École Nationale de la Statistique et de l'Analyse Économique (ENSAE)

Université du Québec à Montréal (UQAM)

RESEARCH THESIS

Submitted in partial fulfillment of the requirements for the Diplôme d'Ingénieur

Boosting for Time Series and an Application to Value at Risk Forecasting

Abdou Hamid ALAGBE
Student Engineer in Statistics and Economics, ENSAE Dakar

Rasmane BAMOGO
Student Engineer in Statistics and Economics, ENSAE Dakar

Under the supervision of:

Prof. Cédric Beaulac
Professor of Statistics, Université du Québec à Montréal (UQAM)

May 2026

Boosting for Time Series and an Application to Value at Risk Forecasting

Abdou Hamid ALAGBE[†]
alagbehamid@gmail.com

Rasmane BAMOGO[†]
bamogo6370@gmail.com

Prof. Cédric Beaulac[‡]
beaulac.cedric@gmail.com

A thesis submitted in partial fulfillment of the requirements for the Diplôme d'Ingénieur.

[†] Ingénieur Statisticien Économiste, ENSAE Dakar
[‡] Professor of Statistics, Université du Québec à Montréal (UQAM)

May 25, 2026

Contents

Abstract

We introduce BlockBoost[1], a boosting algorithm for time series forecasting with provable generalization guarantees. Standard boosting algorithms inherit the i.i.d. assumption of the PAC learning framework, which temporal dependence violates by design. BlockBoost addresses this at the algorithmic level by replacing instance-wise weight updates with block-wise updates over contiguous temporal segments, constraining the weight sequence to a block-constant subspace of the probability simplex. We prove that this modification drives the block-wise empirical risk to zero, introduces implicit regularization against noise memorization, and yields a generalization bound for stationary β-mixing sequences. We further prove that locally stationary processes – a class encompassing most financial return series – can be approximated in L^2 by a strictly stationary β-mixing process under mild regularity conditions, extending the learning guarantees to this broader setting. A regression variant, BlockBoost.R2, is derived and applied to Value at Risk forecasting on four equity indices across three markets: the WAEMU regional exchange, the CAC 40, and the S&P 500. Evaluated over 2014–2026 via standard backtesting procedures, BlockBoost achieves superior unconditional coverage at the 5% confidence level across all markets, with the advantage most pronounced in frontier markets where parametric distributional assumptions are misspecified. At the 1% level, GARCH models with heavy-tailed innovations retain an edge, attributable to their capacity to extrapolate into the extreme tail from limited data. These results establish BlockBoost as a theoretically grounded and empirically competitive alternative to the GARCH family for financial risk measurement.

[1]All related empirical testing procedures, codes, data and results are available at `https://github.com/Guerzoniansu/BlockBoost`

1

Introduction

The collapse of Barings Bank in 1995, the collapse of Long-Term Capital Management in 1998, and the subprime crisis of 2008 are all painful episodes that reminded the financial world of a fundamental truth: misjudging risk means suffering the consequences. These successive crises profoundly reshaped financial institutions' approach to market risk management and accelerated the adoption of more rigorous quantitative measures, foremost among which is Value at Risk (VaR). Defined as the maximum potential loss of a portfolio over a given time horizon and for a fixed confidence level, VaR is now the essential benchmark tool in the risk management systems of banking and financial institutions around the world [56].

It was precisely under the impetus of the Basel Accords that VaR became established as an international prudential standard. The amendment to Basel I in 1996 formalised its use for calculating regulatory capital requirements for market risk, allowing banks to use internal models subject to validation by supervisory authorities [6]. Basel II (2004) consolidated this architecture by relying on a VaR calculated at a 99 per cent confidence level and a ten-day rolling horizon [7]. In response to the shortcomings revealed by the 2008 crisis, particularly the underestimation of tail losses and the procyclicality of capital requirements, Basel III and the fundamental review of the trading book gradually shifted regulation towards Expected Shortfall while maintaining VaR at the heart of operational practices [11]. These developments reflect a growing demand for precision in risk modelling and, therefore, an ongoing race to find better methods for forecasting VaR.

In the West African Economic and Monetary Union (WAEMU) zone, this international dynamic has not gone unnoticed. The Central Bank of West African States (BCEAO), as part of its mission to ensure regional financial stability, has embarked on a process of gradual convergence towards the prudential standards set out in the Basel Accords. The revised prudential framework, governed by Instruction No. 026-11-2016 on the capital adequacy of credit institutions and financial companies in the WAEMU, has raised the solvency ratio and introduced stricter risk governance requirements. The BCEAO's 2021–2025 Strategic Plan explicitly reaf-

firms the need to strengthen the resilience of the regional financial system to exogenous shocks, in a context where the financial stability of the WAEMU zone remains exposed to significant structural vulnerabilities. In this changing regulatory environment, the ability of financial institutions in the zone to produce reliable VaR estimates is not only a prudential requirement but also a real strategic challenge. It is in this context that the following question becomes central: which tools enable accurate and efficient measurement of VaR, and do the models selected work in a comparable way on financial markets that are fundamentally different in nature?

To answer this question, this study sets itself a general objective: to provide a rigorous and effective method for measuring VaR in contrasting financial markets, combining the contributions of financial econometrics and machine learning. This general objective is broken down into three specific objectives. First, to apply GARCH models and their extensions to calculate VaR and evaluate their performance on the WAEMU, US, and French markets. Second, to develop a new method for calculating VaR based on a novel time series variant of AdaBoost: BlockBoost. Third, to apply this boosting approach in order to evaluate and compare its performance with that of GARCH models in these same markets, thereby enabling an informative intercontinental comparative analysis. This work thus aims to make a twofold contribution. On the one hand, it makes a methodological contribution by developing an original boosting approach for VaR forecasting and rigorously comparing it with GARCH models and their extensions, all evaluated using recognized backtesting tests. On the other hand, it makes an empirical and regional contribution by producing results on the BRVM that are currently lacking in French-language and African scientific literature on quantitative financial risk management, thereby responding to a real need among practitioners and regulators in the WAEMU zone in a context of gradual compliance with Basel standards.

In the broader landscape of VaR forecasting, a wide range of methods have been explored. On the one hand, classical econometric approaches, notably GARCH-type models and their numerous extensions, have long been the workhorse for volatility-based VaR calculations. Their strength lies in their interpretability and their ability to capture stylized facts of financial returns such as volatility clustering. On the other hand, the rise of machine learning has seen a surge in the application of black-box models, including Long Short-Term Memory networks (LSTMs), recurrent neural networks (RNNs), gradient boosting frameworks such as XGBoost and LightGBM, all of which have demonstrated competitive predictive performance. Yet, despite their empirical success, these machine learning approaches are often applied in a heuristic manner, with limited theoretical guarantees regarding their generalization ability in the context of time series data. This stands in contrast to the long-standing tradition of financial econometrics, where models are typically grounded in rigorous asymptotic theory. The question then arises: can we design a machine learning method for VaR forecasting that is not only empirically effective but also comes with strong theoretical guarantees?

This is precisely the direction we pursue in this paper. We approach the VaR forecasting problem from a boosting standpoint, but with a distinct emphasis on theoretical underpinnings. Boosting algorithms, particularly AdaBoost, have been a significant development in machine learning theory. A simple description in a single sentence of the underlying principle of boosting would "be transforming weak learners into strong learner" hence the name "boosting", with very specific definition of what makes a learner "weak" or "strong" in the context. In 1990, Robert Schapire formulated a constructive proof for the famous problem defined in simple terms "Is weak learnability equivalent to strong learnability?" [90] posed by Valiant and Kearns in [58], essentially answering in the affirmative. In [44], Freund and Schapire proposed the adaptive learning algorithm namely AdaBoost which showed how we can efficiently transform weak learners into strong ones. Many works followed the proposal of AdaBoost as it revealed itself to have excellent capacities to resist overfitting while also giving good prediction qualities and generalization error. However, it has been proven that while AdaBoost resists overfitting

well, it is not immune to it [48]. Specifically, in noisy environments, AdaBoost is known to eventually overfit as we run it indefinitely i.e. the number of training rounds $T \to \infty$ while the number of training examples n is fixed. This behavior eventually inspired the investigation of regularization techniques for AdaBoost in order to prevent overfitting in noisy environments [72, 103, 94, 86]. Furthermore, the overfitting behavior in AdaBoost occurs because of the loss function which is used for the minimization problem i.e. exponential loss. Considering this, robust regularizations of AdaBoost have been proposed by investigating the change of the loss function and/or the update rule defined [45, 43].

Crucially, the theoretical guarantees of AdaBoost rest on the assumption that observations are independent and identically distributed (i.i.d.). In the context of time series, particularly financial returns, this assumption is clearly violated. Yet, the analysis of temporality in statistical learning theory is not a new field; numerous works have established generalization bounds under specific mixing conditions [72, 64, 74, 75]. Lozano et al. [72] established the consistency of regularized boosting for β-mixing processes. While their result is asymptotic, this validates the use of boosting for stationary β-mixing time series. More recent work by Gao et al. [46] and Vidyasagar [99] has further explored the equivalence of weak and strong learnability for stationary mixing sequences, providing a theoretical foundation for applying boosting in such settings. This line of work suggests that if one is willing to assume that the data generating process is stationary and β-mixing, then a boosting algorithm can be designed with provable learning guarantees.

In this paper, we build on these theoretical foundations to construct a boosting algorithm for VaR forecasting that is not a mere heuristic but comes with learning guarantees. We call this algorithm BlockBoost. The core idea is to respect the temporal structure of the data by enforcing weight constancy over time blocks, thereby reducing the Rademacher complexity of the weight sequence and making the algorithm amenable to theoretical analysis in the stationary β-mixing setting. However, we recognize that the assumption of strict stationarity is often too strong for financial time series, which may exhibit trends, structural breaks, or time-varying volatility regimes. To bridge this gap, we relax the stationarity assumption by considering locally stationary processes, a class that encompasses many real-world financial series. We prove that under mild conditions, a locally stationary process with independent innovations can be approximated in L^2 by a stationary β-mixing process. This approximation result, formalized in Theorem 4.1, allows us to extend the learning guarantees from the stationary setting to the locally stationary one, provided the algorithm is used for one-step-ahead prediction.

At this point, BlockBoost as a classifier provides a theoretically grounded method. But for VaR forecasting, we need a regressor that can predict the conditional quantile of returns. We therefore construct a regression version, BlockBoost.R2, by extending the heuristic of AdaBoost.R2 [37], which, despite its heuristic origin, is known to have strong empirical performance. This extension retains the theoretical guarantees of the classification version while adapting to the regression setting.

What about the case where the data is so non-stationary that the locally stationary approximation fails? For such completely non-stationary sequences with wild behavior, we cannot rely on the same approximation. In this scenario, we instead adopt an online learning perspective following the framework of Kuznetsov and Mohri [64]. We show that the blocking strategy in BlockBoost introduces a trade-off: depending on the nature of the sequence, the block size can either diminish or increase the generalization bound. This analysis provides a practical guideline for choosing the block size in highly volatile environments.

Finally, we empirically validate BlockBoost.R2 on three contrasting markets: the WAEMU regional stock exchange (BRVM), the US market, and the French market. We compare its performance against traditional GARCH-type models, demonstrating that our theoretically grounded boosting approach achieves competitive, and often superior, predictive accuracy while

providing the kind of theoretical guarantees that have been lacking in previous machine learning applications to VaR forecasting. This work thus contributes both methodologically, by introducing a novel boosting algorithm with learning guarantees for time series, and empirically, by providing a comprehensive evaluation across developed and emerging markets.

2

Background and Preliminaries

2.1 Boosting algorithms

In the classical PAC setting, we assume a domain $\mathcal{X}$, a label space $\mathcal{Y}$, and a fixed but unknown distribution $\mathcal{P}$ over $\mathcal{X}$. The learner receives a sample $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ drawn independently and identically distributed ($X \sim \mathcal{P}$, $Y = c(X)$) from $\mathcal{P}$. The goal is to select a hypothesis h from a class $\mathcal{H}$ such that its generalization error $R(h) = \mathbb{P}_{X\sim\mathcal{P}}[h(X) \neq c(X)]$ is minimized. We denote by $c : \mathcal{X} \to \mathcal{Y}$ the target function or concept, that is $c(X) = Y$. The concept class $\mathcal{C}$ is the set of possible target functions.

Definition 2.1 (Strong PAC-learnability)**.** A concept class $\mathcal{C}$ is strongly PAC-learnable if there exists an algorithm $\mathcal{A}$ such that, for any $c \in \mathcal{C}$, any distribution $\mathcal{P}$ over $\mathcal{X}$, for all $\varepsilon > 0$ and all $\delta \in (0, 1)$, if $\mathcal{A}$ is given access to an oracle $EX(c, \mathcal{P})$ returning labeled examples $(x, c(x))$, then it outputs a hypothesis h for which, with probability at least $1 - \delta$,

$$R(h) = \mathbb{P}_{X\sim\mathcal{P}}[h(X) \neq c(X)] \leq \varepsilon$$

Moreover, the algorithm runtime must be polynomial in $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$ (and also polynomial in the dimension of the input d, though we'll forego mentioning this latter requirement as it is orthogonal to our current study).

We define weak learnability by enforcing two additional weakening requirements.

Definition 2.2 (Weak PAC-learnability)**.** A concept class $\mathcal{C}$ is weakly PAC-learnable if there exists an algorithm $\mathcal{A}, \gamma > 0$ and $\delta_0 \in (0, 1)$ such that, for any $c \in \mathcal{C}$, any distribution $\mathcal{P}$ over $\mathcal{X}$, if $\mathcal{A}$ is given access to an oracle $EX(c, \mathcal{P})$ returning $n_0 = N(\gamma, \delta_0)$ labeled examples $(x, c(x))$, then it outputs a hypothesis h for which, with probability at least $1 - \delta_0$,

$$R(h) = \mathbb{P}_{X\sim\mathcal{P}}[h(X) \neq c(X)] \leq \frac{1}{2} - \gamma$$

Within this framework, Kearns and Valiant [58] posed a fundamental question regarding the hierarchy of learnability: Does the existence of a "weak" learner (an algorithm that performs only slightly better than random guessing) imply the existence of a "strong" learner that can achieve arbitrary accuracy? Schapire [90] answered this question in the affirmative and provided a constructive proof that laid the groundwork for boosting. The result implies that by iteratively re-weighting the training distribution to focus on "hard" examples, a combination of weak hypotheses can be boosted into a strong classifier with arbitrarily low error. This theoretical breakthrough led to the development of AdaBoost (Freund and Schapire [43]), which realizes this amplification efficiently.

2.1.1 The adaptive boosting algorithm AdaBoost

Here, we describe the general learning scenario considered in most theoretical exploration of AdaBoost [44].

Let $\mathcal{X}$ denote the feature space and $\mathcal{Y} = \{-1, +1\}$ the output space. We assume we are given a weak learning algorithm WeakLearn which weak learns $\mathcal{C}$ with edge γ and success probability δ_0 and a training sample $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$. AdaBoost proceeds in rounds, constructing one weak hypothesis h_t in each of T rounds. In round t of AdaBoost, the weak learner WeakLearn is called on a reweighted version of the training sample with weights specified by a distribution D_t over examples $\{1, \ldots, n\}$, $D_t \in \Delta_n$ where Δ_n is the n-probability simplex. The weights are chosen carefully to give more emphasis to examples which were misclassified in the past. By the weak learning assumption, the weak learner will return some hypothesis h_t for which $\varepsilon_t := \hat{R}(h_t) = \mathbb{P}_{j \sim D_t}[h_t(x_j) \neq y_j] \leq \frac{1}{2} - \gamma_t$ for some $\gamma_t \geq \gamma$ (with probability at least δ_0). The following pseudo-code (in Algorithm 1) describes the adaptive boosting algorithm.

Algorithm 1 AdaBoost

Initialize $D_1(j) = \frac{1}{n}$ for $j = 1, \ldots, n$
for $t = 1, \ldots, T$ **do**
    Call base learner WeakLearn on distribution D_t, obtaining hypothesis h_t
    Compute error $\varepsilon_t = \sum_{j=1}^n D_t(j) \mathbb{1}(h_t(x_j) \neq y_j)$
    Set $\alpha_t = \dfrac{1}{2} \log\left(\dfrac{1 - \varepsilon_t}{\varepsilon_t}\right)$
    for $j = 1, \ldots, n$ **do**
        $D_{t+1}(j) = \dfrac{D_t(j) \exp(-\alpha_t y_j h_t(x_j))}{Z_t}$ where $Z_t = \sum_{j=1}^n D_t(j) \exp(-\alpha_t y_j h_t(x_j))$
    end for
end for
Output hypothesis:

$$h(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Friedman et al. [45] showed that AdaBoost implicitly minimizes the exponential loss $L(x, y) = \exp(-yf(x)))$, a convex surrogate for the non-differentiable 0-1 loss. The empirical risk is $\hat{\mathcal{L}}(f) = \sum_i \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i)\right)$. At each iteration t, AdaBoost performs greedy stagewise minimization: given ensemble f_{t-1}, it solves

$$(\alpha_t, h_t) = \arg\min_{\alpha, h} \sum_i D_t(i) \exp(-\alpha y_i h(x_i))$$

where $D_t(i) \propto \exp(-y_i f_{t-1}(x_i))$. The optimal h_t minimizes weighted error ε_t; the optimal step

size is $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$. Its convergence and generalization analysis was subsequently done by Freund and Schapire [44], and Schapire et al. [91].

Theorem 2.1 (Freund and Schapire, 1997, Theorem 6 [44]). *Let ε_t be the weighted error of the hypothesis h_t at round t and let $\gamma_t = \frac{1}{2} - \varepsilon_t$ be the edge of the learner. The empirical training error of the final ensemble $h(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$ is bounded by:*

$$\mathbb{P}_{i\sim D_1}(h(x_i) \neq y_i) \leq \prod_{t=1}^{T} Z_t \leq \exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$$

where $\mathbb{P}_{i\sim D_1}(h(x_i) \neq y_i) = \frac{1}{n}\sum_i \mathbb{1}(h(x_i) \neq y_i)$ and $Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$ is the normalization factor at round t.

Theorem 2.2 (Freund, Schapire, Bartlett, Lee, 1998, Theorem 1 [91]). *Let $\mathcal{S}$ be a sample of n examples chosen independently at random according to identical distribution $\mathcal{D}$. Assume that the base hypothesis space $\mathcal{H}$ is finite and let $\delta > 0$. Then with probability at least $1-\delta$ over the random choice of the training set $\mathcal{S}$, every weighted average function $f \in \mathbf{C}$ satisfies the following bound for all $\theta > 0$:*

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[Yf(X) \leq 0\right] \leq \hat{\mathbb{P}}_{\mathcal{S}}\left[yf(x) \leq \theta\right] + O\left(\frac{1}{\sqrt{n}}\left(\frac{\ln n \ln|\mathcal{H}|}{\theta^2} + \ln\left(\frac{1}{\delta}\right)\right)^{\frac{1}{2}}\right)$$

More generally, for finite or infinite $\mathcal{H}$ with VC-dimension d, the following bounds holds as well:

$$\mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[Yf(X) \leq 0\right] \leq \hat{\mathbb{P}}_{\mathcal{S}}\left[yf(x) \leq \theta\right] + O\left(\frac{1}{\sqrt{n}}\left(\frac{d\ln^2\left(\frac{n}{d}\right)}{\theta^2} + \ln\left(\frac{1}{\delta}\right)\right)^{\frac{1}{2}}\right)$$

where $\hat{\mathbb{P}}_S[yf(x) \leq \theta] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(y_i f(x_i) \leq \theta)$ is the empirical fraction of training examples with margin less than θ.

Theorem 2.3 (Freund, Schapire, Bartlett, Lee, 1998, Theorem 1 [91]). *Suppose the base learning algorithm, when called by AdaBoost, generates hypotheses with weighted training error $\varepsilon_1, \ldots, \varepsilon_T$. Then, for any θ, we have that*

$$\hat{\mathbb{P}}_S[yf(x) \leq \theta] \leq 2^T \prod_{t=1}^{T}\sqrt{\varepsilon_t^{1-\theta}(1-\varepsilon_t)^{1+\theta}}$$

2.1.2 Robust variants of AdaBoost

AdaBoost's exponential loss $\phi(z) = e^{-z}$ assigns unbounded penalty to misclassified examples. Consider a single mislabeled instance k where the true margin is $K > 0$ but the observed margin is $-K$. The weight at iteration $t+1$ satisfies $D_{t+1}(k) \propto e^K$. As the ensemble improves on clean data ($K \to \infty$), the algorithm exponentially upweights this single outlier, distorting the decision boundary to accommodate noise. Long and Servedio [69] formalized this failure:

Theorem 2.4 (Long and Servedio 2010 [69]). *For any convex potential function ϕ and noise rate $0 < \eta < 1/2$, the corresponding booster B_ϕ does not tolerate random classification noise at rate η; neither via global minimization, early stopping, nor L_1 regularization.*

Prior work proposed two mitigation strategies: (1) Surrogate losses like LogitBoost [45] replaces exponential loss with logistic loss $\ln(1 + \exp(-2yf(x)))$, which grows linearly (not exponentially) for negative margins, bounding outlier influence. (2) Non-monotonic weighting like BrownBoost [43] models margins as Brownian particles with finite capacity to fix errors, assigning weight $w_i(t) = \exp\left(-\frac{(r_i(t)+c-t)^2}{c-t}\right)$ that decays for both very confident correct predictions and unlearnable outliers. These methods improve noise robustness but do not address temporal dependence in time series.

2.2 Value at Risk

Value at Risk (shortly, VaR) is one of the most widely adopted downside risk measures in financial risk management, introduced formally by Jorion (1997) [56] and later institutionalized through the Basel Accords as a regulatory benchmark for market risk capital requirements [6]. At its core, VaR quantifies the maximum expected loss on a portfolio or asset over a given holding period h, at a specified confidence level $\alpha \in (0, 1)$.

Formally, let r_t denote the return of a financial asset or portfolio at time t, and let $F_r(\cdot)$ denote its cumulative distribution function (CDF). The VaR at confidence level α over horizon h is formally defined as the negative of the $(1 - \alpha)$-quantile of the return distribution:

$$\mathrm{VaR}_{\alpha,h} = -\inf\{x \in \mathbb{R} : F_r(x) > 1 - \alpha\} = -F_r^{-1}(1 - \alpha),$$

so that the probability of a loss exceeding $\mathrm{VaR}_{\alpha,h}$ is bounded by the tail probability $1 - \alpha$:

$$\Pr(r_t < -\mathrm{VaR}_{\alpha,h}) = 1 - \alpha.$$

In practice, confidence levels of $\alpha = 0.95$ or $\alpha = 0.99$ are standard, the latter being mandated under the Basel II and Basel III frameworks.

The most classical and widely used approach to VaR estimation in the empirical finance literature relies on the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) framework introduced by Bollerslev [13], which extends the ARCH model of Engle [39].

2.2.1 Theoretical foundations of the GARCH model

Introduced by Bollerslev (1986) [13], the Garch model is formulated to mathematically capture conditional heteroscedasticity and volatility persistence in financial time series [8]. The model decomposes the asset return r_t into a conditional mean μ_t and an innovation ϵ_t:

$$r_t = \mu_t + \epsilon_t, \quad \epsilon_t = \sigma_t z_t$$

where the sequence $z_t \sim i.i.d.(0, 1)$ is typically assumed to follow a standard normal or Student's t-distribution to accommodate the heavy tails characteristic of financial returns [62, 98].

The core dynamic of the Garch(p, q) model governs the conditional variance σ_t^2, defining it as a linear function of a constant ω, past squared innovations (Arch terms), and past conditional variances (Garch terms) [62, 105, 67]:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

To guarantee a strictly positive conditional variance and wide-sense stationarity of the process, the parameters are strictly bounded by the following constraints [62, 105]:

$$\omega > 0, \quad \alpha_i \geq 0, \quad \beta_j \geq 0, \quad \text{and} \quad \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1.$$

2.2.2 Asymmetric extensions of GARCH models

The standard GARCH formulation assumes symmetric responses to shocks. To capture the empirical "leverage effect"–where negative shocks amplify future volatility more than positive shocks of equal magnitude [21, 59]–several asymmetric extensions have been developed.

EGARCH The EGARCH model [79] models the logarithm of the conditional variance, intrinsically guaranteeing its positivity without parameter constraints [61, 60]:

$$\ln(\sigma_t^2) = \omega + \beta \ln(\sigma_{t-1}^2) + \alpha Z_{t-1} + \gamma(|Z_{t-1}| - \mathbb{E}[|Z_{t-1}|])$$

where $Z_t = \epsilon_t/\sigma_t$. The leverage effect is present when $\alpha < 0$ and $\alpha + \gamma > 0$, indicating that negative shocks generate greater volatility impacts than positive shocks [61].

GJR-GARCH The GJR-GARCH model [47] introduces asymmetry via an indicator function $I_{t-1} = \mathbb{I}_{\{\epsilon_{t-1}<0\}}$ [59, 31]:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \gamma I_{t-1}\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

A strictly positive γ confirms the leverage effect. The total variance impact is $(\alpha + \gamma)$ for negative shocks ($I_{t-1} = 1$) and α for positive ones ($I_{t-1} = 0$) [59].

APARCH The APARCH model [35] is a flexible generalization incorporating a power transformation $\delta > 0$ and an asymmetry parameter $\gamma_i \in (-1, 1)$ [87, 38, 22]:

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i(|\epsilon_{t-i}| - \gamma_i\epsilon_{t-i})^\delta + \sum_{j=1}^{p} \beta_j\sigma_{t-j}^\delta$$

This generalized structure encompasses several specifications depending on parameter restrictions. For instance, setting $\delta = 1$ and $\gamma_i = 0$ reduces the formulation to the TS-GARCH model [97, 93, 28].

Estimation method

Parameter estimation (θ) primarily utilizes the Quasi-Maximum Likelihood (QML) method due to its robustness against conditional distribution misspecifications [41, 19]. Assuming Gaussian innovations, the objective log-likelihood function is [41, 36]:

$$L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\left(\ln\sigma_t^2 + \frac{\epsilon_t^2}{\sigma_t^2}\right)$$

Under standard regularity conditions, optimizing this function yields QML estimators $\hat{\theta}$ that are convergent and asymptotically normal, ensuring reliable inference even when actual returns deviate from strict normality [41].

2.2.3 Value at Risk calculation

Within the framework of GARCH family models, the one-step-ahead Value at Risk (VaR) at a confidence level $1-\alpha$ is formulated as a linear function of the conditional moments [81, 83]:

$$\mathrm{VaR}_{\alpha,t+1} = \mu_{t+1} + Z_\alpha \sigma_{t+1}$$

The components of this equation are strictly defined as follows:

- $\mu_{t+1} = \mathbb{E}[r_{t+1}|\mathcal{F}_t]$ is the conditional expected return for period $t+1$ given the information set $\mathcal{F}_t$ available at time t. In many high-frequency financial applications, this term is assumed to be zero.
- $\sigma_{t+1} = \sqrt{\mathrm{Var}(r_{t+1}|\mathcal{F}_t)}$ is the conditional standard deviation forecast, derived deterministically from the specific volatility equation of the selected model (e.g., Standard GARCH, EGARCH, GJR-GARCH, or APARCH) using $\mathcal{F}_t$.
- $Z_\alpha = \inf\{z \in \mathbb{R} : F_Z(z) \geq \alpha\}$ is the α-quantile of the standardized innovation distribution F_Z. While the standard normal distribution $Z \sim \mathcal{N}(0,1)$ is common (yielding $Z_{0.05} \approx -1.645$ for $\alpha = 0.05$), heavy-tailed distributions such as the Student's t are frequently substituted to better model empirical financial residuals [81].

By definition, this VaR establishes the threshold loss such that $\mathbb{P}(r_{t+1} \leq \mathrm{VaR}_{\alpha,t+1}|\mathcal{F}_t) = \alpha$.

2.3 Value at Risk backtesting

Rigorous evaluation of Value at Risk (VaR) models relies on backtesting procedures that systematically analyze the sequence of violation indicators (or "hits"), defined as $I_t(\alpha) = \mathbb{I}_{\{r_t < -\mathrm{VaR}_t(\alpha)\}}$, where r_t is the asset return [80]. Under the null hypothesis of a correctly specified model, the sequence $\{I_t(\alpha)\}_{t=1}^T$ must form an independent and identically distributed (i.i.d.) Bernoulli process with success probability α [20]. To formally test this dual condition of correct unconditional coverage and temporal independence, we employ three standard likelihood-ratio (LR) tests alongside a comparative predictive accuracy test.

A comprehensive derivation of these test statistics, their likelihood functions, and asymptotic properties is relegated to Appendix D.

2.3.1 Absolute Performance: Coverage and Independence Tests

The statistical validity of the hit sequence is evaluated through three complementary tests:

- **Unconditional Coverage (Kupiec, 1995) [63]:** Evaluates whether the empirical violation rate $\hat{p} = \frac{1}{T}\sum_{t=1}^T I_t(\alpha)$ statistically equates to the theoretical level α, without considering temporal dynamics. The likelihood-ratio statistic $\mathrm{LR}_{\mathrm{UC}}$ follows a χ^2_1 distribution under $H_0 : p = \alpha$ [82, 52, 92, 68, 78, 85].
- **Independence (Christoffersen, 1998) [26]:** Models the hit sequence as a first-order Markov chain to test against temporal clustering. Let $\pi_{ij} = \mathbb{P}(I_t = j|I_{t-1} = i)$. Under the null hypothesis of independence $H_0 : \pi_{01} = \pi_{11}$, the test statistic $\mathrm{LR}_{\mathrm{IND}}$ follows a χ^2_1 distribution [24, 25, 42].

- **Conditional Coverage (Christoffersen, 1998) [26]:** A joint test of the i.i.d. Bernoulli property. Exploiting the additivity of likelihood ratios, the statistic is exactly $\mathrm{LR}_{\mathrm{CC}} = \mathrm{LR}_{\mathrm{UC}} + \mathrm{LR}_{\mathrm{IND}}$. Under the joint null hypothesis $H_0 : \pi_{01} = \pi_{11} = \alpha$, $\mathrm{LR}_{\mathrm{CC}}$ converges in distribution to χ^2_2 [82, 2, 89].

2.3.2 Relative Performance: Diebold and Mariano Test

While the Christoffersen and Kupiec frameworks evaluate absolute model adequacy, the Diebold-Mariano (DM) test (1995) [34, 33] assesses the relative predictive accuracy of two competing models.

This is achieved by analyzing the loss differential $d_t = \mathrm{L_t}(\mathrm{VaR}_{1,t}(\alpha)) - \mathrm{L_t}(\mathrm{VaR}_{2,t}(\alpha))$. In financial risk management, asymmetric loss functions are preferred to heavily penalize risk underestimation [3, 27], such as the tick loss function proposed by Lopez (1999) [70, 71]:

$$\mathrm{L_t}(\mathrm{VaR}_t(\alpha)) = \begin{cases} 1 + (r_t + \mathrm{VaR}_t(\alpha))^2 & \text{if } r_t < -\mathrm{VaR}_t(\alpha) \\ 1 & \text{otherwise} \end{cases}$$

Under the null hypothesis of equal predictive accuracy ($\mathbb{E}[d_t] = 0$), the DM statistic is constructed using the empirical mean $\bar{d}$ and a Newey-West robust variance estimator $\widehat{\mathrm{Var}}(\bar{d})$ to account for potential autocorrelation in the loss differentials [82, 88]. Asymptotically, the statistic follows a standard normal distribution:

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{under } H_0$$

2.4 Time series and learning theory

Time series forecasting plays a significant role in a number of domains ranging from weather forecasting, earthquake prediction to applications in economics and finance. The classical statistical approaches to time series analysis are based on generative models such as the autoregressive moving average (ARMA) models, or their integrated versions (ARIMA) and several other extensions [39, 13, 18, 16, 50]. Most of these models rely on strong assumptions about the noise terms, often assumed to be i.i.d. random variables sampled from a Gaussian distribution, and the guarantees provided in their support are only asymptotic. An alternative non-parametric approach to time series analysis consists of extending the standard i.i.d. statistical learning theory framework to that of stochastic processes. In this endeavor, most works focus on two of the most important properties of stochastic processes (stationarity and dependence) which are then used as starting assumptions for said extension.

In this section, we aim to provide a taxonomy of these properties and review the progress made in establishing rigorous learning guarantees for non-i.i.d. environments. Essentially, we describe the different mixing conditions of stochastic processes and introduce Rademacher complexity as a measure of hypothesis richness. We then define notions of weak and strict stationarity and finally cite and/or define some key tools we will use for our future proofs.

Throughout this section, the probability space is $(\Omega, \mathcal{F}, \mathbb{P})$.

2.4.1 Measures of dependence (mixing) and stationarity

Most of the material in this paragraph (except the stationarity definition) can be found in Bradley [17] (for furher information).

In what follows, expressions such as $\sup_{q\in Q, s\in S} h(q,s)$ will often be written as $\sup h(q,s), q \in Q, s \in S$. For any two σ-fields $\mathcal{A}, \mathcal{B} \in \mathcal{F}$, define the following measures of dependence.

$$\alpha(\mathcal{A},\mathcal{B}) := \sup |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|, \quad A \in \mathcal{A}, B \in \mathcal{B}$$
$$\phi(\mathcal{A},\mathcal{B}) := \sup |\mathbb{P}(B|A) - \mathbb{P}(B)|, \quad A \in \mathcal{A}, B \in \mathcal{B}, \mathbb{P}(A) > 0$$
$$\beta(\mathcal{A},\mathcal{B}) := \sup \sum_{i=1}^{I}\sum_{j=1}^{J} |\mathbb{P}(A_i \cap B_j) - \mathbb{P}(A_i)\mathbb{P}(B_j)|$$

where the supremum is taken over all pairs of (finite partitions) $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of $\Omega/\ A_i \in \mathcal{A}\ \forall i, B_j \in \mathcal{B}\ \forall j$.

Suppose $X := (X_k, k \in \mathbb{Z})$ is a (not necessarily stationary) sequence of random variables. The notation $\sigma(\ldots)$ means the σ-field $\subset \mathcal{F}$ generated by $(\ldots)$. For $-\infty \le J \le L \le +\infty$, define the σ-field

$$\mathcal{F}_J^L := \sigma(X_k, J \le k \le L(k \in \mathbb{Z}))$$

For each $n \ge 1$, define the following dependence coefficients

$$\alpha(n) := \sup_{j\in\mathbb{Z}} \alpha(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{+\infty}), \quad \beta(n) := \sup_{j\in\mathbb{Z}} \beta(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{+\infty}), \quad \phi(n) := \sup_{j\in\mathbb{Z}} \phi(\mathcal{F}_{-\infty}^{j}, \mathcal{F}_{j+n}^{+\infty})$$

Definition 2.3 (Strong mixing conditions)**.** The random sequence X is said to be
"strongly mixing" or α-mixing if $\alpha(n) \to 0$ as $n \to +\infty$;
"absolutely regular" or β-mixing if $\beta(n) \to 0$ as $n \to +\infty$;
ϕ-mixing if $\phi(n) \to 0$ as $n \to +\infty$

Definition 2.4 (Strict and weak stationarity)**.** The random sequence X is said to be strictly stationary if $\forall n \in \mathbb{N}, \forall (t_1, \ldots, t_n) \in \mathbb{Z}^n$ such that $t_1 < \ldots < t_n, \forall h \in \mathbb{Z}$,

$$(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h})$$

The random sequence X is said to be weakly stationary if $\mathbb{E}(X_t) = \mu \in \mathbb{R}, \mathbb{E}(X_t^2) = \mu_2 < \infty, \mathrm{Cov}(X_t, X_{t+h}) = \gamma(h)\ , \forall t \in \mathbb{Z}$.

