RISK ANALYSIS OF RECURRENT NEURAL NETWORKS FOR FINANCIAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting financial time series is a fundamental task for modern quantitative finance. Recurrent neural networks (RNNs) have been widely applied in this domain, yet their statistical behavior under the structural features of financial data remains poorly understood. We study the excess risk of RNNs for price movement prediction with k-line chart data. To obtain a financially interpretable risk characterization, we introduce a tractable data-generating model and decompose excess risk into approximation and estimation errors. The approximation analysis separates finite temporal representation, nonlinear function approximation, and normalization-induced information loss. Under binary cross-entropy loss, we prove that the finite-width term decays as $O((m-2p-2)^{-2})$. We further show that normalized observations inherit the β-mixing property of the underlying dynamics, yielding an estimation bound based on Rademacher complexity for dependent data. Combining the two error bounds, we reveal how financially interpretable factor into risk: the signal-to-noise ratio (SNR) proxy B_s^2/σ^2 governs approximation difficulty, while the persistence parameter ρ controls effective statistical information. Notably, a larger SNR proxy can increase approximation error and hence excess risk. We conduct controlled experiments on simulated data. The results exhibit trends consistent with the theoretical analysis, and empirically reveal the approximation–estimation tradeoff across model capacity and sample size.

1 INTRODUCTION

Forecasting financial time series is a central pillar of modern quantitative finance (Abu-Mostafa & Atiya, 1996). Recurrent neural networks (RNNs) process sequential data by recursively updating a hidden state (Elman, 1990; Qin et al., 2017), a mechanism that drives their widespread success on forecasting various financial variables, including asset returns (Fischer & Krauss, 2018), volatility (Bucci, 2020), and macroeconomic indicators (Barkan et al., 2023). However, theoretical understanding lags behind practical applications. While recent studies provide fundamental bounds for RNNs (Li et al., 2020; Cheng et al., 2025), they rely on abstract settings that fail to capture the distinct structural and dynamic features of financial time series. This fundamental gap motivates our work: a rigorous risk analysis designed to explicitly link the statistical behavior of the RNNs hypothesis class to the key structural features of financial time series.

Specifically, we study the statistical risk of RNNs for k-line chart data, which is the most widely used financial time series data that summarizing the price variation with open-high-low-close (OHLC) price sequences. Consider the fundamental problem of predicting the direction of the next-period price movement from historical OHLC observations (Hu et al., 2018; Gunduz, 2021). The directional target naturally casts the forecasting problem as a binary classification task, allowing prediction performance to be studied through classification risk. At the same time, the OHLC setting remains tractable enough for explicit analysis while retaining statistical features specific to financial time series data.

Several structural properties of financial time series make this problem challenging to analyze. First, it is unclear which characteristics of the underlying financial process determine the statistical difficulty of RNNs prediction and how they factor into the excess risk. Second, financial observations can vary substantially in scale, making normalization important both in practice and for theoretical analysis. Yet normalization may discard information relevant for prediction, raising the question of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

how its statistical cost should be quantified. Third, predictive information may be distributed over a non-trivial historical horizon, while RNNs have only finite representational capacity. A longer history can therefore improve the available information, but also requires more capacity to retain it and may increase the statistical cost of learning. These considerations create a coupled tradeoff among the financial data distribution, available history, and model capacity.

Our goal is therefore to characterize the excess risk of RNN predictors and understand how it depends on the model characteristics and properties of the financial time series. To make this connection explicit, we introduce a stylized data-generating model for OHLC sequences, in which the open and close dynamics follow a finite-order vector autoregression (VAR), while the intraperiod high and low are generated through a Brownian-bridge conditional on the endpoints. We study empirical risk minimization (ERM) over RNNs trained on window-wise normalized histories. The normalization places inputs on a bounded scale, which facilitates the theoretical analysis and remaining consistent with our experiment setup. However, it may also remove scale information for prediction. Accordingly, we measure excess risk relative to the Bayes predictor with access to the history with full information.

Our analysis proceeds through an approximation–estimation decomposition of the excess risk. For a surrogate loss satisfying suitable regularity conditions, the approximation bound separates the effect of finite temporal representation, nonlinear function approximation, and normalization-induced information loss. We further show that normalized OHLC sequence inherits the β-mixing property of the VAR process, allowing an estimation bound based on Rademacher complexity for dependent observations. Combining the two analyses yields an explicit surrogate risk bound; for binary cross-entropy (BCE) loss, a calibration inequality further translates the surrogate excess risk bound to the 0–1 classification risk. The resulting bound identifies several financially interpretable quantities, including a proxy of signal-to-noise ratio (SNR) B_s^2/σ^2 and persistence parameter ρ. A notable finding is that increasing the SNR proxy can increase the approximation error and, in turn, the excess risk. Controlled experiments with BCE-trained RNNs reproduce the predicted qualitative trends across network width, history length, predictability, and sample size.

Our contributions are fourfold:

- We design a financially interpretable data-generating process for OHLC sequences and formulate a distribution-dependent risk framework for RNN-based directional prediction.
- We derive explicit approximation and estimation bounds for RNN predictors. The approximation error contains an $O((m - 2p - 2)^{-2})$ finite-width term and an information-loss term, while the estimation bound accounts for temporal dependence through β-mixing and Rademacher complexity.
- We identify financially interpretable quantities that govern statistical difficulty. In particular, the SNR proxy B_s^2/σ^2 controls the approximation difficulty, while the persistence parameter ρ determines the effective statistical information.
- Controlled experiments on trained RNNs reproduce the main qualitative predictions of the theory, including the effects of width, history length, and predictability on approximation risk, and the approximation–estimation tradeoff across network capacity and sample size.

2 Related Work

2.1 Approximation Theory for RNNs

Early work established universal approximation properties of RNNs for dynamical systems and state-space models (Funahashi & Nakamura, 1993; Schäfer & Zimmermann, 2007). More recent work has moved beyond universality to study quantitative approximation bounds and their dependence on network structure and temporal memory. Li et al. (2021) derive approximation rates for continuous-time linear RNNs and identify the "curse of memory", where slower memory decay requires substantially more width for approximation. Bishop & Bonilla (2023) derive approximation bounds for recurrent representations of Bayesian filters with time-uniform guarantees, while Wang et al. (2024b) extend the memory-dependent approximation limits to nonlinear RNNs through an inverse approximation result.

2.2 Generalization and Risk Bounds for RNNs

A line of work studies generalization of RNNs through norm-based control. Tu et al. (2020) derive generalization bounds using matrix and Fisher–Rao norms, while Chen et al. (2020) establish non-asymptotic generalization bounds for several recurrent architectures based on Rademacher complexity and spectral norms. Cheng et al. (2025) further extend this line of analysis to estimation-error bounds using local Rademacher complexity under a Bernstein condition. Mitarchuk et al. (2024) present length-independent PAC-Bayes bounds under spectral-norm conditions on the recurrent weight matrix and hidden states saturation. For dependent sequential data, Eringis et al. (2024) derive PAC-Bayes generalization bounds for stable dynamical systems under non-i.i.d. time series observations.

2.3 Financial Time Series and OHLC Sequences

Financial time series often exhibit temporal dependence across multiple horizons. Autoregressive and vector autoregressive models provide a parsimonious framework for representing such lagged dependence and have been widely used in return modeling and forecasting (DeMiguel et al., 2014; Gupta et al., 2020). Beyond close-to-close returns, high and low prices summarize interperiod price variation and contain information about latent volatility. Classical range-based estimators exploit this information using high–low range or the full OHLC observation (Parkinson, 1980; Yang & Zhang, 2000). More generally, high-low observations can be viewed as interval-valued time series, for which machine-learning methods have been developed beyond financial applications (Wang et al., 2024a), including meteorology (Maia et al., 2008) and energy (Xiong et al., 2017). This perspective connects the OHLC setting to a broader class of interval-valued sequential learning problems.

3 Preliminary

Problem Settings. We consider a sequence of OHLC observations $X_t^\tau = (X_{t-\tau+1}, \cdots, X_t)$, where $X_t = (O_t, H_t, L_t, C_t)$ represents the log-transformation open, high, low, close values over interval t. The classification label is defined as $Y_t = \mathbf{1}\{C_{t+1} - C_t \geq 0\}$. Denote $\eta_\tau(x) := \mathbb{P}(Y_t = 1 | X_t^\tau = x)$ as the conditional class probability given history of length τ and $\eta_\infty := \mathbb{P}(Y_t = 1 | X_t^\infty)$ given full history.

