bounds on the first hitting time of general classes of contractive nonlinear time series models that can be linked to mean reverting processes. As an application to finance, we translate our results to a pairs trading strategy with probabilistic guarantees on its returns.

1. Introduction

The proliferation of learning-based time series estimation techniques increases the need for widely applicable theoretical tools for understanding generic nonlinear time series models. Here, first hitting time guarantees and mean reversion of learning-based nonlinear time series models are properties of particular importance in a great many application domains, including in finance, econometrics, control and dynamical systems.

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 are then computed numerically (Basak and Ho, 2004) (Di Nardo, 2008) or can be lower bounded analytically in the case of the AR(1) model (Novikov, 1991). This approach has been explored in various domains; in statistical arbitrage and quantitative finance for optimal thresholds setting (Krauss, 2017)(Puspaningrum et al., 2010), for predicting population extinction and time to extinction in ecology (Ferguson and Ponciano, 2014), signal detection and surveillance analysis (Frisén and Sonesson, 2006) or structural health monitoring (Mollineaux and Rajagopal, 2015) (Noh et al., 2009). 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 (Lipton and Kaushansky, 2018) (Martin et al., 2019) (and references therein) under some additional assumptions. Applications are numerous and involve, for example, hydrology (Fisher et al., 2014), neuroscience (Lánskỳ and Smith, 1989) or quantitative finance (Bertram, 2010)(Zeng and Lee, 2014). 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.

What unifies these threads of works is that they provide an understanding of hitting times for time series whose transition functions conform to a specific structure. However, when those functions are identified by black-box machine learning algorithms, existing results are not applicable. What is needed are theoretical bounds that can be computed for time series models whose transition functions have been black-box identified with general classes of popular machine learning methods, such as neural networks or non-parametric models.

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 bounds are shown to translate to probabilistic bounds on the return of the ensuing trading strategy, provided the time series of the asset pair satisfies the required contractive Lipschitz conditions.

2. Model assumptions

Time series model. Let d $\in \mathbb{N}$, $\psi : \mathbb{R}^d \to \mathbb{R}$ and $a \in \mathbb{R}^d$. We assume our time series is modelled by the (stochastic) nonlinear auto-regressive (NAR) process $(y_t)_{t\in\mathbb{N}}$ with transition function ψ and initial conditions $a \in \mathbb{R}^d$ defined as follows:

[1]Oxford-Man Institute of Quantitative Finance, University of Oxford, Oxford, United Kingdom. Correspondence to: Julien Huang <julien.huang@sjc.ox.ac.uk>.

ICML 2021 Time Series Workshop,

$$y_{t+1} := \begin{cases} \psi(y_t, ..., y_{t-(d-2)}, y_{t-(d-1)}) + \epsilon_{t+1} \text{ for } t > d \\ y_i = a_i, a_i \in \mathbb{R}\ \forall i \in \{1, ..., d\}. \end{cases} \tag{1}$$

Here, the noise process $(\epsilon_t)_{t\in\mathbb{N}}$ is a stochastic process (s.p.) satisfying :

Assumption 2.1. *The joint distribution of any finite sequence of consecutive noise variables; $\epsilon_{1:T} := (\epsilon_1, ..., \epsilon_T)$ has a probability density function denoted by $f_{\epsilon_{1:T}}$.*

Note, the commonly encountered white noise process would satisfy our assumption.

Assumption on contractive transition maps. To establish our bounds in subsequent sections, we need to assume the transition function ψ to be a contraction relative to a weighted norm we introduce next:

Definition 2.2 (α^*-norm)**.** *Let $\alpha^* \in \mathbb{R}^d_{\geq 0}$, The α^*-norm $\|\cdot\|_{\alpha^*} : \mathbb{R}^d \to \mathbb{R}$ is defined as the following weighted l_1 norm $\|x\|_{\alpha^*} = \sum_{i=1}^d \alpha_i^* |x_i|\ \forall x \in \mathbb{R}^d$.*

We rehearse the definition of Lipschitz continuity;

Definition 2.3 (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.*

Definition 2.4 (α^*-contracting process)**.** *Let $\mathcal{D} \subseteq \mathbb{R}^d$. An auto-regressive 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 $\alpha^* \in \triangle_+ := \{x \in \mathbb{R}^d_{\geq 0} | \sum_{i=1}^d x_i < 1\}$.*

Assumption 2.5. *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 $\mathcal{D} = \mathbb{R}^d$.