2.4.2 Rademacher complexity

The central objective of supervised learning is to identify a hypothesis within a function class $\mathcal{H}$ that minimizes the generalization error, regardless of most aspects of the learning setting (classification or regression). Theoretical guarantees for this objective typically rely on uniform convergence bounds (most notably in the PAC model), which limit the maximal deviation between the empirical error observed on the training set and the true error of the hypothesis and ensure uniform convergence of the empirical error to the true error. Classically, these bounds have been derived using combinatorial measures of complexity, most notably the VC-dimension, and interestingly showing that by controlling the complexity of the hypothesis class $\mathcal{H}$, we can improve generalization. While foundational, VC-based bounds possess two significant limitations: they are distribution-independent, often yielding loose, worst-case estimates that ignore benign data distributions; and they are primarily restricted to binary classification tasks. To overcome these limitations, modern statistical learning theory relies on Rademacher complexity. Unlike combinatorial measures, Rademacher complexity is distribution-dependent, providing tighter bounds that reflect the statistical properties of the specific data generating

process. Furthermore, it is naturally defined for classes of real-valued functions, making it a versatile tool for analyzing both regression and multiclass classification problems.

Given a space $\mathcal{X}$ and a fixed distribution $\mathcal{P}$ over $\mathcal{X}$, let $S = \{x_1, \ldots, x_n\}$ be a set of examples drawn i.i.d. from $\mathcal{P}$. Let $\mathcal{H}$ be a class of functions $h : \mathcal{X} \to \mathbb{R}$.

Definition 2.5. The empirical Rademacher complexity of $\mathcal{H}$ is defined to be

$$\hat{\Re}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right) \right]$$

where $\sigma_1, \ldots, \sigma_n$ are independent uniform random variables $\{-1, +1\}$-valued random variables.

Definition 2.6. The Rademacher complexity of $\mathcal{H}$ is defined as

$$\Re_n(\mathcal{H}) = \mathbb{E}_{S \sim \mathcal{P}^n} \left[\hat{\Re}_S(\mathcal{H}) \right]$$

Intuitively, the supremum measures, for a given sample S, Rademacher vector σ and for all functions $h \in \mathcal{H}$, the maximum correlation between σ_i and $h(x_i)$. Therefore, by taking the expectation over the Rademacher vector, the empirical Rademacher complexity of $\mathcal{H}$ measures the ability of function of $\mathcal{H}$ (when applied to a fixed sample S) to fit random noise (in the worst case). The Rademacher complexity of $\mathcal{H}$ then measures the expected noise-fitting-ability of $\mathcal{H}$ over all samples $S \in \mathcal{X}^n$ that could be drawn according to the distribution $\mathcal{P}$.

2.4.3 The extension of statistical learning theory to stochastic processes

The fundamental problem in extending statistical learning theory to stochastic processes is the failure of concentration inequalities (Hoeffding and McDiarmid). Classical results, such as McDiarmid's inequality, rely heavily on the independence of random variables. Below is a statement of said inequality.

Theorem 2.5 (Hoeffding's inequality [53])**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Consider the sum of these random variables $S_n = X_1 + \ldots + X_n$.*

Then, for all $\varepsilon > 0$, we have

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$
$$\mathbb{P}(S_n - \mathbb{E}(S_n) \leq -\varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

Theorem 2.6 (McDiarmid's inequality)**.** *Consider independent random variables $X_1, \ldots, X_n \in \mathcal{X}$ and a mapping $\phi : \mathcal{X}^n \to \mathbb{R}$. If there are constants $c_1, \ldots, c_n$ such that for all $i \in \{1, \ldots, n\}$, and for all $x_1, \ldots, x_n \in \mathcal{X}$, the function ϕ satisfies*

$$\sup_{x_i' \in \mathcal{X}} |\phi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - \phi(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

then for any $\varepsilon > 0$

$$\mathbb{P}(\phi(X_1, \ldots, X_n) - \mathbb{E}[\phi(X_1, \ldots, X_n)] \geq \varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right)$$

$$\mathbb{P}(\phi(X_1,\ldots,X_n) - \mathbb{E}[\phi(X_1,\ldots,X_n)] \leq -\varepsilon) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

The earliest and most prominent approach assumes the data generating process is stationary and satisfies a mixing condition, ensuring that the dependence between events decays as their temporal separation increases. The cornerstone of this analysis is the *independent block technique*, formalized by Yu [102] to extend uniform convergence results to β-mixing processes.

Consider a sample $X_1^n = (X_1,\ldots,X_n)$ from a stationary β-mixing sequence of random variables taking values in $\mathcal{X}$. The purpose of the blocking technique used by Yu [102] is to transform a sequence of dependent variables X_1^n into a subsequence of nearly i.i.d. ones. Let a_n and μ_n be non-negative integers such that $2a_n\mu_n = n$. Now divide X_1^n into $2\mu_n$ blocks, each of length a_n. Identify the blocks as follows:

$$U_j = \{X_i : 2(j-1)a_n + 1 \leq i \leq (2j-1)a_n\}$$
$$V_j = \{X_i : (2j-1)a_n + 1 \leq i \leq 2ja_n\}$$

Let $\mathbf{U_n} = \{U_j : j = 1,\ldots,\mu_n\}$ be the entire sequence of odd blocks U_j and let $\mathbf{V_n} = \{V_j : j = 1,\ldots,\mu_n\}$ be the sequence of even blocks V_j. Finally, let $\mathbf{U'_n}$ be a sequence of blocks which are independent of X_1^n but such that each block has the same distribution as a block from the original sequence:

$$U'_j \overset{D}{=} U_j \overset{D}{=} U_1$$

The blocks $\mathbf{U'_n}$ are now an i.i.d. block sequence so standard results apply. (See [102] for a more rigorous analysis of blocking.)

Lemma 2.7 (Yu, Lemma 4.1, [102])**.** *Let $\mathcal{Q}$ be the distribution of $\mathbf{U_n}$. Let $\mathcal{Q}'$ be the distribution $\mathbf{U'_n}$. For any measurable function h on $\mathbb{R}^{a_n\mu_n}$ with bound M,*

$$\left|\mathbb{E}_{\mathcal{Q}}[h(\mathbf{U_n})] - \mathbb{E}_{\mathcal{Q}'}[h(\mathbf{U'_n})]\right| \leq M(\mu_n - 1)\beta(a_n)$$

This lemma effectively allows us to transfer results from the i.i.d. setting to the mixing setting, paying a penalty proportional to the mixing rate $\beta(a_n)$.

2.5 Limitations of boosting in time series and motivations

As we have discussed in the introduction of this chapter, boosting originates from the PAC model and, therefore, inherits from its limitations for time series. Specifically, the theoretical guarantees provided by standard boosting algorithms (AdaBoost particularly) rely on the assumption that the generating process is stationary and independent, i.e., sample examples are drawn i.i.d. according to a fixed but unknown distribution. In time series forecasting, this assumption is violated by design; observations exhibit temporal dependence, and the underlying data-generating process may drift over time (non-stationarity). We have discussed the consequences of random noise on the generalization error of regular boosting algorithms such as AdaBoost. As established by Long and Servedio [69], convex potential boosters are notoriously brittle in the presence of noise (a pervasive feature of many stochastic sequences especially in financial markets). In an attempt to satisfy the margin condition on stochastic outliers, the algorithm exponentially upweights these noise points, corrupting the generalization performance. Furthermore, modern learning theory bounds the generalization error in terms of the empirical risk and the Rademacher complexity of the hypothesis class (See Bartlett, Mendelson [5]). The

validity of these bounds hinges on the concentration of measure (McDiarmid's inequality) which requires that changing a single data point (x_i, y_i) alters the empirical error by at most $O(\frac{1}{n})$. In dependent time series, this condition fails, a single "innovation" or "shock" at time t propagates to future observations. Applying i.i.d. bounds to such data leads to overly optimistic bounds, causing the model to underestimate the true risk and overfit temporal correlations that will not persist in the future.

Despite these fundamental theoretical flaws, the application of boosting to time series in the literature has largely remained an engineering endeavor rather than a theoretical one. Approaches often involve augmenting the input space with lagged features or technical indicators while treating the underlying boosting algorithm as a "black box" (e.g., Chang et al. [23]). While empirically useful, these methods do not address the breakdown of the convergence proofs. They rely on the hope that the algorithm will inherently manage dependence, without providing a mechanism to guarantee it. This paper seeks to bridge this gap. We argue that to rigorously apply boosting to time series, one must move beyond feature engineering and intervene in the algorithmic structure itself. We propose a modification, BlockBoost, designed to respect the mixing properties of the stochastic process, thereby recovering the theoretical guarantees of generalization that standard methods forfeit.

3

Boosting for Time Series Prediction

The theoretical underpinnings of boosting trace back to the Probably Approximately Correct (PAC) learning framework, where the central question of whether weak learnability implies strong learnability was first formulated by Kearns and Valiant [58] and later answered in the affirmative by Schapire [90]. This equivalence, however, was established under the assumption of independent and identically distributed data. Extensions to dependent processes have since been developed, with Vidyasagar [99] and Gao et al. [46] demonstrating that weak and strong learnability remain equivalent for stationary β-mixing sequences. Building on these results, we construct BlockBoost, a boosting algorithm designed for stationary β-mixing time series, and provide its learning guarantees. Whether BlockBoost achieves optimality in this setting is a question we leave for future work. Finally, we provide empirical validation of the block boosting algorithm demonstrating its practical effectiveness.

3.1 Boosting with the block weighting scheme

In the context of time series, due to the temporal dependence, the "hardness" of an example is often the property of surrounding examples (temporal regime) rather than the individual instance. For example, in financial market, a "hard" example typically corresponds to a high volatility regime and therefore if a model struggles to predict at a given time t, there is a high probability it will also struggle to predict at time $t + 1$. Naively applying AdaBoost to such data leads to weight collapsing, where the algorithm overfits to specific short-term trends by exponentially upweighting consecutive correlated errors.

Based on the established theory of learning for time series, we propose a, we believe, new boosting algorithm that incorporates a slight, yet powerful modification of the original algorithm: BlockBoost. This new algorithm is a theoretical solution for AdaBoost designed to overcome its failures in the context of time series forecasting. The intuition of BlockBoost is basically moving the "atomic unit" of learning from individual observations (x_t, y_t) to con-

tiguous time blocks. By assessing performance and updating weights at the block level, the algorithm learns to identify and correct difficult regimes while preserving the internal temporal structure necessary for the weak learner to capture local dynamics. Unlike AdaBoost which uses point-wise weighting, BlockBoost uses block-wise weight update. That is, the algorithm groups time dependent observations into contiguous blocks and weights blocks instead of points.

The main motivation for this algorithm roots from the works of Lozano et al. [72] (which proved the consistency of regularized boosting for stationary β-mixing sequences) and Yu [102] (with the blocking technique in lemma 2.7 originating from Bernstein [10]).

Let $\mathcal{X}$ and $\mathcal{Y}$ denote respectively the feature space and the output space. We first assume we are in a binary classification scenario i.e. $\mathcal{Y} = \{-1; +1\}$. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ the product space and assume we have a stationary β-mixing stochastic process $(Z_t = (X_t, Y_t))_{t\in\mathbb{Z}}$ in $\mathcal{Z}$ such that $Z_t \sim \mathcal{D}, \forall t$. Consider the general learning scenario where the learner receives a sequence $Z_1^T = (Z_1, \ldots, Z_T)$ realizations of the stochastic process $(Z_t)_t$. Note that, in the context of time series, the feature vector can be the collection of past d output observations, that is $X_t = (Y_{t-1}, \ldots, Y_{t-d})$ for some d. The goal of the learner at each round of boosting is to select out of a specified family H, a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that achieves the smallest error given current distribution. We then proceed by boosting to derive the hypothesis $f \in \mathcal{H}$ that minimizes the expected block risk and achieves the smallest generalization error with respect to the stationary distribution. Let $0 < a < T$ be a non-negative and non-zero integer and consider the sequence $Z_1^a = (Z_1, \ldots, Z_a)$. The objective function is defined by the following.

$$\mathcal{L}(f) = \mathbb{E}_{Z_1^a \sim \mathcal{D}_a} L[f(X_1), \ldots, f(X_a), Y_1, \ldots, Y_a]$$

where $L : \mathcal{Y}^a \times \mathcal{Y}^a \to [0, M]$ is a loss function with bound $M > 0$. To abbreviate the notation, we will often write $L(f, z_1^a) = L[f(x_1), \ldots, f(x_a), y_1, \ldots, y_a]$ for any $z_1^a = (z_1, \ldots, z_a) \in \mathcal{Z}^a$. Note that $\mathcal{D}_a$ is the joint distribution of the sequence Z_1^a and $\mathcal{D}_1 = \mathcal{D}$. We denote by $\mathcal{H}$ the convex hull of H as the set of hypotheses that can be generated by weighted average of hypotheses in H.

$$\mathcal{H} = \left\{ f : x \mapsto \sum_{h\in H} \alpha_h h(x) : \alpha_h \geq 0, \sum_h \alpha_h = 1 \right\}$$

where it is understood that only finitely α_h's may be nonzero, therefore the minimization problem $f^* = \arg\min_{f\in\mathcal{H}} \mathcal{L}(f)$. In what follows, we will denote by $z_1^T = (z_1, \ldots, z_T)$ where $z_t = (x_t, y_t)$ the realizations of the sequence Z_1^T. Let a_T and μ_T be non-negative integers such that $2a_T\mu_T = T, a_T \to \infty$ and $\frac{a_T}{T} \to 0$ as $T \to \infty$. We divide the index sequence $1, \ldots, T$ using the blocking technique presented earlier and construct $K_T = 2\mu_T = \left\lfloor \frac{T}{a_T} \right\rfloor$ blocks:

$$\mathrm{B_k} = \{t : (k-1)a_T < t \leq ka_T\}, \quad k = 1, \ldots, K_T$$

BlockBoost operates at two weighting levels: the block level $\mathrm{W_k}$ and the instance level $\mathrm{D_t}$. The difference being that it adapts block weights while keeping $\mathrm{D_t}$ uniform within blocks (preserving local structure).

Below is the pseudo-code of BlockBoost where for precision and simplicity purposes, we denote $L \equiv a_T$ and $K \equiv K_T$ and we assume that T is a multiple of L such that there is no remainder to the division $\frac{T}{L}$. From this pseudo-code, it is easy to remark that BlockBoost generalizes AdaBoost with the introduction of blocking.

Algorithm 2 BlockBoost

Require: Sequence of T time-indexed examples $((x_1, y_1), \ldots, (x_T, y_T))$
Number L of contiguous examples in a block
Integer M specifying the number of boosting rounds
Weak learning algorithm WeakLearn

Partition $\{1, \ldots, T\}$ into $K = \left\lfloor \frac{T}{L} \right\rfloor$ contiguous blocks

$$B_k = \{t : (k-1)L < t \leq kL\}, \quad k = 1, \ldots, K$$

Initialize: Block weights $W_k^{(1)} = \frac{1}{K}$ for $k = 1, \ldots, K$

Instance weights $D_1(t) = \frac{1}{T}$ for $t = 1, \ldots, T$

for $m = 1, \ldots, M$ **do**

  Call WeakLearn, providing it with distribution D_m; get back hypothesis $h_m = \arg\min_{h \in H} \sum_{t=1}^{T} D_m(t) \mathbb{1}(h(x_t) \neq y_t)$

  Compute block-wise errors $r_k^{(m)} = \frac{1}{|B_k|} \sum_{t \in B_k} \mathbb{1}(h_m(x_t) \neq y_t)$

  Compute weighted block error $\varepsilon_m = \sum_{k=1}^{K} W_k^{(m)} r_k^{(m)}$

  if $\varepsilon_m \geq 1/2$ or $\varepsilon_m = 0$ **then**

    break

  end if

  Set $\alpha_m = \frac{1}{2} \log\left(\frac{1 - \varepsilon_m}{\varepsilon_m}\right)$

  Update and normalize block weights $W_k^{(m+1)} \propto W_k^{(m)} \exp\left(\alpha_m (2 r_k^{(m)} - 1)\right)$

  Distribute block weights to instances $D_{m+1}(t) = \frac{W_k^{(m+1)}}{|B_k|}, \quad \forall t \in B_k$

end for

Output ensemble predictor

$$f(x) = \operatorname{sgn}\left(\sum_{m=1}^{M} \alpha_m h_m(x)\right)$$

3.2 Analysis

The theoretical analysis of BlockBoost naturally extends the framework of AdaBoost, pivoting from instance-level to block-level error minimization. Fundamentally, BlockBoost operates as a regularized variant of its predecessor. It restricts the algorithm's weight space from the unconstrained probability T-simplex $\Delta_T = \{w \in \mathbb{R}^T : w \geq 0, \sum_i w_i = 1\}$ to a block-constant subspace $\Delta_T^B \subset \Delta_T$, formally defined as:

$$\Delta_T^B := \{w \in \Delta_T : \forall k = 1, \ldots, K_T, \forall i, j \in B_k, w_i = w_j\}$$

This restriction explicitly alters the bias-variance trade-off of the boosting procedure. While AdaBoost searches the full simplex to aggressively maximize the margin, the constrained optimization of BlockBoost prevents full margin maximization. Consequently, BlockBoost achieves a strictly greater (or equal) empirical training error for a fixed number of rounds, serving as a structural regularization against overfitting. The satisfiability of the weak learning assumption is preserved. Because any block distribution D_m generated by BlockBoost satisfies $D_m \in \Delta_T^B \subset \Delta_T$, the standard unconstrained weak learning guarantees universally bound

the restricted space from below:

$$\inf_{D\in\Delta_T^B}\sup_{h\in\mathcal{H}}\left(\frac{1}{2}-\mathbb{E}_{x\sim D}[\mathbb{1}(h(x)\neq y)]\right)\geq\inf_{D\in\Delta_T}\sup_{h\in\mathcal{H}}\left(\frac{1}{2}-\mathbb{E}_{x\sim D}[\mathbb{1}(h(x)\neq y)]\right)\geq\gamma$$

where $\gamma\in(0,1/2)$. To accommodate this block-level optimization, the standard exponential loss function $L_{\exp}(f,z)=\exp(-yf(x))$ minimized by AdaBoost is generalized to the block exponential loss function:

$$L(f,z_1^{a_T})=\exp\left[-\frac{1}{a_T}\sum_{t=1}^{a_T}y_tf(x_t)\right]$$

To formalize the convergence properties under this loss, let $r_k(f):=\frac{1}{a_T}\sum_{t\in\mathrm{B_k}}\mathbb{1}[f(x_t)\neq y_t]$ denote the empirical instance error rate of hypothesis f within block k. We distinguish the standard instance-wise empirical error $\hat{R}(f)$ from the newly introduced block-wise empirical error $\hat{R}_B(h)$:

$$\hat{R}(f)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}[f(x_t)\neq y_t],\quad\hat{R}_B(h)=\frac{1}{K_T}\sum_{k=1}^{K_T}\mathbb{1}\left[\sum_{t\in\mathrm{B}_k}y_th(x_t)\leq 0\right]$$

The following proposition establishes the algorithmic convergence of BlockBoost. By demonstrating that the block exponential loss serves as a strictly convex surrogate for the block-wise error $\hat{R}_B(h)$, we prove that minimizing this loss effectively drives the block-level misclassification rate to zero.

Proposition 3.1. *Suppose the weak learning algorithm WeakLearn, when called by BlockBoost, generates hypotheses with errors $\varepsilon_1,\ldots,\varepsilon_M$ (as defined in algorithm 2. Then the block-wise error $\hat{R}_B(h)$ of the ensemble $h=\sum_m h_m$ is bounded above by*

$$\hat{R}_B(h)\leq\prod_{m=1}^{M}2\sqrt{\varepsilon_m(1-\varepsilon_m)}$$

If we, furthermore, assume $\varepsilon_m\leq\frac{1}{2}-\gamma_m$ where $\gamma_m\geq\gamma>0$ is the edge of WeakLearn at step m, then, we have

$$\hat{R}_B(h)\leq\exp\left(-2\sum_{m=1}^{M}\gamma_m^2\right)\leq\exp\left(-2M\gamma^2\right)$$

Proof. The proof follows the classical structure of convergence analysis by Freund and Schapire [44]. See Appendix A.1 for the full proof. □

This proposition essentially tells us that during learning, BlockBoost drives the block error $\hat{R}_B(h)$ to zero. Note however that convergence of block error to zero does not imply convergence of instance error to zero. This property can be viewed as a good thing or in another way as a flaw in the algorithm. The good thing arising from this is that, because we know BlockBoost cannot drive instance wise error to zero and because of the block averaging, we may suspect (to some extent) that the algorithm cannot overfit its data conditional on the presence of noise (which reduces the Bayes risk itself). For example, in presence of noise at a rate 10%, the maximum obtainable accuracy in the training process is the Bayes optimal accuracy 90%. But for the algorithm to be able to avoid overfitting, we would need the weak learner to

be weak enough to only pick up the signal and avoid noise. Consider now a worse case scenario where at the very first boosting round, the weak learner is able to find true labels of slightly higher than half of observations in every single block such that $\forall k, r_k^1 \approx 0.49$ (a_T would need to be large enough for such case to emerge). In such a case, $\varepsilon_1 \approx 0.49$ and therefore $\alpha_1 \approx 0.04$ which is indeed extremely small and will result in very small weight updates, this means that on the next round, we will make very little improvement in the boosting process. Recursively, this would lead to a particular case in which BlockBoost takes only microscopic steps at each round. In fact, considering a_T large, if the weak learner satisfies the weak learning assumption ($\varepsilon_1 \approx 0.49$), by the law of large numbers, the instance error in each block converges to the global average which leads to $r_k^{(1)} \approx 0.49$ and therefore to the microscopic update mentioned earlier. This suggests that we would want the block size to be as small as possible to not fall into the law of large numbers scope during execution. In what follows, we will prove that we also need the block size to be large enough to ensure learning properties to hold in particular generalization.

3.2.1 Instance wise learning and robustness

The following lemma traces the idea of the robustness of BlockBoost. Consider again the stationary β-mixing stochastic process $(Z_t = (X_t, Y_t))_{t\in\mathbb{Z}}$ and the sample sequence $z_1^T, z_t = (x_t, y_t)$ which is the observed sequence of the random sequence Z_1^T and define $\tilde{z}_1^T, \tilde{z}_t = (x_t, \tilde{y}_t)$ where $\tilde{y}_t = (1 - 2\xi_t)y_t$, $\xi_t \sim \mathcal{B}(\eta)$ i.i.d. and independent of the process (random classification noise). If f is the final hypothesis generated by BlockBoost after training on sequence $\tilde{z}_1^T$ for a given number of rounds, the value $\hat{R}(f) = \frac{1}{T}\sum_{t=1}^T \mathbb{1}[f(x_t) \neq \tilde{y}_t]$ defined as earlier refers the instance-wise error of f. We then define $\hat{R}^*(f) = \frac{1}{T}\sum_{t=1}^T \mathbb{1}[f(x_t) \neq y_t]$ which refers to the instance-wise error with respect to the true classification labels.

Lemma 3.2 (Instance error convergence)**.** *Let $z_1^T, z_t = (x_t, y_t)$ be a stationary β-mixing sequence with coefficients $\beta(k) \to 0$, and let $\tilde{y}_t = (1 - 2\xi_t)y_t$ where $\xi_t \sim \mathcal{B}(\eta)$ i.i.d. and independent of the process. Let f_M be the final hypothesis generated by BlockBoost after M rounds with block size a_T on a noisy data $(x_t, (1 - 2\xi'_t)y_t)$ and evaluated on noisy data $\tilde{z}_1^T, \tilde{z}_t = (x_t, \tilde{y}_t)$ with $\{\xi'_t\}_t \perp\!\!\!\perp \{\xi_t\}_t, \xi'_t \sim \mathcal{B}(\eta)$ i.i.d. and independent of the process. Then, with probability at least $1 - \delta$:*

$$\left|\hat{R}(f_M) - \eta - (1 - 2\eta)\hat{R}^*(f_M)\right| \leq \sqrt{\frac{\log(2/\delta)}{2T}}$$

Proof. See Appendix A.1 □

Lemma 3.2 is a standard result in learning theory in the presence of noise. Essentially, this result gives a convergence property of the instance error on noisy data to that of the clean data associated and completes the result suggesting the convergence of the block error to zero.

Corollary 3.3. *Under the same conditions as in Lemma 3.2, we have that with probability at least $1 - \delta$:*

$$\left|\limsup_{M\to\infty} \hat{R}(f_M) - \eta - (1 - 2\eta)\limsup_{M\to\infty} \hat{R}^*(f_M)\right| \leq \sqrt{\frac{\log(2/\delta)}{2T}}$$

The latter equation is a direct consequence of the lemma. Essentially it gives us a better characterization of the noise robustness of BlockBoost. Ideally, we would like the upper limit of the true instance error to be equal to 0 meaning that the clean problem is learnable and is

captured perfectly by the hypothesis class and the algorithm does not overfit the noise. This would require hypothesis class of sufficiently low complexity. Alternatively to such ideal result, we would want that at any noise rate $\eta \in [0, 1/2)$, $\limsup_{M\to\infty} \hat{R}^*(f_M)$ is strictly less than $1/2$, otherwise, we fall into the very well known "bad" result of the collapse of convex boosters under random classification noise (Long and Servedio[69]). Note that the convergence result only tells us about the plateau value, it says nothing about whether BlockBoost can even avoid catastrophic overfitting during training.

The next section is devised to prove an implicit regularization due to the blocking technique applied in BlockBoost under sufficiently low hypothesis class complexity.

3.2.2 Implicit regularization

The following proposition claims the existence of a fixed lower bound on the error achievable by a weak hypothesis at any round during the training process. We start by assuming the existence of a "hard" block B_{k^*} i.e. there exists at least one block $\inf_{h\in H} r_{k^*}(h) \geq \frac{1}{L}$ for L large enough.

Proposition 3.4. *Let H be a hypothesis class and let the training data be corrupted by random classification noise with rate η. If there exists at least one block B_{k^*} such that $\inf_{h\in H} r_{k^*}(h) \geq \frac{1}{a_T}$, then the weighted block error ε_m in BlockBoost is strictly bounded away from zero by ϵ^*, where $\epsilon^*(1-\epsilon^*)^{a_T-1} = 2^{-a_T}$.*

Proof. We analyze the evolution of the weight of the unrealizable block k^*:

$$\frac{W_{k^*}^{(m+1)}}{W_{k^*}^{(m)}} = \frac{\exp(\alpha_m(2r_{k^*}^{(m)} - 1))}{2\sqrt{\varepsilon_m(1-\varepsilon_m)}} \geq \frac{\exp\left(\alpha_m\left(\frac{2}{a_T} - 1\right)\right)}{2\sqrt{\varepsilon_m(1-\varepsilon_m)}} = \frac{\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right)^{\frac{1}{a_T}-\frac{1}{2}}}{2\sqrt{\varepsilon_m(1-\varepsilon_m)}} = \frac{1}{2}\frac{(1-\varepsilon_m)^{\frac{1}{a_T}-1}}{\varepsilon_m^{\frac{1}{a_T}}}$$

Assume now that $\liminf_{m\to\infty} \varepsilon_m = 0$. We have $\exists (m_j)_{j\geq 1} / \lim \varepsilon_{m_j} = 0$. By analysing the function $x \mapsto (1 - \frac{1}{a_T})\log(1-x) + \frac{1}{a_T}\log x + \log 2$, we prove that for sufficiently small δ,

$$\begin{aligned}
&\frac{W_{k^*}^{(m_j+1)}}{W_{k^*}^{(m_j)}} \geq \frac{1}{2}\frac{(1-\delta)^{\frac{1}{a_T}-1}}{\delta^{\frac{1}{a_T}}} =: c \geq 1, \forall m_j \geq m_1 \\
&\Rightarrow W_{k^*}^{(m_j)} \geq W_{k^*}^{(m_1)} c^{m_j - m_1} \to \infty
\end{aligned}$$

which is a contradiction since $W_{k^*}^{(m)} \leq 1, \forall m$. Therefore the sequence ε_m cannot converge to zero. There exists ϵ^* such that

$$\frac{1}{2}\frac{(1-\epsilon^*)^{\frac{1}{a_T}-1}}{(\epsilon^*)^{\frac{1}{a_T}}} = 1 \Rightarrow \epsilon^*(1-\epsilon^*)^{a_T-1} = 2^{-a_T}$$

For $a_T = 2, \epsilon^* = \frac{1}{2}$ and for $a_T > 2, \epsilon^* \in (0, \frac{1}{a_T})$ exists and $\liminf_{m\to\infty} \varepsilon_m \geq \epsilon^* > 0$

□

The implications of Proposition 3.4 are foundational, as it establishes a strict upper bound on the learning step size α_m. By ensuring that $\varepsilon_m \geq \epsilon^*$, the algorithm intrinsically avoids the degenerate overfitting regime where $\alpha_m \to \infty$, yielding the global bound:

$$\frac{1}{2}\log\left(\frac{1+2\gamma}{1-2\gamma}\right) \leq \alpha_m \leq \frac{1}{2}\log\left(\frac{1-\epsilon^*}{\epsilon^*}\right)$$

The validity of this regularization mechanism, however, hinges entirely on the satisfiability of the assumption $\inf_{h \in H} r_{k^*}(h) \geq \frac{1}{a_T}$. The likelihood of encountering such a strictly unrealizable block B_{k^*} is a direct function of the block size a_T, the complexity of the hypothesis class H, and the noise rate η.

To understand the behavioral spectrum of this assumption, consider the degenerate case where $a_T = 1$, which reduces BlockBoost to standard AdaBoost. In this regime, the assumption demands that $\inf_{h \in H} \mathbb{1}(h(x_{k^*}) \neq y_{k^*}) \geq 1$. This requires the existence of a singleton instance that cannot be correctly classified by any hypothesis in H. For any standard hypothesis class (including those with $\text{VCdim}(H) = 0$, provided they contain basic constant classifiers), a single point can always be shattered. Consequently, the assumption trivially fails for $a_T = 1$. This theoretical failure perfectly aligns with the empirical observation that standard AdaBoost exhibits unconstrained step sizes and is highly susceptible to memorizing noisy instances. Conversely, consider the minimal non-trivial blocking case of $a_T = 2$. The assumption requires a block where $\inf_{h \in H} r_{k^*}(h) \geq \frac{1}{2}$, meaning no hypothesis can correctly classify both instances simultaneously. According to Proposition 3.4, this yields $\epsilon^* = \frac{1}{2}$, which tightly constrains the maximum step size to $\alpha_m \leq \frac{1}{2}\log(1) = 0$. This extreme penalization typically occurs if the block contains two identical, or geometrically proximate, instances with contradictory labels; a scenario whose probability scales with the noise rate η. Although it strictly prevents overfitting, setting $a_T = 2$ may impose an overly aggressive regularization that prematurely halts the boosting process, as the algorithm cannot allocate positive weight to perfectly resolve other regions of the feature space.

Therefore, the assumption $\inf_{h \in H} r_{k^*}(h) \geq \frac{1}{a_T}$ is most theoretically sound and practically natural in the regime where $1 < \text{VCdim}(H) < a_T \ll T$. By the Sauer-Shelah lemma, whenever the block size strictly exceeds the VC-dimension of the weak learner, H can no longer shatter the block. In the presence of classification noise $\eta > 0$, the probability that a block of size a_T contains a topologically unrealizable labeling configuration approaches 1 exponentially fast as a_T increases. Under these conditions, the existence of an unrealizable block is no longer a restrictive assumption but a guaranteed high-probability property of the sampled data. This shows that BlockBoost directly leverages the localized finite capacity of H to intrinsically bound the step size, thereby converting the structural limitations of the weak learner into a guaranteed uniform stability bound for the ensemble.

These results suggest that BlockBoost is noise-robust at the instance level, and the blocking mechanism intrinsically prevents the degenerate overfitting behavior of AdaBoost by bounding the step sizes. What remains to be asked is whether this robustness carries over to the final hypothesis when evaluated on unseen data.

It is easy to verify that BlockBoost does not verify the conditions necessary to observe Long and Servedio's failure[69] mainly because it violates the independent instance-wise optimization. But this does not tell us it doesn't eventually collapse. In this paper, we only provide empirical evidence of BlockBoost's robustness. As for whether or not it eventually falls in the population-wise collapse as proved by Long and Servedio [69] is left as future work.

3.3 Generalization

Up to now, we have discussed the capacity of BlockBoost to drive the block risk to zero and the robustness guarantees it provides. Here, we discuss the error of the final hypothesis outside the training set. The Proposition 3.1 guarantees that the block risk in the sample is small and Lemma 3.2 gives us the guaranties of robust learning, but the quantity that interests us is the generalization error of f, which is the error of f over the whole instance space $\mathcal{X}$. In what follows, we require that the quantity $K_T \beta(a_T) \to 0$ as $T \to \infty$.

Theorem 3.5 (Generalization bound)**.** *Let $z_1^T, z_t = (x_t, y_t)$ be a stationary β-mixing sequence with coefficients $\beta(k) \to 0$ and stationary marginal μ. Let f_M be the final hypothesis generated by* BlockBoost *after M rounds with block size a_T, where $K_T = \frac{T}{a_T}$. Then, with probability at least $1 - \delta - K_T\beta(a_T)$,*

$$\left|\hat{R}^*(f_M) - R^*(f_M)\right| \leq 2\Re_{K_T}(H, \mu) + \sqrt{\frac{a_T \log(2/\delta)}{2T}}$$

Proof. See Appendix A.2 □

Lemma 3.6. *Assume the label noise rate satisfies $\eta \in [0, 1/2)$. With probability at least $1 - \delta - \delta' - K_T\beta(a_T)$, the empirical noisy risk $\hat{R}(f_M)$ and the true clean risk $R^*(f_M)$ satisfy:*

$$\left|\hat{R}(f_M) - \eta - (1-2\eta)R^*(f_M)\right| \leq \sqrt{\frac{\log(2/\delta)}{2T}} + (1-2\eta)\left(2\Re_{K_T}(H, \mu) + \sqrt{\frac{a_T \log(2/\delta')}{2T}}\right)$$

Notice that, on the contrary of AdaBoost, the Rademacher complexity employed here is based on the effective sample size $K_T = \frac{T}{a_T}$. This is the direct consequence of the "cost" one pays because of the dependence between the sample examples. We recall the proper formulation of said Rademacher complexity and derive the bound in terms of VC dimension.

$$\Re_{K_T}(H, \mu) = \mathbb{E}_{z_1'^T \sim \mu^T}\left[\mathbb{E}_\sigma\left[\sup_{h \in \mathcal{H}}\left(\frac{1}{K_T}\sum_{k=1}^{K_T}\sigma_k \bar{h}\left(\mathrm{B_k}\right)\right)\right]\right], \bar{h}\left(\mathrm{B_k}\right) = \frac{1}{a_T}\sum_{t \in \mathrm{B_k}} \mathbb{1}(h(x_t') \neq y_t')$$

where $z_1'^T, z_t' = (x_t', y_t')$ is the ghost sample derived from the sequence z_1^T constructed in proof of Theorem3.5. If H has VC-dimension d, by Massart's lemma followed by Sauer-Shelah cardinality bound

$$\mathbb{E}_\sigma\left[\sup_{h \in \mathcal{H}}\left(\frac{1}{K_T}\sum_{k=1}^{K_T}\sigma_k \bar{h}\left(\mathrm{B_k}\right)\right)\right] \leq \sqrt{\frac{2d\log(eT/d)}{K_T}} = \sqrt{\frac{2da_T\log(eT/d)}{T}}$$

The new bounds with VC-dimension follow directly from this last statement. The following statement provides an upper bound on the VC-dimension of the class of final hypotheses generated by AdaBoost in terms of the weak hypothesis class. Note that this is a direct consequence of [44] Theorem 8.