Normalized Input. In accordance with the preprocessing used in training, we apply per-window min-max normalization. Denote $\overline{H} = \max_{t-\tau+1 \leq j \leq t} H_j$, $\underline{L} = \min_{t-\tau+1 \leq j \leq t} L_j$ and the range $d_t = \overline{H} - \underline{L}$. Then we normalizing X_t^τ elementwise $\widetilde{X}_t^\tau = \frac{X_t^\tau - \underline{L}}{d_t^\tau}$.

Data-Generating Process. We design a time series model to simulate the OHLC sequence by two component. The first is a recursive vector autoregression (VAR) model to simulate open and close,

$$\boldsymbol{B}\boldsymbol{r}_t = \boldsymbol{\mu} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{r}_{t-i} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\boldsymbol{r}_t = [O_t - C_{t-1} \quad C_t - O_t]^\top$, $\boldsymbol{\mu}$, $\boldsymbol{B}$ and $\boldsymbol{\Phi}_i$ are the parameters describing the temporal dependency, $\boldsymbol{\varepsilon}_t$ is a two-dimensional Gaussian variable describing the innovation. Second component is a Brownian Bridge to simulate the high and low given interval's open and close.

RNN Class. The RNN function class is defined as $\mathcal{F}_{RNN} = \{f_\theta : W \in \mathbb{R}^{m \times m}, U \in \mathbb{R}^{m \times 4}, v \in \mathbb{R}^m, b_h \in \mathbb{R}^m, b_o \in \mathbb{R}\}$, where

$$f_\theta(\widetilde{X}_t^\tau) = v^\top h_\tau + b_o, \qquad h_j = \sigma\left(W h_{j-1} + U \widetilde{x}_{t-\tau+j} + b_h\right), \ \ j = 1, \ldots, \tau, \qquad h_0 = 0, \tag{2}$$

and $\sigma(\cdot) = \max\{\cdot, 0\}$ is the ReLU function. The induced classifier is $g(\widetilde{X}_t^\tau) = \mathbf{1}\left\{f_\theta(\widetilde{X}_t^\tau) \geq 0\right\}$. The total number of parameters $P = m^2 + 6m + 1$

4 RISK ANALYSIS OF RNN ON VAR-BB

In this section, we study how the statistical performance of RNN predictors is jointly governed by the model capacity, sample size, available history, and temporal dependence of the underlying process. We first formulate the learning problem and decompose the excess risk into approximation and estimation components. We then analyze the two terms separately under the VAR-BB model.

4.1 EXCESS RISK DECOMPOSITION AND LEARNING SETTINGS

Consider the binary classification problem. For any RNN function $f \in \mathcal{F}$, given a sample set $S = \{(\widetilde{X}_i^\tau, Y_i)\}_{i=1}^n$ of size n generating from our DGP, the (expected) risk and empirical risk is defined as $L_\tau(f) := \mathbb{E}_{(\widetilde{X}^\tau, Y)\sim\mathcal{D}}\left[\ell(f(\widetilde{X}^\tau), Y)\right]$ and $\widehat{L}_\tau(f) = \frac{1}{n}\sum_{i=1}^n \ell\left(f(\widetilde{X}_i^\tau), Y_i\right)$ respectively. Specifically, denote $L_\infty^* = \min_{f \text{ measurable}} L_\tau(f)$ the Bayes risk given full information of history X_t^∞.

Since the underlying DGP is ambiguous in practice, we learn over the RNN class by empirical risk minimization (ERM). Our goal is to study the excess risk of the ERM learner $\hat{f}_S \in \arg\min_{f\in\mathcal{F}} \widehat{L}_\tau(f)$. What we ultimately care about in binary classification is the 0-1 risk or the misclassification error $L_\tau^{01}(f) = \mathbb{P}(g(X^\tau) \neq Y)$. However, since 0-1 loss is nonconvex and discontinues, we train RNN using a surrogate loss ϕ. As standard in the risk analysis, we decompose the excess risk into approximation error and estimation error

$$L_\mathcal{D}(f_S) - L_\mathcal{D}^* = \left(L_\mathcal{D}(f_S) - \min_{f\in\mathcal{F}} L_\mathcal{D}(f)\right) + \left(\min_{f\in\mathcal{F}} L_\mathcal{D}(f) - L_\mathcal{D}^*\right) := \epsilon_{\text{est}} + \epsilon_{\text{app}}. \tag{3}$$

To characterize the two terms, we impose the following assumptions.

Assumption 4.1. *The norms of weight matrices are bounded,* $\|W\|_2 \le B_W$, $\|U\|_2 \le B_U$, $\|v\|_2 \le B_v$, $\|b_h\|_2 \le B_h$, $|b_o| \le B_o$.

Assumption 4.2. *The surrogate loss function* $\phi : \mathbb{R}\times\{0,1\} \to \mathbb{R}_+$ *is locally Lipschitz. For* $B > 0$, *define the Lipschitz coefficient as* $\rho_\phi(B) = \max_{y\in\{0,1\}} \sup_{|x|,|x'|\le B} \frac{|\phi(x,y)-\phi(x',y)|}{|x-x'|} < \infty$.

4.2 APPROXIMATION ERROR ϵ_{app}

We study the approximation error under a single regime. Under our DGP, the conditional distribution of Y_t depends on the past through a finite-dimensional state

$$s_t = \mu_0 + \sum_{i=0}^{p-1} a_i O_{t-i} + \sum_{i=0}^{p} b_i C_{t-i}. \tag{4}$$

By Gaussianity, the conditional class probability $\eta(s) = \Phi(\frac{s}{\sigma})$. For a surrogate loss ϕ, define the conditional risk $C(a,s) = \eta(s)\phi(a,1)+(1-\eta(s))\phi(a,0)$, where a denote the prediction score, e.g., the output of RNN. To derive the approximation rate explicitly, we impose the additional assumption.

Assumption 4.3. *There exists* $\psi : \mathbb{R} \to \mathbb{R}$ *such that* $\psi(s) \in \arg\min_{a\in\mathbb{R}} C(a,s)$ *and* $\psi \in C_{loc}^{1,1}(\mathbb{R})$.

The RNN, however, operates on the normalized history $\widetilde{X}_t^\tau$, introducing both representation and capacity limitations. We use an effective history length $\tau' \le \tau$ to characterize the temporal information represented by the recurrent state, while the variability of the window range d_t quantifies the information loss from normalization.

Thm 4.4 (OHLC Approximation)**.** *Suppose* $\tau > p$. *Let* $d_0 = \mathbb{E}[d_t^\tau]$, $\Delta_k = \sum_{i=k}^{p-1}|a_i| + \sum_{i=k}^{p}|b_i|$ *and* $\Gamma_\tau = |1 - \frac{d_0}{d_t}||s_t - \mu_0|$, *where* $\Delta_k = 0$ *if* $k > p$. *Then*

1. *Under 0-1 risk, for* $\tau' = \min\{\tau, \lfloor\frac{m}{2}\rfloor\}$, *the approximation error satisfies*

$$\epsilon_{app}^{01} \le \sqrt{\frac{2}{\pi\sigma^2}}(d_0\Delta_{\tau'} + \mathbb{E}\Gamma_\tau). \tag{5}$$

Especially, $\epsilon_{\text{app}} = 0$ *when* $\mu_0 = 0$ *and* $\tau' > p$.

2. *For any surrogate loss ϕ satisfying assumption 4.2 and 4.3, for $1 \le \tau' \le \min\{\tau, \lfloor \frac{m-2}{2} \rfloor\}$ and $M_s > 0$,*

$$\epsilon_{app}^{\phi} \le \inf_{\tau', M_s} \left\{ \frac{C_1}{(m-2\tau')^2} + C_2(d_0 \Delta_{\tau'} + \mathbb{E}\Gamma_\tau) + C_3 \mathbb{P}[A_{M_s}^c] \right\} \tag{6}$$

where $A_{M_s} = \{|s_t| \vee |\tilde{s}_t^{\tau'}| \le M_s\}$, and C_1, C_2, C_3 are the constants depends on M_s.

The proof can be found at Appendix B.1. The two bounds reflect different mechanisms. Under the 0-1 loss, the approximation error comes from finite-width temporal representation ($d_0 \Delta_{\tau'}$), and the information loss from normalization ($\mathbb{E}\Gamma_\tau$). In contrast under surrogate risk, it additionally requires approximating the nonlinear function ψ, which yields a capacity allocation between temporal history representation and function approximation. The tail term accounts for the boundary region where the local regularity of ψ does not provide a uniform bound.

Then we specify Thm 4.4 under the binary cross-entropy (BCE) loss $\phi(x, y) = \log(1 + e^x) - xy$, for which $\psi(s) = \log \frac{\Phi(s/\sigma)}{1-\Phi(s/\sigma)}$.