3. First hitting time guarantees

We will now state bounds on first hitting times of our time series. We assume all definitions and assumptions introduced in Sec. 2 hold.

Appealing to Banach's fixed point theorem one can show the existence of a unique fixed point $y^* = \psi(y^*, \ldots, y^*)$.

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.

Definition 3.1 (First hitting time (f.h.t.))**.** *For $a \in \mathbb{R}^d$ with $a_d > y^*$ and $\gamma \in [0, a_d - y^*[$, we define the upper first hitting time of $(y_t)_{t\in\mathbb{N}}$:*

$$\tau_\gamma^+ := \inf\{t \in \mathbb{N} | y_{t+d} - y^* < \gamma\}\ .$$

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

$$\tau_\gamma^- := \inf\{t \in \mathbb{N} | y_{t+d} - y^* > \gamma\}\ .$$

Initial value a_d 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 (Wise, 1955), one can show our first first principal result:

Theorem 3.2. *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 (4), $f_{\epsilon_{1:T}}$ is defined in Assumption 2.1 and $b_i := \left(B^i(a - y^*\mathbf{1}_d)\right)_1 - \gamma$ for $i = 1, \ldots, T$ where B is defined in (5). We have:

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

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

Remark 3.3. *Analogous bounds can be derived for $\mathbb{P}\big(\tau_\gamma^- > T\big)$ and $\mathbb{E}[\tau_\gamma^-]$.*

Remark 3.4. *Some comments on the behaviour $\mathfrak{I}^+_{(\alpha^*, y^*)}(T)$:*

(1) $\forall T \in \mathbb{N}$, $\mathfrak{I}^+_{(\alpha^*, y^*)}(T)$ is decreasing in γ.
(2) If $\forall\, i$, $\alpha_i^* < \beta_i^*$ then $\mathfrak{I}^+_{(\alpha^*, y^*)}(T) < \mathfrak{I}^+_{(\beta^*, y^*)}(T)$.
(3) $\forall T \in \mathbb{N}$: $\lim_{\|\alpha^*\|_1 \to 0} \mathfrak{I}^+_{(\alpha^*, y^*)}(T) = \frac{1}{2^T}$.

The integral stated in $\mathfrak{I}^+_{(\alpha^*, y^*)}(T)$ corresponds to the computation of orthant probabilities and can be done numerically. The following result gives a condition under which $\mathbb{E}[\tau_\gamma^+]$ is finite.

Proposition 3.5. *If instead of Assumption 2.1, $(\epsilon_t)_{t\in\mathbb{N}}$ is assumed to be a white noise process then $\mathbb{E}[\tau_\gamma^+] < \infty$.*

Remark 3.6. *Prop. 3.5 implies that $\mathbb{P}\big(\tau_\gamma^+ < \infty\big) = 1$.*

