The Nonstationarity-Complexity Tradeoff in Return Prediction*

Agostino Capponi† Chengpiao Huang‡ J. Antonio Sidaoui§
Kaizheng Wang¶ Jiacheng Zou‖

This version: August 13, 2026

Abstract

Does more data improve return prediction? In non-stationary financial markets, longer training windows improve prediction of complex models but incorporate outdated economic regimes, whereas simpler models require less data and are less vulnerable to changes in economic conditions. We formally characterize this nonstationarity-complexity tradeoff, showing that model complexity and training window length must be jointly optimized. We propose an adaptive selection procedure with formal performance guarantees. Over three decades of U.S. equity markets, our method improves out-of-sample R^2 on industry portfolios by 14% relative to fixed-window and regime-switching benchmarks, with large gains during recessions.

Keywords: Non-stationarity, Model complexity, Return prediction, Model selection, Machine learning, Adaptive window selection

Authors are listed in alphabetical order. Code is available at https://github.com/ch3702/ATOMS.

*We are grateful to the seminar participants of the Tech 4 Finance conference, Fordham Quant 2026 Conference, Imperial College, and Voleon Research group. We acknowledge the constructive comments from Guillaume Coqueret, Marcin Kacperczyk, Serge Darolles, Blair Bilodeau, Yash Deshpande, Francesco Sangiorgi, Johannes Muhle-Karbe. Agostino Capponi's research was supported in part through grants of the Global Risk Institute and of Fi-Tek. Chengpiao Huang and Kaizheng Wang's research is supported by National Science Foundation grants DMS-2210907 and DMS-2515679.

†Department of IEOR and Columbia Business School, Columbia University. Email: ac3827@columbia.edu.

‡Department of IEOR, Columbia University. Email: chengpiao.huang@columbia.edu.

§Department of IEOR, Columbia University. Email: j.sidaoui@columbia.edu.

¶Department of IEOR and Data Science Institute, Columbia University. Email: kaizheng.wang@columbia.edu.

‖Email: jiachengzou@alumni.stanford.edu.

1 Introduction

Does more data improve return predictions? The machine learning models now standard in this literature (Gu et al., 2020; Freyberger et al., 2020; Kelly and Xiu, 2023; Kelly et al., 2024b) are heavily parameterized, and estimating them well requires ample samples, which in turn requires long periods of observation whenever the sampling frequency is fixed.[1] Yet old data reflect economic conditions that may no longer hold: risk premia vary over the business cycle (Hamilton, 1989), policy regimes change, and the returns to known predictors decline after those predictors become widely known (McLean and Pontiff, 2016).

Classical bias-variance tradeoff centers on model complexity and sample size under stationary conditions, so additional observations reduce estimation variance at no cost. Non-stationarity complicates this logic. A complex model has low misspecification error but high estimation variance. Controlling that variance and the overall prediction error requires training data drawn from economic conditions similar to those prevailing at the prediction date. Hence, a longer training window of observations help only if they come from the same regime rather than an outdated one. Otherwise, a simple model estimated on a short, recent window can outperform a complex model estimated on the full history when market conditions are most volatile. We refer to the resulting tension as the *nonstationarity–complexity tradeoff*. Resolving this tension requires answering:

How to choose model complexity and training window jointly under non-stationarity?

Existing studies treat these as two separate decisions: a modeler first commits to a model class, then to a training window, typically an expanding window that uses all available history or a rolling window of fixed length. Empirically, we find that this separation could be costly when it matters economically. In doing so, we complement the machine learning in asset pricing literature, including the documented success of complex models and large factor zoos in return prediction (Kelly et al., 2022, 2024a; Didisheim et al., 2024), by considering markets in which economic conditions shift over time in ways that cannot be known in advance.

To answer this, we first derive a finite-sample bound that decomposes prediction error into model misspecification, statistical uncertainty, and a cost of non-stationarity, and use it to characterize the model class and training window that jointly minimize prediction error. We

[1] For instance, if the firm characteristics are constructed from quarterly disclosures of accounting information (Freyberger et al., 2020), researchers cannot easily choose to increase the observation frequency.

then develop an adaptive data-driven procedure that selects both without advance knowledge of when the economic environment will shift, and prove it performs nearly as well as the best choice one could have made in hindsight. Applied to 17 U.S. industry portfolios from 1990–2016, the procedure raises out-of-sample R^2 by 14% and the mean-variance monthly Sharpe ratio from 0.30 to 0.33 relative to the best fixed-window alternative, with the largest gains concentrated in the three NBER recessions in our sample.

Our paper makes methodological and empirical contributions. We prove a finite-sample bound on prediction error that separates three sources of error: the misspecification of the model class, the statistical uncertainty from estimating the model on a finite training sample, and a cost of non-stationarity that reflects how far the economic conditions spanned by the training window have drifted from those prevailing at the time of prediction. The bound makes precise how model complexity and training-window length jointly affect the prediction error: a richer model class lowers misspecification error but raises statistical uncertainty, which can only be reduced with a longer window, and a longer window in turn raises exposure to outdated regimes. Jointly, these forces pin down the model class and window length that minimize prediction error, and the optimal choice depends on how severe the prevailing non-stationarity is.

We develop an adaptive procedure that selects the model class and training window jointly, without requiring prior knowledge of when or how the economic environment will shift. The procedure compares candidate models pairwise, choosing a separate look-back window for each comparison to balance stale data against too little data, and combines these comparisons through a sequential elimination tournament that stays fast even as more candidate models are added. We prove that the resulting procedure selects a model class and training window whose performance is nearly as good as the best choice one could have made in hindsight, without knowing in advance how market conditions would evolve. The procedure requires no parametric model of the return-generating process and imposes only mild regularity conditions on how that process drifts over time, so it accommodates any set of candidate models, whatever their model class or training horizon.

We evaluate this procedure on daily returns of 17 U.S. industry portfolios from Kenneth French's data library, with models estimated on daily data and predictive accuracy measured at the monthly frequency, over an out-of-sample period from January 1990 to November 2016. The procedure raises the average out-of-sample R^2, computed against a zero-forecast

benchmark, to 0.049, a 14% improvement over the strongest fixed-window alternative and more than double the performance of a benchmark validated only on recent data. In an implementable mean-variance strategy across the 17 portfolios, the same procedure raises the monthly Sharpe ratio from 0.30 for the best fixed-window alternative to 0.33 over the full sample, without higher turnover.

These gains concentrate exactly where the tradeoff predicts they should: in the three recessions identified by the National Bureau of Economic Research (NBER)[2] that fall within our sample, when non-stationarity is most severe. During the 1990 Gulf War recession, our procedure delivers the highest out-of-sample R^2 (0.027); the medium- and long-window benchmarks and the regime-switching model all turn negative, and even the shortest fixed window, restricted to only the most recent data, manages just 0.009. During the 2001 recession, which combined the dot-com repricing with the September 11 attacks, our procedure attains an R^2 of 0.125 — more than five percentage points above cross-validation's 0.071, and narrowly ahead of the best long-window benchmark, which reaches 0.117. During the 2008 financial crisis, a longer-lived downturn that gave fixed-window methods more time to catch up, our procedure's advantage narrows but does not disappear: its R^2 of 0.041 is slightly above the best long-window benchmark's 0.039, and its mean-variance Sharpe ratio of 0.24 remains well above cross-validation's 0.05. This pattern holds across all 17 industries rather than concentrating in a few, and the resulting portfolio gains are not offset by higher trading costs.

The rest of the paper is organized as follows. Section 2 describes the problem setup. Section 3 investigates the nonstationarity-complexity tradeoff. Section 4 presents the adaptive model selection algorithm. Section 5 illustrates our algorithm on real datasets. Section 6 concludes the paper and discusses future directions. Mathematical proofs are deferred to the appendices. The code for implementing the algorithms in the paper is available at https://github.com/ch3702/ATOMS.

1.1 Related Literature

Our paper relates to two strands of research in asset pricing: one that documents the gains from model complexity in predicting stock returns, and one that studies how much historical

[2]NBER Business Cycle Dating, https://www.nber.org/research/business-cycle-dating.

data investors should use to estimate return models as economic conditions change.

The integration of machine learning into return prediction and asset pricing was initially driven by the "multidimensional challenge", that is, the need to identify which of the hundreds of proposed firm characteristics provide independent information for expected returns.

Early influential work by Gu et al. (2020) demonstrated that nonlinear interactions missed by traditional regressions are a primary source of predictive gains, identifying trees and neural networks as superior methods. Freyberger et al. (2020) used adaptive group LASSO to show that only a small subset of characteristics provides incremental information when nonlinearities are properly accounted for. Choi et al. (2025) applied machine learning to 32 international markets, concluding that market-specific neural networks achieve stronger results than global models by capturing local return-characteristic relationships. We refer to Kelly and Xiu (2023) for an excellent survey on financial machine learning.

Building on these empirical successes, a series of theoretical research have formalized the virtue of complexity, proving that out-of-sample forecast accuracy and portfolio Sharpe ratios can be strictly increasing in model complexity. This phenomenon occurs because high complexity induces "implicit shrinkage", which reduces prediction variance without the heavy bias costs associated with explicit shrinkage. This line of research advocates for the largest approximating model one can compute, because the gains from better approximation of the unknown truth dominate the statistical costs of heavy parameterization. Foundational works include Kelly et al. (2022); Kelly and Malamud (2025); Kelly et al. (2024b). They focus on time-series return prediction and market timing, resolving the "double limit" problem of growing parameters and observations to show that complexity captures unknown nonlinearities that improve Sharpe ratios. A recent study by Didisheim et al. (2024) extends these insights to the cross-section of returns, tackling a "three infinities" problem involving a simultaneously large number of assets, parameters, and observations. Their work proposes using random Fourier features to generate vast numbers of nonlinear factors, shifting the statistical objective from pure return prediction to minimizing pricing errors and constructing a high-complexity stochastic discount factor that reflects the true drivers of investors' marginal rates of substitution.

Kelly and Malamud (2025) explicitly acknowledge the limitations of the large-T asymptotic framework when applied to real-world financial data. In their discussion of finite-sample properties, they note that "likely non-stationarities in financial markets data ... begs an

agenda to understand how machine learning models can aid in the development of models that are robust to structural breaks". Our work directly addresses this question by developing a framework that accounts for non-stationarity in the joint selection of model complexity and training windows. While Kelly and Malamud (2025) establish that complexity can be virtuous under stationarity, we demonstrate that realizing this virtue in practice requires careful calibration to the local non-stationary environment, particularly during structural breaks such as recessions, in which, as we show below, simple models with a short training window can outperform complex models estimated over longer histories.

A separate literature studies how to detect changes in market conditions, such as structural breaks and regime changes (Banerjee and Urga, 2005). Foundational works such as Chow (1960); Andrews (1993); Bai and Perron (1998); Chib (1998) developed methods to identify these changes, which have been applied to financial time series including realized volatility (Liu and Maheu, 2007) and speculative bubbles (Homm and Breitung, 2012). While these studies focus on identifying when breaks occur, our work addresses how to optimally use non-stationary data for estimation. Rather than pinpointing change points, we determine the optimal training window to minimize prediction error in the presence of non-stationarity.