Cor 4.5 (Approximation under BCE Loss)**.** *Denote $B_s = |\mu_0| + d_0 \left(\sum_{i=0}^{p-1} |a_i| + \sum_{i=0}^{p} |b_i| \right)$. Under the BCE loss, if $m \ge 2p + 4$, then*

$$\epsilon_{app}^{BCE} \le \frac{c B_s^2}{\sigma^2 (m - 2p - 2)^2} + \frac{\mathbb{E}\Gamma_\tau^2}{2\sigma^2}, \tag{7}$$

where c is a universal constant.

The approximation construction is also compatible with the bounded-norm RNN class used in the subsequent estimation analysis, as formalized by the following proposition.

Prop 4.6. *For a fixed single-regime DGP, there exists an RNN attaining the approximation bound above whose parameter norms are finite and independent of the network width m.*

4.3 Estimation Error ϵ_{EST}

The OHLC observations $\{Z_t := (\widetilde{X}_t^\tau, Y_t)\}_{t=1}^n$ are temporally dependent. To control the estimation error, we first introduce two standard tools for learning with dependent observations.

Def 4.7 (Empirical Rademacher Complexity)**.** *Given a sample set $S = (z_1, \ldots, z_n)$, the empirical Rademacher complexity of a set of real-valued functions $\mathcal{F}$ with respect to S is defined as*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \frac{1}{n} \mathbb{E}_\sigma [\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(z_i)], \tag{8}$$

where $\{\sigma_i\}_{i=1}^n$ are i.i.d. Rademacher random variables such that $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$.

Def 4.8 (β-Mixing)**.** *Let $\{Z_t\}_{t=-\infty}^{\infty}$ be a stationary sequence. For any $i, j \in \mathbb{Z} \cup \{-\infty, +\infty\}$, let $\mathcal{G}_i^j$ denote the σ-algebra generated by the random variables Z_k, $i \le k \le j$. Then for any integer $k \ge 1$, the β-mixing coefficient of the stochastic process $\{Z_t\}_{t=-\infty}^{\infty}$ is defined as*

$$\beta(k) = \sup_n \mathbb{E} \left[\sup_{A \in \mathcal{G}_{n+k}^{\infty}} \left| \mathbb{P}[A | \mathcal{G}_{-\infty}^n] - \mathbb{P}[A] \right| \right]. \tag{9}$$

$\{Z_t\}_{t=-\infty}^{\infty}$ is said to be β-mixing if $\beta(k) \to 0$.

Under stability and regularity condition, the Gaussian VAR process is stationary geometrically β-mixing (Cordoni & Sancetta, 2024; Meitz & Saikkonen, 2021). The following lemma shows that these properties can be inherited by normalized OHLC sequence.

Lemma 4.9. *Suppose the VAR process $\{\boldsymbol{r}_t\}$ is stationary and β-mixing, and the Brownian-bridge innovations are i.i.d. across t and independent of $\{\boldsymbol{r}_t\}$. Then for any fixed τ and $k > \tau$, the observations of normalized OHLC sequence are stationary and β-mixing such that $\beta_Z(k) \le \beta_r(k - \tau)$. In particular, if the VAR process satisfies $\beta_r(k) \le C_\beta \rho^k$, $\rho \in (0, 1)$, then $\beta_Z(k) \le C_\beta \rho^{k-\tau}$.*

For the dependent data with stationary and β-mixing, the standard approach for generalization bound is the independent block technique (Mohri & Rostamizadeh, 2008). Here we adopt a slightly different version for normalized OHLC and estimation bound, the proof can be found at appendix B.2

Lemma 4.10. *Let $\phi \circ \mathcal{F}$ be the loss function class such that $0 \leq \phi(f(\widetilde{x}), y) \leq M$. For any $S = (z_1, \ldots, z_n)$ drawn from a stationary β-mixing process, and for any $a, \tilde{n}$ with $n = 2a\tilde{n}$, suppose $\hat{f}_S$ is the ERM minimizer, then for any $5(\tilde{n}-1)\beta(a) < \delta < 1$, with probability at least $1-\delta$,*

$$L(\hat{f}_S) - \inf_{f \in \mathcal{F}} L(f) \leq 2\widehat{\mathfrak{R}}_{S_{\tilde{n}}}(\phi \circ \mathcal{F}) + 4M\sqrt{\frac{\log(5/\delta')}{2\tilde{n}}}, \tag{10}$$

where $\delta' = \delta - 5(\tilde{n}-1)\beta(a)$ and $S_{\tilde{n}} = (Z_1, Z_{2a+1}, \ldots, Z_{2(\tilde{n}-1)a+1})$.

Lemma 4.11. *Under the assumption 4.1, let $S = (z_1, \ldots, z_n)$ be a sample of normalized OHLC sequences, and define $b_\tau = \sum_{i=0}^{\tau-1} B_W^i$, $a_\tau = \sum_{i=1}^{\tau-1} iB_W^i$. The empirical Rademacher complexity is bounded by*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq cB_v(2B_U + B_h)\sqrt{b_\tau \max\{b_\tau, a_\tau\}}\sqrt{\frac{m^2+6m}{n}} + B_o\sqrt{\frac{1}{n}} \tag{11}$$

where c is a universal constant.

By lemma B.1, $|f_\theta(\tilde{x})| \leq B_f = B_v(2B_U + B_h)b_\tau + B_o$ for $f_\theta \in \mathcal{F}$. Under assumption 4.2, the loss is uniformly bounded by $B_\phi = \max_{y \in \{0,1\}} \phi(0, y) + \rho_\phi(B_f)B_f$.

Thm 4.12. *Let $S = (z_1, \ldots, z_n)$ the sample set generated by the stationary VAR-BB process. Under assumptions 4.1 and 4.2, for any $a, \tilde{n}$ with $n = 2a\tilde{n}$, suppose $\hat{f}_S$ is the ERM minimizer of ϕ, then for any $5(\tilde{n}-1)\beta_Z(a) < \delta < 1$, with probability at least $1-\delta$,*

$$\epsilon_{est}^{\phi} \leq c\rho_\phi B_v(2B_U + B_h)\sqrt{b_\tau \max\{b_\tau, a_\tau\}}\sqrt{\frac{m^2+6m}{\tilde{n}}} + c\rho_\phi \frac{B_o}{\sqrt{\tilde{n}}} + 4B_\phi\sqrt{\frac{\log(5/\delta')}{2\tilde{n}}}, \tag{12}$$

where $\delta' = \delta - 5(\tilde{n}-1)\beta_Z(a)$. Additionally, if the VAR process is geometric β-mixing with $\beta(k) \leq c_\beta\rho^k$, then for $a > \tau$, $\delta' \geq \delta - 5(\tilde{n}-1)c_\beta\rho^{a-\tau}$.

Under geometric β-mixing, choose the block length such that $5c_\beta(\frac{n}{2a} - 1)\rho^{a-\tau} \leq \frac{\delta}{2}$. It is sufficient to set $a = \left\lceil \tau + \frac{\log(10c_\beta n/\delta)}{|\log\rho|} \right\rceil$. Consequently, up to norm-dependent factors,

$$\epsilon_{\text{est}}^{\phi} = O\left(\rho_\phi \mathcal{C}(m, \tau)\sqrt{\frac{\tau + \log(n/\delta)/\log|p|}{n}}\right), \tag{13}$$

where $\mathcal{C}(m, \tau) = \sqrt{(m^2+6m)b\tau\max\{b_\tau, a_\tau\}}$. $\mathcal{C}(m, \tau)$ reflects the familiar role of the existing recurrent spectral norm in controlling sensitivity to the long history. When $B_W < 1$, both b_τ and a_τ are bounded, and $\mathcal{C}(m, \tau) = O(m)$; at $B_W = 1, \mathcal{C} = O(m\tau^{3/2})$ grows polynomially; and when $B_W > 1$, the recurrent sensitivity can grow exponentially with $O(m\sqrt{\tau}B_W^\tau)$. Furthermore, in our setting, a longer history could incur an additional cost because adjacent normalized OHLC windows overlap. Through $\beta_Z(k) \leq c_\beta\rho^{k-\tau}$, increase τ also reduce the effective sample size.

4.4 Excess Risk and Statistical Tradeoffs

We now combine the above results to study the excess risk. First, we show the general excess risk bound for the surrogate loss. Denote $\epsilon_{\text{app}}^{\phi}(m, \tau, \mathcal{D})$ and $\epsilon_{\text{est}}^{\phi}(m, \tau, n, \mathcal{D})$ the upper bounds in (6) and (12) respectively, where $\mathcal{D}$ represents the relevant quantities for the financial DGP.