Theorem 3.7. *If the hypotheses generated by* WeakLearn *are chosen from a hypothesis class of VC-dimension $d \geq 2$, then the final hypothesis generated by* BlockBoost *after M iterations belong to a class of VC-dimension at most $2(d+1)(M+1)\log_2[e(M+1)]$ (where e is the base of the natural logarithm)*

We redirect the reader to Freund and Schapire [44] for a complete proof of this result. As for the choice of the block size a_T, we recall that we want large enough block size such that to guarantee independence of blocks but we also want small enough block size such that to avoid falling into the law of large numbers. One may choose (as a theoretical value of) the optimal block size such as to minimize the upper bound of the generalization bound, that is minimizing the worst case risk. That been said, the optimal theoretical value may be loose as it indeed is a worst case scenario and we may have at hand a scenario that is not as bad and therefore requires larger or smaller block size.

The theoretical objective that results from the generalization bound is defined by

$$J(a_T) = (1-2\eta)\left(\sqrt{\frac{8da_T\log(eT/d)}{T}} + \sqrt{\frac{a_T\log(2/\delta)}{2T}}\right) + K_T\beta(a_T)$$
$$= C_\eta\sqrt{\frac{a_T}{T}} + \frac{T\beta(a_T)}{a_T} \text{ where } C_\eta = (1-2\eta)\left(\sqrt{8d\log(eT/d)} + \sqrt{\frac{\log(2/\delta)}{2}}\right)$$

the goal being to minimize $J(a)$ under the constraint that $1 \leq a \ll T$. This optimal value is only useful if the hypothesis generated is quite good at the task it was assigned since minimizing $J(a)$ only guarantee that the empirical instance error of the hypothesis is probably close to the true error. Note that in a real world setting, this is fairly hard to compute as it requires at least knowing the noise rate η and the mixing rate $\beta(a_T)$.

3.4 Empirical experimentation

In this section, we present the materials and methods used to evaluate BlockBoost. First, we will define the prediction task. Second, the time series data sets are described. We then formalize the methodology employed for performance estimation. The experimental design is in Appendix C.1.

3.4.1 Predictive task

In this first experimental case, we evaluate BlockBoost on a time series classification problem. We consider that we are given a temporal sequence of labels $Y = \{y_1, \ldots, y_T\}, y_t \in \{-1, +1\}$ and covariates $X = \{x_1, \ldots, x_T\}, x_t \in \mathbb{R}^d$. Similarly as in classical learning theory, the objective here is to predict y_{T+l} knowing x_{T+l} for some l, the objective function (final hypothesis) being estimated as a function of past occurrences of both sequences X, Y.

Note that feature vector x_t may be the past d occurences of a given process $(X_t)_t$. In a general form, the dataset can be formalized as the following matrix

$$Y_{[T,d]} = \left[\begin{array}{ccccc|c} x_{1,1} & x_{1,2} & \ldots & x_{1,d-1} & x_{1,d} & y_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{t,1} & x_{t,2} & \ldots & x_{t,d-1} & x_{t,d} & y_t \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{T,1} & x_{T,2} & \ldots & x_{T,d-1} & x_{T,d} & y_T \end{array}\right]$$

3.4.2 Time series data

We use synthetic data sampled from two types of data generating processes: one auto-regressive with lag 1 (**DGP$_1$**) and the other, a Long-Servedio type dataset designed to stress AdaBoost and observe its collapse (**DGP$_2$**). The data generating processes are all stationary and are designed as follows:

DGP$_1$: 10 (seeds) random samples of an AR(1) sequence of dimension d. The training sample size is $T_{train} = 2000$ and the test sample size is $T_{test} = 1000$. If we denote by $(X_t)_{t\in\mathbb{Z}}$ the said AR sequence then, we have with $X_{0,j} \sim \mathcal{N}(0, \sigma^2), j \in \{1, \ldots, d\}$:

$$X_t = \rho X_{t-1} + \sigma\sqrt{1-\rho^2}\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \mathbb{I}_5)$$

Thus, each coordinate is a stationary AR(1) sequence with $\mathrm{Corr}(X_{t,j}, X_{t-k,j}) = \rho^k$. Denote by S_t the linear signal in components (features) of the AR process and defined by $S_t = X_{t,1} + 0.5X_{t,2} - 0.3X_{t,3}$. We introduce temporal dependence by creating a new AR sequence Z_t defined by $Z_0 = S_0, Z_t = S_t + 0.2Z_{t-1}, t \geq 1$. Finally, we derive the labels as the sign of Z_t.

$$Y_t = \mathrm{sgn}(Z_t)$$

We set $\sigma = 0.3$ and $\rho = 0.95$. The latter was set to be close enough to 1 in order to simulate extreme temporal dependence between two consecutive observations.

DGP$_2$: 20 (seeds) random samples of $N = 10000$ i.i.d. observations in $\mathbb{R}^2$ drawn from a three-component Gaussian mixture: $N_A = 2500$ samples from $\mathcal{N}([1,0]^\top, \sigma^2\mathbb{I}_2)$ (Cluster A), $N_B = 5000$ from $\mathcal{N}([\gamma, -\gamma]^\top, \sigma^2\mathbb{I}_2)$ (Cluster B) and $N_C = 2500$ from $\mathcal{N}([\gamma, 20\gamma]^\top, \sigma^2\mathbb{I}_2)$ (Cluster C), with $\gamma = 0.15$ and $\sigma = 0.2$. All clean labels satisfy $y = +1$. We used the first $N_{train} = 1000$ samples as the training set and the remaining $N_{test} = 9000$ as a clean holdout. Training labels received i.i.d. symmetric label flips at rate $\eta = 0.1$ (100 random flips based on the seed).

3.4.3 Empirical results

Figure 3.1 details the comparative generalization trajectories of BlockBoost ($a_T = 10$) and AdaBoost.M1 across 1500 boosting rounds under varying classification noise rates $\eta \in \{0, 0.1, 0.2, 0.3\}$, rigorously validating the theoretical bias-variance trade-offs induced by spatial block-averaging on dependent sequences. In the strictly noiseless regime ($\eta = 0$), AdaBoost.M1 behaves as an optimal variance-minimizing algorithm on a clean, learnable problem, achieving near-perfect interpolation on the training set (accuracy ≈ 1.0) and a superior clean test accuracy of ≈ 0.97, whereas the structural regularization of BlockBoost imposes a deliberate, albeit unnecessary, bias that bounds its test accuracy at ≈ 0.94. Note however that the gap between the train and test set in this setting is far less visible in BlockBoost than it is in AdaBoost.

However, as the corruption probability η increases, the exponential weight amplification of standard boosting becomes pathological; under moderate noise ($\eta = 0.1, 0.2$), AdaBoost.M1 greedily chases mislabeled instances, leading to a widening train-test divergence, whereas BlockBoost tightly couples its training and testing curves to strictly control variance. At $\eta = 0.2$, this mechanism enables BlockBoost to dominate with a clean test accuracy of ≈ 0.91 (accompanied by substantially tighter confidence bands) compared to ≈ 0.89 for AdaBoost.M1. This structural advantage persists into highly corrupted regimes ($\eta = 0.3$), where, despite the natural theoretical degradation of both algorithms as they approach the uninformative boundary of $\eta = 0.5$, BlockBoost maintains a clean test accuracy of ≈ 0.89 against AdaBoost.M1's ≈ 0.86 while entirely avoiding the latter's severe overfitting. Ultimately, these dynamics formalize an explicit empirical bias-variance trade-off: by mathematically aggregating empirical errors at the block rather than the sample level, BlockBoost dilutes the influence of isolated noise and ensures consistency across initializations, thereby sacrificing peak interpolative performance under idealized conditions in exchange for robust, strictly bounded generalization under realistic distributional corruption.

Figure 3.2 and Table 3.1 present results of the data generating process 2 evaluation for block size $a_T = 10$. The contrast in the behavior of both algorithms gives us an empirical validation of the earlier proposition 3.4. AdaBoost.M1's noisy train accuracy climbs steadily toward 0.98, confirming it is actively memorizing the corrupted training set. BlockBoost's noisy

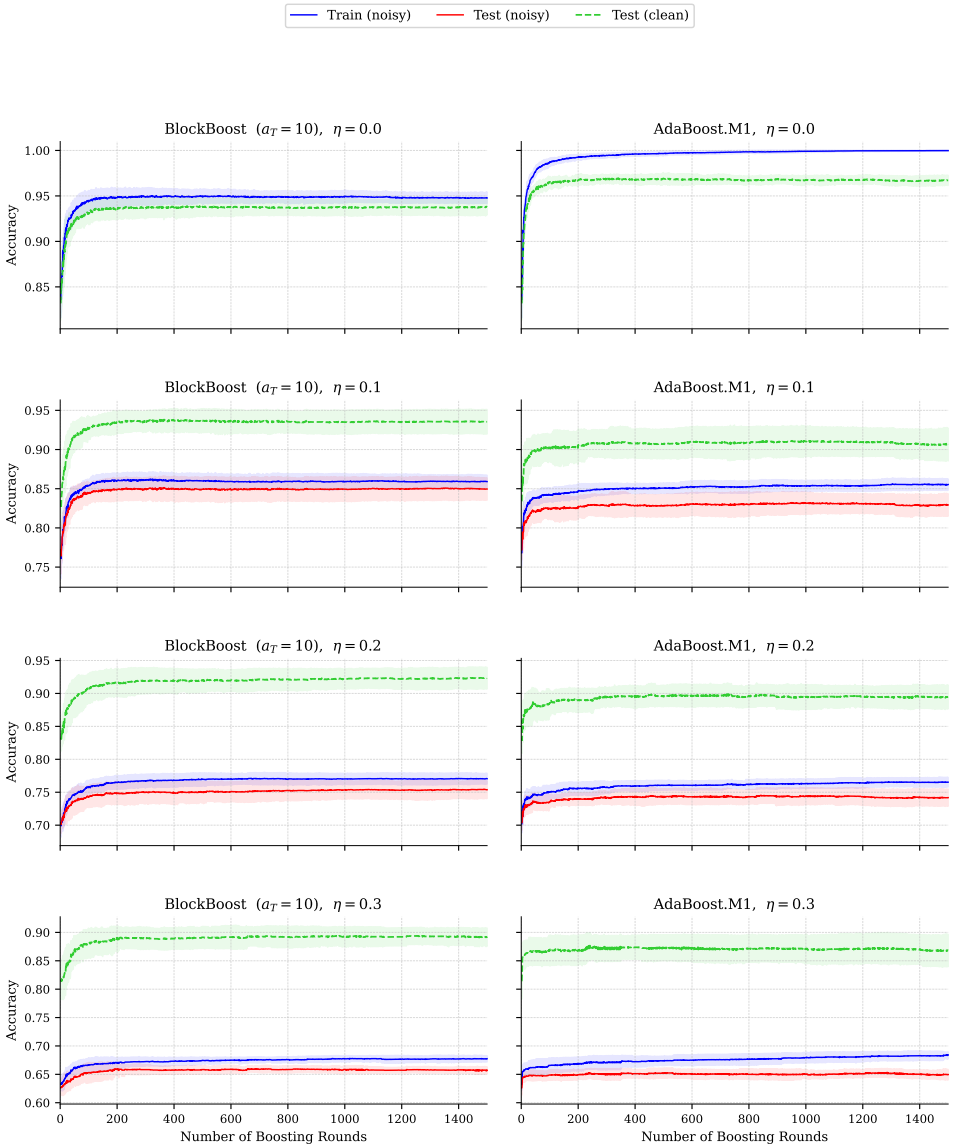


Figure 3.1: Generalization Trajectories under Random Classification Noise. Performance of BlockBoost ($a_T = 10$) versus AdaBoost.M1 on the AR(1) sequence ($\mathbf{DGP_1}$) across noise rates η. Curves denote accuracy on corrupted training (solid blue), corrupted test (solid red), and clean holdout (dashed green) sets. In the noiseless regime ($\eta = 0$), BlockBoost's spatial pooling introduces a mild approximation bias. However, under stochastic corruption ($\eta > 0$), this block-averaging acts as a strict structural regularizer. By filtering high-frequency outliers, BlockBoost completely circumvents the catastrophic noise memorization and margin explosion endemic to AdaBoost, preserving robust generalization.

train accuracy remains flat near 0.90 throughout (essentially avoiding overfitting, consistent with noise level $\eta = 0.1$) which directly explains its superior clean generalization.

On the clean test, BlockBoost maintains stable generalization throughout, with clean holdout accuracy declining only marginally from 0.997 to 0.991 over 5000 boosting rounds. AdaBoost.M1 degrades monotonically and severely, dropping from 0.997 to 0.965, with widening confidence bands indicating increasing instability. The divergence is progressive and accelerating, consistent with runaway overfitting to corrupted labels.

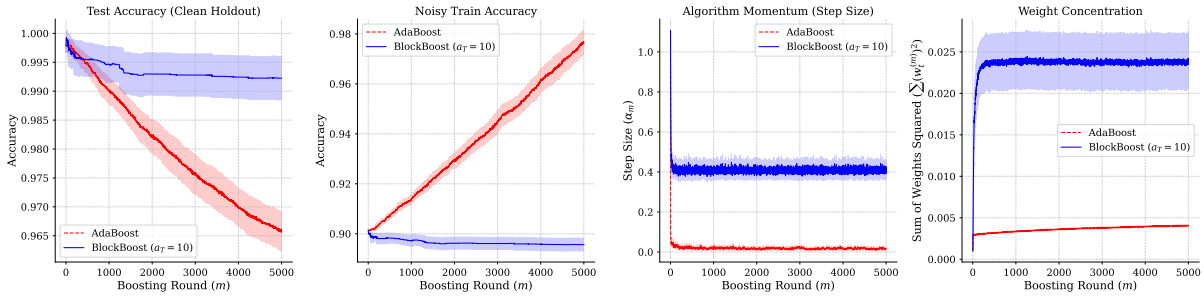


Figure 3.2: Optimization Dynamics under Adversarial Geometry ($\mathbf{DGP_2}$). Comparative evaluation of BlockBoost ($a_T = 10$) and standard AdaBoost over $M = 5000$ rounds with $\eta = 0.1$ training label corruption. The unconstrained AdaBoost actively memorizes the random classification noise (rising noisy train accuracy), triggering continuous generalization decay on the clean holdout and a catastrophic collapse of the step size ($\alpha_m \to 0$). In contrast, BlockBoost physically manifests the theoretical guarantees of Proposition 3.4: the spatial filter bounds the algorithmic momentum away from zero and strictly caps empirical risk minimization at the true signal threshold (≈ 0.90), ensuring robust and stable generalization.

Athough these results may not show a direct collapse of the convex booster AdaBoost.M1 in the Long and Servedio sense, they do show a consistent and slow degradation of the adaptive boosting algorithm while its blocked counterpart shows slow degradation across the same number of boosting rounds.

Table 3.1: Comparison of AdaBoost.M1 and BlockBoost on the adversarial DGP for different block sizes a_T. Values report mean test accuracy ± standard deviation over all runs, where test accuracy is reported at the last boosting round i.e. the 5000th one. Statistical significance is evaluated using a one-sided Wilcoxon signed-rank test (H_1: BlockBoost > AdaBoost.M1) AdaBoost.M1 corresponds to the special case $a_T = 1$.

Block size a_T	Test Accuracy (mean ± sd)	Wilcoxon statistic	p-value
1	0.9659 ± 0.0072		–
10	$\mathbf{0.9922 \pm 0.0085}$	210.00	4.42×10^{-5}***
15	0.9875 ± 0.0201	190.00	3.54×10^{-4}***
20	0.9891 ± 0.0159	201.00	3.15×10^{-5}***
25	0.9802 ± 0.0271	163.50	1.45×10^{-2}*
30	0.9855 ± 0.0208	184.00	1.59×10^{-3}**

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

4

Extensions of the Block Boosting Algorithm

In the previous chapter, we introduced the block boosting algorithm as a classifier method, theoretically justified for stationary β-mixing sequences. But, in practice this particular type of generating process and time series classification problems are rather rare. Most application of time series are regression type (with or without covariates). Furthermore, the assumption that the data generating process is stationary β-mixing sequence is hard to verify in real world data. For example, a dominating characteristics of financial time series in general is their weak stationarity. And thus, one would want to extend theoretical guarantees for BlockBoost into the case where the series is weakly stationary. But this is hard to obtain as the discrepancy between a strongly stationary and one that is weakly stationary is "huge" theoretically. In this paper, we approach this problem through the lense of a generalized version of locally stationary sequences [30] and propose a novel approach to bound the generalization error when the generating sequence is locally stationary with some mild assumptions.

This chapter is dedicated to the extension of BlockBoost to real world case, first regression tasks and second to locally stationary sequences. The remaining of this chapter is organized as follows. Section 4.1 shows that under some mild assumptions, a locally stationary sequence can be L^2-approximated by a stationary (β-mixing) sequence with a sample decaying cost that we will define. Section 4.2 proposes the extension of BlockBoost to regression tasks and its generalization bounds for both stationary β-mixing and locally stationary sequences.

4.1 Approximation of locally stationary sequences

In practice, the stationary and β-mixing condition may be too strong. In particular, some processes are only weakly stationary (which is the case of most financial return time series) and therefore break the assumption completely even considered the β-mixing condition holds. Therefore, one can extend these definitions to the case of locally stationary sequences (Dahlhaus [29]). In particular, in this paper, we focus on the more technical part of the work, that is,

proving that given any sequence from the class of locally stationary β-mixing sequences, under some assumptions, we can couple it with a non-trivial stationary β-mixing sequence.

Under the equivalence between weak and strong learnability in the stationary β-mixing scenario, we conjecture that this also implies weak-strong equivalence for locally stationary β-mixing sequences and work with that principle in the rest of this work.

4.1.1 Locally stationary sequences

In Dahlhaus (1997) [30], a sequence of stochastic processes $\{X_{t,T}\}_{t=1,\dots,T}$ is called locally stationary with transfer function A^0 and trend μ if there exists a representation:

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} \exp(i\lambda t) A^0_{t,T}(\lambda) d\xi(\lambda)$$

where the following holds

(i) $\xi(\lambda)$ is a stochastic process on $[-\pi, \pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ with some other regular assumptions, we refer the reader to [30] for precision.

(ii) There exists a constant K and a 2π-periodic function $A : [0,1] \times \mathbb{R} \to \mathbb{C}$ with $A(u, -\lambda) = \overline{A(u,\lambda)}$ and

$$\sup_{t,\lambda} \left| A^0_{t,T}(\lambda) - A\left(\frac{t}{T}, \lambda\right) \right| \leq KT^{-1}$$

for all T; $A(u,\lambda)$ and $\mu(u)$ are assumed to be continuous in u.

In the standard Dahlhaus definition, the stochastic integrator $d\xi(\lambda)$ is only required to be a process with orthogonal increments: $\mathbb{E}[d\xi(\lambda)] = 0$, $\mathbb{E}[d\xi(\lambda_1)\overline{d\xi(\lambda_2)}] = 0$ for $\lambda_1 \neq \lambda_2$ and $\mathbb{E}[|d\xi(\lambda)|^2] = d\lambda$. We start here by noting the reader that these conditions are only sufficient to prove that the locally stationary framework is designed to extend weak stationarity into a more general case (or at least naturally). In fact, considering the degenerate case where the transfer function $A(u,\lambda) = A(\lambda)$ and the mean $\mu(u) = \mu$, that is, they do not depend on time, we have:

$$X_{t,T} = \mu + \int_{-\pi}^{\pi} A(\lambda) \exp(i\lambda t) d\xi(\lambda)$$

It is easy to see that $\mathbb{E}(X_{t,T}) = \mu = \text{cste}$ and $\gamma(h) = \text{Cov}(X_{t+h,T}, X_{t,T}) = \int |A(\lambda)|^2 \exp(i\lambda h) d\lambda$ (which depends only on h) and therefore that the sequence is weakly stationary. Another representation of this sequence is given by the following (See [29])

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} a_{t,T}(j)\varepsilon_{t-j} \quad \text{where} \quad A(u,\lambda) = \sum_{j=-\infty}^{\infty} a(u,j)\varepsilon_{t-j}$$
$$\text{and} \quad \sup_j \left| a_{t,T}(j) - a\left(\frac{t}{T}, j\right) \right| \leq K$$

$\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a sequence satisfying $\mathbb{E}(\varepsilon_t) = 0$, $\mathbb{E}(\varepsilon_t^2) = 1$ i.e. a weak white noise.

4.1.2 Approximation by the stationary sequence

Ideally, to achieve the extension from stationary β-mixing sequences to that of locally stationary β-mixing sequences, considering that locally stationary sequences are generalized cases of weakly stationary sequences, we would want a general bound that gives a good approximation of a locally stationary β-mixing sequence by a stationary β-mixing sequence.

But, because weakly stationary sequences are parts of the space of locally stationary sequences, that would mean that we would want to approximate a weakly stationary sequence by one that is stationary. There exists some cases where this is trivial, especially those where weak stationarity implies strong learnability, for example gaussian processes. And there are also trivial (and somewhat naive) ways to approximate a weakly stationary sequence, that is, by taking any stationary gaussian sequence with the same mean and variance and couple it independently but the bound derived from this is just a crude and loose one as it does not capture meaningful information about the underlying sequence. However, in a general case, this is much more difficult and non-trivial to get as this requires additional structure and/or assumptions on the weakly stationary sequence itself.

Of course, with the additional assumption that $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is a strong white noise, the approximation is not useful because, the degenerate case is not only weakly stationary anymore but rather stationary. And therefore, one can easily construct the stationary approximation as in [29]:

$$\tilde{X}_t(u) = \mu(u) + \sum_{j=-\infty}^{\infty} a(u,j)\varepsilon_{t-j}$$

In this paper, we give some of these conditions and give under those conditions a proof of the construction of such a non-trivial stationary sequence. Let us consider that the sequence of stochastic processes $X_{t,T}$ has the more general representation where the innovation sequence is itself a triangular array, that is:

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} a_{t,T}(j)\varepsilon_{t-j,T} \tag{4.1}$$

$\{\varepsilon_t\}_{t\in\mathbb{Z},T\geq 1}$ is a sequence of mutually independent, real-valued random variables. We now elaborate on the assumptions required for the approximation.

(A1) There exists a function $a : [0,1]\times\mathbb{Z}\to\mathbb{R}$ such that:

$$\text{(i)} \sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} |a(u,j)| < \infty, \text{ (ii)} \sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} |j||a(u,j)| < \infty, \text{ (iii)} \sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} |\partial_u a(u,j)| < \infty$$
$$\text{and (iv)} \left|a_{t,T}(j) - a\left(\frac{t}{T},j\right)\right| \leq b_j T^{-1} \text{ with } \sum_j |b_j| < \infty$$

(A2) The process $\varepsilon_{t,T}$ has finite second-moments $\sup_{t\in\mathbb{Z}} \mathbb{E}(\varepsilon_{t,T}^2) < \infty$

(A3) Let $P_{t,T} = \mathbb{L}(\varepsilon_{t,T})$ the distribution of $\varepsilon_{t,T}$. There exists $L_\varepsilon > 0$ such that

$$W_2(P_{t,T}, P_{s,T}) = \left(\inf_{\pi\in\Pi(P_{t,T},P_{s,T})} \int ||x-y||^2 d\pi(x,y)\right)^{1/2} \leq L_\varepsilon T^{-1}|t-s|, \quad \forall t,s$$

where $\Pi(P_{t,T}, P_{s,T})$ is the set of all joint distributions with marginals $P_{t,T}$ and $P_{s,T}$, and $L_\varepsilon > 0$ is a Lipschitz constant for the noise distributions. $W_2(P_{t,T}, P_{s,T})$ is the Kantorovich, or Wasserstein distance.

(A4) For each $u \in [0,1]$ of the form $\frac{t}{T}, t \in \{1, \dots, T\}$, there exists an i.i.d. sequence $\{\tilde{\varepsilon}_t(u)\}_{t\in\mathbb{Z}}$ with law $P_{uT,T}$

(A5) The trend function $\mu(\cdot)$ is Lipschitz continuous on $[0,1]$ with Lipschitz parameter $L_\mu > 0$.

We state some remarks on the conditions. Conditions (A1), (A2) and (A5) are standard regularity assumptions in the theory of locally stationary processes, ensuring the smooth evolution of the linear filter and the existence of finite variance. The requirement in (A1) that the first moment of the filter coefficients is uniformly bounded is crucial to dampen the cumulative approximation error originating from the infinite past. Condition (A3) is the central assumption of our framework. By imposing a Wasserstein-Lipschitz continuity on the marginal laws of the innovations, we strictly relax the classical assumption of Dahlhaus (1997), which requires the driving noise to be independent and identically distributed (i.i.d.) or strictly stationary. Our condition trivially encapsulates the classical i.i.d. scenario – where the Wasserstein drift is exactly zero – but extends the theory to non-stationary innovations whose underlying probability measures are allowed to smoothly mutate over rescaled time. This relaxation is particularly useful for modeling financial time series, where the physical nature of market shocks frequently transitions from light-tailed Gaussian behaviors during stable macroeconomic periods to heavy-tailed, asymmetric distributions during liquidity crises. Finally, condition (A4) simply establishes the existence of the strictly stationary reference variables necessary to execute the L^2 optimal transport coupling without violating the marginal distributions governed by (A3).

The following theorem is the key result of the approximation.

Theorem 4.1 (Stationary approximation of locally stationary sequences)**.** *Assume we have a generalized locally stationary sequence $\{X_{t,T}\}_{t\in[T]}$ defined as in Equation 4.1. Under conditions (A1)-(A5), for each $u \in [0,1]$, there exists a strictly stationary process coupling $\{\tilde{X}_t(u)\}_{t\in[T]}$ defined by $\tilde{X}_t(u) = \mu(u) + \sum_{j=-\infty}^{\infty} a(u,j)\tilde{\varepsilon}_{t-j}(u)$ where $\{\tilde{\varepsilon}_t(u)\}$ is i.i.d. sequence defined as in (A4) such that*

$$|| X_{t,T} - \tilde{X}_t(u) ||_2 \leq c_1 \left| \frac{t}{T} - u \right| + c_2 \frac{1}{T} \text{ and therefore that}$$
$$\mathbb{E}\left[|X_{t,T} - \tilde{X}_t(u)|^2 \right] \leq 2c_1^2 \left| \frac{t}{T} - u \right|^2 + 2c_2^2 \frac{1}{T^2}$$

where c_1 and c_2 are constants independent of t, T and u.

Proof. The proof is based on transport argument from Villani [100] induced by the assumptions. We use optimal transport to find the coupling innovation variables $\tilde{\varepsilon}_1(u)$ that link to the Wassertein distance in assumption (A3). The rest of the proof follows directly from basic norm calculations. See Appendix A.3. □

We recall the reader that the consider representation encapsulates both strongly stationary sequences and "well behaved" weakly stationary sequences under the assumptions denoted before. Essentially, Theorem 4.1 demonstrates that any locally stationary sequence – and by extension, any "well-behaved" weakly stationary sequence – satisfying conditions (A1)-(A4) can be L^2-approximated by a strictly stationary sequence. Furthermore, if the innovation laws possess a continuous probability density, this strongly stationary approximant is guaranteed to be β-mixing [76].

It is easy to extend these results to the multivariate case provided one rearrange the assumptions stated earlier. The following corollary is the multivariate approximation extension of Theorem 4.1. Consider the multivariate representation

$$\mathbf{X}_{t,T} = \boldsymbol{\mu}\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} \mathbf{A}_{t,T}(j)\boldsymbol{\varepsilon}_{t-j,T} \tag{4.2}$$

where $\mathbf{X}_{t,T}$, $\boldsymbol{\mu}(u) \in \mathbb{R}^d$, $\mathbf{A}_{t,T}(j) \in \mathbb{R}^{d\times d}$ and $\boldsymbol{\varepsilon}_{t,T} \in \mathbb{R}^d$ with the same properties as earlier. The assumptions (A1)-(A5) are then modified to their matrix-valued versions.

(A1) There exists a function $\mathbf{A} : [0,1] \times \mathbb{Z} \to \mathbb{R}^{d\times d}$ such that:

(i) $\sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} ||\mathbf{A}(u,j)|| < \infty$, (ii) $\sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} |j|||\mathbf{A}(u,j)|| < \infty$,

(iii) $\sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} ||\partial_u \mathbf{A}(u,j)|| < \infty$ and (iv) $\left|\left|\mathbf{A}_{t,T}(j) - \mathbf{A}\left(\frac{t}{T}, j\right)\right|\right| \leq b_j T^{-1}$ with $\sum_j |b_j| < \infty$

Here $||\cdot||$ denotes the induced spectral norm (or Frobenius norm).

(A2) The process $\boldsymbol{\varepsilon}_{t,T}$ has finite second-moments $\sup_{t\in\mathbb{Z}} \mathbb{E}(||\boldsymbol{\varepsilon}_{t,T}||^2) < \infty$

(A3) Let $P_{t,T} = \mathbb{L}(\boldsymbol{\varepsilon}_{t,T})$ the d-dimensional distribution of $\boldsymbol{\varepsilon}_{t,T}$. There exists $L_\varepsilon > 0$ such that

$$W_2(P_{t,T}, P_{s,T}) = \left(\inf_{\pi\in\Pi(P_{t,T},P_{s,T})} \int ||x-y||^2 d\pi(x,y)\right)^{1/2} \leq L_\varepsilon T^{-1}|t-s|, \quad \forall t,s$$

where $\Pi(P_{t,T}, P_{s,T})$ is the set of all joint distributions with marginals $P_{t,T}$ and $P_{s,T}$, and $L_\varepsilon > 0$ is a Lipschitz constant for the noise distributions. $W_2(P_{t,T}, P_{s,T})$ is the 2-Kantorovich, or 2-Wasserstein distance on $\mathbb{R}^d$.

(A4) For each $u \in [0,1]$ of the form $\frac{t}{T}, t \in \{1,\ldots,T\}$, there exists an i.i.d. sequence $\{\tilde{\boldsymbol{\varepsilon}}_t(u)\}_{t\in\mathbb{Z}}$ with law $P_{uT,T}$

(A5) The trend function $\boldsymbol{\mu}(\cdot)$ is Lipschitz continuous on $[0,1]$ with Lipschitz parameter $L_\mu > 0$ coordinate wise.

Corollary 4.2. *Assume we have a generalized locally stationary sequence $\{\mathbf{X}_{t,T}\}_{t\in[T]}$ defined as in Equation 4.2. Under conditions (A1)-(A5) as stated in the multivariate form, for each $u \in [0,1]$, there exists a strictly stationary process coupling $\{\tilde{\mathbf{X}}_t(u)\}_{t\in[T]}$ defined by $\tilde{\mathbf{X}}_t(u) = \boldsymbol{\mu}(u) + \sum_{j=-\infty}^{\infty} \mathbf{A}(u,j)\tilde{\boldsymbol{\varepsilon}}_{t-j}(u)$ where $\{\tilde{\boldsymbol{\varepsilon}}_t(u)\}$ is i.i.d. sequence defined as in (A4) such that*

$$|| \mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u) ||_2 \leq c_1 \left|\frac{t}{T} - u\right| + c_2 \frac{1}{T} \text{ and therefore that}$$
$$\mathbb{E}\left[||\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)||^2\right] \leq 2c_1^2 \left|\frac{t}{T} - u\right|^2 + 2c_2^2 \frac{1}{T^2}$$

where c_1 and c_2 are constants independent of t, T and u.

Proof. The proof follows the same structure as Theorem 4.1; See Appendix A.3 for a detailled proof of the statement. □

Remark 4.1. Any subsequent claim in the univariate case can be extended to the multivariate space with little to no difficulty.

Note that assumption (A1)-(iii) $\sum_{j\in\mathbb{Z}} \sup_{u\in[0,1]} |\partial_u a(u,j)| < \infty$ assumes that $a(\cdot, j)$ is differentiable. This can be a strong assumption. One can relax this assumption by replacing it by a $L_a(j)$-Lipschitz continuity or α-Hölder continuity for some exponent $\alpha \in (0, 1]$.

Corollary 4.3. *Theorem 4.1 holds if we replace (A1)-(iii) by: (A1)-(iii) There exist constants $L_a(j) \geq 0$ such that $|a(u,j) - a(v,j)| \leq L_a(j)|u - v|$ for all $u, v \in [0,1]$, and $\sum_{j\in\mathbb{Z}} L_a(j) < \infty$*

Proof. The proof follows the same structure as for Theorem 4.1, but with $C_{\partial a} = \sum_{j=-\infty}^{\infty} L_a(j) < \infty$ □

Corollary 4.4. *If we replace (A1)-(iii) by: (A1)-(iii) There exist constants $L_a(j) \geq 0$ and $\alpha \in (0, 1]$ such that $|a(u,j) - a(v,j)| \leq L_a(j)|u - v|^\alpha$, then*

$$|| X_{t,T} - \tilde{X}_t(u) ||_2 \leq c_1 \left|\frac{t}{T} - u\right|^\alpha + c_2 \left|\frac{t}{T} - u\right| + c_3 \frac{1}{T} \text{ and therefore that}$$
$$\mathbb{E}\left[|X_{t,T} - \tilde{X}_t(u)|^2\right] \leq 3\left(c_1^2 \left|\frac{t}{T} - u\right|^{2\alpha} + c_2^2 \left|\frac{t}{T} - u\right|^2 + c_3^2 \frac{1}{T^2}\right)$$

where c_1, c_2 and c_3 are constants independent of t, T and u.

Proof. The proof yet again follows the same structure as for Theorem 4.1, but with $C_{\partial a} = \sum_{j=-\infty}^{\infty} L_a(j) < \infty$ and small changes in the intermediate steps. □

The following lemma links back this approximation with learning theory by bounding the difference of expectations between $h(X_{t,T})$ and $h(\tilde{X}_t(u))$ where h is a measurable L_h-Lipschitz continuous function.

Lemma 4.5. *Assume we have a generalized locally stationary sequence $\{X_{t,T}\}_{t\in[T]}$ defined as in Equation 4.1 and its strictly stationary L^2-approximation $\tilde{X}_t(u) = \mu(u) + \sum_{j=-\infty}^{\infty} a(u,j)\tilde{\varepsilon}_{t-j}(u)$ as in Theorem 4.1. For any measurable function h on $\mathbb{R}$ that is L_h-Lipschitz continuous,*

$$\left|\mathbb{E}[h(X_{t,T})] - \mathbb{E}[h(\tilde{X}_t(u))]\right| \leq \mathbb{E}\left|h(X_{t,T}) - h(\tilde{X}_t(u))\right| \leq L_h\left(c_1\left|\frac{t}{T} - u\right| + c_2\frac{1}{T}\right)$$

where constants c_1 and c_2 are defined by Theorem 4.1.