A complementary literature examines optimal training window selection under non-stationarity but with fundamentally different assumptions than this paper. Pesaran and Timmermann (2007) showed that under structural breaks, optimal window selection should balance the bias from including pre-break data against the variance from using only post-break data. Subsequent work explored various selection criteria, including minimizing estimation loss functions (Pesaran and Timmermann, 2007; Inoue et al., 2017) and aggregating predictions across multiple windows (Pesaran and Pick, 2011). However, these approaches typically assume linear models and specific non-stationary structures, such as single breaks or random walks. Our contribution differs along two dimensions. We take a model-free and distribution-free approach that imposes no parametric assumptions on either the prediction function or the non-stationary dynamics, allowing us to handle more general patterns; and we extend the bias-variance tradeoff identified by Pesaran and Timmermann (2007) to the modern machine learning context by jointly optimizing model complexity and training window length with formal finite sample error bound, whereas prior work typically selects a window for a model that is fixed in advance.

Notation. We introduce the mathematical notation used throughout the paper. Let $\mathbb{Z}_+ = \{1, 2, ...\}$ be the set of positive integers. For $n \in \mathbb{Z}_+$, define $[n] = \{1, 2, ..., n\}$. For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $x \in \mathbb{R}$, let $x_+ = x \vee 0$. For non-negative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = O(b_n)$ if there exists $C > 0$ such that for all $n \in \mathbb{Z}_+$, $a_n \leq C b_n$. We write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Unless otherwise stated, $a_n \lesssim b_n$ also represents $a_n = O(b_n)$, and $a_n \asymp b_n$ also represents $a_n = \Theta(b_n)$. For a finite set S, we use $|S|$ to denote its cardinality.

2 Framework for Return Prediction under Non-Stationarity

We describe the return prediction problem common in asset pricing. At modeling time t, the goal is to estimate the conditional mean $\mathbb{E}_t[y \mid \boldsymbol{x}]$ of an asset return $y \in \mathbb{R}$ given predictors $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. The model is trained on historical data from periods $\tau = 1, \ldots, t-1$, where for each past period τ we observe a batch B_τ[3] of observations $\mathcal{D}_\tau = \{(\boldsymbol{x}_s, y_s)\}_{s \in B_\tau}$. The key challenge is non-stationarity: the joint distribution P_τ of $(\boldsymbol{x}, y)$ varies over time, with $P_\tau \neq P_{\tau'}$ for $\tau \neq \tau'$ in a fully general sense.

The data points $\mathcal{D}_\tau = \{(\boldsymbol{x}_s, y_s)\}_{s \in B_\tau}$ are independent and identically distributed (i.i.d.) following the period-τ distribution P_τ, where $s \in B_\tau$ describes the period-τ batch of data. We use this notation to reflect that the distribution P_τ could change in unknown manners as its time index τ changes, but has the typical statistical similarity within a given period. That is, throughout our paper, we assume that observations within each batch are i.i.d., whereas batches from different time periods are independent but may follow distinct distributions, hence allowing for non-stationarity. Formally,

Assumption 2.1 (Independent data)**.** For each $\tau \leq t-1$, the dataset $\mathcal{D}_\tau$ consists of independent and identically distributed samples. The datasets $\{\mathcal{D}_\tau\}_{\tau=1}^{\infty}$ are independent across time periods.

While returns and many other financial time series inherently exhibit temporal dependence, Assumption 2.1 is a standard simplification in the theoretical analysis of machine

[3]We use a batch B_τ to refer to a general collection of observations at a time τ. This could represent, for example, (i) multiple observations of the same asset from intraday intervals; (ii) a cross-sectional collection of returns from different assets. Our framework accommodates (i) and (ii) as well as any other sampling and observation scheme, so long as they satisfy the no-look-ahead-bias constraint of $\tau \leq t-1$ at the given time τ. With a slight abuse of notation, we also use B_τ to denote its own cardinality.

learning (Kelly et al., 2024b; Didisheim et al., 2024). Adopting this independence assumption allows us to isolate the effect of non-stationarity, without introducing additional technicalities from temporal dependence. More generally, our theoretical analysis extends to certain dependent settings. In particular, suppose there exists a latent state process $\{z_\tau\}_{\tau=1}^{\infty}$, which may be temporally dependent. In each period τ, conditioned on z_τ, the dataset $\mathcal{D}_\tau$ is drawn i.i.d. from a distribution $Q_\tau(\cdot \mid z_\tau)$, independent of the past states $\{z_{\tau'}\}_{\tau' \leq \tau-1}$ and data $\{\mathcal{D}_{\tau'}\}_{\tau' \leq \tau-1}$. In this case, the datasets $\{\mathcal{D}_\tau\}_{\tau=1}^{\infty}$ are independent conditioned on the latent states $\{z_\tau\}_{\tau=1}^{\infty}$, and our theoretical results should be interpreted as being conditioned on the realized latent states. Unconditionally, however, the datasets may remain dependent.

Our goal is to use the historical data $\{\mathcal{D}_\tau\}_{\tau=1}^{t-1}$ to construct a prediction model $f_t : \mathcal{X} \to \mathbb{R}$ that performs well on the current, unobserved distribution P_t. The performance of a model $f : \mathcal{X} \to \mathbb{R}$ with respect to the data distribution P_t is measured by the mean squared error (MSE):

$$L_t(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim P_t}\left[(f(\boldsymbol{x}) - y)^2\right]. \tag{2.1}$$

In line with the empirical finance literature, we also use the R^2 metric to evaluate the performance of a given modeling procedure or algorithm `alg` that produces a prediction model $f_t^{\texttt{alg}}$ at each time t. The out-of-sample R^2 for the algorithm `alg` over an evaluation period $[t_1, t_2]$ is computed as[4]

$$R^2_{[t_1,t_2]}(\texttt{alg}) = 1 - \frac{\sum_{t=t_1}^{t_2}\sum_{s\in B_t}\left(f_t^{\texttt{alg}}(x_s) - y_s\right)^2}{\sum_{t=t_1}^{t_2}\sum_{s\in B_t} y_s^2}. \tag{2.2}$$

For completeness, in the appendices we also present results using the statistical R^2 metric

$$R^2_{[t_1,t_2],\mathbf{s}}(\texttt{alg}) = 1 - \frac{\sum_{t=t_1}^{t_2}\sum_{s\in B_t}\left(f_t^{\texttt{alg}}(x_s) - y_s\right)^2}{\sum_{t=t_1}^{t_2}\sum_{s\in B_t}(y_s - \bar{y})^2}, \tag{2.3}$$

where $\bar{y}$ is the mean of the samples $\{y_s : s \in B_t,\, t \in [t_1, t_2]\}$.

In a stationary environment where $P_\tau = P$ for all τ, the standard approach for learning

[4]We note that in the R^2 metric (2.2), the denominator is the sum of the squared responses y_s^2 *without demeaning*. In other words, we are benchmarking against a forecast of zero rather than the historical mean as in the statistical R^2 metric. As noted by Gu et al. (2020), predicting future excess stock returns with historical averages can be problematic and is not assumption-free, because the historical mean is estimated with significant noise, often performing worse than a forecast of zero.

a model f consists in choosing a model class $\mathcal{F}$ (e.g., linear model, random forest) and then fitting a model $\widehat{f} \in \mathcal{F}$ by minimizing the empirical loss over the training data. The choice of the model class $\mathcal{F}$ involves a classic bias-variance trade-off. A simple class may exhibit high bias due to model misspecification, while a complex class may suffer from high estimation variance.

When the environment is non-stationary (that is, $P_\tau \neq P_{\tau'}$ for $\tau \neq \tau'$), the problem becomes significantly more complicated. One must now make two critical choices simultaneously: the model class $\mathcal{F}$ and the amount of historical data used for training. Data from the distant past may no longer be representative of the current environment, and can be misleading for model training. This creates the core tension of our paper: complex models require more data to reduce estimation variance, but using more data may introduce stronger non-stationarity that increases bias. Thus, it is possible for simple models with less training data to outperform complex models trained on more data. Our goal is to develop an approach to jointly choose the model class and training window size.

3 The Nonstationarity-Complexity Tradeoff

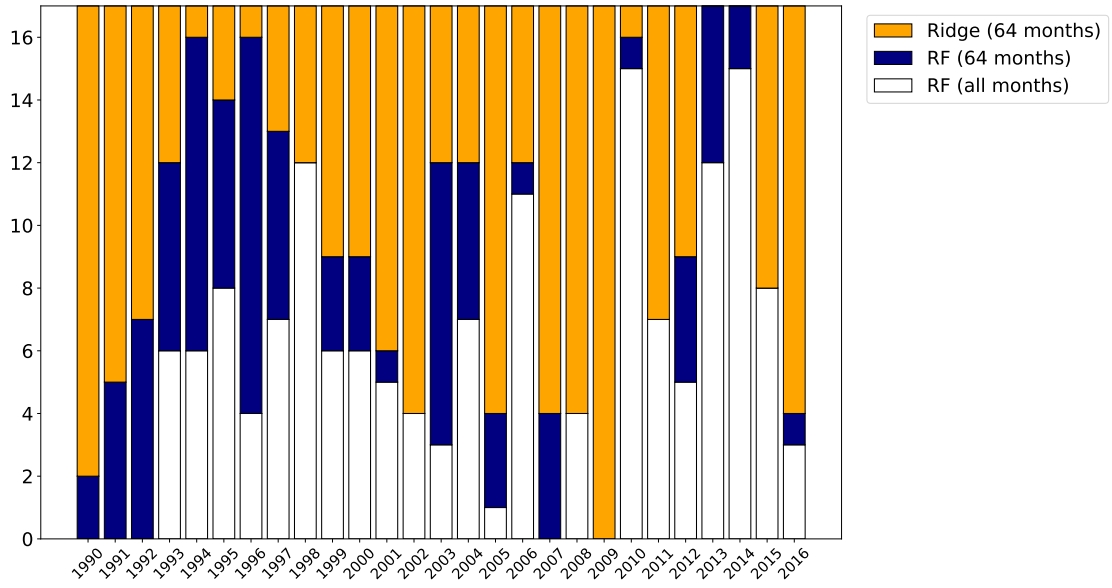
In this section, we formally investigate the tension created by the joint choice of model class and training window. We begin by providing empirical evidence for this tradeoff, and then formalize it through a prediction error bound as a new theoretical result of learning theory, which allows for a wide range of machine learning models and non-stationarity.

3.1 Empirical Evidence

We begin with an empirical illustration that highlights the challenges of jointly choosing a model class and a training window under non-stationarity. The task is to forecast the excess returns of 17 industry portfolios from Kenneth French's data library using a set of covariates, with training data starting from September 1987 and ending in October 2016.[5] We highlight that our data spans several recessions documented in NBER Business Cycle Dating: the 1990 Gulf War recession, the 2001 dot-com bubble bust and the 9/11 attack, and the 2008 Financial Crisis. To show that model performance is fundamentally linked

[5]More details about the dataset are provided in Section 5.