Thm 4.13. *Under the assumptions of Theorems 4.4 and 4.12. Let $\hat{f}_S$ be the ERM of loss ϕ. Then, with probability at least $1-\delta$,*

$$\mathcal{R} = L_\tau^\phi(\hat{f}_S) - L_\infty^{\phi,*} \leq \epsilon_{app}^{\phi}(m, \tau, \mathcal{D}) + \epsilon_{est}^{\phi}(m, \tau, n, \mathcal{D}). \tag{14}$$

We next specialize the excess surrogate risk bound and the relative 0-1 risk bound to the BCE loss.

Cor 4.14 (Excess Risk under BCE Loss). *Under the assumptions of Theorems 4.4 and 4.12, consider the binary cross-entropy loss. Suppose $m \geq 2p+4$, let $n = 2a\tilde{n}$ and $\delta' = \delta - 5(\tilde{n}-1)\beta_Z(a) > 0$. Then with probability at least $1-\delta$,*

$$\begin{aligned}\mathcal{R} = L_\tau^{\mathrm{BCE}}(\hat{f}_S) - L_\infty^{\mathrm{BCE},*} \leq \frac{c_1 B_s^2}{\sigma^2(m-2p-2)^2} + \frac{\mathbb{E}\Gamma_\tau^2}{2\sigma^2} \\ + c_2 B_v(2B_U + B_h)\sqrt{b_\tau \max\{b_\tau, a_\tau\}}\sqrt{\frac{m^2+6m}{\tilde{n}}} \\ + c_2\frac{B_o}{\sqrt{\tilde{n}}} + 4\log(1+e^{B_f})\sqrt{\frac{\log(5/\delta')}{2\tilde{n}}},\end{aligned} \tag{15}$$

where c_1, c_2 are universal constants, $B_s = |\mu_0| + d_0\left(\sum_{i=0}^{p-1}|a_i| + \sum_{i=0}^{p}|b_i|\right)$ and $B_f = B_v(2B_U + B_h)b_\tau + B_o$. Moreover, for the induced classifier $\hat{g}_S(x) = \mathbf{1}\{\hat{f}_S(x) \geq 0\}$, the calibration inequality for BCE loss yields $L_\tau^{01} - L_\infty^{01,*} \leq \sqrt{2\left(L_\tau^{\mathrm{BCE}}(\hat{f}_S) - L_\infty^{\mathrm{BCE},*}\right)}$

The excess risk bound characterizes both the effects of network capacity, sample size, available history, and the distribution-dependent quantities of the financial DGPs. We highlight three consequences below.

Network Capacity. First we consider the how network capacity affect the excess risk. Fixing the other variable, the width-dependent part of the bound is $\mathcal{R}(m) \lesssim \frac{\tilde{c}_1}{(m-2p-2)^2} + \tilde{c}_2\sqrt{\frac{m^2+6m}{\tilde{n}}}$, where we absorb all quantities independent on m into $\tilde{c}_1$ and $\tilde{c}_2$. To minimize the excess risk, the stationary point of m satisfies $\frac{2\tilde{c}_1}{(m-2p-2)^3} = \frac{\tilde{c}_2}{\sqrt{\tilde{n}}}\frac{m+3}{\sqrt{m^2+6m}}$. Thus, the balancing width scale given by the upper bound is

$$m^* = O\left(2p + \left(\frac{B_s^2/\sigma^2}{B_v(2B_U+B_h)\sqrt{b_\tau\max\{b_\tau,a_\tau\}}}\right)^{1/3}\tilde{n}^{1/6}\right) \tag{16}$$

Available History. Once $\tau > p$, all predictive information are available, so increasing τ does not further reduce truncation error. Its remains affecting approximation through the normalization quantities d_0 and Γ_τ.

Distribution-Dependent Finance Quantities. The bound can be interpreted by several quantities from a financial perspective. First, we consider the signal-to-noise ratio (SNR), which measures the power of the target signal to the power of the unpredictable noise. Here we define a proxy of SNR as

$$\mathrm{SNR} = \frac{B_s^2}{\sigma^2}, \quad B_s = |\mu_0| + d_0\left(\sum_{i=0}^{p-1}|a_i| + \sum_{i=0}^{p}|b_i|\right). \tag{17}$$

Intuitively, a larger SNR corresponds to the greater intrinsic predictability, but does not necessarily imply a smaller excess risk. In our bound, SNR enters the finite-width approximation term directly: a larger signal scale requires RNN to approximate a more extreme condition risk minimizer ψ. In addition, window-wise normalization removes the window range d_t, resulting more information loss as SNR increases. The second quantity is persistence, represented by the mixing coefficient ρ. A larger ρ leads to more persistent dynamics and requires a larger blocking gap to obtain approximately independent observations. As a result, two financial datasets with the same calender length can have different statistical information when their return dynamics exhibit different persistence.

5 Experiments

We conduct controlled experiments under the proposed OHLC DGP to empirically examine the qualitative implications of our risk analysis. Our goal is to test whether the mechanisms and tradeoffs identified by the theory are reflected in trained RNNs. Concretely, we ask the following questions.

Q1: Does the approximation analysis capture how network width, available history, and predictability affect the excess risk of trained RNN?

Q2: How do network width and sample size jointly shape excess risk, and is the approximation–estimation tradeoff predicted by the theory visible empirically?

We generate OHLC sequences from the single-regime VAR-BB model with $p = 6$. We train the RNN class defined in Eq.(2) using BCE loss and Adam with learning rate 10^{-3}. Models are trained for 100 epochs with early stopping, and results are averaged over five independently generated datasets. Excess risk is evaluated with the full-history Bayes predictor on an independent test set. All experiments are implemented in PyTorch and run on a NVIDIA GeForce RTX 5070 Ti GPU.

Approximation Trends (Q1).

We first examine Q1. Figure 1 shows that the trained RNN broadly follows the trends predicted by the approximation bound for network capacity, available history, and predictability of the sequence measured by SNR. One notable feature is the sharp drop of the theoretical bound at $m = 16$. Under Theorem 4.4, this is the first width that can represent the full effective history, which leads to a clear reduction in the approximation error bound. We also observe that excess risk increases with the SNR proxy B_s^2/σ^2, showing that stronger financial predictability does not necessarily make approximation easier under normalized inputs. The same trend result holds for the variance-based SNR $\mathrm{Var}(s_t)/\sigma^2$, see in Appendix C. Overall, although the numerical results remain conservative, the approximation analysis captures the main trends observed in the trained RNNs.

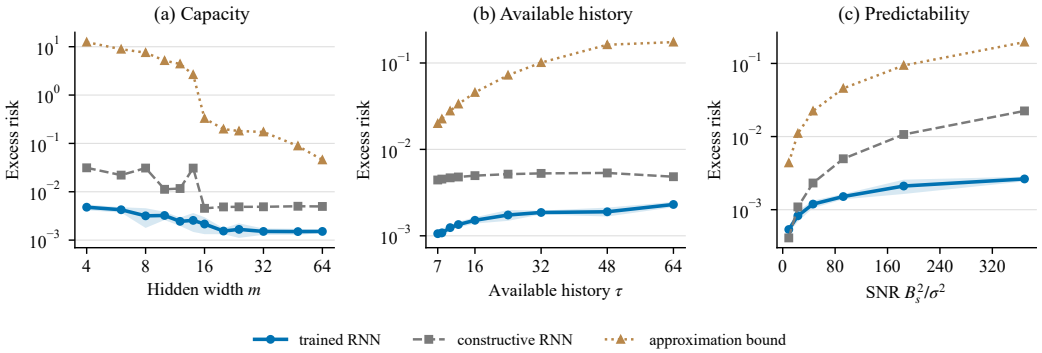


Figure 1: Approximation trends under the VAR-BB model. We use $n = 2^{18}$ training sample and vary (a) the hidden width m with $\tau = 16$, (b) the available history τ with $m = 64$, (c) the predictability B_s^2/σ^2. Curves show the trained RNN, the constructive RNN from the approximation analysis, and the approximation bound, taken as the smaller of the bounds in Theorem 4.4 and Corollary 4.5; for $m < 2p + 4$, we use Theorem 4.4. Further details are given in Appendix C

.

Approximation–Estimation Tradeoff (Q2).

We next examine Q2 by jointly varying the hidden width m and sample size n. Directly evaluating the theoretical estimation bound is not informative at the scale considered here: the spectral norms of trained recurrent matrices W are frequently larger than one, making the resulting norm-based upper bounds highly conservative. We therefore examine the tradeoff empirically. Figure 2 shows the excess risk across network width at three representative sample size. For small n, increase width provides little improvement, whereas for large n, excess risk decreases more clearly with model capacity. This suggests that additional capacity becomes beneficial only when supported by sufficient data. This pattern is qualitatively consistent with the balancing width m^*, which suggests that larger samples support wider networks.