Proof. This is a direct consequence and trivial derivation from the previous theorem. □

4.1.3 Generalization bound

Let us now get back in a learning context. Assume we have a sequence $\mathbf{Z}_{t,T} = (\mathbf{X}_{t,T}, Y_{t,T}); \mathbf{X}_{t,T} \in \mathbb{R}^d, Y_{t,T} \in \mathbb{R}$ of T samples from a locally stationary sequence as defined in Equation 4.2 and verifying conditions (A1)-(A5) (in their multivariate form).

$$\mathbf{Z}_{t,T} = \boldsymbol{\mu}\left(\frac{t}{T}\right) + \sum_{j=-\infty}^{\infty} \mathbf{A}_{t,T}(j)\boldsymbol{\varepsilon}_{t-j,T}$$

where $\mathbf{Z}_{t,T}$, $\boldsymbol{\mu}(u) \in \mathbb{R}^{d+1}$, $\mathbf{A}_{t,T}(j) \in \mathbb{R}^{(d+1)\times(d+1)}$ and $\boldsymbol{\varepsilon}_{t,T} \in \mathbb{R}^{d+1}$ is a sequence of mutually independent. Here, we are interested in a one-step-ahead prediction. The previous section gave us a good way to approximate any locally stationary sequence with a stationary (β-mixing) one.

Within this framework, the common empirical method known as "rolling-window" estimation makes sense since to predict the next step ahead, we would want to use only recent data in the learning process to avoid too much cost of approximation to the coupled stationary approximation. As for the choice of the parameter u, for recent window data approximation, it is easy to see that it is best to always choose $u = 1$ which corresponds to fixating most approximation on the latest observed sample.

We assume that we have a window size W, a class of functions H and a γ-Lipschitz continuous loss function in its second argument uniformly over $h \in H$ $L : H \times \mathbb{R}^{d+1} \to \mathbb{R}$. For any measurable function h on $\mathbb{R}^d$, we denote respectively by $R_{T+1}(h)$ and $\hat{R}_{T+1}(h)$ the true risk of h at time $T+1$ and its empirical counterpart and defined by

$$R_{T+1}(h) = \mathbb{E}L(h, \mathbf{Z}_{T+1,T}) \quad \text{and} \quad \hat{R}_{T+1}(h) = \frac{1}{W}\sum_{t=T-W+1}^{T} L(h, \mathbf{Z}_{t,T}) \equiv \frac{1}{W}\sum_{t\in\{W\}} L(h, \mathbf{Z}_{t,T})$$

In this framework, we want to bound the value $|R_{T+1}(h) - \hat{R}_{T+1}(h)|$. Let us denote by $\tilde{\mathbf{Z}}_t(1) = (\tilde{\mathbf{X}}_t(1), \tilde{Y}_t(1))$ the stationary coupled sequence (as in Corollary 4.2).

$$\tilde{\mathbf{Z}}_t(1) = \boldsymbol{\mu}(1) + \sum_{j=-\infty}^{\infty} \mathbf{A}(1,j)\tilde{\boldsymbol{\varepsilon}}_{t-j}(1)$$

where $\{\tilde{\boldsymbol{\varepsilon}}_t(u)\}$ is i.i.d. sequence. We symmetrically define risk measures for the ghost sequence.

$$\tilde{R}_{T+1}(h) = \mathbb{E}L(h, \tilde{\mathbf{Z}}_{T+1}(1)) \quad \text{and} \quad \hat{\tilde{R}}_{T+1}(h) = \frac{1}{W}\sum_{t\in\{W\}} L(h, \tilde{\mathbf{Z}}_t(1))$$

$$|R_{T+1}(h) - \hat{R}_{T+1}(h)| \le |R_{T+1}(h) - \tilde{R}_{T+1}(h)| + |\tilde{R}_{T+1}(h) - \hat{\tilde{R}}_{T+1}(h)| + |\hat{\tilde{R}}_{T+1}(h) - \hat{R}_{T+1}(h)|$$
$$|R_{T+1}(h) - \tilde{R}_{T+1}(h)| = \left|\mathbb{E}L(h, \mathbf{Z}_{T+1,T}) - \mathbb{E}L(h, \tilde{\mathbf{Z}}_{T+1}(1))\right| \le \gamma\,(c_1 + c_2)\,\frac{1}{T} \text{ by } 4.5$$

$|\tilde{R}_{T+1}(h) - \hat{\tilde{R}}_{T+1}(h)|$ is the generic generalization bound for the coupling variables. It therefore remains to bound $|\hat{\tilde{R}}_{T+1}(h) - \hat{R}_{T+1}(h)|$.

Proposition 4.6. *The difference between the empirical risk of the true sequence $\hat{R}_{T+1}(h)$ and the empirical risk of the ghost sequence $\hat{\tilde{R}}_{T+1}(h)$ is bounded as follows*

$$|\hat{\tilde{R}}_{T+1}(h) - \hat{R}_{T+1}(h)| \le \gamma W^{-1} \sum_{t\in[W]} \left|\left|\mathbf{Z}_{t,T} - \tilde{\mathbf{Z}}_t(1)\right|\right|_{\mathbb{R}^{d+1}}$$

We denote it by $\hat{\Delta}_W$, that is, $\hat{\Delta}_W = W^{-1}\sum_{t\in\{W\}}\Big(L(h, \mathbf{Z}_{T+1,T}) - L(h, \tilde{\mathbf{Z}}_{T+1}(1))\Big)$. The expectation of $\hat{\Delta}_W$ can be bounded as follows

$$\mathbb{E}\left|\hat{\Delta}_W\right| \le \gamma\left(c_1\frac{W-1}{2T} + c_2\frac{1}{T}\right)$$

Proof. This first bound follows directly from the triangle inequality and the γ-Lipschitz continuous assumption on the loss function L. As for the second one, it follows from basic algebraic computation and probability theory

$$\mathbb{E}\left|\hat{\Delta}_W\right| \leq \gamma W^{-1} \sum_{t\in\{W\}} \mathbb{E}\left|\left|\mathbf{Z}_{t,T} - \tilde{\mathbf{Z}}_t(1)\right|\right|_{\mathbb{R}^{d+1}} \leq \gamma W^{-1} \sum_{t\in\{W\}} || \mathbf{Z}_{t,T} - \tilde{\mathbf{Z}}_t(1) ||_2$$
$$\implies \mathbb{E}\left|\hat{\Delta}_W\right| \leq \gamma W^{-1} \sum_{t\in\{W\}} \left\{ c_1 \left|\frac{t}{T} - 1\right| + c_2 \frac{1}{T} \right\} = \gamma \left(c_1 \frac{W-1}{2T} + c_2 \frac{1}{T} \right)$$

□

From Proposition 4.6, we now have a bound on the difference between the two empirical measures of risk. Basically, one can view this difference as a way of characterizing the bias introduced by the temporal changing behavior of the sequence. We know from the central limit theorem applied to stationary β-mixing sequences that the empirical risk $\hat{\bar{R}}_{T+1}(h)$ concentrates around its expectation at a rate of $O(T^{-1/2})$, provided the mixing coefficients decay sufficiently fast. However, the presence of non-stationarity introduces an additional bias term, captured precisely by $\hat{\Delta}_W$ which reflects the discrepancy between the true data-generating process and its ghost counterpart. As the bound in Proposition 4.6 shows, this bias vanishes as $T \to \infty$ at a rate governed by the window size W and the constants c_1, c_2. And thus, for large enough T, the ghost sequence provides an asymptotically faithful surrogate for the true sequence.

It remains to provide concentration inequalities in order to bound this quantity with high probability. For this matter, remark that the bound is characterized by the quantity we denote as S and define as

$$\begin{aligned} S &= \sum_{t\in\{W\}} \left|\left|\mathbf{Z}_{t,T} - \tilde{\mathbf{Z}}_t(1)\right|\right|_{\mathbb{R}^{d+1}} \\ &= \sum_{t\in\{W\}} \left|\left| \boldsymbol{\mu}\left(\frac{t}{T}\right) - \boldsymbol{\mu}(1) + \sum_{j=-\infty}^{\infty} (\mathbf{A}_{t,T}(j)\boldsymbol{\varepsilon}_{t-j,T} - \mathbf{A}(1,j)\tilde{\boldsymbol{\varepsilon}}_{t-j}(1)) \right|\right|_{\mathbb{R}^{d+1}} \end{aligned}$$

The bound then is the simple $|\hat{\Delta}_W| \leq \gamma W^{-1} S$. By assumption on the innovations $\boldsymbol{\varepsilon}_{t,T}$ (mutual independence) and by construction of the coupling innovations $\tilde{\boldsymbol{\varepsilon}}_t(1)$, the pairs $(\boldsymbol{\varepsilon}_{t,T}, \tilde{\boldsymbol{\varepsilon}}_t(1))$ form a mutually independent sequence and thus both $\mathbf{Z}_{t,T}$ and $\tilde{\mathbf{Z}}_t(1)$ are linear functions of the independent sequence $\{\eta_s\}_{s\in\mathbb{Z}}$ where $\eta_s = (\boldsymbol{\varepsilon}_{s,T}, \tilde{\boldsymbol{\varepsilon}}_s(1))$. In what, we will consider S as a function of the sequence $\boldsymbol{\eta} = \{\eta_s\}_{s\in\mathbb{Z}}$. We start by proving that $S(\boldsymbol{\eta})$ is L^2-Lipschitsz in the innovation sequence.

Proposition 4.7. *Let $\boldsymbol{\eta} = \{\eta_s\}_{s\in\mathbb{Z}}$ and $\boldsymbol{\eta}' = \{\eta'_s\}_{s\in\mathbb{Z}}$ be two different innovation sequences. $S(\boldsymbol{\eta})$ is L^2-Lipschitz in the innovation sequence, with Lipschitz constant finite by assumption (A1) and the following equation holds*

$$\begin{aligned} |S(\boldsymbol{\eta}) - S(\boldsymbol{\eta}')| &\leq \sum_{s\in\mathbb{Z}} \left\{ \sum_{t\in\{W\}} (||\mathbf{A}_{t,T}(t-s)||^2 + ||\mathbf{A}(1,t-s)||^2)^{1/2} \right\} \cdot ||\eta_s - \eta'_s|| \\ &\leq \left\{ \sum_{t\in\{W\}} \left(\sum_{j\in\mathbb{Z}} ||\mathbf{A}_{t,T}(j)||^2 + ||\mathbf{A}(1,j)||^2 \right)^{1/2} \right\} \cdot ||\boldsymbol{\eta} - \boldsymbol{\eta}'|| \end{aligned}$$

Proof. The proof rely on basic norm calculation properties and Cauchy-Schwarz inequality. See Appendix A.3 for the detailled proof. □

Up to this point, we have constructed the math mostly in a general sense, with only minimal assumptions on the innovations. Precisely, we have only assumed the innovations are independent and have finite second moment. However, for the purpose of getting a (tight)

generalization bound for a learning algorithm, we would need to make further assumptions on the sequence. In fact, remark that the innovations are not assumed to be bounded variables and therefore changing them may change the sum S in arbitrarily large ways. One may think about using concentration inequalities generalized for unbounded summands but these require a tail behavior or high probability bounding statement on the variables themselves.

To move forward, we adopt the standard econometric practice of assuming that the innovations are sub-Gaussian. (See [14], Section 2.3 for more information.) This is a mild strengthening of the second-moment condition. The following theorem gives the generalization bound in the locally stationary scenario when our locally stationary sequences satisfies assumptions (A1)-(A5) and innovations are sub-Gaussian.

Theorem 4.8. *If $\boldsymbol{\eta} = \{\eta_s\}_{s\in\mathbb{Z}}$ are sub-Gaussian with variance proxy σ_s^2, that is, for all $\lambda \in \mathbb{R}^{2(d+1)}$, $\mathbb{E}[\exp(\lambda^\top(\eta_s - \mathbb{E}\eta_s))] \leq \exp(\frac{\sigma_s^2||\lambda||^2}{2})$. Then the following bound holds*

$$\mathbb{P}(|S(\boldsymbol{\eta}) - \mathbb{E}S(\boldsymbol{\eta})| \geq t) \leq 2\exp\left\{\frac{-t^2}{2\sum_{s\in\mathbb{Z}}\kappa_s^2\sigma_s^2}\right\}$$

where $\kappa_s = \sum_{t\in\{W\}}(||\mathbf{A}_{t,T}(t-s)||^2 + ||\mathbf{A}(1,t-s)||^2)^{1/2} < \infty$ under (A1)

Proof. The proof uses Bobkov and Götze theorem [12] and tensorization of the T_1 transport cost inequality. See Appendix A.3 for the complete proof. □

Lemma 4.9. *Under the assumptions of Theorem 4.8, the following bound holds with probability at least $1-\delta$*

$$|\hat{\Delta}_W| \leq \gamma\left(c_1\frac{W-1}{2T} + c_2\frac{1}{T}\right) + \frac{\gamma}{W}\sqrt{2\log\left(\frac{2}{\delta}\right)\sum_{s\in\mathbb{Z}}\kappa_s^2\sigma_s^2}$$

Proof. The bound follows from the decomposition $|\hat{\Delta}_W| \leq \gamma W^{-1}S = \gamma W^{-1}\mathbb{E}S + \gamma W^{-1}(S - \mathbb{E}S)$, Theorem 4.8 and Proposition 4.6. □

Optimizing the upper bound of the previous lemma with respect to the window size W yields

$$W^* = \sqrt{\frac{2T}{c_1}}\left(2\log\left(\frac{2}{\delta}\right)\sum_{s\in\mathbb{Z}}\kappa_s^2\sigma_s^2\right)^{1/4}$$

Thus, the optimized generalization bound is given by the following corollary

Corollary 4.10. *Under the assumptions of Theorem 4.8, the following bound holds with probability at least $1-\delta$*

$$|\hat{\Delta}_W| \leq \frac{\gamma}{T}\left(c_2 - \frac{c_1}{2}\right) + \gamma\sqrt{\frac{2c_1}{T}}\left(2\log\left(\frac{2}{\delta}\right)\sum_{s\in\mathbb{Z}}\kappa_s^2\sigma_s^2\right)^{1/4}$$

4.2 BlockBoost for regression problems

So far, we have restricted our attention to binary classification problems in which the set of labels $\mathcal{Y}$ contains only two elements. In this section, we describe the extension of BlockBoost for regression problems. As before, the learner receives examples $z_t = (x_t, y_t)$ observations

of a stationary β-mixing sequence $(Z_t)_{t\in\mathbb{Z}}$ and its goal is to find an hypothesis $h : \mathcal{X} \to \mathcal{Y}$ which, given some value x_t, predicts approximately the value y_t that is likely to be seen. It is important to note our method is a derivation of AdaBoost.R2 and especially, like its classification counterpart, it coincides with AdaBoost.R2 [37] when the block size is set to 1. And as such, it inherits the characteristic of heuristic of BlockBoost(Algorithm 2).

Algorithm 3 presents BlockBoost.R2, regression extension of AdaBoost.R2 where block size is denoted as L instead of a_T. Note that in this algorithm, the loss function used is $\ell(h, z) = |h(x) - y|, z = (x, y)$. In general, one could write this algorithm using a different loss function, for example a squared loss function $\ell(h, z) = (h(x) - y)^2$.

Algorithm 3 BlockBoost.R2

Require: Sequence of T time-indexed examples $((\mathrm{x}_1, \mathrm{y}_1), \ldots, (\mathrm{x}_\mathrm{T}, \mathrm{y}_\mathrm{T}))$
Number L of contiguous examples in a block
Integer M specifying the number of boosting rounds
Weak learning algorithm WeakLearn

Partition $\{1, \ldots, \mathrm{T}\}$ into $\mathrm{K} = \left\lfloor \frac{\mathrm{T}}{\mathrm{L}} \right\rfloor$ contiguous blocks

$$\mathrm{B_k} = \{\mathrm{t} : (\mathrm{k}-1)\mathrm{L} < \mathrm{t} \leq \mathrm{kL}\}, \quad \mathrm{k} = 1, \ldots, \mathrm{K}$$

Initialize: Block weights $\mathrm{W_k^{(1)}} = \frac{1}{\mathrm{K}}$ for $\mathrm{k} = 1, \ldots, \mathrm{K}$

Instance weights $\mathrm{D}_1(\mathrm{t}) = \frac{1}{\mathrm{T}}$ for $\mathrm{t} = 1, \ldots, \mathrm{T}$

for $\mathrm{m} = 1$ to M **do**

Call WeakLearn, providing it with distribution $\mathrm{D_m}$; get back hypothesis $\mathrm{h_m} = \arg\min_{\mathrm{h}\in\mathrm{H}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{D_m(t)}\,|\mathrm{h(x_t)} - \mathrm{y_t}|$

Compute blockwise mean error $\mathrm{r_k^{(m)}} = \frac{1}{|\mathrm{B_k}|} \sum_{\mathrm{t}\in\mathrm{B_k}} \frac{|\mathrm{h_m(x_t)} - \mathrm{y_t}|}{\max_{\mathrm{t}'} |\mathrm{h_m(x_{t'})} - \mathrm{y_{t'}}|}$

Compute weighted block error $\varepsilon_\mathrm{m} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{W_k^{(m)} r_k^{(m)}}$

Compute estimator confidence $\beta_\mathrm{m} = \frac{\varepsilon_\mathrm{m}}{1 - \varepsilon_\mathrm{m}}$

Update and normalize block weights $\mathrm{W_k^{(m+1)}} \propto \mathrm{W_k^{(m)}} \cdot \beta_\mathrm{m}^{1-\mathrm{r_k^{(m)}}}$

Update instance weights $\mathrm{D_{m+1}(t)} = \frac{\mathrm{W_k^{(m+1)}}}{|\mathrm{B_k}|}$

end for

Output ensemble predictor

$$\mathrm{f(x)} = \inf \left\{ \mathrm{y} \in \mathcal{Y} : \sum_{\mathrm{m}:\mathrm{h_m(x)}\leq\mathrm{y}} \log \frac{1}{\beta_\mathrm{m}} \geq \frac{1}{2} \sum_{\mathrm{m}} \log \frac{1}{\beta_\mathrm{m}} \right\}$$

Under the assumption on the process (Z_t), one can easily deduce the generalization properties of the previous algorithm by analogy with what was previously done with Theorem 3.5 in the classification case.

4.3 Empirical experiments

In this section, we aim to provide empirical evidence of success of the extensions of the block boosting algorithm BlockBoost. We use the same structure as in section 3.4. Our experiments

are devised to test BlockBoost.R2 against its main time series model competitors AR and AR+GARCH. The predictive task stays the same as previously and the experimental design is in Appendix C.2.

4.3.1 Time series data

We evaluate BlockBoost.R2 on a synthetic time series generated from a Markov regime-switching AR(1)-GARCH(1,1) process with Student-t innovations, designed to replicate the key stylized facts of financial return series: volatility clustering, fat tails, and structural regime changes.

DGP$_3$: This data generating process is designed to evaluate BlockBoost.R2.
Regime dynamics: The latent regime process $\{s_t\}_{t\geq 1}$ takes values in $S = \{0, 1, 2\}$, corresponding to normal, volatile, and crisis regimes respectively. The regime evolves as a first-order Markov chain with transition matrix:

$$\mathbf{P} = \begin{pmatrix} 0.985 & 0.010 & 0.005 \\ 0.050 & 0.930 & 0.020 \\ 0.080 & 0.120 & 0.800 \end{pmatrix}$$

where $P_{ij} = \mathbb{P}(s_t = j \mid s_{t-1} = i)$. The expected regime durations implied by this matrix are approximately $1/(1 - P_{ii})$, giving 67 days in the normal regime, 14 days in the volatile regime, and 5 days in the crisis regime. The stationary distribution concentrates heavily on the normal regime, reflecting realistic market conditions.
Observation model: Conditional on the regime sequence, the observation y_t follows a weak AR(1) process:

$$y_t = c + \phi y_{t-1} + \varepsilon_t, \qquad \varepsilon_t = \sigma_{s_t} \cdot \sigma_t^{\text{GARCH}} \cdot z_t$$

with drift $c = 0.0005$, autoregressive coefficient $\phi = 0.08$, and regime-specific scale parameters $\sigma_0 = 0.01$, $\sigma_1 = 0.02$, $\sigma_2 = 0.04$ for the normal, volatile, and crisis regimes respectively. The innovations $z_t \overset{\text{i.i.d.}}{\sim} \mathcal{T}(6)$ follow a standardized Student-t distribution with 6 degrees of freedom, capturing the fat-tailed behavior of financial returns.
Volatility dynamics: The base volatility component σ_t^{GARCH} evolves independently of the regime scaling according to a unit-variance GARCH(1,1) process:

$$\left(\sigma_t^{\text{GARCH}}\right)^2 = \omega + \alpha \tilde{z}_{t-1}^2 + \beta \left(\sigma_{t-1}^{\text{GARCH}}\right)^2$$

where $\tilde{z}_{t-1} = \varepsilon_{t-1}/(\sigma_{s_{t-1}} \cdot \sigma_{t-1}^{\text{GARCH}})$ denotes the standardized innovation net of regime scaling, and the parameters $(\omega, \alpha, \beta) = (0.05, 0.10, 0.85)$ satisfy $\omega + \alpha + \beta = 1$, targeting unit base variance. This two-component volatility structure (where the regime multiplier σ_{s_t} modulates the level and the GARCH component σ_t^{GARCH} drives the persistence) ensures that regime changes and volatility clustering are captured as distinct but interacting phenomena.

The complete observation model therefore is

$$y_t = c + \phi y_{t-1} + \sigma_{s_t} \cdot \sigma_t^{\text{GARCH}} \cdot z_t, \qquad z_t \overset{\text{i.i.d.}}{\sim} t_6$$

From the generated sequence $\{y_t\}_{t=1}^N$ with $N = 3500$, we construct a lag-feature matrix with $p = 21$ lags. The feature vector at time t is:

$$\mathbf{x}_t = (|y_t|, |y_{t-1}|, \ldots, |y_{t-p+1}|)^\top \in \mathbb{R}^p$$

and the regression target is the absolute return one step ahead:

$$\hat{y}_{t+1} = |y_{t+1}|$$

The use of absolute returns as both features and target reflects the volatility forecasting nature of the task. The dataset is split chronologically with 70% of observations reserved for training and 30% for out-of-sample evaluation, yielding $N_{\text{train}} = 2435$ and $N_{\text{test}} = 1044$ observations respectively.

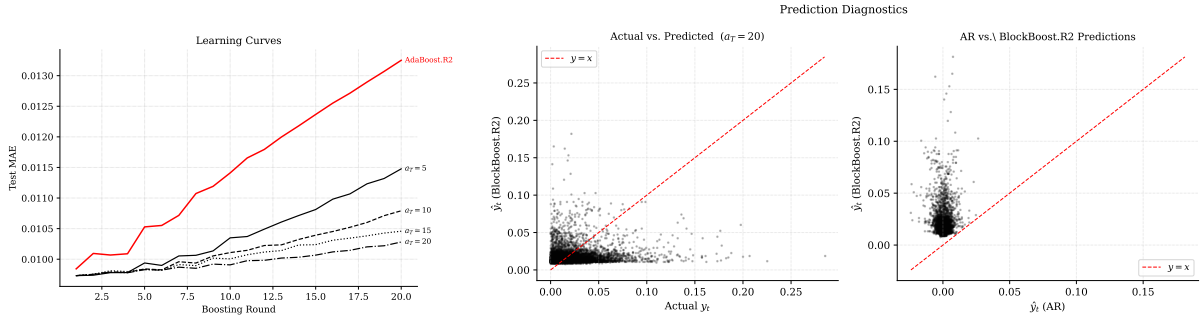
To assess robustness, all experiments are repeated across 20 independent random seeds controlling the initialization of the Markov chain and the innovation sequence. Results are reported as mean ± standard deviation across seeds. Out-of-sample performance is evaluated using the Diebold-Mariano test [33] against the AR(p) baseline, with p selected by AIC on the training set.

4.3.2 Empirical results

Table 4.1 reports aggregate out-of-sample performance across 20 independent random seeds. Figure 4.1 presents the learning curves and prediction diagnostics for BlockBoost.R2 at $L = 20$.

Table 4.1: Aggregate out-of-sample performance across 20 random seeds. Results are reported as mean ± standard deviation. DM p-values are two-sided tests against the AR(p^*) and AR(p^*)+GARCH(1,1) baselines respectively, using MAE loss differentials. A p-value of 0 denotes a value below machine epsilon ($\approx 2.2 \times 10^{-16}$).

Model	Test MAE ($\times 10^{-2}$)	Test MSE ($\times 10^{-4}$)	DM p-val vs. AR	DM p-val vs. AR+GARCH
AR(p^*)	1.34 ± 0.14	4.19 ± 1.08	—	—
AR(p^*)+GARCH(1,1)	1.32 ± 0.13	4.14 ± 1.05	1.76×10^{-102}	—
AdaBoost.R2	1.33 ± 0.12	2.91 ± 0.65	0.298	0.773
BB.R2 ($L = 5$)	1.15 ± 0.09	2.57 ± 0.70	1.16×10^{-95}	8.30×10^{-83}
BB.R2 ($L = 10$)	1.08 ± 0.08	2.47 ± 0.66	1.15×10^{-197}	4.07×10^{-178}
BB.R2 ($L = 15$)	1.05 ± 0.09	2.49 ± 0.69	1.25×10^{-266}	9.52×10^{-244}
BB.R2 ($L = 20$)	$\mathbf{1.03 \pm 0.09}$	$\mathbf{2.47 \pm 0.74}$	0	6.02×10^{-287}



(a) Aggregate test MAE learning curves across 20 seeds.

(b) Prediction diagnostics for BlockBoost.R2 at $L = 20$.

Figure 4.1: Learning dynamics and prediction diagnostics for BlockBoost.R2 on the simulated AR(1)-RS-GARCH(1,1) sequence. Results are aggregated over 20 random seeds.

Overall performance. BlockBoost.R2 achieves consistent and statistically significant improvements over all baselines across all block sizes tested. At $L = 20$, the best-performing configuration, the MAE reduction relative to AR(p^*) is 23.1% (1.03×10^{-2} versus 1.34×10^{-2}), and

relative to AR(p^*)+GARCH(1,1) is 21.9%. The corresponding DM p-values are effectively zero, indicating that the forecast accuracy differences are statistically significant at any conventional level and robust to the autocorrelation structure of loss differentials.

A notable finding is that AdaBoost.R2 fails to significantly outperform the AR(p^*) baseline, with DM p-value 0.298, despite achieving a lower MSE (2.91×10^{-4} versus 4.19×10^{-4}). This apparent inconsistency is explained by the asymmetry between MAE and MSE on a near-zero-mean, heavy-tailed target: AdaBoost.R2 controls large errors effectively – as reflected in its MSE – but does not improve average absolute deviation relative to the AR baseline, which already achieves near-mean prediction. This confirms the theoretical concern raised in Section 2.5: the instance-level reweighting of AdaBoost.R2 is poorly suited to temporally structured data where difficulty is regime-clustered rather than instance-specific.

Table 4.1 reveals a clear and monotone improvement in MAE as L increases from 5 to 20: 1.15, 1.08, 1.05, 1.03 (all $\times 10^{-2}$). The gains diminish as L grows, consistent with a regime whose expected normal-regime duration is approximately 67 steps. The MSE ordering is less clean: $L = 10$ and $L = 20$ both achieve 2.47×10^{-4} but the latter exhibits a wider confidence interval ($\pm 0.74 \times 10^{-4}$ versus $\pm 0.66 \times 10^{-4}$), suggesting that for extreme value control, a moderate block size may be preferable. The cross-seed standard deviations increase slightly with L, indicating a marginal variance-bias tradeoff as block granularity decreases.

Figure 4.1a further shows that all BlockBoost.R2 configurations converge within approximately 10 boosting rounds and plateau well below the AdaBoost.R2 curve at every round, confirming that the block structure provides a systematic rather than incidental advantage established from the very first boosting rounds.

Figure 4.1b reveals an important limitation that aggregate metrics alone do not fully expose. Predictions are compressed toward the empirical mean of absolute returns: large realized values are systematically underpredicted. Formally, $\hat{\sigma}_{\text{BB}} = 7.08 \times 10^{-3}$ versus $\hat{\sigma}_y = 1.55 \times 10^{-2}$, indicating that BlockBoost.R2 captures 45.7% of the target's standard deviation, with bias 1.58×10^{-3}. The correlation between actual and predicted values is $\rho(y_t, \hat{y}_t) = 0.206$ for BlockBoost.R2 versus $\rho = 0.008$ for AR(p^*), with a regression slope of 0.451. While the signal extracted by BlockBoost.R2 is statistically non-trivial and markedly superior to the AR baseline, the model remains a conservative mean-predictor on extreme observations. This behavior is expected given the near-unpredictability of the absolute return target ($\phi = 0.08$) and the dominance of idiosyncratic noise; it does not invalidate the MAE/MSE comparisons but qualifies the practical forecasting utility of any model in this setting.

5

Value at Risk Prediction

5.1 Data and sources

The data used in this study covers the period from January 25, 2014, to January 28, 2026. It consists of stock market indices from three different geographical areas, namely the West African regional market, the European market, and the US market.

For the regional financial market, two indices from the Regional Stock Exchange (BRVM) were selected. The BRVM Composite Index, which reflects overall market trends, is composed of 2990 observations and two variables, corresponding to the date and value of the index. The BRVM Finance Index, which represents the financial sector, comprises 2928 observations, also structured around two variables.

With regard to international markets, the study includes the CAC 40 index, representative of the French stock market, and the S&P 500 index, reflecting developments in the US stock market. The data for these two indices are grouped together in a single database comprising 3,100 observations and three variables, corresponding to the date and values of the two indices.

The differences observed in the number of observations between the different series are mainly due to the specific characteristics of each market's trading calendar, in particular public holidays and periods when trading is suspended.

Daily closing price data for the BRVM Composite and BRVM Finance indices were retrieved from richbourse.com via its public API endpoint, covering previously mentioned period (from 25 January 2014 to 28 January 2026). Data for the CAC 40 (`^FCHI`) and S&P 500 (`^GSPC`) were obtained from Yahoo Finance using the `yfinance` Python library over the same period. All series correspond to unadjusted daily closing prices, from which continuously compounded returns were computed as $r_t = 100 \times \ln(P_t/P_{t-1})$, where P_t denotes the closing price on day t.

5.1.1 Data exploration

First, we shall make a statement about the data preparation process. Essentially, note that the raw price series for the four indices were merged into a single dataset and each series was then processed independently. For each index, observations with missing values were removed prior to any computation – a step of particular relevance for the BRVM indices, whose trading calendar is less dense than that of developed markets and may exhibit non-trading days absent from the other series. Each cleaned series was sorted chronologically, and continuously compounded daily returns were computed as $r_t = 100 \times \ln(P_t/P_{t-1})$, where P_t is the closing price on day t. The first observation, undefined after the log-difference, was subsequently discarded. The four return series were treated as separate objects throughout the analysis to preserve index-specific trading calendars and avoid any artificial alignment that could introduce spurious observations.

Figure 5.1 depicts the evolution of daily returns for the four stock market (BRVM Composite, BRVM Finance, CAC 40, and S&P 500) over the period from 2014 to 2026. Analysis of said figure shows typical financial returns evolution behavior. At first glance, the four series fluctuate constantly around an average close to zero, which is consistent with the generally accepted properties of financial returns. However, visual inspection reveals several important structural phenomena. The first, and most immediately noticeable, is volatility clustering: periods of intense market turmoil tend to cluster together in time, so that large variations are followed by other large variations, whether positive or negative. This phenomenon is particularly pronounced for the CAC 40 and the S&P 500 around the year 2020, a period corresponding to the onset of the COVID-19 pandemic, where we observe oscillations of exceptional amplitude, reaching more than ten percentage points in a single session, followed by a gradual return to low volatility from 2021 onwards. The two Regional Stock Exchange indices behave in a significantly different manner: their ranges of variation are much more contained over the entire period. The shock of 2020 remains relatively invisible in the graph, suggesting that the BRVM remains relatively isolated from the turbulence shaking global financial markets

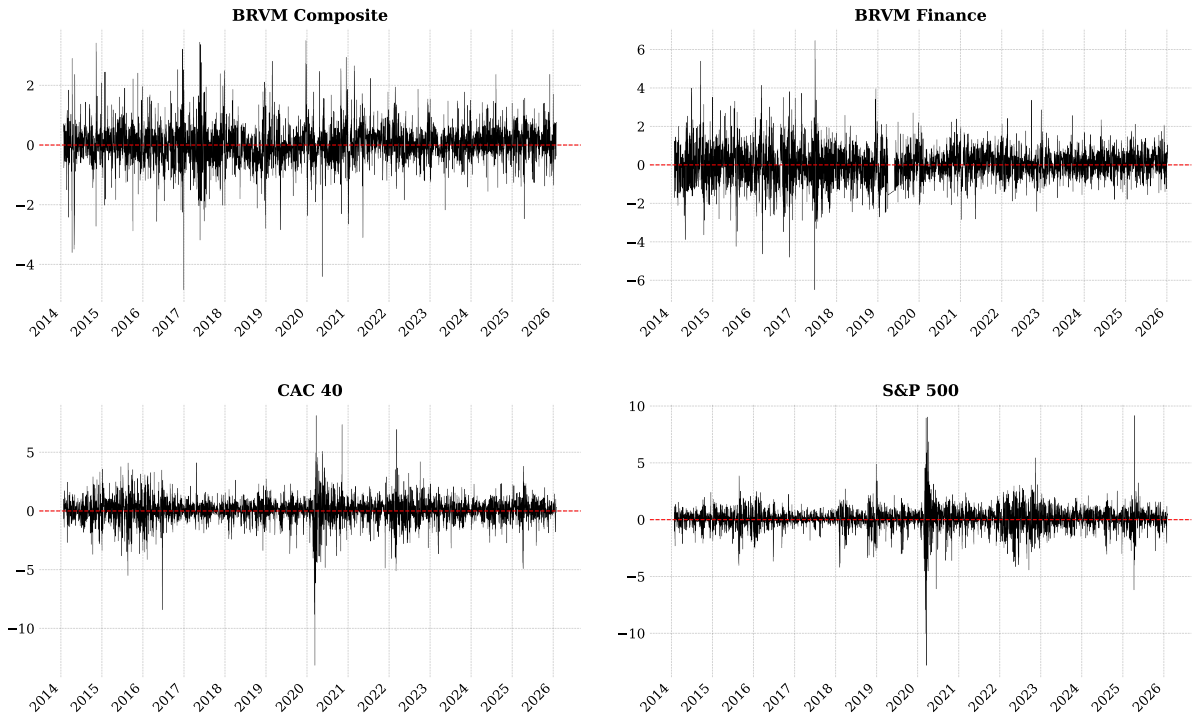
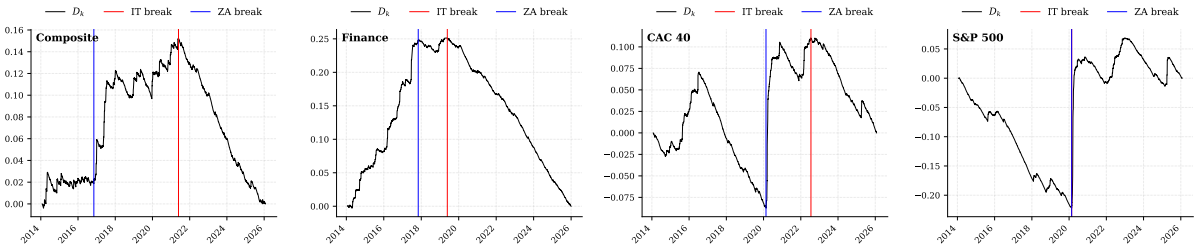


Figure 5.1: Daily log-returns for the BRVM Composite, BRVM Finance, CAC 40, and S&P 500 indices over the period 2014–2026. The red dashed line marks the zero axis. Volatility clustering and occasional extreme observations are visible across all series, with marked turbulence around the COVID-19 shock of early 2020.