Figure 1: Number of Industries where Each Model Attains the Highest Annual Out-of-Sample R^2.



This figure reports the relative performance of three models in predicting the excess returns of the 17 industry portfolios. The three models are: (1) a linear model trained by ridge regression on the most recent 64 months of data (orange), (2) a random forest trained on the most recent 64 months of data (blue), and (3) a random forest trained on all historical data (white). For each year from 1990 to 2016, we compute the models' annual out-of-sample R^2 in each of the 17 industries, and then count, for each model, the number of industries in which it outperforms the other two models in terms of the annual out-of-sample R^2.

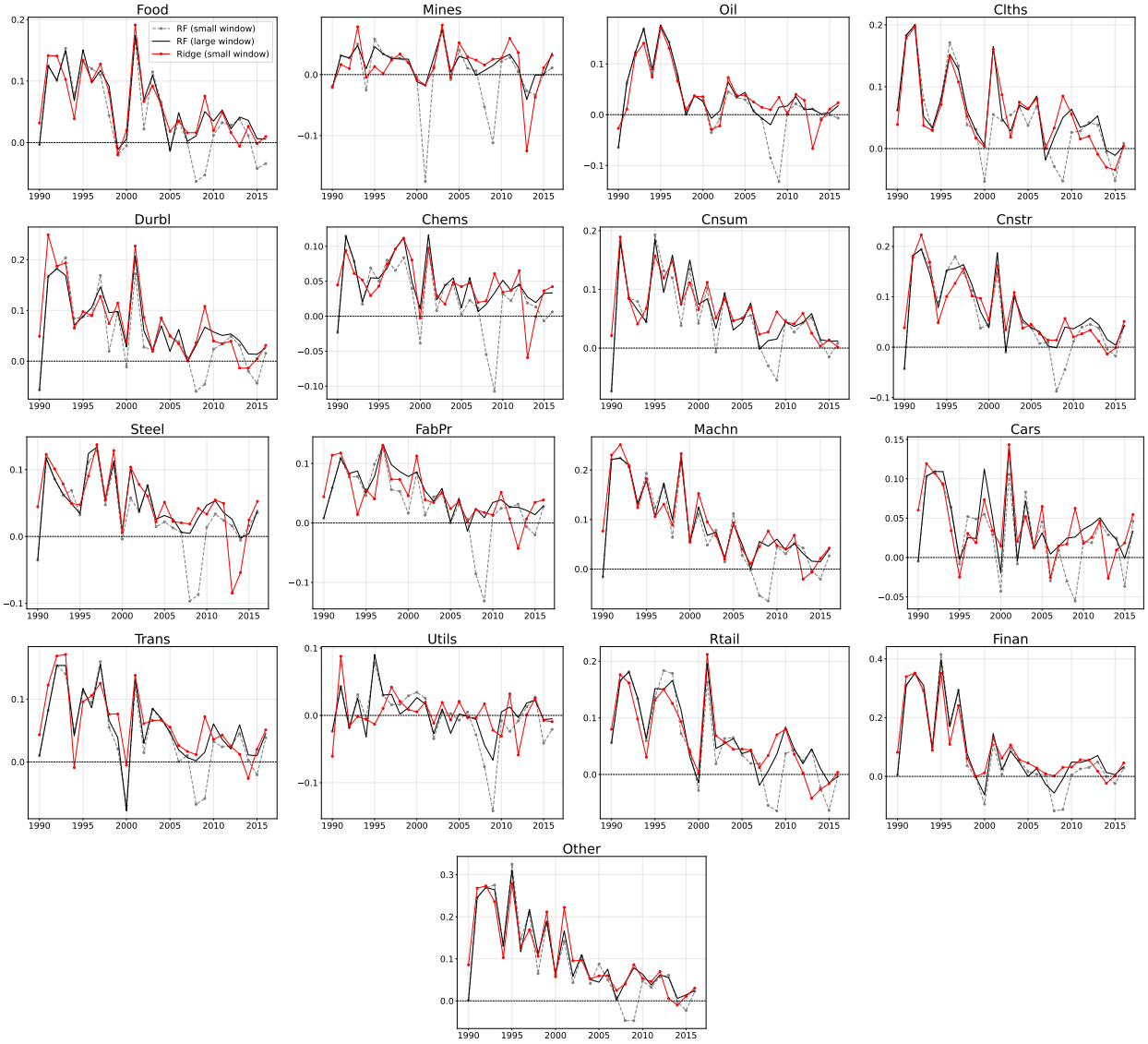
to non-stationarity, we document a simple ranking of linear and nonlinear models in each period, across the different industries.

In each month t, for each industry, we fit three prediction models: (1) a linear model trained by ridge regression using the most recent 64 months of data, (2) a random forest trained on the most recent 64 months of data, and (3) a random forest trained on all historical data. More details of the experiments are given in Appendix F.1.

We compute each model's annual out-of-sample R^2 for every industry. To visualize the models' relative performance across industries, we count the number of industries in which each model achieves the highest out-of-sample R^2 for a given year. Figure 1 summarizes the result. To provide a more granular understanding of the models' performance, Figure 2 further plots the annual out-of-sample R^2 of the models for the 17 industries to provide detailed accuracies in addition to ranking information.

We make two key observations. First, within the same model class, using less training

Figure 2: Annual Out-of-Sample R^2 of Random Forest and Ridge Models for 17 Industry Portfolios.



This figure reports, for each of the 17 industry portfolios, the annual out-of-sample R^2 from 1990 to 2016 for three models: (1) a linear model trained by ridge regression on the most recent 64 months of data (red), (2) a random forest trained using the most recent 64 months of data (gray), and (3) a random forest trained using all available historical data up to that year (black). In periods of strong non-stationarity, such as 1990-1991, 2001-2002 and 2008-2009, the linear model trained on a small window constantly outperforms the more complex random forest trained on a large window. The labels in each figure is the Kenneth French acronym for the industries. For full names of these industries, please refer to Table 10.

data may lead to better performance. For example, in several years including 1994, 1996 and 2003, the random forest trained on the most recent 64 months of data outperforms the random forest trained on all historical data for at least half of the industries.

Second, and more strikingly, a simple model trained on a short window can outperform a complex model trained on a long window. In particular, during the three NBER-designated recessions, the simple linear model trained on 64 months of data outperforms the more complex random forest model trained on all historical data for over half of the industries. This consistent pattern shows that the advantage of a more expressive model class can be completely negated by the non-stationarity in the training data. In Section 4, we propose a data-driven approache to select the best-performing model under non-stationarity.

These empirical findings highlight that in a non-stationary environment, the model complexity and training data size are intricately linked with each other. We call this phenomenon the *nonstationarity-complexity tradeoff*. Crucially, the optimal choice of model class and training window size is not fixed; instead, it generally varies with the non-stationarity.

3.2 Formal Characterization of the Tradeoff

We now provide theoretical support for the nonstationarity-complexity tradeoff, by deriving a finite-sample bound on a model's prediction error under non-stationarity. The bound decomposes the prediction error into three key components: model misspecification error, statistical uncertainty, and non-stationarity, and shows how each of them interacts with the choice of model class and training window size. We note that this bound is derived for theoretical analysis only, and does not explicitly characterize the optimal choice of the model class and training window. In the next section, we will develop a data-driven approach to adaptively select a near-optimal choice of model class and training window.

Consider a model $\widehat{f}$ trained from a model class $\mathcal{F}$ by minimizing the empirical loss over training data from the last k periods, denoted by $\left\{\mathcal{D}_\tau^{\text{tr}}\right\}_{\tau=t-k}^{t-1}$, where $\mathcal{D}_\tau^{\text{tr}} = \{(\boldsymbol{x}_s, y_s)\}_{s\in B_\tau^{\text{tr}}}$ is the training data in period τ, with $B_\tau^{\text{tr}} \subseteq B_\tau$. That is,

$$\widehat{f} = \underset{f\in\mathcal{F}}{\operatorname{argmin}} \frac{1}{B_{t,k}^{\text{tr}}} \sum_{\tau=t-k}^{t-1} \sum_{s\in B_\tau^{\text{tr}}} \left[f(\boldsymbol{x}_s) - y_s\right]^2, \tag{3.1}$$

where $B_{t,k}^{\text{tr}} = \sum_{\tau=t-k}^{t-1} B_\tau^{\text{tr}}$ is the number of training data points in $\left\{\mathcal{D}_\tau^{\text{tr}}\right\}_{\tau=t-k}^{t-1}$.

Define the Bayes optimal least squares estimator $f_t^*(\cdot) = \mathbb{E}_{(\boldsymbol{x},y)\sim P_t}[y \mid \boldsymbol{x} = \cdot]$, which minimizes the MSE $L_t(f)$ over all possible prediction models $f : \mathcal{X} \to \mathbb{R}$. Our bound will be stated in terms of the *excess risk*

$$\mathcal{E}_t(f) = L_t(f) - L_t(f_t^*),$$

which compares the prediction error of a model f against that of f_t^*. To facilitate analysis, we make the following boundedness assumption.

Assumption 3.1 (Boundedness)**.** There exists a constant $M > 0$ such that for all models f in the class $\mathcal{F}$, $(\boldsymbol{x}, y) \sim P_\tau$ and $\tau \in \mathbb{Z}_+$, we have $|f(\boldsymbol{x})| \leq M$ and $|y| \leq M$. Without loss of generality, we assume $M \geq 1$.

To quantify the effective complexity of the model class $\mathcal{F}$ relative to the training window size k, we employ a measure $r_{t,k}(\mathcal{F})$ derived from the well-known *local Rademacher complexity* (Bartlett et al., 2005) commonly adopted in statistical learning theory. Given the technical nature of this measure, we defer its formal definition to Appendix B. Intuitively, the local Rademacher complexity measures the ability of the near-optimal models in $\mathcal{F}$ to fit random noise using data within the training window k. A higher local Rademacher complexity indicates a richer model class that is capable of approximating complex patterns, but also signals a higher estimation variance and thus a higher risk of overfitting.

Examples of Local Rademacher Complexity in Finance. To provide concrete illustrations, we now present the local Rademacher complexity measure $r_{t,k}(\mathcal{F})$ for several common model classes. These results are proved in Appendix B.8.

Example 3.1 (Linear class)**.** Recall $\mathcal{X} \subseteq \mathbb{R}^d$. For every $\boldsymbol{\theta} \in \mathbb{R}^d$, define $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$ by $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$. Suppose $\mathcal{F} \subseteq \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$. Then, $r_{t,k}(\mathcal{F}) \leq cd/B_{t,k}^{\mathrm{tr}}$ for some constant c.

We note two relevant examples of linear class functionals in empirical asset pricing. Classical asset pricing fixed-dimension models such as Fama and French (1996, 2015) fall clearly into this category. Ridge regression used in Gu et al. (2020) corresponds to $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\|_2 \leq R\}$, which is also in the model class described by this example.