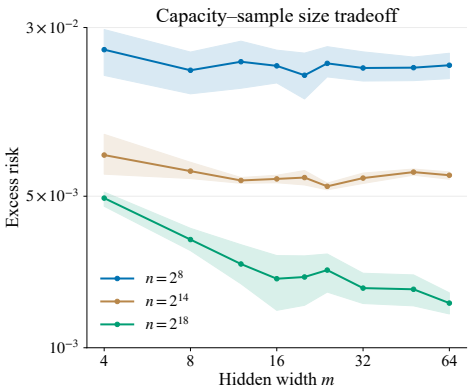


Figure 2: Capacity–Sample Size Tradeoff for selected sample sizes.

To examine the source of this tradeoff, for each width we use the true conditional probabilities provided by the DGP as training targets to construct a large-sample proxy for the population BCE minimizer within the corresponding RNN class. Figure 3 reports empirical error balance over the full (m, n) grid. At small sample size, the estimation error dominates. As n increases, the balance shifts toward approximation, particularly for narrow networks, and larger widths become more effective in reducing excess risk. This indicates that additional capacity is beneficial only when supported by sufficient data, consistent with the tradeoff shown in the theory.

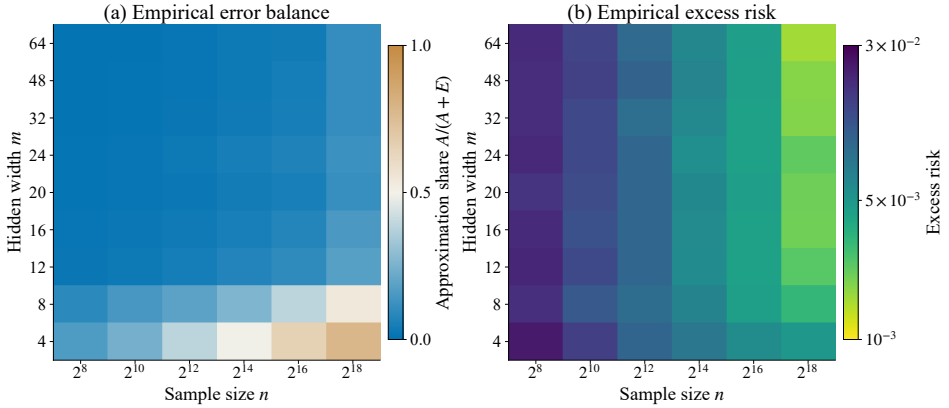


Figure 3: Empirical approximation–estimation tradeoff. We vary the hidden width m and sample size n jointly. Panel (a) indicates whether the excess risk is dominated by approximation or estimation error; panel (b) shows the total excess risk. Here A and E are empirical proxies of approximation and estimation error.

6 Conclusion

Our results suggest that the statistical behavior of RNNs for financial time series is governed by a set of coupled tradeoffs rather than by model capacity or data availability alone. Increasing network width reduces the approximation error, but its benefit for excess risk emerges only when supported by sufficient data, while extending the input history beyond the predictive order need not improve excess risk. Likewise, stronger predictability does not necessarily make the learning problem easier: under normalized inputs, a larger SNR proxy can increase the finite-capacity approximation error. Temporal persistence introduces a different difficulty by reducing the effective amount of statistical information available from a fixed-length sample. Together, these results indicate that RNN performance should be understood jointly in terms of model capacity, preprocessing, historical horizon, and characteristics of the underlying financial process.

The current scope also suggests several natural directions for extending the framework. Our analysis intentionally focuses on a simple, stylized financial DGP in order to obtain explicit and interpretable risk characterization. An important direction for future work is to incorporate richer financial stylized facts and quantify the additional statistical difficulty induced by different market phenomena. On the estimation side, our current analysis is based on global Rademacher complexity control. Sharper bounds may be possible through localized complexity measures, although such analyses typically require additional structural or regularity conditions. More broadly, these extensions would help quantify how richer financial stylized facts translate into statistical difficulty for recurrent predictors.

AI Use Statement

We used generative AI tools to assist with language polishing and writing, and to support literature retrieval and discovery during the preparation of this manuscript. The authors independently reviewed and verified all AI-assisted text and retrieved references, and take full responsibility for the accuracy and integrity of the final manuscript.

REFERENCES

Yaser S Abu-Mostafa and Amir F Atiya. Introduction to financial forecasting. *Applied intelligence*, 6(3):205–213, 1996.

Oren Barkan, Jonathan Benchimol, Itamar Caspi, Eliya Cohen, Allon Hammer, and Noam Koenigstein. Forecasting cpi inflation components with hierarchical recurrent neural networks. *International Journal of Forecasting*, 39(3):1145–1162, 2023.

Adrian N Bishop and Edwin V Bonilla. Recurrent neural networks and universal approximation of bayesian filters. In *International Conference on Artificial Intelligence and Statistics*, pp. 6956–6967. PMLR, 2023.

Andrea Bucci. Realized volatility forecasting with neural networks. *Journal of financial econometrics*, 18(3):502–531, 2020.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1233–1243, 2020.

Xuewei Cheng, Ke Huang, and Shujie Ma. Generalization and risk bounds for recurrent neural networks. *Neurocomputing*, 616:128825, 2025.

Francesco Cordoni and Alessio Sancetta. Consistent causal inference for high-dimensional time series. *Journal of Econometrics*, 246(1):105902, 2024. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2024.105902. URL https://www.sciencedirect.com/science/article/pii/S0304407624002537.

Carl De Boor and Carl De Boor. *A practical guide to splines*, volume 27. springer New York, 1978.

Victor DeMiguel, Francisco J Nogales, and Raman Uppal. Stock return serial dependence and out-of-sample portfolio performance. *The Review of Financial Studies*, 27(4):1031–1073, 2014.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Deividas Eringis, John Leth, Zheng-Hua Tan, Rafael Wisniewski, and Mihály Petreczky. Pac-bayes generalisation bounds for dynamical systems including stable rnns. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11901–11909, 2024.

Thomas Fischer and Christopher Krauss. Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, 270(2):654–669, 2018.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural networks*, 6(6):801–806, 1993.

Hakan Gunduz. An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination. *Financial innovation*, 7(1): 28, 2021.

Rangan Gupta, Florian Huber, and Philipp Piribauer. Predicting international equity returns: Evidence from time-varying parameter vector autoregressive models. *International Review of Financial Analysis*, 68:101456, 2020.

Hongping Hu, Li Tang, Shuhua Zhang, and Haiyan Wang. Predicting the direction of stock markets using optimized neural networks with google trends. *Neurocomputing*, 285:188–195, 2018.

Zhong Li, Jiequn Han, Qianxiao Li, et al. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. *arXiv preprint arXiv:2009.07799*, 2020.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*, 2021.

André Luis S Maia, Francisco de AT de Carvalho, and Teresa B Ludermir. Forecasting models for interval-valued time series. *Neurocomputing*, 71(16-18):3344–3352, 2008.

Mika Meitz and Pentti Saikkonen. Subgeometric ergodicity and β-mixing. *Journal of Applied Probability*, 58(3):594–608, 2021.

Volodimir Mitarchuk, Clara Lacroce, Rémi Eyraud, Rémi Emonet, Amaury Habrard, and Guillaume Rabusseau. Length independent pac-bayes bounds for simple rnns. In *International Conference on Artificial Intelligence and Statistics*, pp. 3547–3555. PMLR, 2024.

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Michael Parkinson. The extreme value method for estimating the variance of the rate of return. *Journal of business*, pp. 61–65, 1980.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*, 2017.

Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

Zhuozhuo Tu, Fengxiang He, and Dacheng Tao. Understanding generalization in recurrent neural networks. In *International Conference on Learning Representations*, 2020.

Piao Wang, Shahid Hussain Gurmani, Zhifu Tao, Jinpei Liu, and Huayou Chen. Interval time series forecasting: A systematic literature review. *Journal of forecasting*, 43(2):249–285, 2024a.

Shida Wang, Zhong Li, and Qianxiao Li. Inverse approximation theory for nonlinear recurrent neural networks. In *International Conference on Learning Representations*, volume 2024, pp. 27794–27826, 2024b.

Tao Xiong, Chongguang Li, and Yukun Bao. Interval-valued time series forecasting using a novel hybrid holti and msvr model. *Economic Modelling*, 60:11–23, 2017.

Dennis Yang and Qiang Zhang. Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, 73(3):477–492, 2000.