To provide further evidence of the structural breaks in the series, we run two additional tests: the Zivot-Andrews [104] and the Inclan and Tiao cusum tests [55]. See Appendix E.1 and E.2 for details about these tests. Figure 5.2 confirms and sharpens the picture derived in Figure 5.1 through formal structural break analysis.



Asset	ZA Statistic	ZA p-value	ZA Break Date	IT Statistic	IT p-value	CUSUM Break Date
Composite	-12.698^{***}	$< 10^{-5}$	2016-11-07	5.901^{***}	$< 10^{-5}$	2021-05-26
Finance	-14.452^{***}	$< 10^{-5}$	2017-11-03	9.613^{***}	$< 10^{-5}$	2019-05-24
CAC 40	-10.795^{***}	$< 10^{-5}$	2020-02-18	4.324^{***}	$< 10^{-5}$	2022-07-19
S&P 500	-9.542^{***}	$< 10^{-5}$	2020-02-20	8.630^{***}	$< 10^{-5}$	2020-02-21

$^{***}p < 0.001$. ZA critical values (Model C): -5.57 (1%), -5.08 (5%). IT critical values: 1.628 (1%), 1.358 (5%).

Figure 5.2: Structural break analysis of daily squared log-returns for the four assets over 2014–2026. *Top*: Inclan-Tiao CUSUM process $D_k = C_k/C_T - k/T$; the blue dashed line marks the Zivot-Andrews break date and the red solid line the IT CUSUM break date $\hat{T}_b = \arg\max_k |D_k|$. *Bottom*: test statistics and identified break dates for both procedures. All statistics reject the null at the 0.001% level.

The Zivot-Andrews and Inclan-Tiao tests reject the null of constant variance at the 10^{-3}% level

for all four series, providing statistical confirmation of the regime changes visually apparent in Figure 5.1. For CAC 40 and S&P 500, both tests converge on February 2020 to within two days of each other, pinpointing the COVID-19 shock as a clean, instantaneous structural event. The near-perfect agreement between the two procedures on these assets is itself informative – it indicates that the break is sharp rather than gradual, consistent with the sudden amplitude spike observed in the return series. The second structural break in the CAC 40 series occuring at July 2022 likely pinpoints consequences of the Russo-Ukrainian war which itself started on February 2022. For the BRVM assets, the two tests diverge, identifying economically distinct events: the Zivot-Andrews break in 2016-2017 marks the onset of a sustained increase in unconditional volatility associated with the deepening of regional market activity, while the IT break in 2019-2021 captures its peak and subsequent contraction. This divergence reflects the more gradual, development-driven nature of the BRVM structural shift relative to the exogenous shock that dominates the developed market series.

Table 5.1 complements the graphical analysis by providing a precise quantitative characterization of the return series. The daily averages are extremely low for all indices, ranging from 0.01% for the BRVM Composite to 0.05% for the S&P 500, which is fully consistent with financial theory. In terms of dispersion, the CAC 40 has the highest standard deviation of 1.15), followed by the S&P 500 at 1.10, while the BRVM Composite has the lowest unconditional volatility of 0.70, graphically confirming the relative moderation of fluctuations in this market. The skewness coefficients provide statistical confirmation of the asymmetry perceived visually: the CAC 40 (-0.83) and the S&P 500 (-0.66) show significant negative asymmetry, revealing a thicker left tail and therefore a higher probability of observing extremely negative returns. The BRVM indices, with skewness close to zero (-0.07 and 0.03), suggest more symmetrical distributions. The most striking phenomenon is undoubtedly the excess kurtosis: with a kurtosis of 16.29 for the S&P 500 and 10.91 for the CAC 40, the tails of the distributions are much thicker than those of a normal distribution, attesting to the high frequency of extreme events in these markets.

BRVM Finance has the lowest kurtosis of 3.15, very close to the normal level (3.00), indicating a more regular distribution of shocks. These findings are formally confirmed by the Jarque-Bera test, whose p-value is less than 0.1% for the four indices, leading to the unambiguous rejection of the hypothesis of normality of returns. With regard to stationarity, the ADF test rejects the unit root hypothesis for all series ($p < 0.001$), confirming their stationary nature. The results of the KPSS test are broadly consistent for the CAC 40 and the S&P 500 ($p = 0.100$), but the BRVM indices marginally reject the null hypothesis of stationarity ($p < 0.05$), which could indicate the presence of occasional structural breaks, without calling into question their stationarity in the broad sense.

5.2 Methodology

The core idea here is to apply BlockBoost in the process of estimation of Value at Risk. But, we first require the user to notice that the block boosting algorithm is not primarily designed for quantile prediction such as Value at Risk prediction here. Thus, we will first define forecasting strategies to implement. Essentially, in this paper, we will be pinpointing two strategies for the task. We will be benchmarking the new estimator i.e. BlockBoost against its traditional competitors (Garch models with two distributional assumptions; normal and student).

Suppose we have a continuous stochastic process $(Y_t)_{t\in\mathbb{Z}}$. In financial risk management, Y_t is (generally) the negative log-return of a portfolio. Suppose also that we have an explanatory stochastic process $(X_t)_{t\in\mathbb{Z}}$. A typical example X_t is a( ) lagged value(s) of $\{Y_t\}$. We have access to a T-sequence of the realization of these stochastic processes $\{(X_t, Y_t)\}_{t=1}^{T}$. In what follows,

Table 5.1: Descriptive Statistics and Stationarity Tests for Daily Stock Index Returns

Statistic	**BRVM Composite**	**BRVM Finance**	**CAC 40**	**S&P 500**
Mean	0.01	0.02	0.02	0.05
Std. Dev.	0.70	0.97	1.15	1.10
Min	-4.84	-6.48	-13.10	-12.77
Max	3.48	6.43	8.06	9.09
Skewness	-0.07	0.03	-0.83	-0.66
Kurtosis	4.20	3.15	10.91	16.29
ADF p-value	$< 10^{-300***}$	$< 10^{-300***}$	$9.45 \times 10^{-20***}$	$3.34 \times 10^{-21***}$
KPSS p-value	$1.00 \times 10^{-2**}$	$1.44 \times 10^{-2**}$	1.00×10^{-1}	1.00×10^{-1}
Jarque-Bera p-value	$< 10^{-300***}$	$9.54 \times 10^{-263***}$	$< 10^{-300***}$	$< 10^{-300***}$

Notes: ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. ADF: Augmented Dickey–Fuller test; H_0: unit root. KPSS: Kwiatkowski–Phillips–Schmidt–Shin test; H_0: stationarity. Jarque–Bera: H_0: normality.

we present the different strategies we believe may help in tuning BlockBoost for Value at Risk prediction.

5.2.1 Similarity-weighted historical simulation

Let us start by giving a motivation for this approach. We recall that standard historical simulation (HS) assigns equal weight $\frac{1}{T}$ to all historical returns. Weighted HS variants improve this by assigning time-decaying weight [15] or volatility-adjusted weights [54]. We propose using BlockBoost's learned similarity structure to determine relevance weights. Following the theoretical guarantees of Kato [57], we assume the process $\{(X_t, Y_t), t = 1, 2, \ldots\}$ is stationary and α-mixing, which is argued by Kato [57] to cover many commonly used financial time series.

Our basic idea to estimate the value at risk is described as follows.

Training phase The first step is training BlockBoost on recent history to predict returns or volatility $|Y_t|$ or Y_t^2 and we get the ensemble that we denote as $\hat{f}$.

Kernel estimation Next, we estimate the conditional distribution function with a non parametric approach. We specifically use the Nadaraya-Watson (NW) estimator [77, 101] :

$$\hat{F}_{Y|X=x_0}(y) = \frac{\sum_{t=1}^{T} K_h(X_t - x_0)\mathbb{1}(Y_t \leq y)}{\sum_{t=1}^{T} K_h(X_t - x_0)}$$

where $K(\cdot)$ is a kernel function, h is a bandwidth and $K_h(u) = h^{-1}K(u/h)$, X_t is a scalar. Note that this estimator is the particular case of the Weighted Nadaraya-Watson (WNW) estimator (which is the main aim of the analysis of Kato [57]) where the sequence of weights is $p_t(x_0) = 1/T, \ \forall \ t$.

Denote by $\hat{f}$ the final ensemble generated by BlockBoost.R2. We define the estimator of the conditional cumulative distribution function by:

$$\hat{F}_{Y|\hat{f}(X)=\hat{f}(x_0)}(y) = \frac{\sum_{t=1}^{T} K_h(\hat{f}(X_t) - \hat{f}(x_0))\mathbb{1}(Y_t \leq y)}{\sum_{t=1}^{T} K_h(\hat{f}(X_t) - \hat{f}(x_0))}$$

Remark 5.1. Ideally we would want to estimate the true conditional distribution $\hat{F}_{Y|X}$ using NW (or WNW) estimator. But these estimations are asymptotic and guarantees apply only for

sufficiently large sample size T. In the special case where X_t is a lagged value of Y_t (a scalar), the asymptotic guarantees of NW estimator are reasonable. For large dimensional X_t (that is if we denote by p the dimension of X_t, we have $p \gg 1$), the estimator suffers from the so-called "curse of dimensionality" which implies that the convergence rate $((n(\ln n)^{-1}h^d)^{-1/2})$ degrades quickly as the dimension grows [51]. Therefore, here we use $\hat{f}$ to estimate a sufficient scalar index such that the conditional distribution can be approximated as $\hat{F}_{Y|\hat{f}(X)} \approx \hat{F}_{Y|X}$, that is, we assume $\hat{f}(X)$ is a sufficient statistic for the conditional $Y_t \mid X_t$ distribution:

$$Y \mid X \overset{D}{\approx} Y \mid \hat{f}(X)$$

Under stationarity and α-mixing assumption, Fan and Masry [40] proved asymptotic normality of NW, Hall, Wolff and Yao [49] extended NW for the conditional cumulative density function and extended to WNW and Kato [57] provided asymptotic bias properties of the estimator.

As for the choice of the bandwidth h and the Kernel function, one may use more sophisticated methods to get insights on the data at hand for precision, for example varying bandwidth [1]. Here, we use a Gaussian kernel function and use standard Silverman bandwidth rule of thumb [96]:

$$K(x) = \exp(-x^2/2), \qquad h = 1.06 \cdot \hat{\sigma}(\hat{f}(x)) \cdot n^{-1/5}$$

where $\hat{\sigma}(\hat{f}(x))$ is the estimated variance of the in-sample ensemble predictions.

VaR calculation We extract the value at risk at level α

$$\mathrm{VaR}_\alpha(x_0) = -\hat{F}^{-1}_{Y|\hat{f}(X)}(\alpha \mid X = x_0) = -\inf\{y \in \mathbb{R} : \hat{F}_{Y|\hat{f}(X)}(y \mid X = x_0) \geq \alpha\}$$

Typically α takes values between 0.01 (1%) and 0.05 (5%)

5.2.2 Quantile block boosting

The second strategy to value at risk forecasting discussed is the well-known "quantile" approach which requires us to modify the loss function of the algorithm to the pinball loss denoted as $\ell_\tau : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and defined by:

$$\ell_\tau(y, h) = \begin{cases} \tau(y - h) & \text{if } y \geq h \\ (\tau - 1)(y - h) & \text{if } y < h \end{cases}$$

where τ is the VaR confidence level. The process is quite simple. We either train BlockBoost.R2 (with pinball loss ℓ_τ) on recent history to predict Y_t or $-Y_t$.

In the case, we trained the algorithm on predicting Y_t, then to find value at risk at confidence level $1-\alpha$ (that is level α, typically 0.01 or 0.05), we set $\tau = 1-\alpha$ to find the $(1-\alpha)$th percentile of the distribution of Y_t and we have if we denote by $\hat{Y}_t$ the predicted value

$$\mathrm{VaR}_\alpha(X_t) = -\hat{Y}_t$$

Symmetrically, if we trained the algorithm on predicting $-Y_t$, then to find value at risk at confidence level $1 - \alpha$ (that is level α, typically 0.01 or 0.05), we set $\tau = 1 - \alpha$ to find the

$(1-\alpha)$th percentile of the distribution of $-Y_t$ and we have if we denote by $\hat{Y}_t$ the predicted value

$$\mathrm{VaR}_\alpha(X_t) = \hat{Y}_t$$

5.3 Estimation method and procedure

All models are estimated using a rolling window of $T = 1000$ trading days, that is, the model is re-evaluated at each time step on the most recent 1000 observations.

For GARCH models, at each step t, the parameters are re-estimated on the most recent 1000 observations by maximum likelihood, and the resulting estimates are used to produce a one-step-ahead conditional variance forecast $\hat{h}_{t+1|t}$. The predicted VaR at confidence level α is then given by

$$\widehat{\mathrm{VaR}}_{\alpha,t+1|t} = -\hat{q}_\alpha(D)\sqrt{\hat{h}_{t+1|t}}$$

where $\hat{q}_\alpha(D)$ is the α-quantile of the assumed innovation distribution D. Under the Gaussian assumption, $\hat{q}_\alpha(D) = \Phi^{-1}(\alpha)$; under the Student-$t$ specification, $\hat{q}_\alpha(D) = F_\nu^{-1}(\alpha)$, where the degrees of freedom ν are treated as a free parameter estimated jointly with the conditional variance parameters by MLE. This procedure is iterated over the entire out-of-sample period, yielding a sequence of daily VaR forecasts for each model-distribution pair. All models are specified at order $(p, q) = (1, 1)$ and estimated using the `arch` library [95] in Python. The GJR-GARCH is implemented via the asymmetric GARCH parameterisation with one leverage term (`o=1`), and the APARCH model via the dedicated `APARCH` volatility process.

As for the block boosting algorithm, the block size parameter is used here as a tunable hyperparameter hence a hyperparameter selection by proxy on the actual performance of the algorithm with said parameter value. This is done because in this specific context, the stationary β-mixing is likely to not be true as we have only tested weak stationarity. The best one can do therefore is to assume the sequence is locally stationary. Under said assumption and some other mild assumptions as proposed in Section 4.1, by Theorem 4.1, we can find a stationary β-mixing approximation of the sequence. But this sequence is unknown and therefore, we cannot find the "optimal" block size in the theoretical sense. We therefore test block sizes $L \in \{1, 5, 10, 15, 20, 25, 30, 35, 40\}$. For features construction, we use 21 lagged features. The models here are also estimated using a rolling window approach with $T = 1000$ trading days. As weak learner we use a decision tree (CART) with maximum depth 4 (to capture temporal relations). For computational reasons (due to the number of assets and the dataset itself), the number of boosting rounds was set to 10.

VaR forecasts are evaluated through three backtesting procedures, all implemented from scratch to ensure full control over the test specifications. The unconditional coverage test of Kupiec [63] examines whether the empirical violation rate is consistent with the nominal level α. The conditional coverage test of Christoffersen [26] extends this by testing for independence of violations over time, thereby detecting volatility clustering in the exception process. Finally, the predictive accuracy of competing models is compared using the test of Diebold-Mariano [33], applied to the loss differential between violation indicators.

5.4 Results and analysis

This section is divided in three main parts, each answering a different aspect of the problem of value at risk prediction with the block boosting algorithm and traditional GARCH models.

In Subsection 5.4.1, we compare both families in general divising insights on their performances and tradeoff. However, this analysis using BlockBoost is fundamentally flawed. In fact, in Section 4.1, we have extended our framework for locally stationary sequences (well-behaved ones). Here, due to the structural break we mentioned and proved with Figure 5.2, this assumption of local stationarity cannot hold. Thus, any use of BlockBoost in the general case does not follow our learning guarantees. We therefore analyze the results on sub-samples of the time series with the structural break assumed to be at start 2022 in general[1]. This ensures that the local stationarity can be safely assumed at least for the post-2022 period. In Subsection 5.4.2, we analyze the results period-wise.

5.4.1 General performance and observations

The empirical backtesting results (Table 5.2) present a clear and nuanced comparison of the BlockBoost (BB) framework against traditional GARCH models. The data reveals that model performance is highly contingent on the chosen confidence level (α) and the specific market structure.

The results establish a clear empirical trade-off without inflating the utility of either approach. The BlockBoost architecture is highly effective at managing moderate tail risk (5% VaR), completely avoiding the severe underestimation bias that plagues standard GARCH models in less liquid or emerging markets. Conversely, when modeling extreme tail events (1% VaR), the traditional parametric GARCH framework – particularly when paired with asymmetric volatility and heavy-tailed distributions – retains a slight but measurable edge in precision and conditional independence.

Performance at 5% confidence level At the $\alpha = 0.05$ threshold, BlockBoost models exhibit a distinct advantage in unconditional coverage, achieving empirical violation rates ($\hat{p}$) close to the nominal target. This is most pronounced in the frontier markets (BRVM Composite and BRVM Finance), where traditional GARCH models significantly underestimate risk. For instance, in the BRVM Finance index, the optimal GARCH model yields $\hat{p} = 3.37\%$, leading to a decisive rejection of the null hypothesis for unconditional coverage (Kupiec p-value < 0.01). In contrast, BlockBoost.R2 maintains a violation rate of 4.47% and passes both the Kupiec and Christoffersen tests.

In the developed markets (CAC 40 and S&P 500), both modeling frameworks achieve valid unconditional coverage. However, both equally struggle with serial independence at this confidence level; the Christoffersen p-values fall below 0.05 across both models for these indices, indicating that while the total number of VaR violations is correct, these violations tend to cluster consecutively.

Performance at 1% confidence level Moving to the extreme tail ($\alpha = 0.01$), the comparative advantage shifts toward the GARCH family. GARCH models demonstrate a slightly tighter calibration to the 1% nominal rate in three out of the four analyzed indices. For the CAC 40, the GJR-GARCH model with Student-t innovations achieves a near-perfect violation rate of

[1]Note that , one can further divide the analysis and customize it per sequence. Here, for simplicity and focus, we consider a general break even though the tests in Figure 5.2 suggest otherwise. The choice of start 2022 is a general consensus based on the results of the structural break tests.

Table 5.2: Backtesting performance of the best BlockBoost and GARCH models

	BRVM Composite		BRVM Finance		CAC 40		S&P 500	
Metric	BB	GARCH	BB	GARCH	BB	GARCH	BB	GARCH
(a) *VaR at the 5% confidence level*								
Violation (%)	**4.88**	3.77	**4.47**	3.37	**4.88**	5.41	**5.26**	**5.26**
Kupiec p-val	0.80	0.01	0.27	0.00	0.80	0.39	0.60	0.60
Christ. p-val	0.90	0.49	0.93	0.03	0.00	0.00	0.03	0.03
Best Model	R2(1)	G-N	R2(10)	EG-N	QBB(10)	EG-N	QBB(10)	G-N
(b) *VaR at the 1% confidence level*								
Violation (%)	1.06	**1.01**	**0.93**	0.88	1.06	**1.01**	1.14	**1.04**
Kupiec p-val	0.80	0.98	0.77	0.60	0.77	0.95	0.54	0.85
Christ. p-val	0.50	0.20	0.56	0.58	0.00	0.21	0.03	0.00
Best Model	QBB(5)	G-N	R2(15)	GJR-N	QBB(5)	GJR-S	QBB(1)	G-S

Notes: This table presents the best-performing models from the BlockBoost (BB-R2 or Q-BB) and GARCH families over the entire period. Model selection is based on minimizing the absolute deviation $|\hat{p} - \alpha|$ between the empirical violation rate and the nominal confidence level (5% or 1%). The strictly closest violation rate between the two modeling frameworks is highlighted in **bold** (ties are jointly bolded). The corresponding Kupiec and Christoffersen p-values assess the correct unconditional coverage and serial independence, respectively (a p-value ≥ 0.05 indicates failure to reject the null). The *Best Model* row specifies the optimal configuration for each context: the specific BlockBoost architecture (R2 for BB-R2 and QBB for Quantile BlockBoost) with optimal block size b (in parentheses), and the leading GARCH variant with Normal (-N) or Student-t (-S) innovations (G for GARCH, EG for EGARCH, GJR for GJR-GARCH and AP for APARCH).

1.01% and passes the test for serial independence (Christoffersen p-value = 0.21), whereas the BlockBoost equivalent exhibits violation clustering (p-value = 0.00). It is notable that at the 1% level, both BlockBoost and GARCH consistently pass the Kupiec test across all indices. This demonstrates that both architectures are fundamentally capable of sizing the magnitude of severe tail risks, even if traditional models offer marginally higher precision in this specific domain.

5.4.2 Period-wise performance and observations

This analysis is made in addition to the first one in section 5.4.1. Here we aim to respect our theoretical guarantees for locally stationary sequences. We divide the evaluation of violation rates and all subsequent tests in two periods (pre-2022 and post-2022). Note that most structural breaks occured in the 2019-2020 period and therefore the locally stationary assumption can be safely made only in the post-2022 period (and not in the pre-2022 period). We thus expect BlockBoost to provide even better estimates on the post-2022 period compared to those of the pre-2022 period. The main empirical finding in Table 5.3 is that BlockBoost demonstrates superior calibration stability across structural shifts at the moderate (5%) risk level, whereas traditional GARCH models exhibit an early advantage in extreme (1%) tail forecasting that BlockBoost only matches after the 2022 market transition.

Performance at 5% confidence level The results at the 5% confidence level are reflective of what we have previously observed in the general setting. Across the pre-2022 period (Panel A), BlockBoost generally provides tighter unconditional coverage than GARCH, though both models perform adequately in most indices. However, the post-2022 regime (Panel B) reveals

Table 5.3: Period-wise backtesting performance of the best BlockBoost and GARCH models

	BRVM Composite		BRVM Finance		CAC 40		S&P 500	
Metric	BB	GARCH	BB	GARCH	BB	GARCH	BB	GARCH
Panel A: Pre-2022 Period								
(a) *VaR at the 5% confidence level*								
Violation (%)	**4.31**	3.59	**3.97**	3.76	**5.35**	4.57	**5.22**	5.62
Kupiec p-val	0.31	0.03	0.14	0.07	0.61	0.52	0.76	0.38
Christ. p-val	0.89	0.53	0.67	0.10	0.00	0.00	0.06	0.04
Model Spec.	QBB(40)	G-N	R2(10)	EG-N	QBB(5)	AP-S	QBB(5)	G-N
(b) *VaR at the 1% confidence level*								
Violation (%)	**1.03**	**1.03**	**0.86**	**0.86**	1.36	**0.97**	1.71	**1.00**
Kupiec p-val	0.93	0.93	0.66	0.66	0.27	0.93	0.04	0.99
Christ. p-val	0.09	0.65	0.71	0.71	0.01	0.66	0.29	0.00
Model Spec.	QBB(20)	EG-N	R2(15)	EG-N	QBB(1)	G-S	QBB(1)	EG-S
Panel B: Post-2022 Period								
(a) *VaR at the 5% confidence level*								
Violation (%)	**4.93**	4.04	**5.03**	3.02	**5.00**	4.90	5.10	**5.00**
Kupiec p-val	0.92	0.15	0.97	0.00	0.99	0.88	0.89	1.00
Christ. p-val	0.73	0.79	0.72	0.17	0.02	0.00	0.17	0.15
Model Spec.	QBB(10)	AP-N	R2(15)	EG-N	QBB(30)	G-N	QBB(35)	GJR-N
(b) *VaR at the 1% confidence level*								
Violation (%)	1.18	**0.99**	**1.01**	0.91	**0.96**	0.77	**0.98**	**0.98**
Kupiec p-val	0.57	0.96	0.98	0.76	0.90	0.43	0.95	0.95
Christ. p-val	0.59	0.08	0.65	0.68	0.00	0.05	0.08	0.08
Model Spec.	QBB(5)	EG-N	QBB(15)	GJR-N	QBB(15)	GJR-S	QBB(35)	G-S

Notes: This table assesses the period-wise robustness by comparing the single best BlockBoost model against the best GARCH counterpart for both the pre-2022 (Panel A) and post-2022 (Panel B) structural periods. Selection criteria, formatting, and hypothesis testing interpretations map identically to those established in Table 5.2.

a visible divergence in adaptability. Following the structural shift in 2022, standard GARCH models struggle to properly size risk in frontier markets, leading to severe underestimation (e.g., a 3.02% violation rate for BRVM Finance, resulting in a firm Kupiec rejection with $p = 0.00$).

Conversely, as expected, BlockBoost exhibits better performance in the post-2022 period. It achieves near-perfect 5% calibration across all four markets – ranging from 4.93% to 5.10% – comfortably passing both coverage and independence tests in all indices except the CAC 40, where both frameworks continue to exhibit violation clustering.

Performance at 1% confidence level The results at the 1% confidence level are more interesting as they are where the improvement in performance is most expressive and visible. In fact, in the pre-2022 period, GARCH strictly dominates. It tightly tracks the 1% nominal rate across all markets, while BlockBoost visibly struggles with the developed markets S&P 500 (and CAC 40), yielding a 1.71% (1.36%) violation rate and failing the Kupiec test ($p = 0.04$ for S&P 500). However, in the post-2022 period, BlockBoost somewhat closes the performance gap. It outperforms the optimal GARCH models in the BRVM Finance (1.01% vs. 0.91%) and CAC 40 (0.96% vs. 0.77%) indices, and perfectly ties the GARCH framework in the S&P 500 (0.98%).

5.5 Discussions

The results of the study give a nuanced understanding of the performance of BlockBoost in the context of value at risk forecasting. Of course, as we have discussed earlier, at the 5% confidence level, it is apparent that the block boosting algorithm has a consistent edge on the classical family of GARCH models when analyzed both in the general sequence and pre-post 2022. On the other hand, at the 1% confidence level, the picture seems to be less clear and more nuanced as to which one of the two have better performances. In fact, at this level, GARCH family is arguably better (outperforming) in the global evaluation. But analysis of the pre and post 2022 periods reveals that in the pre-2022 period (associated with structural breaks), GARCH models perform better generally (except for the BRVM Finance asset) while in the post-2022 period, BlockBoost models perform arguably better. The question therefore arises at to *why does BlockBoost perform better than GARCH at 5% and perform less better at 1%? And why is it that analyzing pre and post 2022 periods distinctively seem to improve the performance of BlockBoost in the post-2022 period when it is bad in the pre-2022 period ?.*

At the 5% confidence level, we are modelling a moderate tail risk, that is, events that occur roughly once every 20 trading days. There is *enough* data to learn from empirically. A non-parametric model like BlockBoost in particular here can do well here because it doesn't need to assume a specific distribution on the returns. On the other hand, GARCH models impose a distribution on the returns and if the structure is even slightly misspecified, the bias shows up as an underestimation or overestimation.

At 1% however, we are modelling an extreme tail risk (events that occur roughly once every 100 trading days). In a few years of data, we do not observe of these regimes. Here, parametric models gain the advantage because they extrapolate in their tail using the assumed distribution while non-parametric models need more data to populate the 1% tail.

BlockBoost constructs its VaR estimate by combining weak learners through a weighted majority vote logic. At moderate quantiles, the signal-to-noise ratio is high enough for this ensemble mechanism to work well. At extreme quantiles, the signal essentially disappears due to the lack of data per se. And therefore, the weak learners have no reliable information to offer at that threshold. The ensemble aggregation provides no useful information when every individual learner is equally uninformed about the extreme tail. GARCH with Student-t

or asymmetric specifications (GJR, EGARCH) is explicitly designed to capture fat tails and volatility asymmetry, which are precisely the features that dominate at the 1% level. A more detailed analysis of GARCH models results (in Table F.1) reveals that the heavy-tail nature of the sequence can be observed through the performance of the classical methods (GARCH models) especially when looking at the results for developed markets (CAC 40 and S&P 500) where all Student-t GARCH models outperformed Normal models. This suggests that the heavy-tailed distribution assumption encodes prior knowledge of the markets which are not incorporated in BlockBoost.

The analysis per period shows how much important the validity of the assumptions is for BlockBoost. Like we have enounciated previously, it comes apparent that the assumption of local stationarity is most violated in the pre-2022 period compared to the post-2022 period resulting in better performances post-2022 altogether both at 1% and 5% confidence levels albeit marginal at the 1% confidence level due to the data problem we have pointed out earlier.

A defining parameter in the block boosting algorithm is the block size. The results show blocksizes that had the best overall performance in terms of pure violation rate. Overall, we could see that there is no apparent structure in the block sizes for any of the assets. The question therefore is: *Can we interpret the block sizes ? If yes, how do we interpret them ?*

The selected block size should be understood as a regularization scale controlling how local dependence is handled during boosting. In fact, in the theoretical discussions, we have defined the block size as an inherent property of the time series. One could be tempted to interpret these results as if they were indicating something about the dependence of asset prices in markets. But such interpretation is spurious since we have selected the block size with a technique similar to hyperparameter selection based on the violation rate. Larger blocks enforce stronger temporal smoothing and stronger protection against overfitting while smaller blocks preserve adaptivity and reactivity of the boosting process. The optimized value is therefore not interpreted as a structural parameter of the market, but as the empirically preferred trade-off between bias and variance for the forecasting task. This points out a problem we have enounced earlier in this work: the block size selection problem. It is notoriously difficult to estimate mixing coefficients of a stochastic process (and most notably β-mixing coefficients) due to the curse of dimensionality and the drop in sample size as we increase the block size which renders earlier (small block) mixing coefficients more accurate than the later (bigger block). While this work uses hyperparameter selection to choose block size, it is inherently wrong in the sense that it reduces the nature of the block size to the performance of the algorithm on the forecasting task.

As for the difference in performance between the quantile approach and the similarity-weighted historical simulation approach, most of the time, the quantile approach produces better results which is consistent with the existing litterature on the subject albeit the arguable choice of kernel function and bandwidth rule.

5.6 Future work

In this work, we leave the paradigm of empiricism when we use boosting for time series. Our core contributions are the block boosting algorithm which, unlike black-box models like XGBoost, LightGBM, etc., has provable theoretical guarantees and the extension of its learning guarantees to that of generalized locally stationary sequences which is a class of "well-behaved" non-stationary sequences.

At the end of this paper, the main issue is the incapacity of BlockBoost to extend its learning guarantees to that of a wider class of non-stationary sequences. The problem in the approach here is the fact that the block boosting algorithm (as its parent AdaBoost) is a batch-learning algorithm. And thus, even intuitively, one cannot possibly imagine a batch

learning capable of picking up trends in non-stationary sequences in a satisfying way. The problem of learning from non stationary sequences has been evaluated and learning guarantees have been proved by Kuznetsov and Mohri [64] and their approach of the problem was through online learning theory. In Appendix B, we present the sketch of an extension of the method to non stationary sequences based on the online learning framework developed by Kuznetsov and Mohri [64]. In this sketch, we prove how the block structure might help improve the framework's learning guarantees.

Notice that in BlockBoost, we handle dependency between observations by assigning them equal weight inside their respective blocks. It appears that there is no principled reason choice behind the choice of equal weight allocation aside from simplicity of the algorithm. If further work is done to better this algorithm, it might be of interest to look deeper into how dependency is handled and how improving this might help increase performance.

Another theoretical and practical issue is the choice of the block size when using BlockBoost. Throughout the paper, we struggled to provide a good way to find such parameter but it occurs that β-mixing conditions are simply too strong and too difficult to estimate to provide a good and useful theoretical procedure to find the optimal block size other than hyperparameter selection. The use of mixing coefficients to define dependence in stochastic processes have been for long maintained and has shown failures because of this exact problem. The question therefore, is if we can relax the definition of dependence to functions that are far easier to estimate than mixing coefficients. If possible, one might be able to devise a selection protocol for the optimal blocksize for boosting in time series.

From a more practical standpoint, in this paper, we have focused comparative evaluation of BlockBoost to the classical GARCH family renowned in the literature for being the best estimator for value at risk estimation. A more complete pipeline is the comparison of these methods to other boosting algorithm like XGBoost, LightGBM and other value at risk estimation models like CaViar which is out of the scope of this paper. We have pointed out the previous section the weakness of BlockBoost when it comes to the estimation of extreme tail value at risk (1%). We believe a combination with extreme value theory (EVT) would improve the performance of the block boosting algorithm like it has been done with GARCH models.

6

Conclusion

This work set out to answer a deceptively simple question: can we design a boosting algorithm for time series forecasting that is not only empirically effective but also comes with the kind of theoretical guarantees that have long been the standard in financial econometrics? The answer, we argue, is yes – at least for a broad and practically relevant class of time series processes.

Our starting point was a fundamental tension in the existing literature. On one side, classical econometric models like GARCH and its extensions are theoretically grounded, interpretable, and well-understood. On the other side, machine learning methods (including gradient boosting frameworks) have demonstrated strong empirical performance in forecasting tasks but are typically applied in a heuristic manner, with little theoretical justification for their behavior on dependent data. In the context of Value at Risk forecasting, where model validity has direct regulatory and prudential consequences, the absence of theoretical guarantees is a real problem. A model that works well on average but has no provable generalization properties is a fragile foundation for risk management.

We addressed this tension by introducing BlockBoost, a boosting algorithm designed (from the ground up) for stationary β-mixing time series. The core modification is simple but consequential: instead of reweighting individual observations as AdaBoost does, BlockBoost reweights contiguous blocks of observations, constraining the weight sequence to a block-constant subspace of the probability simplex. This single structural change has several effects simultaneously. It reduces the Rademacher complexity of the weight sequence, making the algorithm amenable to generalization analysis under mixing conditions. It introduces implicit regularization by preventing the algorithm from taking unbounded steps in the presence of noise, as formalized in Proposition 3.4. And it preserves the local temporal structure of the data, allowing the weak learner to capture regime dynamics rather than individual anomalies. We proved that BlockBoost drives the block-wise empirical risk to zero, established its generalization bound for stationary β-mixing sequences in Theorem 3.5, and showed empirically that it avoids the catastrophic noise memorization that eventually afflicts standard AdaBoost

under random classification noise.

However, a defining characteristic of financial return series is that strict stationarity is rarely, if ever, a defensible assumption. Structural breaks, time-varying volatility regimes, and slowly evolving market conditions all push financial data outside the stationary β-mixing framework. We addressed this by developing, in Chapter 4, a general approximation result for locally stationary sequences. Theorem 4.1 shows that under mild regularity conditions – smooth evolution of the linear filter, finite second moments, and Wasserstein-Lipschitz continuity of the innovation laws – any locally stationary sequence can be approximated in L^2 by a strictly stationary β-mixing process, with an approximation error that decays at a controlled rate. This result, which we believe is of independent interest beyond its application to boosting, allows the generalization guarantees established in the stationary setting to be extended to the locally stationary one, at the cost of an additional bias term that vanishes as the sample size grows.

The empirical results across three structurally different markets (the BRVM regional stock exchange, the CAC 40, and the S&P 500) paint a coherent and nuanced picture that is broadly consistent with the theoretical analysis.