Example 3.2 (Kernel class)**.** A kernel class is a nonparametric model class where we fit a linear model in a rich (possibly infinite-dimensional) feature space induced by a kernel. Let

$\mathbb{H}$ be a reproducing kernel Hilbert space (Wahba, 1990) with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|_{\mathbb{H}}$, and $\phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping. For any $\boldsymbol{\theta} \in \mathbb{H}$, define $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}$ by $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle$. Consider the function class $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\|_{\mathbb{H}} \leq R\}$ for some constant $R > 0$. Model fitting in this class can be efficiently implemented through kernel ridge regression, which is a finite-dimensional convex program even if $\mathbb{H}$ and $\mathcal{F}$ are infinite-dimensional.

Suppose there exists a trace-class operator $\boldsymbol{S} : \mathbb{H} \to \mathbb{H}$ such that for any $\tau \in \mathbb{Z}_+$ and $\boldsymbol{v} \in \mathbb{H}$, we have $\mathbb{E}_{(\boldsymbol{x},y)\sim P_\tau} |\langle \phi(\boldsymbol{x}), \boldsymbol{v} \rangle|^2 \leq \langle \boldsymbol{v}, \boldsymbol{S}\boldsymbol{v} \rangle$. Let $\{\mu_j\}_{j=1}^{\infty}$ be the eigenvalues of $\boldsymbol{S}$ sorted in descending order. We have the following results:

- (Exponential decay) If there are constants $c_1, c_2 > 0$ such that $\mu_j \leq c_1 e^{-c_2 j}$ holds for all j, then $r_{t,k}(\mathcal{F}) \leq (C \log B_{t,k}^{\mathrm{tr}})/B_{t,k}^{\mathrm{tr}}$ holds with some constant C.
- (Polynomial decay) If there are constants $c > 0$ and $\alpha \geq 1$ such that $\mu_j \leq c j^{-2\alpha}$ holds for all j, then $r_{t,k}(\mathcal{F}) \leq C\big(B_{t,k}^{\mathrm{tr}}\big)^{-\frac{2\alpha}{2\alpha+1}}$ holds with some constant C.

How does the kernel class relate to asset pricing? Fast (exponential) decay corresponds to a pricing relationship that is effectively spanned by a few dominant directions in the feature space, so the kernel class behaves like a low-complexity model despite its nominal richness. Slow (polynomial) decay corresponds to a pricing relationship in which many directions carry independent predictive content, so the model must be treated as genuinely high-dimensional.

Specifically, recent asset pricing literature has also considered machine learning models related to kernel classes. Didisheim et al. (2024) construct high-dimensional SDFs from random Fourier factors, which transform raw firm characteristics into many nonlinear sine and cosine features, and then estimate a linear SDF over these generated factors. This can be viewed as a finite-dimensional approximation to a kernel model, where the random features provide an explicit feature map for a rich nonlinear kernel class. The construction is also equivalent to a two-layer neural network with random first-layer weights and estimated output weights. More generally, wide neural networks are often connected to kernel classes through their neural tangent kernels (Jacot et al., 2018).

We now proceed to describe our prediction error bound under non-stationarity. In the classical setting where the training data $\{\mathcal{D}_\tau^{\mathrm{tr}}\}_{\tau=t-k}^{t-1}$ is i.i.d., the complexity measure $r_{t,k}(\mathcal{F})$ is a key component in bounding the excess risk of $\widehat{f}$. Standard results in statistical learning

theory (Bartlett et al., 2005; Wainwright, 2019) show that with high probability,

$$\mathcal{E}_t(\widehat{f}) \lesssim \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + \left(r_{t,k}(\mathcal{F}) + \frac{1}{B^{\mathrm{tr}}_{t,k}} \right), \tag{3.2}$$

where $\lesssim$ hides a multiplicative universal constant. In particular, the prediction error on the right-hand side of Eq. (3.2) is decomposed into two terms:

1. *Model misspecification error*

$$\min_{f\in\mathcal{F}} \mathcal{E}_t(f) = \min_{f\in\mathcal{F}} L_t(f) - L_t(f_t^*),$$

   which describes how well the model class $\mathcal{F}$ can approximate the Bayes optimal least squares estimator f_t^* at time t. A more complex model class tends to reduce the model misspecification error.

2. *Statistical uncertainty*

$$r_{t,k}(\mathcal{F}) + \frac{1}{B^{\mathrm{tr}}_{t,k}},$$

   which quantifies the estimation variance of the model $\widehat{f}$. As is discussed above, using a more complex model class $\mathcal{F}$ increases the statistical uncertainty of the fitted model. Consequently, a more complex model typically requires a longer training window k to mitigate its estimation variance.

The classical error bound (3.2) shows that in the i.i.d. case, increasing the training window size k always reduces the statistical uncertainty, thereby lowering the total prediction error. However, we now present our theory to show that under non-stationarity, this logic is incomplete. As we increase the window size k to reduce estimation variance, we inadvertently include older data distributions that differ from the target, introducing a third error component. We formalize this in the following theorem.

Theorem 3.1 (Prediction error bound)**.** *Let Assumptions 2.1 and 3.1 hold, and fix $\delta \in (0,1)$. With probability at least $1-\delta$, the model $\widehat{f}$ defined by (3.1) satisfies*

$$\mathcal{E}_t(\widehat{f}) \lesssim \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(1/\delta)}{B^{\mathrm{tr}}_{t,k}} \right) + M^2 \max_{t-k\le\tau\le t-1} \mathrm{TV}\left(P_\tau, P_t\right).$$

Here $\lesssim$ *hides a universal constant, and* $\mathrm{TV}(P_\tau, P_t) = \max_A |P_\tau(A) - P_t(A)|$ *is the total variation distance.*

Proof of Theorem 3.1. See Appendix B. □

Contrasting with the classical stationary case in (3.2), Theorem 3.1 reveals that the non-stationarity adds a third dimension to the classical prediction error bound (3.2), namely, a *non-stationarity* term

$$\max_{t-k\le\tau\le t-1} \mathrm{TV}\left(P_\tau, P_t\right),$$

which quantifies the largest environment shift between the current modeling time t and any period τ included in the training window. Mathematically, it measures the changes in the joint distribution of covariates and return in terms of the total variation distance.

In markets, non-stationarity naturally arises from shifts in risk premia, credit conditions, investor risk aversion, and the distribution of firm characteristics. These shifts as a high-dimensional event since any of these market indicators may have marginal or joint distributional variations. Whenever a shift happens, historical data becomes progressively less informative for model estimation. The term $\mathrm{TV}(P_\tau, P_t)$ summarizes this mismatch between the past data distribution P_τ and the current distribution P_t. The financial economics content of this bound is most transparent during episodes of sudden market stress.

The error decomposition in Theorem 3.1 formalizes the empirical observations in Section 3.1. It shows that model complexity and training-window size affect different components of prediction error. Using a more complex model class $\mathcal{F}$ can reduce misspecification error $\min_{f\in\mathcal{F}} \mathcal{E}_t(f)$, but it also increases statistical uncertainty $r_{t,k}(\mathcal{F})$. Increasing the training window k reduces statistical uncertainty $r_{t,k}(\mathcal{F}) + \frac{\log(1/\delta)}{B^{\mathrm{tr}}_{t,k}}$ by providing more samples, but is also more likely to include data generated under distributions P_τ far from the current distribution P_t. Table 1 summarizes these effects.

To interpret the tradeoff presented by Theorem 3.1 from an investor's perspective, consider an asset manager who estimated a return-prediction model in mid-1990 using the prior five years of data. The August 1990 oil shock shifted the joint distribution of characteristics and returns in ways that were poorly represented in the pre-shock samples. In the notation of Theorem 3.1, the past distributions P_τ are far from the current distribution P_t, increasing $\mathrm{TV}(P_\tau, P_t)$. At the same time, only a small number of post-shock samples were available,

Table 1: Impacts of Model Complexity and Training Window Size on Prediction Error

	Misspecification	Statistical Uncertainty	Non-stationarity
Model complexity ↗	↘	↗	Intrinsic to window
Training window ↗	Intrinsic to model class	↘	↗

The signs ↗ and ↘ indicate the component increases or decreases as the left-column factor increases. "Intrinsic" entries indicate the component is governed by the other dimension and is not directly controlled by the factor shown.

making very short windows statistically noisy. Theorem 3.1 formalizes this dilemma: use more history to reduce estimation variance, or use less history to avoid outdated regimes.

A similar logic applies to the choice of model class. Parsimonious factor models, such as the Fama-French three- or five-factor model (Fama and French, 1993, 2015), may have a larger misspecification error $\min_{f\in\mathcal{F}} \mathcal{E}_t(f)$, but its low complexity means that the estimation variance $r_{t,k}(\mathcal{F})$ can be controlled with relatively few samples. Lower complexity models can therefore be estimated on a shorter window, limiting exposure to outdated regimes. By contrast, a random forest or neural network using a large factor zoo (Gu et al., 2020) may better approximate nonlinear return relationships and has smaller misspecification error $\min_{f\in\mathcal{F}} \mathcal{E}_t(f)$, but its larger complexity term $r_{t,k}(\mathcal{F})$ requires more data to drive down the estimation variance. This forces the model to reach further into the past, increasing the non-stationarity $\max_{t-k\le\tau\le t-1} \mathrm{TV}\,(P_\tau, P_t)$. This is the mechanism behind the "less can be more" phenomenon observed in Section 3.1: complex models can face a larger non-stationarity cost when their variance reduction requires long historical windows.

Put together, neither greater model complexity nor more training data is uniformly beneficial under non-stationarity. Below we illustrate this phenomenon in a stylized setting.

Example 3.3 (Selection of model class and window under non-stationarity)**.** Let $\eta, \gamma \in [0,1]$ be two small constants. Suppose that at each time t, we collect one sample, and the covariate and response $(x, y) \sim P_t$ satisfy $x \sim \mathrm{Uniform}[0,1]$, $y \sim N(f_t^*(x), 1)$ conditioned on x, and

$$f_t^*(x) = c_t x + \gamma \sin(2\pi x),$$

where $\{c_t\}_{t=1}^{\infty}$ is a deterministic sequence in $[0,1]$ satisfying $|c_{t+1} - c_t| = \eta$. We observe a single sample per period. Consider two model classes: linear class and kernel class with

a first-order Sobolev kernel (see, e.g., Example 12.16 in Wainwright (2019)). This kernel class contains smooth nonlinear functions represented through Fourier sine and cosine bases, which can capture the sinusoidal component in f_t^*, while the linear class cannot.

- If we train a linear model with a training window k, then the three components of the prediction error bound in Theorem 3.1 satisfy

$$\min_{f\in\mathcal{F}} \mathcal{E}_t(f) \asymp \gamma^2, \qquad r_{t,k}(\mathcal{F}) \asymp k^{-1}, \qquad \max_{t-k\le\tau\le t-1} \mathrm{TV}\left(P_\tau, P_t\right) \asymp k\eta.$$

  Optimizing their sum over k yields the optimal window size $k^* \asymp \eta^{-1/2}$, which leads to an $O(\gamma^2 + \eta^{1/2})$ bound on the prediction error.