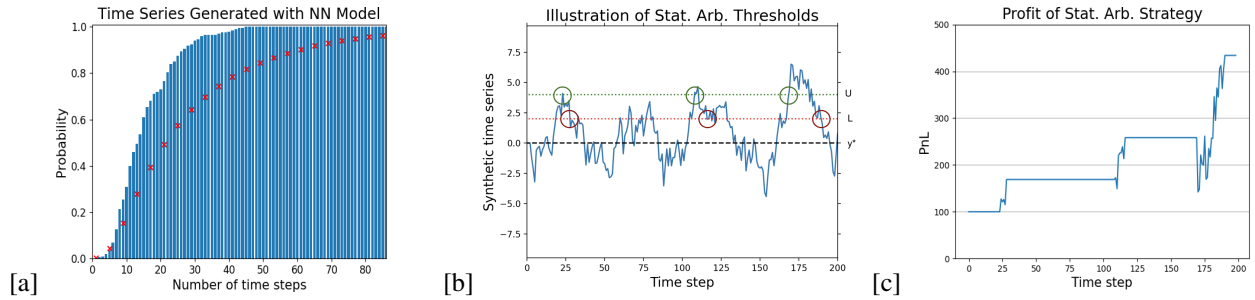

Figure 1: **[a]**:Blue columns represent empirical estimate of CDF of f.h.t. of time series generated by a neural network (4-layers, Relu activation); $\mathbb{P}(\tau_\gamma^+ \leq T)$. Red line is computed from $1 - \mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ in Theorem 3.2. Here, $\alpha^* = (0.7, 0.15, 0.1)$ is computed using the approach described in subsection 3.1. **[b]**: Illustration of 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^*. **[c]**: Cumulative realised $P\&L$ of the trading strategy applied in Figure 1[b] .

Prop. 3.5 implies that the α^*-contracting stochastic process is mean reverting in the sense that it will eventually hit any barrier between the shock and y^* of the time series with probability 1. In particular, since γ can be chosen to be 0, we have that the stochastic process eventually hits y^* with probability 1. A result on the tightness of the bounds given in Theorem 3.2 can also be shown:

Proposition 3.7 (Tightness)**.** *The upper bounds in Theorem 3.2 are tight for all* $\alpha^* \in \triangle_+$.

3.1. Estimation of α^* from machine learning models

A main benefit of the theoretical results obtained in the previous section is the intuitive formulation of the Lipschitz type conditions used. In particular, if ψ is differentiable, we have the following result;

Proposition 3.8. *If* $\mathcal{D} \subseteq \mathbb{R}^d$ *is convex and* $\psi \in C^1(\mathcal{D})$ *then* $\psi \in \mathcal{L}^{\alpha^*}(\mathcal{D})$ *with* $\alpha_i^* = \max_{x\in\mathcal{D}} |\frac{\partial f}{\partial x_i}(x)|$.

From Prop. 3.8, we have that if there exists $\{\lambda_i\}_{i\in\{1,\dots,d\}}$ such that $\max_{x\in\mathcal{D}} |\frac{\partial f}{\partial x_i}(x)| \leq \lambda_i$ for all $i \in \{1, ..., d\}$ and $\sum_{i=1}^d \lambda_i < 1$ then Theorem 3.2 can be applied. While the computation of Lipschitz constants of machine learning models is difficult (with the exception of some nonparametric frameworks (Calliess et al., 2020)), computing gradients of the learned model is generally straightforward.

For nonlinear autoregressive models that rely on neural networks, backpropagation can be used to compute the partial derivatives and existing deep learning libraries (eg. Pytorch or Tensorflow) can be leveraged (see torch.autograd/tf.GradientTape). Alternatively, for several nonparametric machine learning model choices, it is possible to incorporate gradient learning into the model fitting process. This would offer a more direct way of estimating the $\{\lambda_i\}_{i\in\{1,\dots,d\}}$ coefficients. The robustness of the estimation of the $\{\lambda_i\}_{i\in\{1,\dots,d\}}$ coefficients is not discussed in this paper as it is dependent on the choice of the time series forecasting framework. For some of the modelling frameworks mentioned above, research on robust estimation of the gradient/partial derivatives relative to the input of the model can be found (Cardaliaguet and Euvrard, 1992) (Wang et al., 2019).

4. Application to pairs trading

In this section, we apply our theoretical results to statistical arbitrage. In particular, we consider the popular case of pairs trading. Here, one trades a synthetic asset whose price series Z is computed as the difference of two other assets X, Y. That is, one trades $Z_t = X_t - \beta Y_t$. *Hedging coefficient* β is tuned to render Z mean reverting. A pairs trading strategy then aims to profit by leveraging the mean reverting behavior of the synthetic asset. It enters a long trade whenever the price of the synthetic asset reaches a threshold level that is far below the mean. It closes the long trade whenever the asset price has reverted back to the mean level by selling it. Conversely, the strategy goes short trade is initiated the price of Z reaches a level that is far above the mean by short-selling the synthetic asset and closes out the position upon reaching the mean. Typically, the thresholds U, L are found heuristically, or based on restrictive assumptions on the time series model such as being an OU model (Bertram, 2010).