At the moderate tail risk level of 5%, BlockBoost demonstrates a consistent advantage over the GARCH family. This is most visible in the BRVM markets, where traditional parametric models systematically underestimate risk, likely because their distributional assumptions are poorly suited to the specific characteristics of a frontier market with a less dense trading calendar and a different structural volatility profile. The period-wise analysis sharpens this conclusion. In the post-2022 period, where the assumption of local stationarity is most defensible given the structural breaks identified in Figure 5.2, BlockBoost achieves near-perfect calibration at the 5% level across all four markets, with violation rates ranging from 4.93% to 5.10% and passing both the Kupiec and Christoffersen tests in three out of four cases.

At the extreme tail risk level of 1%, the picture is more nuanced. GARCH models, particularly asymmetric specifications with Student-t innovations, retain an advantage in the pre-2022 period and in developed markets. We argue that the almost visible and intuitive non-validity of the local stationarity assumption and the lack of data due to the extreme tail (with window size 1000, only 10 observations approximately) are at cause for this lack of performance. Parametric models compensate for this data scarcity by extrapolating into the tail using their assumed distribution, encoding prior knowledge about the heavy-tailed nature of financial returns that BlockBoost does not possess. In the post-2022 period, BlockBoost closes this gap substantially, outperforming GARCH in two out of four markets at the 1% confidence level and matching it in a third.

Several questions remain open. The most pressing is the block size selection problem. The theoretical optimal block size minimizes an upper bound that depends on the noise rate and the mixing coefficients, both of which are difficult or impossible to estimate reliably in practice. The difficulty of estimating β-mixing coefficients is a known and longstanding problem in the mixing processes literature, one that goes beyond the scope of this work. A promising direction is to relax the definition of dependence entirely – replacing mixing coefficients with quantities that are easier to estimate from data. If such a relaxation is possible while preserving the generalization guarantees, it would allow the block size to be chosen in a principled rather than empirical manner.

A second open question concerns the collapse behavior of BlockBoost under random classification noise. Proposition 3.4 establishes that the block weighting mechanism prevents the algorithm from taking unbounded steps during training, and the empirical results on DGP2 confirm that BlockBoost does not exhibit the runaway overfitting of AdaBoost. However, whether the block boosting algorithm eventually collapses at the population level in the sense of Long and Servedio – that is, whether the final hypothesis converges to a trivial predictor as

the number of rounds goes to infinity – remains unresolved.

From a practical standpoint, the most natural extension of this work is a combination of BlockBoost with extreme value theory (EVT) for the 1% VaR regime. The complementarity is clear: BlockBoost handles the moderate tail well, EVT handles the extreme tail well, and a principled combination could potentially dominate the GARCH family across both confidence levels. A similar combination has been explored for GARCH models with documented success, and the theoretical framework developed here – particularly the locally stationary approximation – could provide the foundation for a rigorous treatment of such a hybrid. A more complete comparative study including CaViaR and other conditional quantile models would also strengthen the empirical case.

This work leaves the paradigm of pure empiricism that has characterized most applications of boosting to financial time series. The argument we have made is not that BlockBoost is the best forecasting tool in every setting. The argument is that it is possible to build a machine learning method for time series and particularly for VaR forecasting that comes with theoretical guarantees, that these guarantees extend to a broad and practically relevant class of non-stationary processes, and that the resulting algorithm performs competitively against the best classical methods.

For practitioners and regulators in the WAEMU zone, the results on the BRVM markets are perhaps the most immediately relevant contribution. The consistent underestimation of risk by standard GARCH models in these markets, and the ability of BlockBoost to correct it at the 5% level, points to a limitation of blindly importing methods calibrated on developed market data to a structurally different financial environment. As the BCEAO continues its convergence toward Basel standards, the availability of methods that are both theoretically grounded and empirically validated on regional data is not a merely academic concern but also a practical one.

A

Proofs of main results

A.1 BlockBoost analysis

Proof of Proposition 3.1. To derive that the empirical block risk is driven to 0 by BlockBoost, we extend the proof of theorem 2.1. At iteration m, the block weights are updated according to

$$W_{m+1}(k) = \frac{W_m(k)\exp\left(\alpha_m(2r_k(h_m)-1)\right)}{Z_m} \text{ where } Z_m = \sum_{k=1}^{K_T} W_m(k)\exp\left(\alpha_m(2r_k(h_m)-1)\right)$$

By iterating the recursion and using the initialization $W_1(k) = \frac{1}{K_T}$,

$$W_{M+1}(k) = \frac{\exp\left(\sum_{m=1}^{M}\alpha_m(2r_k(h_m)-1)\right)}{K_T\prod_{m=1}^{M} Z_m} \text{ and thus}$$

$$\prod_{m=1}^{M} Z_m = \frac{1}{K_T}\sum_{k=1}^{K_T}\exp\left(\sum_{m=1}^{M}\alpha_m(2r_k(h_m)-1)\right) \text{ since } \sum_{k=1}^{K_T} W_{M+1}(k) = 1$$

$$\text{with } 2r_k(h_m)-1 = \frac{\sum_{t\in \mathrm{B_k}}(2\mathbb{1}[f(x_t)\neq y_t]-1)}{a_T} = -\frac{1}{a_T}\sum_{t\in \mathrm{B_k}} y_t h_m(x_t) \text{ implies}$$

$$\prod_{m=1}^{M} Z_m = \frac{1}{K_T}\sum_{k=1}^{K_T}\exp\left(-\frac{1}{a_T}\sum_{t\in \mathrm{B_k}} y_t h(x_t)\right) \text{ where } h(x) = \sum_{m=1}^{M}\alpha_m h_m(x)$$

But we also have $\mathbb{1}\left[\sum_{t\in \mathrm{B_k}} y_t h(x_t)\leq 0\right] \leq \exp\left(-\frac{1}{a_T}\sum_{t\in \mathrm{B_k}} y_t h(x_t)\right)$ which leads to

$$\hat{R}_B(h) \leq \prod_{m=1}^{M} Z_m = \prod_{m=1}^{M} \sum_{k=1}^{K_T} W_m(k) \exp\left(\alpha_m(2r_k(h_m) - 1)\right)$$

We optimize the upper bound by finding the optimal choice of α_m at each boosting step. By the convexity argument $\exp\left(\alpha_m(2r_k(h_m) - 1)\right) \leq [(1 - r_k(h_m)) \exp\left(-\alpha_m\right) + r_k(h_m) \exp\left(\alpha_m\right)]$, we have

$$Z_m \leq \sum_{k=1}^{K_T} W_m(k) \left[(1 - r_k(h_m)) \exp\left(-\alpha_m\right) + r_k(h_m) \exp\left(\alpha_m\right)\right]$$

Taking the derivative of this new upper bound with respect to α_m and setting it to 0, we find the step size as defined in the algorithm $\alpha_m = \frac{1}{2} \log \frac{1-\varepsilon_m}{\varepsilon_m}$. By replacing α_m with its optimal value, we complete the proof. The second inequality is obtained trivially from the assumptions and the previous bound.

□

Proof of Lemma 3.2. We established in the previous proof that the instance error decomposes as:

$$\hat{R}(f_M) = \frac{1}{T} \sum_{t=1}^{T} (1 - q_t)(1 - p_t) + \frac{1}{T} \sum_{t=1}^{T} q_t p_t \tag{A.1}$$

Expanding and rearranging A.1:

$$\hat{R}(f_M) = \frac{1}{T} \sum_{t=1}^{T} (1 - p_t) + \frac{1}{T} \sum_{t=1}^{T} q_t(2p_t - 1) = \hat{R}^*(f_M) + \frac{1}{T} \sum_{t=1}^{T} q_t(2p_t - 1) \tag{A.2}$$

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} q_t(2p_t - 1) \middle| z_1^T, \xi_1', \ldots, \xi_T'\right] = \frac{\eta}{T} \sum_{t=1}^{T} (2p_t - 1) = \eta\left(1 - 2\hat{R}^*(f_M)\right)$$

Substituting into A.2

$$\mathbb{E}\left[\hat{R}(f_M) \middle| z_1^T, \xi_1', \ldots, \xi_T'\right] = \hat{R}^*(f_M) + \eta\left(1 - 2\hat{R}^*(f_M)\right) = \eta + (1 - 2\eta)\hat{R}^*(f_M) \tag{A.3}$$

From A.2 and A.3, we have

$$\begin{aligned}
\hat{R}(f_M) - \mathbb{E}\left[\hat{R}(f_M) \middle| z_1^T, \xi_1', \ldots, \xi_T'\right] &= \frac{1}{T} \sum_{t=1}^{T} q_t(2p_t - 1) - \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} q_t(2p_t - 1) \middle| z_1^T, \xi_1', \ldots, \xi_T'\right] \\
&= \frac{1}{T} \sum_{t=1}^{T} (q_t(2p_t - 1) - \eta(2p_t - 1))
\end{aligned}$$

Conditional on $z_1^T, \xi_1', \ldots, \xi_T'$, the terms of the sum $(q_t - \eta)(2p_t - 1)$ are independent with $|(q_t - \eta)(2p_t - 1)| \leq 1$. By Hoeffding's inequality applied conditionally on $z_1^T, \xi_1', \ldots, \xi_T'$, we get

$$\mathbb{P}\left(\left|\hat{R}(f_M) - \eta - (1 - 2\eta)\hat{R}^*(f_M)\right| > \sqrt{\frac{\log(2/\delta)}{2T}} \middle| z_1^T, \xi_1', \ldots, \xi_T'\right) \leq \delta$$

Therefore, with probability at least $1-\delta$:

$$\left|\hat{R}(f_M) - \eta - (1-2\eta)\hat{R}^*(f_M)\right| \leq \sqrt{\frac{\log(2/\delta)}{2T}}$$

□

A.2 Generalization bounds

Proof of Theorem 3.5. Since $\{z_t\}_t, z_t = (x_t, y_t)$ is stationary and β-mixing, the Berbee coupling lemma [9] gives: for each block $\mathrm{B_k}$ of size a_T, there exists a copy $\{z'_t\}_t, z'_t = (x'_t, y'_t)$ of i.i.d. random variables with the same marginal, such that

$$\mathbb{P}\left(\{z_t\}_{t\in\mathrm{B_k}} \neq \{z'_t\}_{t\in\mathrm{B_k}}\right) \leq \beta(a_T)$$

i.e. the sequence within each block behaves as i.i.d. up to error $\beta(a_T) \to 0$.

Define the event $\mathcal{E}_k = \{\{z_t\}_{t\in\mathrm{B_k}} = \{z'_t\}_{t\in\mathrm{B_k}}\}$, so $\mathbb{P}(\bar{\mathcal{E}}_k) \leq \beta(a_T)$

$$\frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x_t) \neq y_t) = \frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x'_t) \neq y'_t)\cdot\mathbb{1}(\mathcal{E}_k) + \frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x_t) \neq y_t)\cdot\mathbb{1}(\bar{\mathcal{E}}_k)$$

By the triangle inequality,

$$\begin{aligned}&\left|\frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x_t) \neq y_t) - \mathbb{E}[\mathbb{1}(f_M(x) \neq y)]\right| \\ &\leq \left|\frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x'_t) \neq y'_t) - \mathbb{E}[\mathbb{1}(f_M(x) \neq y)]\right| + \mathbb{1}(\bar{\mathcal{E}}_k)\end{aligned}$$

The term $\mathbb{1}(\bar{\mathcal{E}}_k)$ is handled probabilistically: with probability at most $\beta(a_T)$, the coupling fails and $\mathbb{1}(\bar{\mathcal{E}}_k) = 1$ for any block $\mathrm{B_k}$.

$$\forall k, \mathbb{P}(\bar{\mathcal{E}}_k) \leq \beta(a_T) \Rightarrow \mathbb{P}\left(\bigcup_{k=1}^{K_T} \bar{\mathcal{E}}_k\right) \leq \sum_{k=1}^{K_T} \mathbb{P}(\bar{\mathcal{E}}_k) \leq K_T\beta(a_T)$$

The global empirical risk on the clean labels can be decomposed into the block averages:

$$\hat{R}^*(f_M) = \frac{1}{K_T}\sum_{k=1}^{K_T}\left(\frac{1}{a_T}\sum_{t\in\mathrm{B_k}} \mathbb{1}(f_M(x_t) \neq y_t)\right)$$

Let $\hat{R}^*_{ghost}(f_M) = \frac{1}{K_T}\sum_{k=1}^{K_T}\left(\frac{1}{a_T}\sum_{t\in\mathrm{B'_k}} \mathbb{1}(f_M(x'_t) \neq y'_t)\right)$ be the empirical risk evaluated on the independent ghost blocks. Under the event $\mathcal{E}$, we have exact equality: $\hat{R}^*(f_M) = \hat{R}^*_{ghost}(f_M)$. Because the K_T ghost blocks are strictly independent and the block-wise risks take values in $[0,1]$. Denote $\mathcal{F} = \mathrm{sgn}(\mathcal{H})$. By standard symmetrization and Rademacher complexity bounds applied to the independent block averages, we have with probability at least $1-\delta$ over the ghost blocks:

$$\sup_{f\in\mathcal{F}}\left|\hat{R}^*_{ghost}(f)-R^*(f)\right|\leq 2\Re_{K_T}(\mathcal{F},\mu)+\sqrt{\frac{\log(2/\delta)}{2K_T}}$$

where $\Re_{K_T}(\mathcal{F},\mu)$ is the Rademacher complexity of $\mathcal{F}$ computed over the K_T independent blocks.

But the hypothesis $f_M\in\{-1,+1\}$ is the sign of the ensemble $h^M\in\mathcal{H}=\text{conv}(H)$ therefore $\mathcal{F}=\text{sgn}(\mathcal{H})=\text{sgn}(\text{conv}(H))$ and we have

$$\Re_{K_T}(\mathcal{F},\mu)=\Re_{K_T}(\text{sgn}(\text{conv}(H)),\mu)\leq\Re_{K_T}(\text{conv}(H),\mu)=\Re_{K_T}(H,\mu)$$

We thus have, with probability at least $1-\delta-K_T\beta(a_T)$:

$$\left|\hat{R}^*(f_M)-R^*(f_M)\right|\leq 2\Re_{K_T}(H,\mu)+\sqrt{\frac{a_T\log(2/\delta)}{2T}} \tag{A.4}$$

□

A.3 Locally stationary sequence approximation

Proof of Theorem 4.1. Since the state space $\mathbb{R}$ is a Polish space and the marginal laws $P_{t,T}$ and $P_{uT,T}$ have finite second moments (A2), the infimum in the Kantorovich formulation of the W_2 Wasserstein distance is attained. Therefore, by the standard properties of optimal transport (see, e.g., Villani (2009) [100], Theorem 4.1), there exists an optimal transport plan $\pi^*_{t,u}\in\Pi(P_{t,T},P_{uT,T})$. Consequently, we can define a coupling – a pair of random variables $(\varepsilon_t,\tilde{\varepsilon}^{(u)}_t)$ on a common probability space – distributed according to $\pi^*_{t,u}$ such that

$$\mathbb{E}|\varepsilon_{t,T}-\tilde{\varepsilon}_t(u)|^2=\int||x-y||^2d\pi^*_{t,u}(x,y)=W_2^2(P_{t,T},P_{uT,T})\text{ or equivalently}$$
$$||\varepsilon_{t,T}-\tilde{\varepsilon}_t(u)||_2=W_2(P_{t,T},P_{uT,T})\leq L_\varepsilon\left|\frac{t}{T}-u\right|\text{ by assumption (A3)}$$

Define

$$\check{X}_t(u)=\mu(u)+\sum_{j=-\infty}^{\infty}a(u,j)\varepsilon_{t-j,T}$$

It then comes that

$$X_{t,T}-\check{X}_t(u)=\left(\mu\left(\frac{t}{T}\right)-\mu(u)\right)+\sum_{j=-\infty}^{\infty}\left(a_{t,T}(j)-a(u,j)\right)\varepsilon_{t-j,T}$$
$$\left|\left|X_{t,T}-\check{X}_t(u)\right|\right|_2\leq\left|\mu\left(\frac{t}{T}\right)-\mu(u)\right|+\sum_{j=-\infty}^{\infty}|a_{t,T}(j)-a(u,j)|\cdot||\varepsilon_{t-j,T}||_2$$

By assumption (A2), there exists a constant $M_\varepsilon>0$ such that for all $t\in[T],j\in\mathbb{Z}$ and $T\geq 1$:

$$\text{(A2)}\ \Rightarrow||\varepsilon_{t-j,T}||_2=\sqrt{\mathbb{E}[\varepsilon^2_{t-j,T}]}\leq M_\varepsilon\quad\text{and}\quad\text{(A5)}\ \Rightarrow\left|\mu\left(\frac{t}{T}\right)-\mu(u)\right|\leq L_\mu\left|\frac{t}{T}-u\right|$$

By the triangle inequality, we have

$$|a_{t,T}(j) - a(u,j)| \leq \left|a_{t,T}(j) - a\left(\frac{t}{T}, j\right)\right| + \left|a\left(\frac{t}{T}, j\right) - a(u,j)\right|$$

$$\text{(A1)} \;\Rightarrow \left|a_{t,T}(j) - a\left(\frac{t}{T}, j\right)\right| \leq \frac{b_j}{T} \text{ and } \left|a\left(\frac{t}{T}, j\right) - a(u,j)\right| \leq \left(\sup_{v\in[0,1]} |\partial_v a(v,j)|\right)\left|\frac{t}{T} - u\right|$$

$$\text{thus } |a_{t,T}(j) - a(u,j)| \leq \frac{b_j}{T} + \left(\sup_{v\in[0,1]} |\partial_v a(v,j)|\right)\left|\frac{t}{T} - u\right|$$

$$\begin{aligned}\sum_{j=-\infty}^{\infty} |a_{t,T}(j) - a(u,j)| \cdot ||\varepsilon_{t-j,T}||_2 &\leq M_\varepsilon \sum_{j=-\infty}^{\infty}\left[\frac{b_j}{T} + \left(\sup_{v\in[0,1]} |\partial_v a(v,j)|\right)\left|\frac{t}{T} - u\right|\right] \\ &= \frac{M_\varepsilon}{T}\underbrace{\sum_{j=-\infty}^{\infty} b_j}_{C_b<\infty \text{ by (A1)}} + M_\varepsilon\left|\frac{t}{T} - u\right|\underbrace{\sum_{j=-\infty}^{\infty}\sup_{v\in[0,1]}|\partial_v a(v,j)|}_{C_{\partial a}<\infty \text{ by (A1)}}\end{aligned}$$

$$||X_{t,T} - \check{X}_t(u)||_2 \leq (L_\mu + M_\varepsilon C_{\partial a})\left|\frac{t}{T} - u\right| + \frac{M_\varepsilon}{T} C_b \tag{A.5}$$

It remains to bound the expression $\left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2$ from

$$\begin{aligned}&\check{X}_t^{(u)} - \tilde{X}_t(u) = \sum_{j=-\infty}^{\infty} a(u,j)\left(\varepsilon_{t-j,T} - \tilde{\varepsilon}_{t-j}(u)\right) \\ &\Longrightarrow \left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2 \leq \sum_{j=-\infty}^{\infty} |a(u,j)| \cdot ||\varepsilon_{t-j,T} - \tilde{\varepsilon}_{t-j}(u)||_2 \\ &\Longrightarrow \left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2 \leq \sum_{j=-\infty}^{\infty} L_\varepsilon |a(u,j)|\left|\frac{t-j}{T} - u\right| \leq \sum_{j=-\infty}^{\infty} L_\varepsilon |a(u,j)|\left(\left|\frac{t}{T} - u\right| + \frac{|j|}{T}\right) \\ &\Longrightarrow \left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2 \leq L_\varepsilon\left|\frac{t}{T} - u\right|\underbrace{\left(\sum_{j=-\infty}^{\infty}|a(u,j)|\right)}_{C_a<\infty \text{ by (A1)}} + \frac{L_\varepsilon}{T}\underbrace{\left(\sum_{j=-\infty}^{\infty}|j||a(u,j)|\right)}_{C_{aj}<\infty \text{ by (A1)}}\end{aligned}$$

$$\left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2 \leq L_\varepsilon\left|\frac{t}{T} - u\right| C_a + \frac{L_\varepsilon}{T} C_{aj} \tag{A.6}$$

By the triangle inequality,

$$\left|\left|X_{t,T} - \tilde{X}_t(u)\right|\right|_2 \leq \left|\left|X_{t,T} - \check{X}_t(u)\right|\right|_2 + \left|\left|\check{X}_t^{(u)} - \tilde{X}_t(u)\right|\right|_2$$

Which with Equations A.5 and A.6 yields

$$\left|\left|X_{t,T} - \tilde{X}_t(u)\right|\right|_2 \leq (L_\mu + L_\varepsilon C_a + M_\varepsilon C_{\partial a})\left|\frac{t}{T} - u\right| + \frac{M_\varepsilon C_b + L_\varepsilon C_{aj}}{T}$$

It then follows from $\mathbb{E}\left[|X_{t,T} - \tilde{X}_t(u)|^2\right] = \left|\left|X_{t,T} - \tilde{X}_t(u)\right|\right|_2^2$ that

$$\mathbb{E}\left[|X_{t,T} - \tilde{X}_t(u)|^2\right] \leq 2\left((L_\mu + L_\varepsilon C_a + M_\varepsilon C_{\partial a})^2 \left|\frac{t}{T} - u\right|^2 + \frac{(M_\varepsilon C_b + L_\varepsilon C_{aj})^2}{T^2}\right)$$

which completes the proof. □

Proof of Corollary 4.2. Since the state space $\mathbb{R}^d$ is a Polish space and the marginal laws $P_{t,T}$ and $P_{uT,T}$ have finite second moments (A2), the infimum in the Kantorovich formulation of the W_2 Wasserstein distance is attained. By the standard properties of optimal transport (e.g., Villani (2009) [100], Theorem 4.1), there exists an optimal transport plan $\pi^*_{t,u} \in \Pi(P_{t,T}, P_{uT,T})$. Because the innovations are mutually independent, we construct the global coupling as the infinite product of these marginal plans. Consequently, we define a pair of random vectors $(\boldsymbol{\varepsilon}_{t,T}, \tilde{\boldsymbol{\varepsilon}}_t(u))$ on a common probability space distributed according to $\pi^*_{t,u}$ such that

$$\mathbb{E}||\boldsymbol{\varepsilon}_{t,T} - \tilde{\boldsymbol{\varepsilon}}_t(u)||^2 = \int ||x-y||^2 d\pi^*_{t,u}(x,y) = W_2^2(P_{t,T}, P_{uT,T}) \text{ or equivalently}$$
$$||\boldsymbol{\varepsilon}_{t,T} - \tilde{\boldsymbol{\varepsilon}}_t(u)||_2 = W_2(P_{t,T}, P_{uT,T}) \leq L_\varepsilon \left|\frac{t}{T} - u\right| \text{ by assumption (A3)}$$

where $||\cdot||_2$ denotes the probabilistic L^2-norm defined as $||\mathbf{Y}||_2 = \sqrt{\mathbb{E}[||\mathbf{Y}||^2_{\mathbb{R}^d}]}$. Define:

$$\check{\mathbf{X}}_t(u) = \boldsymbol{\mu}(u) + \sum_{j=-\infty}^{\infty} \mathbf{A}(u,j)\boldsymbol{\varepsilon}_{t-j,T}$$

It then follows that

$$\mathbf{X}_{t,T} - \check{\mathbf{X}}_t(u) = \left(\boldsymbol{\mu}\left(\frac{t}{T}\right) - \boldsymbol{\mu}(u)\right) + \sum_{j=-\infty}^{\infty} \left(\mathbf{A}_{t,T}(j) - \mathbf{A}(u,j)\right)\boldsymbol{\varepsilon}_{t-j,T}$$
$$\left|\left|\mathbf{X}_{t,T} - \check{\mathbf{X}}_t(u)\right|\right|_2 \leq \left|\left|\boldsymbol{\mu}\left(\frac{t}{T}\right) - \boldsymbol{\mu}(u)\right|\right|_{\mathbb{R}^d} + \sum_{j=-\infty}^{\infty} ||\left(\mathbf{A}_{t,T}(j) - \mathbf{A}(u,j)\right)\boldsymbol{\varepsilon}_{t-j,T}||_2$$

By the sub-multiplicativity of the compatible matrix norm and assumption (A2), there exists a constant $M_\varepsilon > 0$ such that for all $t \in [T], j \in \mathbb{Z}$ and $T \geq 1$:

$$\text{(A2)} \Rightarrow ||\boldsymbol{\varepsilon}_{t-j,T}||_2 = \sqrt{\mathbb{E}[||\boldsymbol{\varepsilon}_{t-j,T}||^2]} \leq M_\varepsilon \quad \text{and} \quad \text{(A5)} \Rightarrow \left|\left|\boldsymbol{\mu}\left(\frac{t}{T}\right) - \boldsymbol{\mu}(u)\right|\right|_{\mathbb{R}^d} \leq L_\mu \left|\frac{t}{T} - u\right|$$

By the triangle inequality for matrix norms, we have

$$||\mathbf{A}_{t,T}(j) - \mathbf{A}(u,j)|| \leq \left|\left|\mathbf{A}_{t,T}(j) - \mathbf{A}\left(\frac{t}{T}, j\right)\right|\right| + \left|\left|\mathbf{A}\left(\frac{t}{T}, j\right) - \mathbf{A}(u,j)\right|\right|$$

$$\text{(A1)} \Rightarrow \left|\left|\mathbf{A}_{t,T}(j) - \mathbf{A}\left(\frac{t}{T}, j\right)\right|\right| \leq \frac{b_j}{T} \text{ and } \left|\left|\mathbf{A}\left(\frac{t}{T}, j\right) - \mathbf{A}(u,j)\right|\right| \leq \left(\sup_{v\in[0,1]} ||\partial_v \mathbf{A}(v,j)||\right)\left|\frac{t}{T} - u\right|$$
$$\text{thus } ||\mathbf{A}_{t,T}(j) - \mathbf{A}(u,j)|| \leq \frac{b_j}{T} + \left(\sup_{v\in[0,1]} ||\partial_v \mathbf{A}(v,j)||\right)\left|\frac{t}{T} - u\right|$$

Applying this to the infinite sum:

$$\sum_{j=-\infty}^{\infty} ||\mathbf{A}_{t,T}(j) - \mathbf{A}(u,j)|| \cdot ||\boldsymbol{\varepsilon}_{t-j,T}||_2 \leq M_\varepsilon \sum_{j=-\infty}^{\infty} \left[\frac{b_j}{T} + \left(\sup_{v\in[0,1]} ||\partial_v \mathbf{A}(v,j)|| \right) \left| \frac{t}{T} - u \right| \right]$$

$$= \frac{M_\varepsilon}{T} \underbrace{\sum_{j=-\infty}^{\infty} b_j}_{C_b<\infty \text{ by (A1)}} + M_\varepsilon \left| \frac{t}{T} - u \right| \underbrace{\sum_{j=-\infty}^{\infty} \sup_{v\in[0,1]} ||\partial_v \mathbf{A}(v,j)||}_{C_{\partial\mathbf{A}}<\infty \text{ by (A1)}}$$

$$||\mathbf{X}_{t,T} - \check{\mathbf{X}}_t(u)||_2 \leq (L_\mu + M_\varepsilon C_{\partial\mathbf{A}}) \left| \frac{t}{T} - u \right| + \frac{M_\varepsilon}{T} C_b \tag{A.7}$$

It remains to bound the expression $\left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2$ from

$$\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u) = \sum_{j=-\infty}^{\infty} \mathbf{A}(u,j) \left(\boldsymbol{\varepsilon}_{t-j,T} - \tilde{\boldsymbol{\varepsilon}}_{t-j}(u) \right)$$

$$\implies \left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq \sum_{j=-\infty}^{\infty} ||\mathbf{A}(u,j)|| \cdot ||\boldsymbol{\varepsilon}_{t-j,T} - \tilde{\boldsymbol{\varepsilon}}_{t-j}(u)||_2$$

$$\implies \left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq \sum_{j=-\infty}^{\infty} L_\varepsilon ||\mathbf{A}(u,j)|| \left| \frac{t-j}{T} - u \right| \leq \sum_{j=-\infty}^{\infty} L_\varepsilon ||\mathbf{A}(u,j)|| \left(\left| \frac{t}{T} - u \right| + \frac{|j|}{T} \right)$$

$$\implies \left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq L_\varepsilon \left| \frac{t}{T} - u \right| \underbrace{\left(\sum_{j=-\infty}^{\infty} ||\mathbf{A}(u,j)|| \right)}_{C_{\mathbf{A}}<\infty \text{ by (A1)}} + \frac{L_\varepsilon}{T} \underbrace{\left(\sum_{j=-\infty}^{\infty} |j| ||\mathbf{A}(u,j)|| \right)}_{C_{\mathbf{A}j}<\infty \text{ by (A1)}}$$

$$\left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq L_\varepsilon \left| \frac{t}{T} - u \right| C_{\mathbf{A}} + \frac{L_\varepsilon}{T} C_{\mathbf{A}j} \tag{A.8}$$

By the triangle inequality,

$$\left|\left|\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq \left|\left|\mathbf{X}_{t,T} - \check{\mathbf{X}}_t(u)\right|\right|_2 + \left|\left|\check{\mathbf{X}}_t^{(u)} - \tilde{\mathbf{X}}_t(u)\right|\right|_2$$

Which with Equations A.7 and A.8 yields

$$\left|\left|\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)\right|\right|_2 \leq (L_\mu + L_\varepsilon C_{\mathbf{A}} + M_\varepsilon C_{\partial\mathbf{A}}) \left| \frac{t}{T} - u \right| + \frac{M_\varepsilon C_b + L_\varepsilon C_{\mathbf{A}j}}{T}$$

It follows from $\mathbb{E}\left[||\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)||^2\right] = \left|\left|\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)\right|\right|_2^2$ that

$$\mathbb{E}\left[||\mathbf{X}_{t,T} - \tilde{\mathbf{X}}_t(u)||^2\right] \leq 2 \left((L_\mu + L_\varepsilon C_{\mathbf{A}} + M_\varepsilon C_{\partial\mathbf{A}})^2 \left| \frac{t}{T} - u \right|^2 + (M_\varepsilon C_b + L_\varepsilon C_{\mathbf{A}j})^2 \frac{1}{T^2} \right)$$

which completes the proof.

□

Proof of Proposition 4.7. Define the inner vector for each t

$$\mathbf{V}_t(\boldsymbol{\eta}) := \boldsymbol{\mu}\left(\frac{t}{T}\right) - \boldsymbol{\mu}(1) + \sum_{j=-\infty}^{\infty} (\mathbf{A}_{t,T}(j)\boldsymbol{\varepsilon}_{t-j,T} - \mathbf{A}(1,j)\tilde{\boldsymbol{\varepsilon}}_{t-j}(1))$$

so that $S(\boldsymbol{\eta}) = \sum_{t\in\{W\}} ||\mathbf{V}_t(\boldsymbol{\eta})||$. It follows by the triangle inequality and the reverse triangle inequality that

$$|S(\boldsymbol{\eta}) - S(\boldsymbol{\eta}')| \leq \sum_{t\in\{W\}} |||\mathbf{V}_t(\boldsymbol{\eta})|| - ||\mathbf{V}_t(\boldsymbol{\eta}')||| \leq \sum_{t\in\{W\}} ||\mathbf{V}_t(\boldsymbol{\eta}) - \mathbf{V}_t(\boldsymbol{\eta}')||$$

Denote $\boldsymbol{\delta} = \boldsymbol{\eta} - \boldsymbol{\eta}'$ where $\delta_s = \eta_s - \eta'_s = (\boldsymbol{\varepsilon}_{s,T} - \boldsymbol{\varepsilon}'_{s,T}, \tilde{\boldsymbol{\varepsilon}}_s(1) - \tilde{\boldsymbol{\varepsilon}}'_s(1))$. We have

$$\mathbf{V}_t(\boldsymbol{\eta}) - \mathbf{V}_t(\boldsymbol{\eta}') = \sum_{s\in\mathbb{Z}} (\mathbf{A}_{t,T}(t-s)\delta_{s,1} - \mathbf{A}(1,t-s)\delta_{s,2})$$

Denote $\mathbf{R}_{t,s} = \mathbf{A}_{t,T}(t-s)\delta_{s,1} - \mathbf{A}(1,t-s)\delta_{s,2}$. Applying the triangle inequality followed by Cauchy-Schwarz on the pair $(||A_{t,T}(t-s)||, ||\mathbf{A}(1,t-s)||)^\top$ against $(||\delta_{s,1}||, ||\delta_{s,2}||)^\top$ yields

$$||\mathbf{R}_{t,s}|| \leq ||A_{t,T}(t-s)|| \cdot ||\delta_{s,1}|| + ||\mathbf{A}(1,t-s)|| \cdot ||\delta_{s,2}|| \leq \sqrt{||A_{t,T}(t-s)||^2 + ||\mathbf{A}(1,t-s)||^2} \cdot ||\delta_s||$$

Summing over s and applying Cauchy-Schwarz

$$\begin{aligned} ||\mathbf{V}_t(\boldsymbol{\eta}) - \mathbf{V}_t(\boldsymbol{\eta}')|| \leq \sum_{s\in\mathbb{Z}} ||\mathbf{R}_{t,s}|| &\leq \sum_{s\in\mathbb{Z}} \sqrt{||A_{t,T}(t-s)||^2 + ||\mathbf{A}(1,t-s)||^2} \cdot ||\delta_s|| \\ &\leq \left(\sum_{j\in\mathbb{Z}} ||\mathbf{A}_{t,T}(j)||^2 + ||\mathbf{A}(1,j)||^2\right)^{1/2} \cdot ||\boldsymbol{\delta}|| \end{aligned}$$

Summing over $t \in \{W\}$ and replacing $||\boldsymbol{\delta}|| = \left(\sum_{s\in\mathbb{Z}} ||\eta_s - \eta'_s||^2\right)^{1/2}$ completes the proof

□

Proof of Theorem 4.8. With independent sub-Gaussian innovations of variance proxies $\{\sigma_s^2\}_{s\in\mathbb{Z}}$, we apply the transportation method due to Bobkov and Götze [12] and used in [66, 14].

Since each independent innovation η_s is σ_s^2-sub-Gaussian, the Bobkov-Götze theorem [12] guarantees that its marginal probability measure μ_s satisfies the $T_1(\sigma_s^2)$ transportation-cost inequality with respect to the standard Euclidean norm $||\cdot||$.

Let $\mu = \bigotimes_{s\in\mathbb{Z}} \mu_s$ denote the joint probability measure of the innovation sequence $\boldsymbol{\eta}$. By the tensorization property of the T_1 inequality for product measures, μ satisfies a $T_1(1)$ inequality with respect to the weighted ℓ_2-metric d_σ defined on the product space as:

$$d_\sigma(\boldsymbol{\eta}, \boldsymbol{\eta}') = \sqrt{\sum_{s\in\mathbb{Z}} \frac{||\eta_s - \eta'_s||^2}{\sigma_s^2}}$$

By Proposition 4.7 followed by Cauchy-Schwartz inequality,

$$\begin{aligned} |S(\boldsymbol{\eta}) - S(\boldsymbol{\eta}')| &\leq \sum_{s\in\mathbb{Z}} \kappa_s ||\eta_s - \eta'_s|| \\ &\leq \sqrt{\sum_{s\in\mathbb{Z}} \kappa_s^2 \sigma_s^2} \cdot d_\sigma(\boldsymbol{\eta}, \boldsymbol{\eta}') \end{aligned}$$

It follows that $S(\boldsymbol{\eta})$ is Lipschitz continuous with respect to the weighted metric d_σ, with finite Lipschitz constant $\sqrt{\sum_{s\in\mathbb{Z}} \kappa_s^2 \sigma_s^2}$.