- If we use the kernel class, then f_t^* is well-specified. For a training window k, we have

$$\min_{f\in\mathcal{F}} \mathcal{E}_t(f) = 0, \qquad r_{t,k}(\mathcal{F}) \asymp k^{-2/3}, \qquad \max_{t-k\le\tau\le t-1} \mathrm{TV}\left(P_\tau, P_t\right) \asymp k\eta.$$

  The optimal training window is $k^* \asymp \eta^{-3/5}$, which results in a prediction error of $O(\eta^{2/5})$.

We observe that for both classes in the above example, the optimal window size depends on the severity of the drift η, and is in general not the full window size. If one naïvely uses the kernel class with a large window size, then the resulting error scales as $O(k\eta)$, which is linear in k and can be much worse than the above bounds.

As expected, the preferable model class depends on the interplay between misspecification γ and drift η. The kernel class is more expressive but more sensitive to drift. When $\eta = O(\gamma^5)$, drift is relatively mild and the kernel is optimal, consistent with the "virtue of complexity" (Kelly and Malamud, 2025; Kelly et al., 2022, 2024b). However, when $\eta \gg \gamma^5$, severe non-stationarity requires shorter training windows, under which sample sizes are too limited for the kernel estimator to fully exploit its flexibility advantage. In this high-drift regime, the linear class achieves better performance with its shorter optimal window, explaining the "less can be more" phenomenon observed in our experiments.

Unavoidable Cost of Non-stationarity. The non-stationarity term in Theorem 3.1 is not an artifact of our proof technique: it reflects an intrinsic limitation of learning from non-stationary historical data. Formally, consider any class of non-stationary environments

in which every training distribution is within total variation distance V of the current distribution P_t; one can think of V as the magnitude of the local non-stationarity. Over this class, no estimator, however cleverly designed, can achieve a worst-case prediction error better than the same three components identified in Theorem 3.1: the misspecification error of the model class, the statistical uncertainty from the finite training sample, and a term of order V that is the irreducible cost of training on historical distributions that differ from the current one. Thus, even if the model class is correctly chosen and the estimator is optimal, non-stationarity alone can generate an error of order V, confirming that the upper bound in Theorem 3.1 has the correct dependence on non-stationarity and is not simply a loose consequence of our analysis. We state and prove this lower bound formally as Proposition B.1 in Appendix B.7 of the appendix.

A Refined View of Non-stationarity in High Dimensions. Theorem 3.1 characterizes non-stationarity through a single total variation distance $\max_{t-k\leq\tau\leq t-1} \mathrm{TV}\left(P_\tau, P_t\right)$ between the training and current distributions. This single distance bundles together two economically distinct sources of drift. The first is a shift in how characteristics map into expected returns: the same firm characteristic earns a different expected return in a different regime, i.e., a change in $f^*_\tau(\cdot) = \mathbb{E}_{(\boldsymbol{x},y)\sim P_\tau}[y \mid \boldsymbol{x} = \cdot]$, or more generally in the conditional distribution of y given $\boldsymbol{x}$. In the statistics and machine learning literature, this is called *conditional shift* (or *concept shift*). The second is a shift in the covariate distribution of $\boldsymbol{x}$ itself (Heckman, 1979; Sugiyama and Kawanabe, 2012), for instance, the cross-section of firms or the range of characteristic exposures observed in the market changes, even though the underlying pricing relationship is unchanged. This is known as *covariate shift*. While shifts in the pricing relationship are the more classical concern, recent work has shown that covariate shift alone can substantially degrade performance in modern nonparametric and high-dimensional machine learning models (Koh et al., 2021; Kpotufe and Martinet, 2021; Ma et al., 2023; Patil et al., 2024; Wang, 2026): models may be evaluated on combinations of characteristics that are poorly represented in the training data, leading to large out-of-sample error even when the pricing relationship itself has not changed. In Appendix B.2, we prove a more refined prediction error bound (Theorem B.1) that separately decomposes the non-stationarity term into these two sources.

This refined view highlights the crucial role of non-stationarity in modern ML-based

empirical asset pricing. Classical fixed-dimensional factor models such as the Fama-French three- and five-factor models (Fama and French, 1993, 2015) admit small, fixed risk factor sets. For such low-dimensional linear models, the complexity term $r_{t,k}(\mathcal{F})$ is typically small and decreases quickly with the number of samples. This means using a shorter window to reduce non-stationarity does not come with a large estimation cost. If the covariate space is low-dimensional, moderate changes in factor realizations are less likely to shift into a region with little historical support. Therefore, fixed-dimension models can maintain reasonably stable performance under moderate distribution shifts.

However, the tradeoff becomes acute in the high-dimensional regime where the number of covariates d grows with the sample size. This regime covers neural network and random-feature models with growing effective dimension (Kelly et al., 2024b; Didisheim et al., 2024), LASSO-based methods that use a large cross-section of lagged returns (Chinco et al., 2019), regularized SDF estimators over many characteristic-sorted factors (Kozak et al., 2020; Freyberger et al., 2020), and high-dimensional ML models using dozens or hundreds of firm characteristics (Gu et al., 2020). These models can reduce misspecification error, but their complexity term $r_{t,k}(\mathcal{F})$ is larger and typically requires many more samples to control. They therefore benefit more from long windows in stable periods, but are also more heavily affected by non-stationarity when past data come from different regimes. Covariate shift is also more damaging in high dimensions because the current distribution of characteristics may be poorly represented in historical data. This is the setting in which the nonstationarity-complexity tradeoff is most relevant: the model's flexibility that improves approximation also makes it more reliant on careful and time-dependent choices of training window and hyperparameters.

4 Adaptive Model Selection Algorithm ATOMS

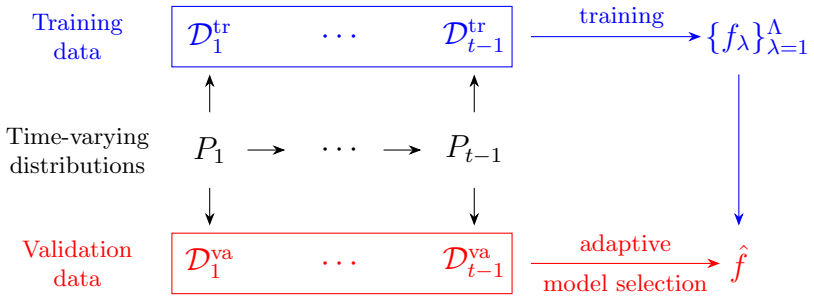
Section 3 shows that the predictive performance of a model depends jointly on its complexity and the size of the training data, and that the optimal choice often varies over time with the non-stationarity. As the non-stationarity is generally unknown *a priori*, the selection of the model class and training window calls for a data-driven approach.

In this section, we develop a novel method that uses historical validation data to select the best model from a set of candidates. These candidate models can come from different model

classes, be trained on different time windows, or use different hyperparameters. The main challenge is that the same non-stationarity that complicates model training also incapacitates standard model selection techniques such as holdout and cross validation. Specifically, in a non-stationary environment, a model that performs well on a validation set from the distant past may not perform as well in the future. Our solution is to adaptively select the relevant validation data that best reflects the current environment, allowing for a more accurate comparison of the candidate models' future performance.

We now formally set up the framework, illustrated in Figure 3. In each period t, we split the available data $\{\mathcal{D}_\tau\}_{\tau=1}^{t-1}$ into a training dataset $\{\mathcal{D}_\tau^{\text{tr}}\}_{\tau=1}^{t-1}$ and a validation dataset $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$. More precisely, each $\mathcal{D}_\tau = \{(\boldsymbol{x}_s, y_s)\}_{s\in B_\tau}$ is stratified into $\mathcal{D}_\tau^{\text{tr}} = \{(\boldsymbol{x}_s, y_s)\}_{s\in B_\tau^{\text{tr}}}$ and $\mathcal{D}_\tau^{\text{va}} = \{(\boldsymbol{x}_s, y_s)\}_{s\in B_\tau^{\text{va}}}$. We use the training data $\{\mathcal{D}_\tau^{\text{tr}}\}_{\tau=1}^{t-1}$ to produce a finite set of candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$. These candidates can come from different model classes, be trained on different data horizons, or use different hyperparameters. We will use the validation data $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$ to select a good model $\widehat{f} = f_{\widehat{\lambda}}$ that achieves near-optimal predictive performance on the current distribution P_t. The selected model $\widehat{f}$ is then tested on the out-of-sample data $\mathcal{D}_t$ at time t.

Figure 3: Our Framework for Model Training and Selection under Non-stationarity.



4.1 Selecting a Model Through Pairwise Comparisons

In this section, we describe our model selection approach, which compares candidate models through a sequential elimination tournament, much like a single-elimination sports bracket. The procedure relies on a pairwise comparison subroutine $\mathcal{A}$, described below, that takes two candidate models f and f' and returns whichever one it judges to have better future

performance, denoted $\mathcal{A}(f, f')$. The tournament proceeds in rounds: in each round, one remaining model is drawn at random as the pivot and is compared, in turn, against every other remaining model. Any model that beats the pivot advances to the next round; if the pivot beats every other remaining model, it is declared the overall winner and the tournament stops. We record the full pseudocode in Appendix C.1 of the appendix, together with the proofs that rely on it.

This tournament structure has two attractive properties. It is computationally cheap:

Lemma 4.1 (Computational complexity)**.** *Algorithm 2 calls the subroutine $\mathcal{A}$ for $\Theta(\Lambda)$ times in expectation.*

Proof of Lemma 4.1. See Appendix C.2. □

It is also statistically efficient: we show later in this section that the tournament preserves the performance guarantee of whichever pairwise comparison subroutine $\mathcal{A}$ it is built on, at the cost of only a logarithmic factor in the number of candidate models Λ.

How Much Validation History to Trust. We now detail the pairwise comparison subroutine $\mathcal{A}$. Comparing two models directly on the full non-stationary validation history $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$ can badly mislead an investor, because validation data from years ago may reflect a market regime that no longer applies. We instead build on the adaptive rolling-window approach of Han et al. (2024), which chooses, separately for each comparison, how far back in the validation history to look.