In contrast, we can harness our theoretical results to inform a pairs trading strategy for wide classes of nonlinear models. For example, we can learn the time series dynamics with an artificial neural network (ANN) or by tuning any econometric model and then inspect the bounds on the partial derivatives to determine whether the synthetic asset is mean reverting in the sense of being α^*-contracting (cf. Prop. 3.8). If it is, we can use our f.h.t. bounds[1] on first hitting times to determine the entry and exit levels U, L such

[1]Of course, the validity of the f.h.t. bounds inferred using the results of this paper depend on the accuracy of the α^* estimate.

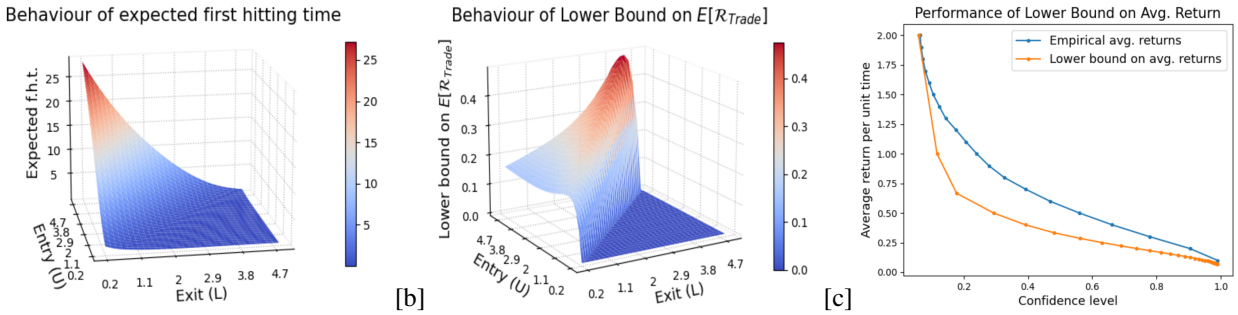

Figure 2: **[a,b]**: Illustration of the dependence of the f.h.t. lower bound guarantee and the expected return lower bound guarantee (Eq. 3) on the entry threshold (U) and exit threshold (L). Here, $\alpha^* = (0.7, 0.15, 0.05)$ and U, L are given in units of noise standard deviation. **[c]**: Setting thresholds $U = 4.4, L = 2.2$ and utilising a lower bound on the return $r(U, L, c) \geq U - L$, the lower bound on the avg. return for various confidence levels is illustrated empirically by computing the return of 5000 opened positions at thresholds (U, L).

that we get a probabilistic guarantee on the return from the trade within a certain time.

An illustrative example is given in Figure 1[b,c]. Here we traded a simulated synthetic asset employing our strategy.

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 4.1 (Informal definition of trading variables)**.**

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

Under common noise assumptions (Gaussian or t-student with finite standard deviation of σ), we can utilise Theorem 3.2 to obtain an upper bound on $T(U, L, c)$ that holds with high probability; $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p) := \min\{T \in \mathbb{N} | \, \mathfrak{I}^+_{(\alpha^*,y^*)}(T) \geq p\}$ 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 avg. return per unit of time;

$$\mathbb{P}\Big(\mathcal{R}_{Trade}(U, L) \geq \frac{r(U, L, c)}{\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)}\Big) \geq p \qquad (2)$$

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

$$\mathbb{E}[\mathcal{R}_{Trade}(U, L)] \geq \frac{r(U, L, c)}{\sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*,y^*)}(T)}. \qquad (3)$$

This result follows from Theorem 3.2 and Jensen's inequality. (2) and (3) can be used to determine trading thresholds that guarantee, either 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 (ie. 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.

Figures 2[a] and 2[b] provide an illustration of the behaviour of the lower bound guarantees on $\mathbb{E}[T(U, L)]$ and $\mathbb{E}[\mathcal{R}_{Trade}(U, L)]$ stated in (3) for various values of U and L. These lower bounds were computed in the context of a simple case of pairs trading $(r(U, L, c) \geq U - L)$ 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 2[c] illustrates the lower bound stated in (2). As expected, for each confidence level p the curve representing the lower bound given in (2) is beneath the curve representing the empirically estimated $(1 - p)$-th quantile of the average return per unit of time.