By the dual equivalence in the Bobkov-Götze theorem, any L-Lipschitz function under a measure satisfying a $T_1(1)$ inequality is exactly L^2-sub-Gaussian. Therefore, $S(\boldsymbol{\eta})$ is sub-Gaussian with variance proxy $\sum_{s\in\mathbb{Z}} \kappa_s^2 \sigma_s^2$. Applying the standard Chernoff bound to the moment generating function of $S(\boldsymbol{\eta})$ directly yields the desired tail inequality

$$\mathbb{P}(|S(\boldsymbol{\eta}) - \mathbb{E}S(\boldsymbol{\eta})| \geq t) \leq 2\exp\left\{\frac{-t^2}{2\sum_{s\in\mathbb{Z}} \kappa_s^2 \sigma_s^2}\right\}$$

which completes the proof.

□

B

Extension to non stationary and/or non mixing sequences

Thus far, the consistency of BlockBoost in both classification and regression has been established under the rather strong assumption that the underlying data-generating process is strictly stationary and β-mixing or locally stationary via the stationary approximation theorem. Here, we aim to provide an extension of the block boosting algorithm with generalization bounds for (truly in the sense of not well behaved) non stationary and/or non-mixing sequences. The extension here is a direct implementation of the work of Kuznetsov and Mohri [64] and thus we move the paradigm to that of online learning.

B.1 Definitions and notations

The goal of the learner to select, out of a specified family H, a specified family H, a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that achieves a small path-dependent generalization error

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) = \mathbb{E}[L(h, Z_{T+1})|\mathbf{Z}_1^T]$$

where $L : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ is a loss function bounded by $M > 0$. We use the notion of discrepancy proposed by [64] (see [65] for the origins) that quantifies the divergence of the target and sample distributions defined by

$$\text{disc}(\mathbf{q}) = \sup_{\mathbf{h} \in \mathcal{H}} \left| \sum_{t=1}^{T} q_t \left(\mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) - \mathcal{L}_{t+1}(h_t, \mathbf{Z}_1^t) \right) \right|$$

where $\mathbf{q} = (q_1, \ldots, q_T)$ is an arbitrary weight vector. The following lemma gives an estimation of this quantity based on data. Assume that the loss function L is Lipschitz, that is,

$|L(y,y') - L(y,y'')| \leq Cd(y',y'')$, where $C > 0$ is a constant and d is the underlying metric. Denote by $\mathcal{F} = \{(z,t) \mapsto L(h_t, z) : h \in \mathcal{H}\}$.

Lemma B.1 (Kuznetsov and Mohri, See [65] for more, Lemma 7)**.** *Let $\mathbf{Z}_1^T$ be a sequence of random variables. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $\alpha > 0$:*

$$disc(\mathbf{q}) \leq \widehat{disc}_H(\mathbf{q}) + \inf_{h^*} \mathbb{E}[d(Z_{T+1}, h^*(X_{T+1}))|\mathbf{Z}_1^T] + 2\alpha + M||\mathbf{q}||_2 \sqrt{2 \log \frac{\mathbb{E}_{z \sim \mathcal{Z}_T}[\mathcal{N}_1(\alpha, \mathcal{F}, z)]}{\delta}}$$

where $\mathcal{N}_1(\alpha, \mathcal{F}, z)$ is the sequential covering number and $\widehat{disc}_H(\mathbf{q})$ is the empirical discrepancy defined by

$$\widehat{disc}_H(\mathbf{q}) = \sup_{h \in H, \mathbf{h} \in \mathcal{H}} \left| \sum_{t=1}^{T} q_t (L(h_t(X_{T+1}), h(X_{T+1})) - L(h_t, Z_{t+1})) \right|$$

We consider the general notion of regret used in on-line learning theory for an on-line learning algorithm playing a sequence of hypotheses $\mathbf{h} = (h_1, \ldots, h_T)$ (See [64] for more)

$$\mathrm{Reg}_T = \sum_{t=1}^{T} L(h_t, Z_t) - \inf_{h^*} \left\{ \sum_{t=1}^{T} L_t(h^*, Z_t) + \mathcal{R}(h^*) \right\}$$

We consider sequences of hypotheses $\mathbf{h} = (h_1, \ldots, h_T)$ adapted to the filtration of $\mathbf{Z}_1^T$, that is, such that h_t is $\mathbf{Z}_1^t$-measurable.

B.2 BlockBoost.2SDM

The idea of the extension is to use BlockBoost as an expert generator. We train sequentially the block boosting algorithm to generate a sequence of ensembles adapted to data up to time t. By using the generalization bounds and the ensemble learning application provided by Kuznetsov and Mohri [64], we derive a learning algorithm that extends BlockBoost for non-stationary and/or non-mixing sequences.

BlockBoost.2SDM is a two stage discrepancy minimization algorithm as proposed by [64]. The following theorem will give a good idea of what is going on and also gives the generalization bound of the algorithm that is yet to be defined.

Theorem B.2 (Kuznetsov and Mohri, See [65] for more, Theorem 2)**.** *Assume L is convex and bounded by M. Let $\mathbf{Z}_1^T$ be any sequence of random variables. Let H^* be a set of sequences of hypotheses that are adapted to $\mathbf{Z}_1^T$ and let $h_1, \ldots, h_T$ be a sequence of hypotheses adapted to $\mathbf{Z}_1^T$. Fix a weight vector $\mathbf{q} = (q_1, \ldots, q_T)$ in the simplex and let h denote $h = \sum_{t=1}^{T} q_t h_t$. Then, for any $\delta > 0$, each of the following bounds holds with probability at least $1 - \delta$*

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^t) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},$$

$$\mathcal{L}_{T+1}(h, \mathbf{Z}_1^T) \leq \inf_{\mathbf{h}^* \in H^*} \left\{ \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t^*, \mathbf{Z}_1^T) + \mathcal{R}(\mathbf{h}^*) \right\}$$
$$+ 2\,\mathrm{disc}(\mathbf{q}) + \mathrm{Reg}_T + M\|\mathbf{q} - \mathbf{u}\|_1 + 2M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},$$

where $\mathbf{u} = (\frac{1}{T}, \ldots, \frac{1}{T}) \in \mathbb{R}^T$

Standard boosting algorithms have as objective the minimization of the empirical loss. In non stationary and/or non mixing scenario, minimizing the empirical is not sufficient as suggested by Theorem B.2 because of the presence of the discrepancy term. The goal of BlockBoost.2SDM therefore is to find an optimum point for the bound (right hand side).

Before going into the algorithm, we first need to discuss the theoretical foundations that originally justified and led to the block boosting algorithm [102, 72]. The primary assumptions that justified the blocking principle where the stationarity and the β-mixing condition, at least from a theoretical standpoint. But under those same assumptions, Lozano et al [72] proved the consistency of regularized boosting which has been observed in Figure 3.1, specifically when the underlying process has high signal to noise ratio (which does not destroy the assumptions may the noise random process be stationary and β-mixing itself). Note that the blocking method comes useful (empirically) when the signal to noise ratio starts to diminish (robustness).

Similarly, from a heuristical standpoint, one could attribute or believe that blocking contiguous observations helps to capture "local regimes" in the data. Note that said argument does not require the data to be stationary or β-mixing. In fact, intuitively, a block in a crisis regime (using financial terms for example) should receive different weight than a block in a normal regime, and BlockBoost's reweighting mechanism does exactly this adaptively, without needing to know the regimes in advance.

By this argument, we can move the analysis from stationary and β-mixing sequences to any sequence of random variables. However, this does not intrinsically change the fact that, at the origin, BlockBoost is designed to capture a fixed dependency function between the predictors and the predicted variable. Hence rendering us to use an on-line learning approach that may help us tune the algorithm to focus more on recent observations (regimes) (since they, most of the time, are more important). The idea of the algorithm is the following: given N examples of a random sequence, we isolate $N - T, T < N$ at least as "old" observations. We train BlockBoost on this sequence of old observations capturing past dynamics including regimes, etc; we get from this the ensemble h_0. Then we train sequentially BlockBoost on data from the past to time $t, t = 1, 2, \ldots, T$, incrementally increasing t until we reach N. We get from this ensembles $\mathbf{h} = (h_1, \ldots, h_T)$. We then create a final ensemble by combining these previous ensembles as suggested by Kuznetsov and Mohri [64]: $h = \sum_{t=0}^{T} q_t h_t$

Algorithm 4 gives the pseudocode of a heuristic version of the algorithm BlockBoost.2SDM. In fact, the use of the "old" set hypothesis h_0 is contrary to the framework devised by Kuznetsov and Mohri [64] as it violates the sequential adaptivity condition of the hypothesis sequence $\mathbf{h}$.

Note that in the algorithm, for consistency concerns, we use L to denote the block size like we always did since the beginning of this paper and we use ℓ to denote the loss function (instead of L). In what follows, the block size is denoted as a_T and the loss function L.

To fully match the theoretical framework, one may set $q_0 = 0$ and $\mathbf{u} = \left(\frac{1}{T}, \ldots, \frac{1}{T}\right)^\top$ and move on by optimizing the sequence of hypotheses that are trained only on the recent window of observations. However, including the base ensemble h_0 may prove useful as it captures past dynamics and may strengthen predictions for long dependence sequences. The choice of whether or not to include the base ensemble is let to be further investigated in future works.

The theoretical advantage of BlockBoost used in the two stage discrepancy minimization framework instead of any other on-line learning algorithm comes from the natural block grouping inside the block boosting algorithm. To demonstrate this, we provide below a slight modification of BlockBoost.2SDM for which our results apply. We base our proofs on the work of Kuznetsov and Mohri [64], specifically on Lemma 1 in their paper which is stated here below.

Algorithm 4 BlockBoost.2SDM

Require: Sequence of N time-indexed examples $(x_1, y_1), \ldots, (x_N, y_N)$
Integer $L \leq T < N$ specifying the size of the sequential calibration window
Integer M specifying the number of boosting rounds
Integer L specifying the number of contiguous examples in a block
Base learning algorithm BlockBoost (or BlockBoost.R2)
Regularization hyperparameter $\lambda_1 \geq 0$
Convex loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, M]$

Initialization: Historical training
Extract the historical training sequence $\mathcal{S}_{\text{old}} = \{(x_t, y_t)\}_{t=1}^{N-T}$
Run BlockBoost on $\mathcal{S}_{\text{old}}$ for M boosting rounds,

$$h_0 = \text{BlockBoost}(\mathcal{S}_{\text{old}}, M, L)$$

Stage 1: Sequential Adaptation (Expanding Window Training)
for $t = 1, 2, \ldots, T$ **do**
    Extract the expanding training sequence $\mathcal{S}_t = \{(x_s, y_s)\}_{s=1}^{N-T+t}$
    Run BlockBoost on $\mathcal{S}_t$ for M boosting rounds

$$h_t = \text{BlockBoost}(\mathcal{S}_t, M, L)$$

end for
Collect the sequence of adapted ensembles $\mathbf{h} = (h_1, h_2, \ldots, h_T)$
Stage 2: Ensemble Aggregation (Discrepancy-Based Weighting)
Estimate the empirical discrepancy $\widehat{\text{disc}}_{\mathcal{H}}(\mathbf{q})$ from the calibration sequence
Solve the convex optimization problem [64]:

$$\mathbf{q}^* = \arg\min_{\mathbf{q} \in \Delta_{T+1}} \widehat{\text{disc}}_{\mathcal{H}}(\mathbf{q}) + \sum_{t=0}^{T-1} q_t \ell(h_t, Z_{N-T+t+1}) \quad \text{subject to} \quad \|\mathbf{q} - \mathbf{u}\|_1 \leq \lambda_1$$

where $\mathbf{u} = \left(\frac{1}{T+1}, \ldots, \frac{1}{T+1}\right)^\top$ is the uniform weight vector and Δ_{T+1} is the T-simplex.
Prediction
For a new observation x_{N+1}, predict:

$$\hat{y}_{N+1} = \sum_{t=0}^{T} q_t^* h_t(x_{N+1})$$

Lemma B.3 (Kuznetsov and Mohri, See [65] for more, Lemma 1)**.** *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $\mathbf{h} = (h_1, \ldots, h_T)$ be any sequence of hypotheses adapted to the filtration of $\mathbf{Z}_1^T$. Let $\mathbf{q} = (q_1, \ldots, q_T)$ be any weight vector. For any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$:*

$$\sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) \leq \sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}},$$
$$\sum_{t=1}^{T} q_t L(h_t, Z_{t+1}) \leq \sum_{t=1}^{T} q_t \mathcal{L}_{T+1}(h_t, \mathbf{Z}_1^T) + \mathrm{disc}(\mathbf{q}) + M\|\mathbf{q}\|_2 \sqrt{2 \log \frac{1}{\delta}}.$$

The requirement that the sequence $\mathbf{h} = (h_1, \ldots, h_T)$ be adapted to $\mathbf{Z}_1^T$ has natural explanation from an on-line learning viewpoint since hypothesis h_t is based on data $\mathbf{Z}_1^t$. An intuitive explanation of what is suggested to be done here is that at each time t, we train a new hypothesis h_t that uses exactly data up to time t ($\mathbf{Z}_1^t$). But one could choose a configuration where we conserve the hypothesis h_1 learnt at the very first iteration through the whole time sequence, that is, $\mathbf{h}^1 = (h_1, \ldots, h_1)$. The sequence $\mathbf{h}^1$ is provably adapted to $\mathbf{Z}_1^T$ because h_1 is $\mathbf{Z}_1^t, \forall t \geq 1$. This is precisely where we position the advantage (nuanced) of BlockBoost.2SDM.

Consider in Algorithm 4 that $T = K_T a_T, K_T \in \mathbb{N}^*$ i.e. T is a multiple of block size a. We modify Algorithm 4 by changing stage 1 (sequential adaptation).

Stage 1: Sequential Adaptation (Expanding Window Training)
for $\mathrm{t} \in \{\mathrm{a_T}, 2\mathrm{a_T}, \ldots, \mathrm{a_T K_T}\}$ **do**
Extract the expanding training sequence $\mathcal{S}_t = \{(\mathrm{x_s}, \mathrm{y_s})\}_{\mathrm{s=1}}^{\mathrm{N-T+t}}$
Run BlockBoost on $\mathcal{S}_t$ for M boosting rounds; $\mathrm{h_t} = \textsc{BlockBoost}(\mathcal{S}_\mathrm{t}, \mathrm{M}, \mathrm{a_T})$

An intuitive explanation here is that we basically retrain the model when we see the totality of a new block. We start with h_0 at $t = 0$ and keep it until we reach $t = a_T$ where we retrain and get h_1, and so on. The sequence $\mathbf{h} = (h_0, \ldots, h_0, h_1, \ldots, h_1, \ldots, h_{K_T-1}, \ldots, h_{K_T-1}, h_{K_T})$ is adapted to $\mathbf{Z}_1^T$.

Formally, define: $h_t = h_{\tau(t)}, t = 1, \ldots, T$ where $\tau(t) = \left\lfloor \frac{t}{a_T} \right\rfloor$ is the most recent block boundary index relative to t. There are $K_T + 1$ distinct hypotheses in the calibration window.

Define the block-aggregated weight vector $\mathbf{Q} \in \mathbb{R}^{K_T+1}$:

$$Q_j = \sum_{t=\tau_j}^{\tau_{j+1}-1} q_t, j = 0, 1, \ldots, K_T$$

where $\tau_0 = 0, \tau_j = j a_T, \tau_{K_T+1} = T + 1$. Note that $\sum_j Q_j = \sum_t q_t = 1$ so $\mathbf{Q} \in \Delta_{K_T+1}$

Lemma B.4. *Let $\mathbf{Z}_1^T$ be any sequence of random variables and let $\mathbf{h} = (h_1, \ldots, h_T)$ be the sequence of BlockBoost hypotheses produced by BlockBoost.2SDM with block size a_T and let $h = \sum_{t=0}^{T} q_t h_t$ be the final ensemble with $\mathbf{q} \in \Delta_{T+1}$. For any $\delta > 0$, each of the following inequalities holds with probability at least $1 - \delta$:*

$$\mathcal{L}_{N+1}(h, \mathbf{Z}_1^N) \leq \sum_{t=0}^{T} q_t \mathcal{L}_{N+1}(h_t, \mathbf{Z}_1^N) = \sum_{j=0}^{K_T} Q_j \mathcal{L}_{N+1}(h_{\tau_j}, \mathbf{Z}_1^N) \leq \sum_{j=0}^{K_T} Q_j L(h_{\tau_j}, Z_{N-T+\tau_j+1})$$
$$+ \mathit{disc}^{\mathrm{BB}}(\mathbf{Q}) + M||\mathbf{Q}||_2 \sqrt{2 \log \frac{1}{\delta}}$$

where the discrepancy is

$$disc^{\mathrm{BB}}(\mathbf{Q}) = \sup_{\mathbf{h}\in\mathcal{H}^{\mathrm{BB}}} \left| \sum_{j=0}^{K_T} Q_j \left(\mathcal{L}_{N+1}(h_{\tau_j}, \mathbf{Z}_1^N) - \mathcal{L}_{N-T+\tau_j+1}(h_{\tau_j}, \mathbf{Z}_1^{N-T+\tau_j}) \right) \right|$$

Proof. By the convexity of L, we can write $\mathcal{L}_{N+1}(h, \mathbf{Z}_1^N) \leq \sum_{t=0}^{T} q_t \mathcal{L}_{N+1}(h_t, \mathbf{Z}_1^N)$ (Jensen's inequality). Since $h_t = h_{\tau_j}$ for all $t \in [\tau_j, \tau_{j+1})$ by the block constant property, we have $\sum_{t=0}^{T} q_t \mathcal{L}_{N+1}(h_t, \mathbf{Z}_1^N) = \sum_{j=0}^{K_T} Q_j \mathcal{L}_{N+1}(h_{\tau_j}, \mathbf{Z}_1^N)$.

By definition of $\mathrm{disc}^{\mathrm{BB}}(\mathbf{Q})$, we have

$$\sum_{j=0}^{K_T} Q_j \mathcal{L}_{N+1}(h_{\tau_j}, \mathbf{Z}_1^N) \leq \sum_{j=0}^{K_T} Q_j \mathcal{L}_{N-T+\tau_j+1}(h_{\tau_j}, \mathbf{Z}_1^{N-T+\tau_j}) + \mathrm{disc}^{\mathrm{BB}}(\mathbf{Q})$$

Define for each j, $A_j = Q_j \left(\mathcal{L}_{N-T+\tau_j+1}(h_{\tau_j}, \mathbf{Z}_1^{N-T+\tau_j}) - L(h_{\tau_j}, Z_{N-T+\tau_j+1}) \right)$

Since h_{τ_j} is $\mathbf{Z}_1^{N-T+\tau_j}$-measurable (adapted by construction), we have:

$$\begin{aligned}
&\mathbb{E}[A_j | \mathbf{Z}_1^{N-T+\tau_j}] \\
&= Q_j \left(\mathbb{E}\left[\mathcal{L}_{N-T+\tau_j+1}(h_{\tau_j}, \mathbf{Z}_1^{N-T+\tau_j}) \middle| \mathbf{Z}_1^{N-T+\tau_j} \right] - \mathbb{E}\left[L(h_{\tau_j}, Z_{N-T+\tau_j+1}) \middle| \mathbf{Z}_1^{N-T+\tau_j} \right] \right) \\
&= Q_j \left(\mathbb{E}\left[L(h_{\tau_j}, Z_{N-T+\tau_j+1}) \middle| \mathbf{Z}_1^{N-T+\tau_j} \right] - \mathbb{E}\left[L(h_{\tau_j}, Z_{N-T+\tau_j+1}) \middle| \mathbf{Z}_1^{N-T+\tau_j} \right] \right) = 0
\end{aligned}$$

So $(A_j)_{j=0}^{K_T}$ is a martingale difference sequence with $|A_j| \leq M|Q_j|$. By Azuma's inequality, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$

$$\sum_{j=0}^{K_T} Q_j \mathcal{L}_{N-T+\tau_j+1}(h_{\tau_j}, \mathbf{Z}_1^{N-T+\tau_j}) \leq \sum_{j=0}^{K_T} Q_j L(h_{\tau_j}, Z_{N-T+\tau_j+1}) + M||\mathbf{Q}||_2 \sqrt{2 \log \frac{1}{\delta}}$$

Thus completing the proof. □

Note that this proof follows the same structure as proof of Lemma 1 in Kuznetsov and Mohri [64]. However, the improvement here comes from two things. First, this new bound worsens the generalization bound compared to the one in B.3 by increasing $||\mathbf{Q}||_2$ and improves it by reducing the discrepancy. To make this precise, let us assume uniform weights, that is, $q_t = \frac{1}{T+1}$. We have $||\mathbf{q}||_2^2 = \frac{1}{T+1}$ and $Q_j = \frac{a_T}{T+1}$ so

$$||\mathbf{Q}||_2^2 = \sum_{j=0}^{K_T} Q_j^2 = \frac{a_T^2(K_T+1)}{(T+1)^2} \implies \frac{||\mathbf{Q}||_2^2}{||\mathbf{q}||_2^2} = \frac{a_T^2(K_T+1)}{T+1} \geq 1 \text{ iff } a_T > 1$$

This implies that if $a_T > 1$, then we are increasing the bound by increasing the euclidian norm of the weights. But notice that the space of block constant hypothesis sequences $\mathcal{H}^{\mathrm{BB}}$ form a strict subset of all adapted hypothesis sequences $\mathcal{H}$. Therefore, by the definition of the supremum, the discrepancy provided by blocking is strictly smaller than the one provided by standard discrepancy, that is,

$$\mathcal{H}^{\mathrm{BB}} \subset \mathcal{H} \implies \mathrm{disc}^{\mathrm{BB}}(\mathbf{Q}) \leq \mathrm{disc}(\mathbf{q})$$

The conclusion we derive is that the BlockBoost.2SDM bound is tighter if and only if the reduction in discrepancy induced by blocking a_T observations outweighs the degradation of the euclidian norm of the weights. Ultimately, this is a characteristic of the sequence itself rather than the algorithm. For hyper non-stationary sequences, one might want to block observations as there is more chance that the reduction in discrepancy will be consequent, blocking which would probably harm performance otherwise.

C

Experimental Designs

C.1 Experimental design for BlockBoost

The experimental design was devised to address the following questions: Is BlockBoost resistant to overfitting as theory suggests ? Does it perform well in (noisy) stationary time series with β-mixing condition verified ? How much does its performance decrease as the number of estimators, i.e. the number of boosting rounds M is increased ? Compared to AdaBoost ?

To test these questions posed by the experimental design, we implement non-regularized versions of BlockBoost and AdaBoost.M1 [44] (for binary classification). We recall that we use the non-regularized version in order to observe brut behavior of algorithms rather than engineered and controlled behavior. Both algorithms use the same weak learner, a decision tree (usually, Cart with maximum depth 1), and the number of boosting rounds is $M = 1500$ (to observe the collapse predicted for AdaBoost in the presence of noise [69]). After training, the models are tested on the test set as an i.i.d. drawn from the sequence, that is we use no rolling window approach here for simplicity. We believe this method makes sense since the entire argument made upon the use of the blocking technique is that distant observations of a stationary β-mixing sequence behave as independent and identically distributed ones.

Data Generating Process 1 ($\mathbf{DGP_1}$)

To empirically validate the noise-robustness and generalization guarantees of BlockBoost against standard AdaBoost.M1, we use $\mathbf{DGP_1}$ i.e. the synthetic stationary autoregressive sequences. This controlled framework allows us to strictly isolate the algorithm's response to random classification noise under severe temporal dependence (recall $\rho = 0.95$).

We inject random classification noise into the training and test labels, holding true test labels (to evaluate generalization capacity). We define the noise rate $\eta \in \{0.0, 0.1, 0.2, 0.3\}$. For each observation, an independent Bernoulli trial determines whether the true label is flipped,

producing the observed noisy label $\tilde{y}_t = (1-2\xi_t)y_t, \xi_t \sim \mathcal{B}(\eta)$. This process is repeated independently across 10 distinct Monte Carlo seeds to ensure statistical significance, resampling both the covariate trajectories and the label noise masks for each iteration.

We use depth-1 decision trees as hypothesis class and run the boosting process for $M = 1500$ rounds to thoroughly capture the asymptotic trajectory. The block size is fixed $a_T = 10$.

Through the learning process, three evaluation metrics are tracked: noisy training and test accuracies and clean test accuracy. The noisy training accuracy evaluates the ensemble's empirical risk on the corrupted training labels to detect noise memorization while the noisy test accuracy evaluates out-of-sample performance on identically corrupted test labels. The clean test accuracy evaluates the true structural generalization of the ensemble by scoring the predictions against the unobservable, perfectly clean labels y_t.

Data Generating Process 2 ($\mathbf{DGP_2}$)

This data generating process was designed explicitly to stress AdaBoost under random classification noise with a construction similar to Long and Servedio [69].

The baseline remains AdaBoost.M1 but here we test it against BlockBoost with different block sizes ($a_T \in \{10, 15, 20, 30\}$) and we instantiate the hypothesis class as decision trees with maximum depth 2 to accommodate the 2-dimensional cluster geometry. The algorithms are trained for $M = 5000$ rounds to guarantee that any theoretical asymptotic collapse will physically manifest. The process is repeated independently across 20 distinct Monte Carlo seeds.

We track the clean test accuracy (generalization), the noisy train accuracy (memorization), the algorithm momentum (α_m) and the weight concentration $\sum_t \left(w_t^{(m)}\right)^2$.

C.2 Experimental design for BlockBoost.R2

The experiments were devised to answer the following questions: First, does the block reweighting mechanism of BlockBoost.R2 provide measurable forecasting gains over AdaBoost.R2 and time series specific baseline models like AR and AR+GARCH on a temporally structured regression target, and are these gains robust across random seeds? Second, how sensitive is BlockBoost.R2 performance to the block size hyperparameter L, and is there an optimal block size consistent with the dependence structure of the $\mathbf{DGP_3}$?

Data Generating Process 3: We first note that the data generating process is β-mixing by construction. All models are evaluated under a strictly out-of-sample, expanding-window protocol. We recall $N_{\text{train}} = \lfloor 0.7 \times N_{\text{eff}} \rfloor$ where $N_{\text{eff}} = N - p = 3500 - 21 = 3479$ is the number of valid observations after lag construction. The training set is denoted as $\mathbf{z}_1^{N_{\text{train}}}$, the testing set as $\mathbf{z}_{N_{\text{train}}+1}^{N_{\text{eff}}}$ contains $N_{\text{test}} = N_{\text{eff}} - N_{\text{train}}$ observations, evaluated strictly out-of-sample.

For boosting models, predictions are generated via a rolling-origin scheme. The test horizon is partitioned into non-overlapping chunks of stride 63 steps (approximately quaterly). At the start of each chunk $\mathbf{c}$, a new model is fitted on the full expanding window $\mathbf{z}_1^{t_\mathbf{c}}$ where $t_\mathbf{c}$ is the first index of chunk $\mathbf{c}$. Predictions for all steps within the chunk are generated from this single fitted model without refitting, that is, for all $t \in \mathbf{c}$, $\hat{y}_t = h^{(c)}(x_t), h^{(c)} = \text{BlockBoost}(\mathbf{z}_1^{t_\mathbf{c}}, M, L)$. This procedure reduces computational cost relative to full rolling-origin refitting while preserving the out-of-sample integrity of all predictions. The boosting models are run independently with $M = 25$ boosting rounds, block size $L \in \{1, 5, 10, 15, 20\}$ (recall that block size 1 coincides with AdaBoost.R2). The base learner WeakLearn is a depth-3 decision tree for both algorithms.

For AR(p) and AR(p)+GARCH(1, 1) baselines, the model is refit every 21 steps (approximately monthly), consistent with standard practice in financial forecasting. At each refitting step, the AR lag order p^* is selected by minimizing the AIC over $p \in \{1, \ldots, 21\}$ on the current expanding window, then the model is fitted via OLS with heteroskedasticity-consistent standard errors. The AR(p)+GARCH(1, 1) model uses Student-t innovations and is estimated via maximum likelihood on scaled returns $\tilde{r}_t = 100 y_t$ (percentage) to improve numerical convergence of the GARCH recursion.

The models are evaluated on the basis of the mean absolute error (MAE) and mean squared error (MSE) both computed on the testing set. Statistical significance of pairwise forecast differences is assessed via the Diebold-Mariano test [33]. All reported p-values are two-sided and computed against the AR(p^*) and AR(p)+GARCH(1, 1) baseline.

D

Value at Risk backtesting

In the current context of financial risk management, rigorous evaluation of the performance of Value at Risk (VaR) models is imperative, both from a regulatory and managerial perspective. Indeed, the reliability of a VaR model directly determines a financial institution's ability to anticipate and manage its exposure to market risks [80].

Backtesting is the cornerstone of this assessment process. This statistical method systematically compares the VaR forecasts generated by a model with the financial losses actually recorded over a given historical period. The fundamental objective of backtesting is to verify whether the model's forecasts are consistent with the observed reality of the financial markets. More specifically, a VaR model can be considered statistically valid when two essential conditions are simultaneously met [20]:

- **Coverage condition**: The number of observed exceptions (situations where the actual loss exceeds the expected VaR) must be statistically consistent with the confidence level chosen when estimating the model. For example, for a confidence level of 95%, approximately 5% of violations should be observed over a sufficiently long period.
- **Independence condition**: Exceptions must not show any significant temporal clustering. In other words, violations must be randomly distributed over time, with no particular concentration during certain periods.

The central concept of VaR violations The formal construction of backtesting tests is based on the analysis of a sequence of indicator variables, commonly referred to as 'hits' or "violations" in the specialised literature. For each day t of the observation period, an indicator variable $I_t(\alpha)$ is defined that captures the occurrence of a VaR exceedance :

$$I_t(\alpha) = \begin{cases} 1 & \text{si } r_t < -\text{VaR}_t(\alpha) \\ 0 & \text{sinon} \end{cases}$$

where r_t represents the return (or negative loss) observed at time t, $\text{VaR}_t(\alpha)$ denotes the Value at Risk forecast for the confidence level α (typically 1% or 5% depending on regulatory and institutional practices). A value of $I_t(\alpha) = 1$ signals the occurrence of a breach, i.e. a situation where the actual loss has exceeded the limit set by the VaR.

From a theoretical standpoint, for a correctly specified VaR model that accurately reflects the underlying distribution of returns, the sequence of hits $\{I_t(\alpha)\}_{t=1}^{T}$ should behave like a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables with a probability of success equal to α [20]. This property forms the theoretical basis for all of the backtesting tests that we will now examine in detail.

D.1 Kupiec's unconditional coverage test

The unconditional coverage test, developed by Kupiec (1995) [63], is one of the first systematic approaches to assessing the validity of a VaR model. This test, also known as the "proportion of failures test", focuses exclusively on the overall frequency of violations, without taking into account their temporal distribution [82, 52]. The underlying intuition is relatively simple: if a VaR model is calibrated for a confidence level of $(1-\alpha)\times 100\%$, then over a sufficiently long period, we should observe approximately $\alpha \times 100\%$ violations. Any significant deviation from this theoretical proportion suggests that the model is poorly specified.

Let N be the total number of violations observed over an evaluation period comprising T trading days. The empirical proportion of violations is then given by $\hat{p} = N/T$. The Kupiec test aims to determine whether this observed proportion is statistically compatible with the theoretical probability of violation α specified when constructing the model.

The hypothetical framework of the test is formulated as follows [92, 68]:

$$H_0 : p = \alpha \quad \text{and} \quad H_1 : p \neq \alpha$$

The construction of the test statistic is based on the principle of likelihood ratio. Under the null hypothesis, the number of violations N follows a binomial distribution $\mathcal{B}(T, \alpha)$, and the associated likelihood function is written as:

$$L(\alpha) = (1-\alpha)^{T-N}\alpha^{N}$$

The likelihood ratio LR_{UC} (Unconditional Coverage) compares the maximised likelihood under the null hypothesis to the maximised likelihood under the alternative hypothesis. Under the alternative hypothesis, the maximum likelihood estimator of p is simply $\hat{p} = N/T$. Kupiec's test statistic is then expressed as [82, 78, 85]:

$$\text{LR}_{\text{UC}} = -2\ln\left[\frac{L(\alpha)}{L(\hat{p})}\right] = -2\ln\left[\frac{(1-\alpha)^{T-N}\alpha^{N}}{\left(1-\frac{N}{T}\right)^{T-N}\left(\frac{N}{T}\right)^{N}}\right]$$

Under the null hypothesis and for a sample of sufficient size, the asymptotic theory of likelihood ratio tests guarantees that the statistic LR_{UC} follows a chi-square distribution (χ^2) with one degree of freedom. The decision rule is as follows:

- If $\text{LR}_{\text{UC}} > \chi^2_{1,1-\beta}$, where $\chi^2_{1,1-\beta}$ denotes the $(1-\beta)$th quantile of the χ^2 distribution with one degree of freedom (with β being the significance level of the test, usually 5%), then the null hypothesis is rejected.

- Otherwise, there is insufficient statistical evidence to reject the model on the basis of the proportion of violations.

Despite its conceptual simplicity and ease of implementation, the Kupiec test suffers from a major limitation that considerably restricts its practical scope. Indeed, this test does not take into account the temporal order or grouping of violations [42].

However, from a risk management perspective, this temporal dimension is of paramount importance. A VaR model may well have an overall proportion of violations that is in line with expectations (and therefore pass the Kupiec test), while generating sequences of consecutive violations during certain periods. Such behaviour generally indicates that the model is failing to capture the dynamics of volatility correctly, particularly during periods of market turbulence when volatility shocks tend to spread over time. For a risk manager, the observation of clustered violations is particularly concerning because it signals that, precisely when market conditions deteriorate (and therefore when risk measurement is most critical), the VaR model systematically underestimates the actual risk. This fundamental flaw in the Kupiec test has led to the development of more sophisticated tests that take into account the temporal dimension of violations.

D.2 Christoffersen's hit independence test

To overcome the shortcomings of the Kupiec test, Christoffersen (1998) [26] developed a test specifically designed to assess the temporal independence of violations. This test is based on the observation that, in a correctly specified VaR model, violations should be distributed independently over time [24, 25]. The temporal clustering of violations is a major red flag for several reasons. First, it suggests that the model fails to adequately capture variations in conditional volatility, a phenomenon that is particularly pronounced during periods of market stress. Second, this clustering indicates that extreme losses tend to occur consecutively, which significantly amplifies the financial institution's risk exposure. Third, from a regulatory perspective (particularly under the Basel Accords), the presence of clustered violations can result in penalties and increased capital requirements.

Christoffersen's independence test models the sequence of hits as a first-order Markov chain. In this approach, we consider that the probability of observing a violation at time t may potentially depend on the state (violation or non-violation) observed at the previous time $t-1$. More formally, we define a transition matrix characterised by four conditional probabilities [24, 25]:

- π_{00}: Probability of not observing a violation today ($I_t = 0$) given that no violation was observed yesterday ($I_{t-1} = 0$)
- π_{01}: Probability of observing a violation today ($I_t = 1$) given that no violation was observed yesterday ($I_{t-1} = 0$)
- π_{10}: Probability of not observing a violation today ($I_t = 0$) given that a violation was observed yesterday ($I_{t-1} = 1$)
- π_{11}: Probability of observing a violation today ($I_t = 1$) given that a violation was observed yesterday ($I_{t-1} = 1$)

By construction, these probabilities satisfy the constraints: $\pi_{00} + \pi_{01} = 1$ and $\pi_{10} + \pi_{11} = 1$. For the purposes of the independence test, we focus mainly on π_{01} and π_{11}.