The idea is as follows. To choose between two models, it suffices to determine the sign of their *performance gap*

$$\Delta_t = L_t(f_1) - L_t(f_2).$$

Indeed, f_2 is better than f_1 if and only if $\Delta_t > 0$. To estimate Δ_t from non-stationary validation data, a natural approach is to fix a look-back window $\ell \in [t-1]$ and average the performance gap over the validation data from the last ℓ periods, $\mathcal{D}_\tau^{\text{va}} = \left\{(\boldsymbol{x}_s, y_s)\right\}_{s \in B_\tau^{\text{va}}}$, $\tau = t-\ell, ..., t-1$, giving a rolling-window estimate

$$\widehat{\Delta}_{t,\ell} = \frac{1}{B_{t,\ell}^{\text{va}}} \sum_{\tau=t-\ell}^{t-1} \sum_{s \in B_\tau^{\text{va}}} u_s, \quad \text{where} \quad u_s = \left[f_1(\boldsymbol{x}_s) - y_s\right]^2 - \left[f_2(\boldsymbol{x}_s) - y_s\right]^2, \tag{4.1}$$

and $B^{\text{va}}_{t,\ell} = \sum_{\tau=t-\ell}^{t-1} B^{\text{va}}_{\tau}$. The accuracy of this estimate hinges on the look-back window ℓ: a longer window pools in more validation observations, but risks pooling in data from a stale market regime. This is exactly the tension identified in Theorem 3.1, now applied to comparing two models' validation performance rather than to fitting a single model. Formally, with probability at least $1-\delta$,

$$\left|\widehat{\Delta}_{t,\ell} - \Delta_t\right| \leq \underbrace{\phi(t,\ell)}_{\text{cost of stale data}} + \underbrace{\psi(t,\ell,\delta)}_{\text{cost of limited data}}. \tag{4.2}$$

Here, the cost of stale data,

$$\phi(t,\ell) = \max_{t-\ell\leq\tau\leq t-1} |\Delta_\tau - \Delta_t|,$$

measures how much the performance gap has moved within the look-back window, and the cost of limited data, $\psi(t,\ell,\delta)$, is a statistical confidence bound that shrinks as more validation observations accumulate:

$$\psi(t,\ell,\delta) = \begin{cases} M', & \text{if } B^{\text{va}}_{t,\ell} = 1 \\ \sigma_{t,\ell}\sqrt{\dfrac{2\log(2/\delta)}{B^{\text{va}}_{t,\ell}}} + \dfrac{2M'\log(2/\delta)}{3B^{\text{va}}_{t,\ell}}, & \text{if } B^{\text{va}}_{t,\ell} \geq 2 \end{cases},$$

where

$$\sigma^2_{t,\ell} = \frac{1}{B^{\text{va}}_{t,\ell}} \sum_{\tau=t-\ell}^{t-1} \sum_{s\in B^{\text{va}}_\tau} \text{var}(u_s) \qquad \text{and} \qquad M' = 8M^2.$$

These are exactly the non-stationarity and statistical uncertainty terms from the bound in Theorem 3.1, now measured for the performance gap between two models rather than for a single fitted model. The ideal look-back window would minimize the sum of the two costs,

$$\ell^* = \operatorname*{argmin}_{\ell\in[t-1]} \left\{\phi(t,\ell) + \psi(t,\ell,\delta)\right\},$$

but both costs depend on the unknown non-stationarity and cannot be computed directly.

To make this choice implementable, we construct data-driven proxies $\widehat{\psi}$ and $\widehat{\phi}$ for these two costs. The proxy for the cost of limited data is built by replacing the unknown variance

$\sigma^2_{t,\ell}$ with the sample variance

$$\widehat{v}^2_{t,\ell} = \frac{1}{B^{\mathrm{va}}_{t,\ell} - 1} \sum_{\tau=t-\ell}^{t-1} \sum_{s \in B^{\mathrm{va}}_{\tau}} \left(u_s - \widehat{\Delta}_{t,\ell}\right)^2, \tag{4.3}$$

which gives

$$\widehat{\psi}(t,\ell,\delta) = \begin{cases} M', & \text{if } B^{\mathrm{va}}_{t,\ell} = 1 \\ \widehat{v}_{t,\ell}\sqrt{\dfrac{2\log(2/\delta)}{B^{\mathrm{va}}_{t,\ell}}} + \dfrac{8M'\log(2/\delta)}{3(B^{\mathrm{va}}_{t,\ell} - 1)}, & \text{if } B^{\mathrm{va}}_{t,\ell} \geq 2 \end{cases}. \tag{4.4}$$

The proxy for the cost of stale data compares the window-ℓ estimate against every shorter window, after netting out each estimate's own statistical noise[6]:

$$\widehat{\phi}(t,\ell,\delta) = \max_{i \in [\ell]} \left(\left|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}\right| - \left[\widehat{\psi}\,(t,\ell,\delta) + \widehat{\psi}\,(t,i,\delta)\,\right] \right)_+ . \tag{4.5}$$

Intuitively, in light of the decomposition (4.2), the quantity

$$\left(\left|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}\right| - \left[\widehat{\psi}\,(t,\ell,\delta) + \widehat{\psi}\,(t,i,\delta)\,\right] \right)_+ \tag{4.6}$$

measures how much the window-ℓ and window-i estimates disagree once their combined statistical noise has been subtracted off: any remaining disagreement must reflect a genuine shift in the performance gap over that stretch of history, rather than sampling noise. The estimated cost of stale data, $\widehat{\phi}(t,\ell,\delta)$, is the largest such disagreement over all shorter windows $i \in [\ell]$.

The procedure then selects a look-back window that minimizes the sum of the two estimated costs,

$$\widehat{\ell} = \operatorname*{argmin}_{\ell \in [t-1]} \left\{ \widehat{\phi}(t,\ell) + \widehat{\psi}(t,\ell,\delta) \right\}, \tag{4.7}$$

and compares the two models using $\widehat{\Delta}_{t,\widehat{\ell}}$: the subroutine selects f_1 if and only if $\widehat{\Delta}_{t,\widehat{\ell}} \leq 0$. We record the full procedure as pseudocode (Algorithm 3) in Appendix C.1 of the appendix.

[6]This construction adapts the Goldenshluger–Lepski method from the nonparametric statistics literature on bandwidth selection (Goldenshluger and Lepski, 2008) to the problem of choosing a look-back window.

Combining this comparison subroutine with the tournament described above yields our model selection procedure, which we call Adaptive Tournament Model Selection, or `ATOMS` for short.

4.2 Performance Guarantees for the Selection Procedure

We now present theoretical guarantees for our model selection framework. Theorem 4.1 establishes a performance bound for the pairwise model comparison subroutine described above.

Theorem 4.1 (Near-optimal model comparison)**.** *Let Assumptions 2.1 and 3.1 hold. Choose $\delta \in (0,1)$ and take $\delta' = \delta/(3t)$ in the comparison subroutine. With probability at least $1-\delta$, the output $\widehat{f}$ satisfies*

$$\mathcal{E}_t(\widehat{f}) \lesssim \min\left\{\mathcal{E}_t(f_1), \mathcal{E}_t(f_2)\right\} + M^2 \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \underbrace{\max_{t-\ell \le \tau \le t-1} \mathrm{TV}(P_\tau, P_t)}_{\text{cost of stale data}} + \underbrace{\frac{1}{B_{t,\ell}^{\mathrm{va}}}}_{\text{cost of limited data}} \right\}. \tag{4.8}$$

Here $\lesssim$ hides a universal constant.

Proof of Theorem 4.1. See Appendix C.3. □

To interpret Theorem 4.1, consider an investor who knows, in hindsight, exactly how much historical data to use to choose a good prediction model. This theoretical result states that our comparison procedure performs nearly as well without that knowledge. It identifies on its own whether the current regime is better described by the last few months or the last several years of validation data. The procedure shortens its look-back window when conditions shift and widens it in stable periods, all without any prior model of how markets evolve.

From a learning theory perspective, Theorem 4.1 gives a finite-sample oracle inequality (4.8). It states that with high probability, the excess risk of $\widehat{f}$ does not exceed that of the better model between f_1 and f_2, up to an additive overhead term that we call the *cost of adaptivity*,

$$\min_{\ell \in [t-1]} \left\{ \max_{t-\ell \le \tau \le t-1} \mathrm{TV}(P_\tau, P_t) + \frac{1}{B_{t,\ell}^{\mathrm{va}}} \right\}.$$

This overhead reflects the same two forces as Theorem 3.1: the cost of stale data, $\max_{t-\ell\le\tau\le t-1}\mathrm{TV}(P_\tau,P_t)$, and the cost of limited data, $1/B^{\mathrm{va}}_{t,\ell}$, associated with the $B^{\mathrm{va}}_{t,\ell}$ validation samples used to make the comparison. The bound takes the minimum over all look-back windows ℓ, meaning that our procedure performs almost as well as an oracle that knows in hindsight which window would lead to the most accurate comparison. This shows that the procedure adaptively chooses a near-optimal look-back window tailored to the local non-stationarity.

Building on this pairwise guarantee, Theorem 4.2 below shows that our model selection algorithm `ATOMS` inherits the same oracle property when selecting from multiple candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$.

Theorem 4.2 (Near-optimal model selection)**.** *Let Assumptions 2.1 and 3.1 hold. Choose $\delta\in(0,1)$ and take $\delta'=\delta/(3\Lambda^2 t)$ in* `ATOMS`*. With probability at least $1-\delta$,* `ATOMS` *outputs a model $\widehat{f}$ satisfying*

$$\mathcal{E}_t(\widehat{f})\lesssim\min_{\lambda\in[\Lambda]}\mathcal{E}_t(f_\lambda)+M^2\log(\Lambda t/\delta)\cdot\min_{\ell\in[t-1]}\Bigg\{\underbrace{\max_{t-\ell\le\tau\le t-1}\mathrm{TV}(P_\tau,P_t)}_{\text{cost of stale data}}+\underbrace{\frac{1}{B^{\mathrm{va}}_{t,\ell}}}_{\text{cost of limited data}}\Bigg\}.$$

Here $\lesssim$ hides a universal constant.

Proof of Theorem 4.2. See Appendix C.4. □

Theorem 4.2 states that with high probability, the performance of the selected model $\widehat{f}$ is at most the performance of the best model in $\{f_\lambda\}_{\lambda=1}^{\Lambda}$, up to the same cost-of-adaptivity term and an extra $O(\log\Lambda)$ factor that represents the price of searching simultaneously over all candidate models rather than committing to one in advance. In other words, `ATOMS` identifies a model whose performance is nearly as good as the best candidate one could have selected in hindsight using the non-stationary validation data. In practice, this resolves the investor's question of whether to use a linear or nonlinear specification this month, and over what history, without any prior knowledge of how market conditions will evolve.

We remark that our model selection framework (Algorithm 2) is general and can be combined with any model comparison subroutine $\mathcal{A}$. In particular, in Appendix C.4, we prove a general reduction lemma (Lemma C.1) that converts any theoretical guarantee of the subroutine $\mathcal{A}$ to a guarantee of Algorithm 2. In Appendix E, we further develop a R^2-

based pairwise comparison subroutine that targets the R^2 metric. When equipped with this R^2-based subroutine, Algorithm 2 enjoys a guarantee with respect to the R^2 metric.

Remark 1 (Benefits of pairwise comparison)**.** As a natural alternative, one may consider estimating the individual performance of the candidate models over a common window, then using the estimated performance for model selection or model aggregation. In a non-stationary environment, estimating the absolute performance of a model can be challenging. In contrast, our procedure makes use of pairwise comparisons and estimates the loss differences between two models. This is statistically advantageous because the two losses are evaluated on the same validation data, and taking their difference may cancel out common fluctuations and non-stationarity. As a consequence, it typically yields a lower-variance estimate of the performance gap than separate estimation of individual model performance. It is also operationally more flexible, since each model pair can be evaluated on its own window rather than forcing all models to share a single window.