A OHLC SIMULATION

Algorithm 1 OHLC Simulation

Require: $(\mu_N, \mu_D) \in \mathbb{R}^2$, $(\sigma_N^t, \sigma_D^t) > 0$, $C_0 > 0$, $T > 0$
Ensure: Realization path $(O_t, H_t, L_t, C_t)_{t=0}^T$

```
t ← 0
while t < T do
    μ̃ ← μ_D^t / σ_D^t
    draw N(0,1)-distributed prn z_1
    O_t ← C_{t-1} + (σ_N · z_1 + μ_N)
    (z, a, b) ← B(μ̃)                    ▷ B : ℝ → ℝ^3: Becker's sampling algorithm
    C_t, L_t, H_t ← O_t + zσ_D^t, O_t + aσ_D^t, O_t + bσ_D^t
    t ← t + 1
end while
```

B PROOF OF SECTION 3

B.1 PROOF OF THM 4.4

Proof. Denote $\boldsymbol{e} = [1 \quad 1]^\top$ and $r_{t+1} = C_{t+1} - C_t$, the VAR model can be rewritten as

$$\begin{aligned} C_{t+1} - C_t &= \boldsymbol{e}^T \boldsymbol{B^{-1}} \boldsymbol{\mu} + \sum_{i=1}^{p} \boldsymbol{e}^T \boldsymbol{B^{-1}} \boldsymbol{\Phi_i} \boldsymbol{r_t} + \boldsymbol{e}^\top \boldsymbol{B^{-1}} \boldsymbol{\varepsilon_t} \\ &= \mu_0 + \sum_{i=0}^{p-1} a_i O_{t-i} + \sum_{i=0}^{p} b_i C_{t-i} + \sigma \varepsilon_t \\ &\overset{\text{set}}{=} s_t + \sigma \varepsilon_t, \ \varepsilon_t \sim N(0,1). \end{aligned} \tag{18}$$

Then the conditional class probability given the full history is $\eta_\infty = \mathbb{P}(Y_t = 1 | X_t^\infty) = \Phi(\frac{s_t}{\sigma})$.

Given the normalized OHLC time series $\widetilde{X}_t^\tau$ with $\tau \geq p+1$, since $\sum_{i=0}^{p-1} a_i + \sum_{i=0}^{p} b_i = 0$, define the quantity $q_t = \sum_{i=0}^{p-1} a_i \widetilde{O}_{t-i} + \sum_{i=0}^{p} b_i \widetilde{C}_{t-i}$, then $s_t = \mu_0 + d_t \left(\sum_{i=0}^{p-1} a_i \widetilde{O}_{t-i} + \sum_{i=0}^{p} b_i \widetilde{C}_{t-i} \right)$, where d_t is the window range. Since $\widetilde{O}_t, \widetilde{C}_t \in [0,1]$, $|q_t| \leq B_q := \sum_{i=0}^{p-1} |a_i| + \sum_{i=0}^{p} |b_i|$. Note that the learner does not observe d_t. We therefore use a proxy constant d_0 to replace d_t and approximate $\tilde{s}_t = \mu_0 + d_0 q_t$. Since the limitation of finite width and history, suppose RNN can utilize history of length τ', then define

$$q_t^{\tau'} = \sum_{i=0}^{\min\{p-1, \tau'-1\}} a_i \widetilde{O}_{t-i} + \sum_{i=0}^{\min\{p, \tau'-1\}} b_i \widetilde{C}_{t-i}, \qquad \tilde{s}_t^{\tau'} = \mu_0 + d_0 q_t^{\tau'}. \tag{19}$$

For the first bound, denote $\tau' = \min\left\{\tau, \lfloor \frac{m}{2} \rfloor\right\}$. We first show that RNN can realize $\tilde{s}_t^{\tau'}$ by arranging the parameters of weight matrices. Set

$$U_{m\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \mathbf{0}_{(m-2)\times 4} \end{bmatrix}, \qquad W_{m\times m} = \begin{bmatrix} & & \mathbf{0}_{2\times m} & & \\ \boldsymbol{I}_2 & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{I}_2 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \boldsymbol{I}_2 & 0 \end{bmatrix}, \qquad b_h = 0 \tag{20}$$

Under this construction, $h_\tau = (\widetilde{O}_t, \widetilde{C}_t, \cdots, \widetilde{O}_{t-\tau'+1}, \widetilde{C}_{t-\tau'+1})$. Setting $v = d_0 \cdot [a_0, b_0, \cdots, a_{\tau'+1}, b_{\tau'+1}, 0, \cdots, 0]^\top$ and $b_o = \mu_0$, then $f_\theta(\widetilde{X}_t^\tau) = v^\top h_\tau = \mu_0 + d_0 q_t^{\tau'} = \tilde{s}_t^{\tau'}$.

Then we analyze the risk gap between the RNN classifier $g(\widetilde{X}_t^\tau) = \mathbf{1}\{\tilde{s}_t^{\tau'} \geq 0\}$ and the Bayes classifier $g^*(X_t^\infty) = \mathbf{1}\{s_t \geq 0\}$. Note that $|s_t| \cdot \mathbf{1}\{g \neq g^*\} \leq |s_t - \tilde{s}_t^{\tau'}|$, then

$$\begin{aligned}
\epsilon_{\text{app}} \leq L_\tau^{01}(g) - L_\infty^{01,*} &= \mathbb{E}\left[|2\eta_\infty - 1| \cdot \mathbf{1}\{g \neq g^*\}\right] \\
&\leq \mathbb{E}\left[2\left|\Phi\left(\frac{s_t}{\sigma}\right) - \Phi(0)\right| \cdot \mathbf{1}\{g \neq g^*\}\right] \\
&\leq \mathbb{E}\left[\sqrt{\frac{2}{\pi\sigma^2}}\,|s_t| \cdot \mathbf{1}\{g \neq g^*\}\right] \\
&\leq \sqrt{\frac{2}{\pi\sigma^2}}\,\mathbb{E}|s_t - \tilde{s}_t^{\tau'}| \\
&\leq \sqrt{\frac{2}{\pi\sigma^2}}\mathbb{E}|d_t q_t - d_0 q_t^{\tau'}| \\
&\leq \sqrt{\frac{2}{\pi\sigma^2}}\left(d_0\mathbb{E}|q_t - q_t^{\tau'}| + \mathbb{E}[|d_t - d_0||q_t|]\right)
\end{aligned} \tag{21}$$

Choose $d_0 = \mathbb{E}[d_t]$ and let $\Delta_{\tau'} = \sum_{i=\tau'}^{p-1}|a_i| + \sum_{i=\tau'}^{p}|b_i|$, $\Gamma_\tau = \mathbb{E}|d_t - d_0||q_t| = \mathbb{E}|1 - \frac{d_0}{d_t}||s_t - \mu_0|$, then the approximation bound of 0-1 risk is

$$\epsilon_{\text{app}}^{01} \leq \sqrt{\frac{2}{\pi\sigma^2}}(d_0\Delta_{\tau'} + \Gamma_\tau) \tag{22}$$

For the second bound, for any effective history length τ' satisfying $1 \leq \tau' \leq \min\{\tau, \lfloor\frac{m-2}{2}\rfloor\}$, we use the first $2\tau'$ hidden units to represent $\tilde{s}_{\tau'}$, while reserving the remaining $m - 2\tau'$ units to approximate ψ_ϕ. We divide the sample space according to the event $A_s := \{|s_t| \vee |\tilde{s}_t^{\tau'}| \leq M_s\}$ and its complement A_δ^c. Denote $f_\theta^{\tau'}$ the constructed RNN representing τ' history, $B_\psi(M_s) = \sup_{|s|\leq M}|\psi(s)|$, $B_{\psi'}(M_s) = \sup_{|s|\leq M_s}|\psi'(s)|$, and $\rho_{\psi'}(M_s) = \sup_{|s|,|s'|\leq M_s}\frac{|\psi'(s)-\psi'(s')|}{|s-s'|}$. Then ϵ_{app} is controlled by three terms,

$$\begin{aligned}
\epsilon_{\text{app}} &\leq L_\tau(f_\theta^{\tau'}) - L_\infty^* \\
&= \mathbb{E}\left[\left(C(f_\theta^{\tau'}, s_t) - H_\phi(s_t)\right)\mathbf{1}_{A_s}\right] + \mathbb{E}\left[\left(C(f_\theta^{\tau'}, s_t) - H(s_t)\right)\mathbf{1}_{A_s^c}\right] \\
&= \mathbb{E}\left[\left(C(f_\theta^{\tau'}, s_t) - C(\psi(\tilde{s}_t^{\tau'}), s_t)\right)\mathbf{1}_{A_s}\right] + \mathbb{E}\left[\left(C(\psi(\tilde{s}_t^{\tau'}), s_t) - H(s_t)\right)\mathbf{1}_{A_s}\right] \\
&\quad + \mathbb{E}\left[\left(C(f_\theta^{\tau'}, s_t) - H(s_t)\right)\mathbf{1}_{A_s^c}\right] \\
&\overset{\text{set}}{=} T_1 + T_2 + T_3.
\end{aligned} \tag{23}$$