The null hypothesis of the independence test posits that violations are independent over time. This independence implies that the probability of observing a violation today is the same, regardless of whether or not a violation was observed yesterday. Formally:

$$H_0 : \pi_{01} = \pi_{11} = \pi \quad \text{and} \quad H_1 : \pi_{01} \neq \pi_{11}$$

where π represents the unconditional probability of violation, which under H_0 does not depend on the previous state.

The LR_{IND} (Independence test) statistic is based on a likelihood ratio comparing the constrained model (under H_0, with a single violation probability π) to the unconstrained model (under H_1, with two distinct probabilities π_{01} and π_{11}). To construct this statistic, we begin by counting the transitions observed in the sequence of hits:

- N_{00}: number of transitions from the "no violation" state to "no violation" state
- N_{01}: number of transitions from the "no violation" state to "violation" state
- N_{10}: number of transitions from the "violation" state to "no violation" state
- N_{11}: number of transitions from the "violation" state to "violation" state

The maximum likelihood estimators of the transition probabilities are:

$$\hat{\pi}_{01} = \frac{N_{01}}{N_{00} + N_{01}}, \quad \hat{\pi}_{11} = \frac{N_{11}}{N_{10} + N_{11}}, \quad \hat{\pi} = \frac{N_{01} + N_{11}}{T}$$

The test statistic is then expressed as [24, 25]:

$$\text{LR}_{\text{IND}} = -2 \ln \left[\frac{(1 - \hat{\pi})^{T-N} \hat{\pi}^{N}}{(1 - \hat{\pi}_{01})^{N_{00}} \hat{\pi}_{01}^{N_{01}} (1 - \hat{\pi}_{11})^{N_{10}} \hat{\pi}_{11}^{N_{11}}} \right]$$

Under the null hypothesis of independence of violations, the statistic LR_{IND} asymptotically follows a chi-square distribution with one degree of freedom (χ_1^2). The standard decision rule applies: the null hypothesis is rejected if LR_{IND} exceeds the appropriate quantile of the χ_1^2 distribution. Rejection of the null hypothesis indicates the presence of significant temporal dependence in the sequence of violations. In practice, when $\hat{\pi}_{11} > \hat{\pi}_{01}$, this suggests a clustering phenomenon: a violation yesterday increases the probability of observing a violation today, which is typical of a model that fails to correctly capture the persistence of volatility.

D.3 Christoffersen's conditional coverage test

The conditional coverage test, also developed by Christoffersen (1998) [26], represents an elegant and powerful synthesis of the two previous approaches. Rather than testing the proportion of violations (Kupiec's test) and their temporal independence (independence test) separately, this joint test simultaneously evaluates the two fundamental properties that a valid VaR model must satisfy [82, 2, 89]. This integrated approach has several methodological advantages. On the one hand, it offers a single, consistent test that reflects the very definition of a correctly specified VaR model. On the other hand, it allows the overall significance level of the test to be controlled appropriately, thus avoiding the multiple testing problems that would arise if the Kupiec test and the independence test were performed separately.

The null hypothesis of the conditional coverage test combines the two previous conditions into a single specification [2]:

$$H_0 : \pi_{01} = \pi_{11} = \alpha \quad \text{and} \quad H_1 : H_0 \text{ est fausse}$$

The construction of the LR_{CC} (Conditional Coverage) test statistic exploits a fundamental property of likelihood ratios: additivity. Indeed, the conditional coverage test can be broken down as the sum of the two previous tests [82], [89]:

$$\text{LR}_{\text{CC}} = \text{LR}_{\text{UC}} + \text{LR}_{\text{IND}}$$

This additive decomposition has an intuitive interpretation: the joint test simultaneously measures the deviation from the expected proportion of violations (captured by LR_{UC}) and the deviation from the independence hypothesis (captured by LR_{IND}).

The additivity property of likelihood ratio statistics is also reflected in the asymptotic distribution. Under the null hypothesis of correct conditional coverage, the statistic LR_{CC} follows a chi-square distribution with two degrees of freedom (χ^2_2). This result follows directly from the fact that the null hypothesis imposes two constraints: $\pi_{01} = \alpha$ and $\pi_{11} = \alpha$.

The standard testing procedure applies:

- If $\text{LR}_{\text{CC}} > \chi^2_{2,1-\beta}$, where $\chi^2_{2,1-\beta}$ is the $(1-\beta)$ quantile of the χ^2_2 distribution, the null hypothesis is rejected.
- A rejection indicates that the VaR model is inadequate, either because it generates an incorrect proportion of violations, or because these violations exhibit temporal dependence, or both simultaneously.

D.4 Diebold and Mariano Test

The Diebold and Mariano test (1995) [34, 33] occupies a distinct place in the VaR backtesting toolbox. Unlike previous tests (Kupiec, Christoffersen) that focus on the statistical properties of violations as a binary sequence (violation or non-violation), the Diebold-Mariano (DM) test takes a comparative perspective: it allows us to assess whether one forecasting model is significantly more accurate than another in terms of loss function [82]. This approach is particularly useful in the following contexts:

- **Model selection**: When a risk manager has several candidate VaR models (e.g., historical VaR, parametric VaR, GARCH VaR, etc.) and wishes to identify the model offering the best predictive performance.
- **Continuous improvement**: To assess whether a modification to an existing model (change in specification, addition of explanatory variables, modification of the estimation window, etc.) significantly improves its predictive accuracy.
- **Asymmetric cost sensitivity**: The DM test explicitly incorporates the fact that, from a risk management perspective, different types of forecasting errors do not have the same cost (underestimating risk is generally more serious than overestimating it).

The DM test is based on the definition of a loss function that quantifies the difference between the VaR forecast and the actual loss. For VaR models, asymmetric loss functions are generally preferred in order to reflect the natural preference for models that avoid underestimating risk [3], [27]. Several loss functions have been proposed in the literature. The tick loss function proposed by Lopez (1999) [70, 71] is particularly popular:

$$\mathrm{L_t}(\mathrm{VaR}_t(\alpha)) = \begin{cases} 1 + (r_t + \mathrm{VaR}_t(\alpha))^2 & \text{si } r_t < -\mathrm{VaR}_t(\alpha) \text{ (violation)} \\ 1 & \text{sinon} \end{cases}$$

This function penalises violations twice: first by the occurrence of the violation itself (constant term), then proportionally to the square of the magnitude of the excess (quadratic term). Non-violations receive only a minimal constant penalty.

Another commonly used function is the asymmetric quadratic loss function:

$$\mathrm{L_t}(\mathrm{VaR}_t(\alpha)) = \begin{cases} \theta(r_t + \mathrm{VaR}_t(\alpha))^2 & \text{si } r_t < -\mathrm{VaR}_t(\alpha) \\ (1-\theta)(r_t + \mathrm{VaR}_t(\alpha))^2 & \text{sinon} \end{cases}$$

where $\theta > 0.5$ is an asymmetry parameter that gives greater weight to violations.

Consider two competing VaR models, denoted (Model) 1 and (Model) 2, generating the forecasts $\mathrm{VaR}_{1,t}(\alpha)$ and $\mathrm{VaR}_{2,t}(\alpha)$ respectively for each day t of the evaluation period. For a chosen loss function $L(\cdot)$, we calculate the losses associated with each model: $L_t(\mathrm{VaR}_{1,t}(\alpha))$ and $L_t(\mathrm{VaR}_{2,t}(\alpha))$.

The loss differential d_t measures the difference in performance between the two models at each point in time:

$$d_t = \mathrm{L_t}(\mathrm{VaR}_{1,t}(\alpha)) - \mathrm{L_t}(\mathrm{VaR}_{2,t}(\alpha))$$

If, on average, $d_t > 0$, this suggests that Model 1 generates greater losses than Model 2, and therefore that Model 2 is preferable. Conversely, if $d_t < 0$ on average, Model 1 is preferable. The null hypothesis of the DM test posits equal predictive performance:

$$H_0 : \mathbb{E}[d_t] = 0 \quad \text{and} \quad H_1 : \mathbb{E}[d_t] \neq 0$$

The DM test statistic is then written as [88]:

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}}$$

where:

- $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ is the empirical mean of the loss differentials
- $\widehat{\mathrm{Var}}(\bar{d})$ is a robust estimate of the variance of $\bar{d}$, usually obtained by the Newey-West estimator, which takes into account the potential autocorrelation in the series of loss differentials.

The appropriate estimation of $\widehat{\mathrm{Var}}(\bar{d})$ is crucial for the validity of the DM test. In the presence of autocorrelation in the series $\{d_t\}$, the naive estimator of the variance (simply the empirical variance of d_t divided by T) would be biased, leading to incorrect inferences. The Newey-West estimator (1987) corrects this problem by taking into account autocovariances up to a certain order l (number of lags) [82], [88]:

$$\widehat{\mathrm{Var}}(\bar{d}) = \frac{1}{T}\left(\gamma_0 + 2\sum_{k=1}^{l-1} w_k \gamma_k\right)$$

where:

- $\gamma_k = \frac{1}{T}\sum_{t=k+1}^{T}(d_t - \bar{d})(d_{t-k} - \bar{d})$ is the autocovariance of order k
- $w_k = 1 - \frac{k}{l}$ are Bartlett's weights, which ensure that the variance estimator is positive
- l is chosen according to the sample size, for example $l = \lfloor T^{1/3} \rfloor$ or $l = \lfloor 4(T/100)^{2/9} \rfloor$

Under the null hypothesis of equal predictive performance and under certain conditions of regularity, Diebold and Mariano (1995) demonstrated that the test statistic asymptotically follows a reduced normal distribution:

$$\text{DM} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{under } H_0$$

The decision rule for a two-tailed test at significance level β is as follows:

- We reject H_0 if $|\text{DM}| > z_{1-\beta/2}$, where $z_{1-\beta/2}$ is the $(1-\beta/2)$th quantile of the standard normal distribution (for example, $z_{0.975} = 1.96$ for $\beta = 5\%$)
- If $\text{DM} > z_{1-\beta/2}$, we conclude that Model 2 is significantly more accurate than Model 1.
- If $\text{DM} < -z_{1-\beta/2}$, we conclude that Model 1 is significantly more accurate than Model 2.

E

Additional statistical tests

E.1 Zivot-Andrews Structural Break Test

The Zivot-Andrews (ZA) test [104] is an endogenous structural break test designed to detect a single unknown break in the intercept, the trend, or both, of a time series under the alternative hypothesis of stationarity. It extends the classical Augmented Dickey-Fuller (ADF) test [32] by allowing the break date to be determined by the data rather than specified a priori, thereby addressing the well-documented sensitivity of standard unit root tests to the presence of structural change [84].

The null hypothesis of the ZA test is that the series $\{y_t\}$ contains a unit root with no structural break:

$$H_0 : y_t = \mu + y_{t-1} + \varepsilon_t$$

The alternative hypothesis admits a single structural break at an unknown date $T_b \in (1, T)$. Three model specifications are considered:

- **Model A** – break in the intercept only:

$$y_t = \hat{\mu} + \hat{\beta}t + \hat{\gamma}DU_t(\lambda) + \hat{\alpha}y_{t-1} + \sum_{j=1}^{k} \hat{d}_j \Delta y_{t-j} + \hat{\varepsilon}_t \tag{E.1}$$

- **Model B** – break in the trend only:

$$y_t = \hat{\mu} + \hat{\beta}t + \hat{\gamma}DT_t(\lambda) + \hat{\alpha}y_{t-1} + \sum_{j=1}^{k} \hat{d}_j \Delta y_{t-j} + \hat{\varepsilon}_t \tag{E.2}$$

- **Model C** – break in both intercept and trend:

$$y_t = \hat{\mu} + \hat{\beta}t + \hat{\gamma}DU_t(\lambda) + \hat{\theta}DT_t(\lambda) + \hat{\alpha}y_{t-1} + \sum_{j=1}^{k} \hat{d}_j \Delta y_{t-j} + \hat{\varepsilon}_t \tag{E.3}$$

where $\lambda = T_b/T \in (0,1)$ denotes the break fraction, $DU_t(\lambda) = \mathbb{1}(t > T_b)$ is a dummy variable indicating the post-break intercept shift, $DT_t(\lambda) = (t - T_b)\,\mathbb{1}(t > T_b)$ captures the post-break trend shift, and k is the number of lagged difference terms included to correct for serial correlation. The lag order k is typically selected by minimizing the Akaike Information Criterion on the augmented regression.

The break date $\hat{T}_b$ is chosen endogenously as the value of T_b that minimizes the t-statistic on the autoregressive coefficient $\hat{\alpha}$ – equivalently, the break date that provides the strongest evidence against the unit root null:

$$\hat{T}_b = \arg\min_{T_b \in \mathcal{T}} t_{\hat{\alpha}}(T_b)$$

where $\mathcal{T} = \{[0.15T], \ldots, [0.85T]\}$ trims the boundary observations to ensure sufficient data in each subsample. The test statistic is the minimized t-ratio:

$$\text{ZA} = \inf_{T_b \in \mathcal{T}} t_{\hat{\alpha}}(T_b)$$

The null is rejected when ZA falls below the critical values tabulated by [104], which differ from those of the standard ADF test due to the data-driven selection of T_b. Under the null, the asymptotic distribution of the test statistic is non-standard and is obtained by simulation. For Model C, the asymptotic 1%, 5%, and 10% critical values are approximately -5.57, -5.08, and -4.82 respectively. Rejection of the null implies that the series is stationary around a single structural break occurring at $\hat{T}_b$, providing both evidence of stationarity and a data-driven estimate of the break date.

An important limitation of the ZA test is that it is designed to detect at most one structural break. When multiple breaks are present, the test may exhibit reduced power or misidentify the dominant break [73]. In such cases, the Bai-Perron test [4] provides a natural extension to the multiple break setting. For the purposes of the present analysis, the single-break assumption is considered adequate given the relatively short post-break stationary windows available for each asset.

E.2 Inclan-Tiao CUSUM Test for Variance Breaks

The Inclan-Tiao (IT) test [55] is a nonparametric cumulative sum (CUSUM) procedure designed to detect sudden changes in the unconditional variance of a time series. Unlike the Zivot-Andrews test, which targets breaks in the level or trend under a unit root framework, the IT test is specifically constructed for the second moment, making it particularly suited to financial return series whose mean is approximately zero but whose variance exhibits pronounced regime-dependent behaviour.

Let $\{a_t\}_{t=1}^{T}$ denote a sequence of mean-zero returns and define the cumulative sum of squared observations:

$$C_k = \sum_{t=1}^{k} a_t^2, \qquad k = 1, \ldots, T$$

with C_T denoting the total sum of squares. The centred and normalised CUSUM process is defined as:

$$D_k = \frac{C_k}{C_T} - \frac{k}{T}, \qquad k = 1, \ldots, T$$

Under the null hypothesis of constant unconditional variance, D_k fluctuates randomly around zero. A structural break in variance causes D_k to deviate systematically – rising when the pre-break variance exceeds the global average, and falling thereafter, or vice versa. The break date is estimated as the point of maximum absolute deviation:

$$\hat{T}_b = \arg \max_{1 \leq k \leq T} |D_k|$$

The IT test statistic is:

$$\text{IT} = \sqrt{\frac{T}{2}} \sup_{1 \leq k \leq T} |D_k| \tag{E.4}$$

Under the null hypothesis of homoskedastic innovations $a_t \overset{i.i.d.}{\sim} (0, \sigma^2)$, the asymptotic distribution of IT converges to the supremum of the absolute value of a standard Brownian bridge:

$$\text{IT} \xrightarrow{d} \sup_{0 \leq s \leq 1} |W^0(s)| \tag{E.5}$$

where $W^0(s) = W(s) - sW(1)$ denotes a standard Brownian bridge and $W(\cdot)$ a standard Wiener process. Critical values at the 1%, 5%, and 10% significance levels are approximately 1.628, 1.358, and 1.224 respectively [55]. The null hypothesis of constant variance is rejected when IT exceeds the relevant critical value.

F

Detailed results and additional analysis of Value at Risk Prediction

F.1 Results and analysis of traditional models

The 95% violation rates range from 1.81% to 3.77% for the BRVM market indices, well below the theoretical threshold of 5%. This result is even more pronounced with models with a student specification. However, for the CAC40 and S&P500 indices, the normal distribution models show rates slightly above the 5% threshold (between 5.26% and 5.75%), reflecting reasonable adequacy at the 95% level. In contrast, Student's distributions produce rates between 3.19% and 3.58%, confirming their tendency.

For the BRVM Composite and BRVM Finance indices, all eight models generate violation rates below or close to the theoretical threshold of 1%. The Student's distribution models produce remarkably low rates (all below 0.31% for the BRVM Composite and 0.33% for the BRVM Finance), indicating an overestimation of risk. Normal distribution models show slightly higher rates (0.96% to 1.06% for the BRVM Composite and 0.78% to 0.88% for the BRVM Finance) but remain within an acceptable range. The results differ radically in developed markets. Normal distribution models produce violation rates of over 2%, more than double the theoretical threshold of 1%, indicating a serious underestimation of risk. In contrast, Student's distribution models maintain rates between 0.77% and 1.04%, remaining in line with the target for both series.

Kupiec test results At the 95% threshold, on the BRVM markets, the normal models have very low p-values, almost zero, indicating a rejection of adequacy. The violation rates are too far from the target threshold of 5%. However, on the CAC40 and S&P500, models with a normal specification pass the test with p-values greater than 0.1, while all Student models fail

Table F.1: Empirical violation rates and Kupiec test p-values at the 5% and 1% confidence levels

	BRVM Composite		BRVM Finance		CAC 40		S&P 500	
Model	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$
(a) *VaR at the 5% confidence level*								
GARCH	**3.77**† (0.01)	1.86 (0.00)	**3.17** (0.00)	2.65 (0.00)	**5.61**† (0.21)	3.58 (0.00)	**5.26**† (0.60)	3.37 (0.00)
APARCH	**3.72**† (0.01)	1.81 (0.00)	**3.22** (0.00)	2.70 (0.00)	**5.46**† (0.34)	3.58 (0.00)	**5.55**† (0.26)	3.37 (0.00)
EGARCH	**3.72**† (0.01)	1.81 (0.00)	**3.37** (0.00)	2.70 (0.00)	**5.41**† (0.39)	3.19 (0.00)	**5.60**† (0.22)	3.47 (0.00)
GJR-GARCH	**3.57** (0.00)	1.81 (0.00)	**3.01** (0.00)	2.60 (0.00)	**5.75**† (0.13)	3.29 (0.00)	**5.65**† (0.19)	3.42 (0.00)
(b) *VaR at the 1% confidence level*								
GARCH	**1.01**† (0.98)	0.20 (0.00)	**0.78**† (0.31)	0.31 (0.00)	2.13 (0.00)	**0.77**† (0.28)	2.53 (0.00)	**1.04**† (0.85)
APARCH	**0.96**† (0.84)	0.30 (0.00)	**0.78**† (0.31)	0.31 (0.00)	2.22 (0.00)	**0.87**† (0.54)	2.58 (0.00)	**1.04**† (0.85)
EGARCH	**1.01**† (0.98)	0.25 (0.00)	**0.83**† (0.44)	0.26 (0.00)	2.17 (0.00)	**0.77**† (0.28)	2.68 (0.00)	**0.84**† (0.47)
GJR-GARCH	**1.06**† (0.80)	0.20 (0.00)	**0.88**† (0.60)	0.31 (0.00)	2.22 (0.00)	**1.01**† (0.95)	2.53 (0.00)	**1.04**† (0.85)

Notes: Each cell reports the empirical violation rate (in %) with the Kupiec likelihood ratio test p-value in parentheses below. $\mathcal{N}$ and $\mathcal{T}$ denote the Normal (Gaussian) and Student-t_ν innovation distributions, respectively. The theoretical violation rates are 5% and 1% for panels (a) and (b). † denotes failure to reject the Kupiec null hypothesis of correct unconditional coverage at the 5% significance level. For each model and index, the better-performing distributional assumption – defined as the higher Kupiec p-value – is displayed in bold.

with p-values that are zero.

The normally distributed models are validated by the Kupiec test for BRVM market indices. They have p-values between 0.80 and 0.98 for BRVM Composite and between 0.31 and 0.60 for BRVM Finance. Thus, the models comply with the violation proportion, which must be 1%. On the other hand, Student's distribution models are unanimously rejected with a p-value of zero, confirming that their violation rates are statistically too low compared to the theoretical threshold of 1%. Conversely, the conclusion is reversed in the two developed markets. Normal models are rejected with a p-value of zero due to their excessive violation rates, while Student's models are validated with a p-value of 0.28 to 0.95 for the CAC 40 and 0.47 to 0.85 for the S&P500.

Christoffersen independence test results At the 95% level, BRVM markets maintain partially satisfactory results for normal models (p = 0.15–0.49), while Student models fail on BRVM Composite (p = 0.00). This suggests that 95% violations under Student distribution exhibit serial dependence in these markets. On CAC40 and S&P500, all models reject the hypothesis of independence, regardless of distribution.

High p-values greater than 0.05 for a large number of models for BRVM Composite and BRVM Finance indicate that violations are temporally independent for all models, regardless

Table F.2: Christoffersen independence test p-values at the 5% and 1% confidence levels

	BRVM Composite		BRVM Finance		CAC 40		S&P 500	
Model	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$	$\mathcal{N}$	$\mathcal{T}$
(b) *VaR at the 5% confidence level*								
GARCH	**0.49**†	0.00	**0.45**†	0.10†	0.00	0.00	0.03	0.03
APARCH	**0.46**†	0.00	0.04	**0.09**†	0.00	0.00	0.03	0.03
EGARCH	**0.20**†	0.17†	0.03	**0.09**†	0.00	0.00	0.00	**0.04**
GJR-GARCH	**0.15**†	0.00	**0.53**†	0.10†	0.00	0.00	0.01	**0.04**
(a) *VaR at the 1% confidence level*								
GARCH	0.20†	**0.90**†	0.63†	**0.85**†	0.00	**0.01**	**0.01**	0.00
APARCH	0.17†	**0.85**†	0.63†	**0.85**†	0.00	**0.01**	0.00	0.00
EGARCH	0.20†	**0.87**†	0.60†	**0.87**†	0.00	**0.11**†	**0.02**	0.00
GJR-GARCH	0.22†	**0.90**†	0.58†	**0.85**†	0.02	**0.21**†	**0.01**	0.00

Notes: Reported values are p-values of the Christoffersen [26] likelihood ratio test for independence of VaR violations. Under the null hypothesis, violations are serially independent. † denotes failure to reject the null at the 5% significance level, indicating no evidence of violation clustering. For each model and index, the better-performing distributional assumption – defined as the higher p-value – is displayed in bold. $\mathcal{N}$ and $\mathcal{T}$ denote the Normal (Gaussian) and Student-t_ν innovation distributions, respectively.

of distribution, except for the GARCH Student, APARCH Student and CJR-GARCH Student models for BRVM Composite and APARCH Normal and EGARCH Normal models for BRVM Finance. For the CAC40 and S&P500, the results are clearly unfavourable, with p-values close to zero for almost all models. This clustering of violations reveals that the models fail to capture the temporal dynamics of extreme risks in developed markets.

Diebold-Mariano test Diebold-Mariano tests conducted at both confidence levels (95% and 99%) show that the distribution of innovations is the determining factor in the quality of VaR forecasts, far more so than the choice of GARCH specification. Models with a normal distribution consistently outperform their Student's t-distribution counterparts, with highly significant DM statistics at the 1% level across all markets and specifications; this result is explained by the tendency of the Student's t-distribution to overestimate risk by producing overly conservative VaR estimates. In contrast, the differences between specifications (GARCH, APARCH, GJR-GARCH, or eGARCH) remain virtually nil and insignificant regardless of the distribution chosen, indicating that asymmetric refinements do not yield any measurable forecasting gains. These differences diminish slightly at the 99% confidence level compared to the 95% level, without, however, altering the hierarchy of the models. Finally, when comparing markets, significant DM statistics reveal that the BRVM Composite is the index most sensitive to the choice of modeling at the 95% confidence level, displaying the highest values among the four markets, while at the 99% level, the S&P 500 becomes the most discriminating index. The CAC40 and BRVM Finance, on the other hand, show more moderate significant statistics at both confidence levels.

Bibliography

[1] Ian S. Abramson. On bandwidth variation in kernel estimates–a square root law. *The Annals of Statistics*, 10(4):1217–1223, 1982.

[2] C. Argyropoulos and E. Panopoulou. Backtesting var and es under the magnifying glass. *International Review of Financial Analysis*, 64:22–37, 05 2019.

[3] K. Avdulaj and J. Baruník. Are benefits from oil stocks diversification gone? new evidence from a\n dynamic copula and high frequency data. *arXiv (Cornell University)*, 07 2013.

[4] Jushan Bai and Pierre Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

[5] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[6] Basel Committee on Banking Supervision. Amendment to the capital accord to incorporate market risks. Report, Bank for International Settlements, Basel, Switzerland, 1996.

[7] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: A revised framework comprehensive version. Report, Bank for International Settlements, Basel, Switzerland, 2006.

[8] R. S. Bautista and J. A. N. Mora. Value-at-risk predictive performance: a comparison between the caviar and garch models for the mila and asean-5 stock markets. *Journal of Economics Finance and Administrative Science*, 26:197–221, 11 2021.

[9] Henry C. P. Berbee. *Random walks with stationary increments and renewal theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum Amsterdam, 1979.

[10] S. N. Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927.

[11] Stefan Best. Minimum capital requirements for market risk: An overview and critical analysis of the standardized approaches under basel iii. Working paper, RheinMain University of Applied Sciences, Wiesbaden Institute of Finance and Insurance, 2021.

[12] Sergey G. Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[13] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

[14] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[15] Jacob Boudoukh, Matthew Richardson, and Robert Whitelaw. The best of both worlds: a hybrid approach to calculating value at risk. *Risk*, 11(5):64–67, 1998.

[16] G. EP Box, G. M. Jenkins, and G. C Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, 3rd edition, 1994. Note: Often cited as Box and Jenkins (1976) for the earlier edition, but 1994 is the standard modern reference with Reinsel.

[17] R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[18] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, 1986.

[19] V. Bucevska. An empirical evaluation of garch models in value-at-risk estimation: Evidence from the macedonian stock exchange. *Business Systems Research Journal*, pages 1–16, 01 2012.

[20] S. Campbell. A review of backtesting and backtesting procedures, 01 2007.

[21] C. L. Chang and M. McAleer. The correct regularity condition and interpretation of asymmetry in egarch. *Economics Letters*, 161:52–55, 09 2017.

[22] C. L. Chang, M. McAleer, and R. Tansuchat. Modelling long memory volatility in agricultural commodity futures returns. *Annals of Financial Economics*, 7:1250010–1250010, 12 2012.

[23] V. Chang, T. Li, and Z. Zeng. Towards an improved adaboost algorithmic method for computational financial analysis. *Journal of Parallel and Distributed Computing*, 137:196–206, 2020.

[24] P. Christoffersen. Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, 2:84–108, 04 2004.

[25] P. Christoffersen and D. Pelletier. Backtesting value-at-risk: A duration-based approach. *SSRN Electronic Journal*, 01 2003.

[26] Peter F. Christoffersen. Evaluating interval forecasts. *International Economic Review*, 39(4):841–862, 1998.

[27] M. P. Clements, A. B. Galvão, and J. H. Kim. Quantile forecasts of daily exchange rate returns from forecasts of realized volatility. *Journal of Empirical Finance*, 15:729–750, 12 2007.

[28] J. Cotter. Uncovering long memory in high frequency uk futures. *SSRN Electronic Journal*, 01 2005.

[29] R. Dahlhaus. Locally stationary processes. *Handbook of statistics*, 30:351–413, 2012.

[30] Rainer Dahlhaus. Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37, 1997.

[31] S. Degiannakis and E. Xekalaki. Autoregressive conditional heteroscedasticity (arch) models: A review, 01 2004.

[32] David A. Dickey and Wayne A. Fuller. Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, 49(4):1057–1072, 1981.

[33] Francis X Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–18, 2015.

[34] Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

[35] Zhuanxin Ding, Clive W. J. Granger, and Robert F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106, 1993.

[36] C. Dritsaki. An empirical evaluation in garch volatility modeling: Evidence from the stockholm stock exchange. *Journal of Mathematical Finance*, 7:366–390, 01 2017.

[37] Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 107–115. Morgan Kaufmann, 1997.

[38] N. G. Emenogu, M. O. Adenomon, and N. O. Nweze. On the volatility of daily stock returns of total nigeria plc: evidence from garch models, value-at-risk and backtesting. *Financial Innovation*, 6, 03 2020.

[39] R. F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

[40] Jianqing Fan and Elias Masry. Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. *Journal of Multivariate Analysis*, 43(2):237–271, 1992.

[41] M. Fałdziński. Peaks over threshold approach with a time-varying scale parameter and range-based volatility estimator for value-at-risk and expected shortfall estimation. *Przegląd Statystyczny Statistical Review*, 2024:23–77, 01 2025.

[42] F. Fei, A. M. Fuertes, and E. Kalotychou. Dependence in credit default swap and equity markets: Dynamic copula with markov-switching. *International Journal of Forecasting*, 33:662–678, 04 2017.

[43] Y. Freund. An adaptive version of the boost by majority algorithm. *Machine Learning*, 43:293–318, 2001.

[44] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[45] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[46] Wei Gao, Xin-Yi Niu, and Zhi-Hua Zhou. Learnability of non-i.i.d. In *Asian Conference on Machine Learning*, pages 158–173. PMLR, 2016.

[47] Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801, 1993.

[48] A. J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[49] Peter Hall, Rodney C. L. Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999.

[50] J. D Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[51] Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

[52] A. Hitaj, C. Mateus, and I. Peri. Lambda value at risk and regulatory capital: A dynamic approach to tail risk. *SSRN Electronic Journal*, 01 2016.

[53] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[54] John Hull and Alan White. Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of Risk*, 1(1):5–19, 1998.

[55] Carla Inclán and George C. Tiao. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[56] Philippe Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk.* McGraw-Hill, New York, 1997.

[57] Kengo Kato. Weighted nadaraya-watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, 10(2):265–291, 2012.

[58] M. Kearns and L. G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 433–444. ACM, 1989.

[59] P. Keith and K. N. Langeland. Forecasting exchange rate volatility: Garch models versus implied volatility forecasts. *International Economics and Economic Policy*, 12:127–142, 05 2014.

[60] E. Khemakhem. Three essays on regulation and taxation of stocks and derivatives. *HAL (Le Centre pour la Communication Scientifique Directe)*, 12 2020.

[61] R. Kovačević. Modelling the exchange rate of the euro against the dollar using the arch/garch models. *Bankarstvo*, 45:20–49, 01 2016.

[62] P. N. Kumar, N. Umeorah, and A. Alochukwu. Dynamic graph neural networks for enhanced volatility prediction in financial markets. *arXiv (Cornell University)*, 10 2024.

[63] Paul H Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995.

[64] V. Kuznetsov and M. Mohri. Time series prediction and online learning. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 49:1–24, 2016.

[65] V. Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems*, pages 541–549, 2015.

[66] Michel Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, Rhode Island, 2001.

[67] Y. Li. *Empirical Analysis of Constructing GARCH Model to Predict Stock Prices with Trading Volume*, pages 589–602. Atlantis Press, 01 2023.

[68] A. Livada and M. C. Anagnostopoulou. Risk measures and inequality. *Frontiers in Applied Mathematics and Statistics*, 5, 12 2019.

[69] P. M. Long and R. A. Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78:287–304, 2010.

[70] Jose A. Lopez. Methods for evaluating value-at-risk estimates. *Economic Review-Federal Reserve Bank of San Francisco*, (2):3–17, 1999.

[71] Jose A. Lopez. Regulatory evaluation of value-at-risk models. *The Journal of Risk*, 1(2):37–64, 1999.

[72] A. Lozano, S. Kulkarni, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *Advances in Neural Information Processing Systems*, 2006.

[73] Robin L. Lumsdaine and David H. Papell. Multiple trend breaks and the unit-root hypothesis. *The Review of Economics and Statistics*, 79(2):212–218, 1997.

[74] M. Mohri and A. Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

[75] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, pages 1097–1104, 2009.

[76] Abdelkader Mokkadem. Mixing properties of arma processes. *Stochastic Processes and their Applications*, 29(2):309–315, 1988.

[77] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[78] T. Ndlovu and D. Chikobvu. Comparing riskiness of exchange rate volatility using the value at risk and expected shortfall methods. *Investment Management and Financial Innovations*, 19:360–371, 07 2022.

[79] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370, 1991.

[80] O. Nieppola. Backtesting value-at-risk models. *Aaltodoc (Aalto University)*, 01 2009.

[81] C. O. Omari, Peter N. Mwita, and A. G. Waititu. Using conditional extreme value theory to estimate value-at-risk for daily currency exchange rates. *Journal of Mathematical Finance*, 7:846–870, 01 2017.

[82] L. Padmakumari and M. Shaik. An empirical investigation of value at risk (var) forecasting based on range-based conditional volatility models. *Engineering Economics*, 34:275–292, 06 2023.

[83] L. L. Pauwels, P. Radchenko, and A. L. Vasnev. High moment constraints for predictive density combination. *RePEc: Research Papers in Economics*, 06 2023.

[84] Pierre Perron. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57(6):1361–1401, 1989.

[85] N. S. Pranav. A review of backtesting var models, 04 2020.

[86] G. Rätsch, T. Onoda, and K. R. Müller. Soft margins for adaboost. *Machine Learning*, 42:287–320, 2001.

[87] G. Rodriguez. Selecting between autoregressive conditional heteroskedasticity models: An empirical application to the volatility of stock returns in peru. *Revista de analisis economico*, 32:69–94, 04 2017.

[88] P. Sadorsky. Stochastic volatility forecasting and risk management. *Applied Financial Economics*, 15:121–135, 01 2005.

[89] M. G. Sampid, H. M. Hasim, and H. Dai. Refining value-at-risk estimates using a bayesian markov-switching gjr-garch copula-evt model. *PLoS ONE*, 13, 06 2018.

[90] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[91] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 1998.

[92] T. C. Schleifer and M. El Asmar. *Measuring Financial Risk*, pages 92–100. Informa, 10 2021.

[93] G. William Schwert. Why does stock market volatility change over time? *The Journal of Finance*, 44(5):1115–1153, 1989.

[94] R. A. Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4:633–648, 2003.

[95] Kevin Sheppard et al. arch: Python 3 data analysis tools, 2023.

[96] Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1986.

[97] Stephen J. Taylor. *Modeling Financial Time Series*. John Wiley & Sons, Chichester, 1986.

[98] A. Vasiljeva, A. Matvejevs, and J. Fjodorovs. Dependence modeling and portfolio optimization with copula-garch: a european investment perspective. *Frontiers in Applied Mathematics and Statistics*, 11, 10 2025.

[99] Mathukumalli Vidyasagar. *A Theory of Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer, London, 1997.

[100] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media, Berlin, Heidelberg, 2009.

[101] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

[102] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

[103] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 2005.

[104] Eric Zivot and Donald W. K. Andrews. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics*, 10(3):251–270, 1992.

[105] B. Égert and Y. Koubaa. Modelling stock returns in the g-7 and in selected cee economies: A non-linear garch approach. *Deep Blue (University of Michigan)*, 02 2004.

Contents