Remark 2 (Comparison with prior work)**.** Our model selection framework builds upon the model comparison method of Han et al. (2024). Below we discuss the main differences between our work and theirs. First, their analysis of the model comparison procedure assumes that the distribution of the covariates $\boldsymbol{x}$ remains fixed across time. Our theory removes this assumption entirely, and covers the general non-stationary setting where the joint distribution of $(\boldsymbol{x}, y)$ can change arbitrarily. Second, for model selection, they propose a single-elimination procedure which performs $\Lambda - 1$ model comparisons, but incurs an additional $O(\log \Lambda)$ factor in the performance bound. In contrast, our approach maintains a linear complexity in Λ in expectation while achieving a sharper bound.

4.3 Jointly Choosing Model Class and Training Window

Finally, we apply Theorem 4.2 to the joint selection of model class and training sample size. Let $\mathscr{F}$ be a finite collection of model classes, e.g., $\mathscr{F} = \{\text{linear model}, \text{random forest of a certain size}\}$. For each model class $\mathcal{F} \in \mathscr{F}$, we train models on different windows $k \in [t-1]$ of the training data $\{\mathcal{D}_\tau^{\mathrm{tr}}\}_{\tau=1}^{t-1}$. Let $\widehat{h}(\mathcal{F}, k)$ denote the model from $\mathcal{F}$ trained on $\{\mathcal{D}_\tau^{\mathrm{tr}}\}_{\tau=t-k}^{t-1}$. Then, the set of candidate models is given by

$$\{f_\lambda\}_{\lambda=1}^{\Lambda} = \left\{\widehat{h}(\mathcal{F}, k) : \mathcal{F} \in \mathscr{F},\, k \in [t-1]\right\}. \tag{4.9}$$

Applying Theorem 4.2 to this set of candidate models yields the following guarantee. For simplicity, we assume that training-validation data splitting ratio is fixed across time.

Assumption 4.1 (Balanced training-validation split)**.** There exists $c > 0$ such that $|\mathcal{D}_\tau^{\mathrm{tr}}|/|\mathcal{D}_\tau^{\mathrm{va}}| = c$ for all $\tau \in \mathbb{Z}_+$.

Theorem 4.3 (Near-optimal model-and-data selection)**.** *Let Assumptions 2.1, 3.1 and 4.1 hold. Suppose the set of candidate models is given by (4.9). Choose $\delta \in (0,1)$ and take $\delta' = \delta/(6|\mathscr{F}|^2 t^3)$ in* ATOMS*. Then, with probability at least $1-\delta$, the output $\widehat{f}$ of* ATOMS *satisfies*

$$\mathcal{E}_t(\widehat{f}) \lesssim \min_{\mathcal{F}\in\mathscr{F},\, k\in[t-1]} \left\{ \underbrace{\min_{f\in\mathcal{F}} \mathcal{E}_t(f)}_{\text{misspecification}} + \underbrace{\left(r_{t,k}(\mathcal{F}) + \frac{1}{B_{t,k}^{\mathrm{tr}}} \right)}_{\text{cost of limited data}} + \underbrace{\max_{t-k\le\tau\le t-1} \mathrm{TV}(P_\tau, P_t)}_{\text{cost of stale data}} \right\}, \tag{4.10}$$

where $\lesssim$ hides the constants M and c and logarithmic factors of t, δ^{-1} and $|\mathscr{F}|$.

Proof of Theorem 4.3. See Appendix C.5. □

The bound on the right-hand side of (4.10) is exactly the model performance bound of the model $h(\mathcal{F}, k)$ in Theorem 3.1: the same misspecification, cost of limited data, and cost of stale data that govern any single model's prediction error also govern the performance of the model that ATOMS selects. Thus, our algorithm selects a near-optimal pair of model class and training window size, up to logarithmic factors, without ever being told which of the two costs dominates at a given point in time.

5 Empirical Evidence: Industry Portfolios Across Economic Regimes

In this section, we evaluate the predictive performance of our method on industry portfolios. The results below are the empirical counterpart to the near-optimality guarantee established in Section 4: without advance knowledge of when the economic environment will shift, ATOMS should track the performance of the best fixed-window choice in hindsight, a pattern we now test directly. First, we describe our comprehensive dataset of firm-characteristic managed

portfolios and industry portfolios. Then, we report our findings that our adaptive method ATOMS achieves superior out-of-sample performance in predicting expected returns compared with fixed-window and regime-switching benchmarks across different economic regimes.

5.1 Data

We examine the pricing of 17 industry portfolio returns from Kenneth French's data library, covering the period from September 1987 to November 2016. Our predictor set combines macroeconomic factors, risk premia from characteristic-sorted portfolios, and lagged returns to capture the complex dynamics driving industry returns, sourced for widely cited public datasets.[7] We provide full details of the dataset construction in Appendix A.1, including data preprocessing and long-short portfolio constructions.

Macroeconomic and Systematic Factors. We incorporate 15 factors from Chen et al. (2024), who estimate a SDF using deep learning while imposing no-arbitrage restrictions. These factors include: (i) the estimated SDF representing the aggregate price of risk; (ii) ten beta-sorted decile portfolios based on firms' SDF exposure; and (iii) four macroeconomic hidden states extracted from 178 macro time series via a generative adversarial network. These monthly observations are assigned to all trading days within each month. We also include the daily Fama-French three factors (market, size, and value) from Fama and French (1993) as benchmark risk factors.

Characteristic-Sorted Portfolios. Following Gu et al. (2020), we construct 94 long-short portfolios sorted on firm characteristics that capture price trends, liquidity, size, and risk measures. For each characteristic, we form decile portfolios using all CRSP-listed stocks and create a long-short strategy that buys the top decile and shorts the bottom decile. This

[7] More specifically, our data combines daily and monthly sources to construct a comprehensive time series of covariates combining macroeconomic and cross-sectional signals. The final sample spans the time period from September 1987 to November 2016. We merge daily CRSP excess returns with monthly characteristics from Gu et al. (2020), which provides 94 standardized characteristics for U.S. equities encompassing valuation ratios, profitability measures, investment activity, liquidity, and past return dynamics. These characteristics have become the canonical set of firm-level predictors in modern empirical asset pricing. We construct daily long-short portfolios by sorting firms into deciles based on each characteristic and taking the difference between the top and bottom decile returns, following the methodology of Gu et al. (2020).

approach transforms firm-level characteristics into interpretable factor returns that isolate the pricing implications of each characteristic.

Predictor Set. Our final predictor set comprises: (i) 15 macroeconomic factors from Chen et al. (2024); (ii) 3 Fama-French factors of Fama and French (1993); (iii) 94 characteristic-sorted long-short portfolio returns of Gu et al. (2020) described above; and (iv) the 17 lagged industry returns. This comprehensive set of 129 predictors combines traditional risk factors with modern high-dimensional representations, allowing us to test whether our adaptive method can effectively navigate the complex, time-varying relationships between these predictors and industry returns.

5.2 Estimation Framework and Benchmarks

We evaluate our adaptive method using candidate models from different specifications with varying parameters and training windows. We take one month as a period, where the data $\mathcal{D}_t = \{(\boldsymbol{x}_s, y_s)\}_{s \in B_t}$ in month t consists of daily covariate-return pairs in that month.

Model Specifications. We consider the following specifications that approximate stochastic discount factors. For a vector of covariates $\boldsymbol{x} \in \mathbb{R}^d$, we write $\widetilde{\boldsymbol{x}} = (\boldsymbol{x}^\top, 1)^\top \in \mathbb{R}^{d+1}$.

1. Non-linear specification using random forests (RF). Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and two parameters, namely the number of trees $n_{\texttt{tree}}$ and the maximum tree depth $d_{\max}$, RF estimates a random forest model.

2. Linear specification estimated with ridge regularization (Ridge). Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and regularization parameter $\alpha > 0$, Ridge estimates a linear model $f(\boldsymbol{x}) = \langle \widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{x}} \rangle$ by
$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{x}}_i \rangle - y_i \right)^2 + \alpha \|\boldsymbol{\theta}\|_2^2 \right\}.$$

3. Linear specification with LASSO regularization (LASSO). Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and regularization parameter $\alpha > 0$, LASSO estimates a linear model $f(\boldsymbol{x}) = \langle \widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{x}} \rangle$ by
$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^n \left(\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{x}}_i \rangle - y_i \right)^2 + \alpha \|\boldsymbol{\theta}\|_1 \right\}.$$

4. Linear specification with elastic net regularization (E-Net). Given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and regularization parameters $\alpha > 0$ and $r \in (0,1)$, E-Net estimates a linear model $f(\boldsymbol{x}) = \langle \widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{x}} \rangle$ by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} (\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{x}}_i \rangle - y_i)^2 + \alpha r \|\boldsymbol{\theta}\|_1 + \frac{\alpha}{2}(1-r)\|\boldsymbol{\theta}\|_2^2 \right\}.$$

For each specification, we consider a grid of hyperparameter values, detailed in Appendix A.3. In each month t, we fit candidate models from these specifications on training windows of $4^k \wedge (t-1)$ months, where $0 \le k \le 5$; more details are provided in Appendix F.2. This creates a factorial candidate set formed by crossing model specifications, their hyperparameters, and training window lengths.

Benchmark Approaches. To verify the adaptivity of our framework, we compare it with the following non-adaptive benchmarks that use a fixed window for training and/or validation. We organize the notations in Table 2 for quick reference as well.

1. Fixed-horizon validation window for model selection (`FH-Short`, `FH-Mid`, and `FH-Long`). This is the non-adaptive fixed-window counterpart of `ATOMS`. In each month t, we train the candidate models above, and then use validation data from the last ℓ periods $\{\mathcal{D}_\tau^{\mathrm{va}}\}_{\tau=t-\ell}^{t-1}$ to perform model selection. The detailed description for each of the fixed window length is given in Algorithm 1. We consider three choices of the validation window ℓ, which leads to three benchmarks as `FH-Short`, `FH-Mid`, and `FH-Long`, where `FH` stands for Fixed-Horizon and the suffix indicates the validation window length. `FH-Short` uses $\ell = 32$ months of validation data (approximately 3 years), capturing only the recent regime. `FH-Mid` uses 128 months (approximately one full business cycle), and `FH-Long` uses all available history. The contrast across these three captures how much history a fixed-horizon practitioner would choose to use. `ATOMS` is designed to adapt to the optimal choice of validation window size ℓ in hindsight at each time.

2. Fixed-horizon cross-validation (`FH-CV`). In each month t, we use data from the last 36 months $\{\mathcal{D}_\tau\}_{\tau=t-36}^{t-1}$ to perform 5-fold cross-validation to estimate and select a model out of the candidate models with the same sets of parameters.

Table 2: Summary of methods

Benchmark	Notation	ℓ (months)	Approx. duration	Economic characterization
Short-horizon	`FH-Short`	32	$\approx$ 3 yrs	Captures only the most recent regime; low non-stationarity cost, high estimation variance.
Medium-horizon	`FH-Mid`	128	$\approx$ 1 business cycle	Spans a full expansion-recession cycle.
Long-horizon	`FH-Long`	≤ 512	All available history	Uses the full in-sample record; implicitly assumes stationarity across all historical regimes.
Cross-validation	`FH-CV`	36 (CV)	$\approx$ 3 yrs	Selects models via 5-fold cross-validation over the most recent 36 months.
Regime-switching	`Regime-switch`	N/A	N/A	Fits a 2-state Markov-switching linear model (Hamilton, 1989); explicitly models non-stationarity but restricts each regime to a fixed linear specification.
Adaptive	`ATOMS`	Data-driven	Time-varying	Selects window ℓ to balance non-stationarity and variance, without prior knowledge of how market conditions evolve.