For T_1, we first approximate ψ on the bounded interval $[-M_s, M_s]$. Let $k = m - 2\tau'$, and partition $[-M_s, M_s]$ into $k-1$ equal subintervals, denoted as $-M_s = \xi_0 < \xi_1 < \cdots < \xi_{k-1} = M_s$. Consider the linear spline function to approximate $\psi(s)$,

$$S_k(s) = \beta_0 + \sum_{i=0}^{k-1}\gamma_i(s - \xi_i)^+ \tag{24}$$

where $\beta_i = \psi(\xi_i)$ for $i = 0, \cdots, k-1$, $\gamma_0 = \frac{\beta_1 - \beta_0}{\xi_1 - \xi_0}$, $\gamma_i = \frac{\beta_{i+1} - \beta_i}{\xi_{i+1} - \xi_i} - \frac{\beta_i - \beta_{i-1}}{\xi_i - \xi_{i-1}}$ for $i = 1, \cdots, k-2$ and $\gamma_{k-1} = -\frac{\beta_{k-1} - \beta_{k-2}}{\xi_{k-1} - \xi_{k-2}}$. To realize S_k, we construct the hidden and input weight matrices as follows. Set

$$U_{m\times 4} = \begin{bmatrix} U_{2\tau'\times 4}^{(1)} \\ U_{k\times 4}^{(2)} \end{bmatrix}, \qquad W_{m\times m} = \begin{bmatrix} W_{2\tau'\times m}^{(1)} \\ W_{k\times m}^{(2)} \end{bmatrix}, \qquad b_h = \begin{bmatrix} b_h^{(1)} \\ b_h^{(2)} \end{bmatrix}, \tag{25}$$

such that we allocate the width to memorize the effective history and represent the spline function by two blocks

$$U^{(1)}_{2\tau'\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \mathbf{0}_{(2\tau'-2)\times 4} \end{bmatrix}, \qquad W^{(1)}_{2\tau'\times m} = \begin{bmatrix} & & \mathbf{0}_{2\times 4\tau'} & & \\ \boldsymbol{I}_2 & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{I}_2 & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \boldsymbol{I}_2 & 0 \end{bmatrix}, \qquad b^{(1)}_h = \mathbf{0}_{2\tau'} \tag{26}$$

$$U^{(2)}_{k\times 4} = cr_1^\top, \qquad W^{(2)}_{k\times m} = cr_2^\top, \qquad b^{(2)}_h = c \odot (\mu_0 \mathbf{1}_k - \xi), \tag{27}$$

where

$$c = \begin{bmatrix} \sqrt{|\gamma_0|} & \cdots \sqrt{|\gamma_{k-1}|} \end{bmatrix}^\top, \qquad r_1 = d_0 \begin{bmatrix} a_0^{(\tau')} & 0 & 0 & b_0^{(\tau')} \end{bmatrix}^\top \tag{28}$$

$$r_2 = d_0 \begin{bmatrix} a_1^{(\tau')} & b_1^{(\tau')} & \cdots a_{\tau'-1}^{(\tau')} & b_{\tau'-1}^{(\tau')} & \mathbf{0}_{2+k} \end{bmatrix}^\top. \tag{29}$$

We introduce the vector c to control the norm of the weight matrix satisfying assumption 4.1. Then set

$$v = \begin{bmatrix} \mathbf{0}_{2\tau'}^\top & \operatorname{sign}(\gamma) \odot c^\top \end{bmatrix}^\top, \qquad b_0 = \beta_0. \tag{30}$$

Under this setting, the constructed RNN satisfies that

$$f_\theta^{\tau'}(\widetilde{X}_t^\tau) = \beta_0 + (\operatorname{sign}(\gamma) \odot c)^\top \left(c \odot (\tilde{s}_t^{\tau'} \mathbf{1}_k - \xi)^+ \right) \tag{31}$$

$$= \beta_0 + \sum_{i=0}^{k-1} \gamma_i \left(\tilde{s}_t^{\tau'} - \xi_i \right)^+ = S_k \left(\tilde{s}_t^{\tau'} \right) \tag{32}$$

In the bounded region $[-M_s, M_s]$, according to (De Boor & De Boor, 1978; DeVore et al., 2021),the linear spline function represented by ReLU network with $k-1$ subintervals can approximate the smooth function at the rate that

$$\begin{aligned} \|S_k - \psi\|_\infty &\le \frac{\rho_{\psi'} M_s^2}{2(k-1)^2} \\ &\le \frac{2\rho_{\psi'} M_s^2}{(m-2\tau')^2}. \end{aligned} \tag{33}$$

Let $C_1 = 2\rho_\phi(B_\psi(M_s))\rho_{\psi'}(M_s)M_s^2$. Then

$$\begin{aligned} T_1 &= \mathbb{E}\left[\left(C_\phi(f_\theta^{\tau'}, s_t) - C_\phi(\psi(\tilde{s}_t^{\tau'}), s_t) \right) \mathbf{1}_{A_{M_s}} \right] \\ &\le \rho_\phi \mathbb{E}\left[|f_\theta^{\tau'} - \psi(\tilde{s}_t^{\tau'})| \mathbf{1}_{A_{M_s}} \right] \le \frac{C_1}{(m-2\tau')^2} \end{aligned} \tag{34}$$

For T_2, let $C_2 = \rho_\phi(B_\psi(M_s))B_{\psi'}(M_s)$. According to assumption 4.2 and 4.3,

$$\begin{aligned} T_2 &= \mathbb{E}\left[\left(C(\psi(\tilde{s}_t^{\tau'}), s_t) - H(s_t) \right) \mathbf{1}_{A_{M_s}} \right] \\ &= \mathbb{E}\Big[\big(\eta(s_t) \left[\phi(\psi(\tilde{s}_t^{\tau'}), 1) - \phi(\psi(s_t), 1) \right] \\ &\quad + (1 - \eta(s_t) \left[\phi(\psi(\tilde{s}_t^{\tau'}), 0) - \phi(\psi(s_t), 0) \right] \big) \mathbf{1}_{A_{M_s}} \Big] \\ &\le \rho_\phi \mathbb{E}[|\psi(\tilde{s}_t^{\tau'}) - \psi(s_t)| \mathbf{1}_{A_{M_s}}] \\ &\le \rho_\phi B_{\psi'} \mathbb{E}[|\tilde{s}_t^{\tau'} - s_t| \mathbf{1}_{A_{M_s}}] \\ &\le C_2(d_0 \Delta_{\tau'} + \Gamma_\tau). \end{aligned} \tag{35}$$

For T_3, by construction, RNN $f_\theta^{\tau'}$ has a natural upper bound $|f_\theta^{\tau'}| \le B_\psi(M_s) < \infty$. According to assumption 4.2, for $|a| \le B_\psi$, $\phi(a,y) \le \max_y \phi(0,y) + \rho_\phi B_\psi$. Denote $C_3 = \max_y \phi(0,y) + \rho_\phi B_\psi$, then

$$\begin{aligned} T_3 &= \mathbb{E}\left[\left(C(f_\theta^{\tau'}, s_t) - H(s_t)\right)\mathbf{1}_{A^c_{M_s}}\right] \\ &\le \mathbb{E}\left[\left(\eta(s_t)\phi(f_\theta^{\tau'}, 1) + (1-\eta(s_t))\phi(f_\theta^{\tau'}, 0)\right)A^c_{M_s}\right] \\ &\le C_3 \mathbb{P}[A^c_{M_s}] \end{aligned} \tag{36}$$

Combining 34, 35 and 36, the approximation error of surrogate loss ϕ is bounded by

$$\epsilon_{\text{app}} \le \inf_{\tau', M_s}\left\{\frac{C_1}{(m-2\tau')^2} + C_2(d_0\Delta_{\tau'} + \Gamma_\tau) + C_3\mathbb{P}[A^c_{M_s}]\right\} \tag{37}$$