5. Conclusions

In this work, we have derived novel first hitting time bounds derivable for general classes of nonlinear time series models. 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 conditions 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 have 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 risk bounds on the returns can be provided. Future work will investigate how successful this approach can be when applied to learning-based trading of real financial assets under risk constraints. Of course the generality of our results might 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

Gopal K Basak and Kwok-Wah Remus Ho. Level-crossing probabilities and first-passage times for linear processes. *Advances in applied probability*, pages 643–666, 2004.

Elvira Di Nardo. On the first passage time for autoregressive processes. 2008.

AA Novikov. On the first passage time of an autoregressive process over a level and an application to a "disorder" problem. *Theory of Probability & Its Applications*, 35 (2):269–279, 1991.

Christopher Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

Heni Puspaningrum, Yan-Xia Lin, and Chandra M Gulati. Finding the optimal pre-set boundaries for pairs trading strategy based on cointegration technique. *Journal of Statistical Theory and Practice*, 4(3):391–419, 2010.

Jake M Ferguson and José M Ponciano. Predicting the process of extinction in experimental microcosms and accounting for interspecific interactions in single-species time series. *Ecology letters*, 17(2):251–259, 2014.

Marianne Frisén and Christian Sonesson. Optimal surveillance based on exponentially weighted moving averages. *Sequential Analysis*, 25(4):379–403, 2006.

M Mollineaux and R Rajagopal. Structural health monitoring of progressive damage. *Earthquake Engineering & Structural Dynamics*, 44(4):583–600, 2015.

Hae Young Noh, K Krishnan Nair, Anne S Kiremidjian, and CH 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*, 5(1):95–117, 2009.

Alexander Lipton and Vadim Kaushansky. On the first hitting time density of an ornstein-uhlenbeck process. *arXiv preprint arXiv:1810.02390*, 2018.

RJ Martin, MJ Kearney, and RV 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*, 52 (13):134001, 2019.

Aiden J Fisher, David A Green, Andrew V Metcalfe, and Kunle Akande. First-passage time criteria for the operation of reservoirs. *Journal of hydrology*, 519:1836–1847, 2014.

Petr Lánskỳ and Charles E Smith. The effect of a random initial value in neural first-passage-time models. *Mathematical biosciences*, 93(2):191–215, 1989.

William K Bertram. Analytic solutions for optimal statistical arbitrage trading. *Physica A: Statistical Mechanics and its Applications*, 389(11):2234–2243, 2010.

Zhengqin Zeng and Chi-Guhn Lee. Pairs trading: optimal thresholds and profitability. *Quantitative Finance*, 14(11):1881–1893, 2014.

J Wise. The autocorrelation function and the spectral density function. *Biometrika*, 42(1/2):151–159, 1955.

Jan-Peter Calliess, Stephen J Roberts, Carl Edward Rasmussen, and Jan Maciejowski. Lazily adapted constant kinky inference for nonparametric regression and model-reference adaptive control. *Automatica*, 122:109216, 2020.

Pierre Cardaliaguet and Guillaume Euvrard. Approximation of a function and its derivative with a neural network. *Neural Networks*, 5(2):207–220, 1992.

WenWu Wang, Ping Yu, Lu Lin, and Tiejun Tong. Robust estimation of derivatives using locally weighted least absolute deviation regression. *The Journal of Machine Learning Research*, 20(1):2157–2205, 2019.

A. Appendix

A.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 where the entries are given by

$$A(T)_{ij} := \begin{cases} 1, & \text{if } i-j=0 \\ -\alpha^*_{(i-j)}, & \text{if } 0 < i-j \leq d \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

for all $i, j \in \{1, ..., T\}$. B is a sparse matrix whose entries are given by:

$$B_{ij} := \begin{cases} 1, & \text{if } i-j=1 \\ \alpha^*_j, & \text{if } i=1 \text{ and } 1 \leq j \leq d \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

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