□

B.2 Proof of Lemma 4.10

B.3 Proof of Lemma 4.11

Lemma B.1. *Under Assumptions 4.1, for the normalized OHLC $\tilde{x} \in [0,1]^4$, denote $b_t = \sum_{i=0}^{t-1} B_W^i$. For $t > 0$, it holds that*

$$\|h_t\|_2 \le (2B_U + B_h)\sum_{i=0}^{t-1} B_W^i$$

and

$$\|h_t - h_t'\|_2 \le (2B_U + B_h)\sum_{i=1}^{t-1} iB_W^{i-1}\|W - W'\|_2 + 2b_t\|U - U'\|_2 + b_t\|b_h - b_h'\|_2$$

Proof. For the first inequality, by definition,

$$\begin{aligned} \|h_t\|_2 &= \|\sigma(Wh_{t-1} + Ux_t + b_h) - \sigma(0)\|_2 \\ &\le \|Wh_{t-1} + Ux_t + b_h\|_2 \\ &\le B_W\|h_{t-1}\|_2 + 2B_U + B_h \\ &\le (2B_U + B_h)\left(1 + \cdots + B_W^{t-1}\right) = (2B_U + B_h)\sum_{i=0}^{t-1} B_W^i \end{aligned}$$

Denote $b_t = \sum_{i=0}^{t-1} B_W^i$. For the second inequality, consider two hidden states h_t and h_t' induced by two groups of weight matrix θ and θ',

$$\begin{aligned} \|h_t - h_t'\|_2 &= \|\sigma(Wh_{t-1} + Ux_t + b_h) - \sigma(W'h_{t-1}' + U'x_t + b_h')\|_2 \\ &\le \|(Wh_{t-1} - W'h_{t-1}')\|_2 + 2\|U - U'\|_2 + \|b_h - b_h'\|_2 \\ &\le \|W - W'\|_2\|h_{t-1}\|_2 + B_W\|h_{t-1} - h_{t-1}'\|_2 + 2\|U - U'\|_2 + \|b_h - b_h'\|_2 \\ &\le (2B_U + B_h)b_{t-1}\|W - W'\|_2 + B_W\|h_{t-1} - h_{t-1}'\|_2 + 2\|U - U'\|_2 + \|b_h - b_h'\|_2 \\ &\le (2B_U + B_h)\sum_{i=0}^{t-1} b_i B_W^{t-i-1}\|W - W'\|_2 + 2\sum_{i=0}^{t-1} B_W^i\|U - U'\|_2 + \sum_{i=0}^{t-1} B_W^i\|b_h - b_h'\|_2 \\ &= (2B_U + B_h)\sum_{i=1}^{t-1} iB_W^{i-1}\|W - W'\|_2 + 2b_t\|U - U'\|_2 + b_t\|b_h - b_h'\|_2 \end{aligned}$$

□

Lemma B.2. *f_θ is Lipschitz to the parameters θ such that*

$$|f_\theta - f_{\theta'}| \le L_v\|v - v'\|_2 + L_W\|W - W'\|_2 + L_U\|U - U'\|_2 + L_h\|b_h - b_h'\|_2 + |b_o - b_o'|$$

where $L_v = (2B_U + B_h)\sum_{i=0}^{\tau-1} B_W^i$, $L_W = (2B_U + B_h)B_v\sum_{i=1}^{\tau-1} iB_W^{i-1}$, $L_U = 2B_v\sum_{i=0}^{\tau-1} B_W^i$, $L_h = B_v\sum_{i=0}^{\tau-1} B_W^i$.

Proof. By definition,

$$\begin{aligned}
|f_\theta(\tilde{x}) - f_{\theta'}(\tilde{x})| &= |(v^\top h_\tau + b_o) - ((v')^\top h'_\tau + b'_o)| \\
&\le |v^\top h_\tau - v^\top h'_\tau| + |v^\top h'_\tau - (v')^\top h'_\tau| + |b_o - b'_o| \\
&\le B_v \|h_\tau - h'_\tau\|_2 + \|h'_\tau\|_2 \|v - v'\|_2 + |b_o - b'_o| \\
&\le (2B_U + B_h) B_v \sum_{i=1}^{\tau-1} i B_W^{i-1} \|W - W'\|_2 + 2b_\tau B_v \|U - U'\|_2 \\
&\quad + b_\tau B_v \|b_h - b'_h\|_2 + (2B_U + B_h) \sum_{i=0}^{\tau-1} B_W^i \|v - v\prime\|_2 + |b_o - b'_o|
\end{aligned}$$

□

Now we formally introduce the definition of covering number.

Def B.3 (Covering Number)**.** *Let (A, d) be a metric space. We say that A is ϵ-covered by a finite set A' if for any $a \in A$, there exists $a' \in A'$ such that $d(a, a') \le \epsilon$. The covering number of A is defined as the cardinality of smallest A'*

$$\mathcal{N}(\epsilon, A, d) = \min\{|A'| : \sup_{a \in A} \inf_{a' \in A'} d(a, a') \le \epsilon\}$$

Lemma B.4 (Dudley's Inequality)**.** *Let $\mathcal{F}$ be a real-valued function class taking values in $[-r, r]$, and assume $0 \in \mathcal{F}$. Let $S = (s_1, \cdots, s_n)$ be a given vector, then*

$$\mathfrak{R}_S(\mathcal{F}) \le \inf_{\alpha > 0} \left(\frac{4\alpha}{\sqrt{n}} + \frac{12}{n} \int_\alpha^{2r\sqrt{n}} \sqrt{\log \mathcal{N}(\mathcal{F}, \epsilon, \|\cdot\|_F)} d\epsilon \right).$$

Given $\epsilon > 0$ and a sample set $S = (x_1, \cdots, x_n)$, consider the Euclidean distance $\|S_1 - S_2\|_2 = \left(\sum_{i=1}^n (S_1(x_i) - S_2(x_i))^2\right)^{1/2}$. If $\|v - v'\|_2 \le \frac{\epsilon}{4\sqrt{n} L_v}$, $\|W - W'\|_2 \le \frac{\epsilon}{4\sqrt{n} L_W}$, $\|U - U'\|_2 \le \frac{\epsilon}{4\sqrt{n} L_U}$ and $\|b_h - b'_h\|_2 \le \frac{\epsilon}{4\sqrt{n} L_h}$, by lemma B.2,

$$\begin{aligned}
|v^\top h_\tau - (v')^\top h'_\tau)| &\le \left(\sum_{i=1}^n \left(\frac{\epsilon}{4\sqrt{n}} + \frac{\epsilon}{4\sqrt{n}} + \frac{\epsilon}{4\sqrt{n}} + \frac{\varepsilon}{4\sqrt{n}} \right)^2 \right)^{1/2} \\
&\le \epsilon,
\end{aligned}$$

which gives the following lemma of covering number bound.

Lemma B.5.

$$\begin{aligned}
\log \mathcal{N}(\epsilon, A, \|\cdot\|_2) \le m^2 \log &\left(1 + \frac{8\sqrt{n} B_W L_W}{\varepsilon}\right) \\
&+ 4m \log \left(1 + \frac{8\sqrt{n} B_U L_U}{\varepsilon}\right) \\
&+ m \log \left(1 + \frac{8\sqrt{n} B_v L_v}{\varepsilon}\right) \\
&+ m \log \left(1 + \frac{8\sqrt{n} B_h L_h}{\varepsilon}\right)
\end{aligned} \tag{38}$$

Lemma B.6. *Under the assumption 4.1, let $S = (z_1, \ldots, z_n))$ be a sample of normalized OHLC sequences, and define $b_\tau = \sum_{i=0}^{\tau-1} B_W^i$, $a_\tau = \sum_{i=1}^{\tau-1} i B_W^i$. The empirical Rademacher complexity is bounded by*

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \le c B_v (2B_U + B_h) \sqrt{b_\tau \max\{b_\tau, a_\tau\}} \sqrt{\frac{m^2 + 6m}{n}} + B_o \sqrt{\frac{1}{n}} \tag{39}$$

where c is a universal constant.

C EXPERIMENTS

To verify that the predictability pattern in Figure 1(c) is not specific to the bound-based proxy B_s^2/σ^2, we repeat the experiment using the conventional variance-based SNR $\mathrm{Var}(s_t)/\sigma^2$. Figure 4 shows the same qualitative increase in excess risk as predictability increases.

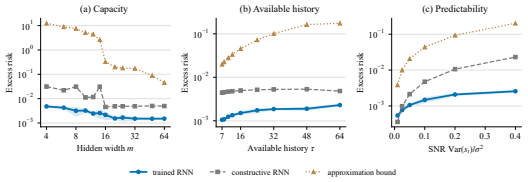


Figure 4: Robustness to an alternative predictability measure, the variance-based SNR $\mathrm{Var}(s_t)/\sigma^2$. The same increasing trend observed with B_s^2/σ^2 in Figure 1(c) is preserved.