This table summarizes the six methods considered in Section 5. `FH` stands for Fixed Horizon; the suffix indicates the validation window length. The three `FH` benchmarks are the fixed-window counterparts of `ATOMS` (Algorithm 1) with window sizes $\ell = 32$, 128, and up to 512 months. `FH-CV` uses 5-fold cross-validation over the most recent 36 months. `Regime-switch` is a 2-state Markov regime-switching linear model re-estimated monthly using all available history. `ATOMS` (Algorithm 2) jointly selects the model class and validation window at each month t without prior knowledge of the non-stationarity structure.

3. Regime-switching model (`Regime-switch`). We also fit a regime-switching model (Hamilton, 1989), which explicitly models the non-stationary dynamics. We consider a 2-state model, each day s is associated with an underlying state $z_s \in \{1, 2\}$, which evolves according to a time-homogeneous Markov chain. Conditioned on $z_s = z$, the covariates $\boldsymbol{x}_s$ and response y_s follow a linear model $y_s = \alpha_z + \boldsymbol{\beta}_z^\top \boldsymbol{x}_s + \varepsilon_s$, where $\varepsilon_s \sim N(0, \sigma_z^2)$. We re-estimate the model at a monthly frequency, using all available historical data.

For all the methods considered in Table 2, we use the same 129-dimensional covariates set of macroeconomic factors, Fama-French factors, characteristic-sorted long-short portfolio returns and lagged industry returns, described in Section 5.1.

As the methods `ATOMS`, `FH-Short`, `FH-Mid`, `FH-Long` and `FH-CV` involve training-validation data splitting, we will evaluate each of them using 20 independent repetitions of a stratified 80%-20% estimation-validation split. Specifically, fixing a random seed, we construct stratified samples in which, for each month t, 80% of trading days are assigned to the training set and the remaining 20% to the validation set. The validation observations are then pooled and ordered chronologically, and the method is applied to this validation set to select a model. Using the selected model, we compute a realized performance metric (e.g., out-of-sample R^2). We repeat this procedure 20 times and report the average performance metric across repetitions. This evaluation design disciplines model selection and reduces the scope for data mining, while respecting the sequential arrival of information faced by investors and preserving the time-series structure of the data, thereby avoiding look-ahead bias. More details can be found in Appendix A.3.

Algorithm 1 Fixed Validation Window for Model Selection

Input: Candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$, validation data $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$, validation window size ℓ. Select

$$\widehat{\lambda} = \underset{\lambda\in[\Lambda]}{\operatorname{argmin}} \sum_{\tau=(t-\ell)\vee 1}^{t-1} \sum_{s\in B_\tau^{\text{va}}} \left[f_\lambda(\boldsymbol{x}_s) - y_s\right]^2 .$$

return $\widehat{f} = f_{\widehat{\lambda}}$.

Performance Metrics. We measure the performance of each approach using the out-of-sample R^2 metric (2.2) that benchmarks against a zero forecast. We compute both the overall out-of-sample R^2 from January 1990 to November 2016, and the annual out-of-sample R^2. The latter provides a more granular understanding of the approaches' performance over time. In Appendix G.2, we also report results for the standard R^2 metric (2.3).

5.3 Return Predictability Across Economic Regimes

We next turn to the empirical analysis of pricing the 17 industry portfolios. The fundamental premise of our adaptive framework is that asset pricing relationships exhibit time-varying dynamics rather than remaining stationary across economic conditions. This non-stationarity is particularly pronounced during economic recessions, when structural breaks in risk premia,

Table 3: Out-of-Sample R^2 Averages Across Industries by Time Period

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	0.049	0.027	0.125	0.041
`FH-Short`	0.022	0.009	0.096	−0.001
`FH-Mid`	0.041	−0.031	0.123	0.035
`FH-Long`	0.043	−0.031	0.117	0.039
`FH-CV`	0.035	−0.009	0.071	0.014
`Regime-switch`	−0.026	−0.441	0.081	−0.008

This table reports out-of-sample (OOS) R^2 averages for return prediction models across all 17 industry portfolios. Full OOS Period refers to our largest available OOS period covering 01/1990∼11/2016. Columns report OOS R^2 averages across all industries and highlight this metric during three recessions, as documented in NBER Business Cycle Dating:

- the 1990 Gulf War recession (06/1990∼10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001∼10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007∼06/2009).

That is, the OOS performance in Gulf War column focuses on model performance comparisons exclusively in the out-of-sample period of 06/1990∼10/1990. All values are calculated using monthly return data. The `FH` benchmarked methods use fixed windows as described in Table 2.

sudden shifts in investor risk aversion, and disruptions to market liquidity mechanisms create environments where long-term historical pricing relationships provide poor guidance for future returns. Recessions therefore serve as a natural laboratory for testing the adaptivity of our framework: if our approach can successfully navigate these turbulent periods when non-stationarity is most severe, it provides compelling evidence for the value of adaptive model selection in asset pricing more generally.

Performance During Recessions. Our adaptive method `ATOMS` improves return predictability across the full 1990–2016 out-of-sample period, with significant gains during the three NBER-dated recessions. Table 3 presents out-of-sample R^2 values across distinct economic regimes. `ATOMS` achieves an out-of-sample R^2 of 0.049 across the full sample period, representing a 14.0% improvement over the best fixed-window benchmark `FH-Long` which has $R^2 = 0.043$. The regime-switching model performs poorly in comparison, achieving a negative R^2 of -0.026.

The 2001 recession consists of two distinct shocks. The dot-com collapse repriced technology-sector earnings, shifting growth-versus-value factor loadings rapidly and persistently. The September 11 attacks then introduced a macro shock with no historical analog. Data in the 1990s reflected an era of technology optimism and subdued risk aversion; neither property held after March 2000. Our method's validation window contracted to focus on recent data, recognizing that the bull-market decade was uninformative for the new regime. During the 2001 recession, `ATOMS` achieves an impressive R^2 of 0.125, outperforming the long window `FH-Long` by 6.8% (0.117) and substantially exceeding the fixed window `FH-CV`'s 0.071 and regime-switching `Regime-switch`'s 0.081.

The 1990 Gulf War recession is the starkest finding of the benefits of adaptivity. Iraq's invasion of Kuwait in August 1990 triggered an oil supply shock with no precedent in the post-1987 sample observations. Jones and Kaul (1996) show that oil shocks transmit to U.S. equity returns through real cash flow effects. Data during 1987–1990 reflected post-crash recovery dynamics and had no exposure to oil-shock co-movements. In terms of Theorem 3.1, $\mathrm{TV}(P_\tau, P_t)$ spiked sharply for times τ before August 1990, and our method's short validation window appropriately discounted that stale history. While `ATOMS` maintains a positive R^2 of 0.027 during this sharp but brief contraction, the fixed-window benchmark `FH-Long` produces a negative R^2 of -0.031, and the regime switching model yields a negative R^2 of -0.441, indicating worse performance than a simple forecast of zero.[8]

The Global Financial Crisis showed a different non-stationarity structure. Systemic credit risk drove cross-industry return correlations toward 1, eliminating the return dispersion that prediction models depend on. For instance, liquidity risk reached historically extreme levels as market stresses occur and empirically cause the liquidity risk premium to drive significant commonality in stock returns (Pástor and Stambaugh, 2003). Leverage and funding costs also shifted to values outside the training distribution (Adrian and Shin, 2014). We found that during the Global Financial Crisis, the adaptive method `ATOMS` achieves an R^2 of 0.041, marginally outperforming the long window `FH-Long` (0.039) and substantially ex-

[8]One possible reason for the significant underperformance of the regime-switching model during the Gulf War is the limited initial training samples. This is a common challenge in latent-state models, which require the unobserved latent states, transition probabilities, and state-dependent parameters to be inferred jointly from the data. With a large 129-dimensional covariates set, our in-sample period, which starts from September 1987, is too short to produce stable and reliable parameter estimates for the out-of-sample period beginning in June 1990.

ceeding cross-validated `FH-CV` (0.014) and regime switching `Regime-switch` (−0.008). The smaller performance gap here reflects the recession's 20-month duration. By the time `ATOMS` is evaluated in late 2008 and 2009, even fixed-window methods have accumulated enough crisis-period observations to reflect the new economic conditions. The long time horizon of the financial crisis out-of-sample period diminishes the relative advantage of our method, as opposed to other shorter period shocks. We consider it as not a failure of our method's adaptivity but rather a natural consequence of a long-lived regime. Nevertheless, the adaptive `ATOMS` maintains its competitiveness throughout this long period, exhibiting robustness across different types of economic contractions.

Why Adaptivity Matters During Downturns. The superior recession performance of `ATOMS` has important implications for asset pricing theory and practice. Traditional asset pricing models assume stationary risk-return relationships, an assumption that becomes particularly problematic during economic downturns when risk aversion typically increases and market liquidity conditions deteriorate. Our adaptive framework explicitly recognizes this non-stationarity by allowing the validation window to expand or contract based on recent predictive performance.

The empirical evidence suggests that during recessions, the optimal window for model selection varies significantly, reflecting the rapid evolution of risk premia. The 1990 Gulf War recession provides the clearest example: its sudden onset and brief duration created an environment where only very recent validation data could accurately capture the new pricing dynamics. Conversely, during more prolonged downturns like the Global Financial Crisis, the optimal window likely expanded gradually as new market conditions became established.

From a theoretical perspective, these findings support the view that stochastic discount factors exhibit time-varying dynamics that are particularly pronounced during economic stress. The adaptive framework's ability to track these dynamics more effectively than fixed-window approaches suggests that the non-stationarity of asset pricing relationships is not merely a statistical artifact but reflects fundamental economic mechanisms that vary with the business cycle.

Why Regime-Switching Benchmark Appears Worse. The relatively weak performance of the regime-switching model is multi-fold. Firstly, we investigate whether our fitted

Markovian two-state regression captures economically meaningful states. In Appendix G.1, we show that the estimated latent states are significantly correlated with the CBOE Volatility Index (VIX), suggesting that the model does capture economically meaningful variations in market conditions that increases as the "fear index" moves upwards. We believe its performance is not simply an artifact of using only two latent states, nor does it reflect a weakened implementation.[9] The limitation is instead the restricted form of conditional linearity: after conditioning on the latent state, the mapping from covariates to return is a fixed linear specification. This restriction arises largely from the need to maintain a tractable likelihood when the transition dynamics of the latent states are modeled explicitly. Our method `ATOMS` avoids this rigid restriction by adaptively selecting over various model classes and training horizons, which allows it to flexibly handle both time-varying economic regimes and model specifications.

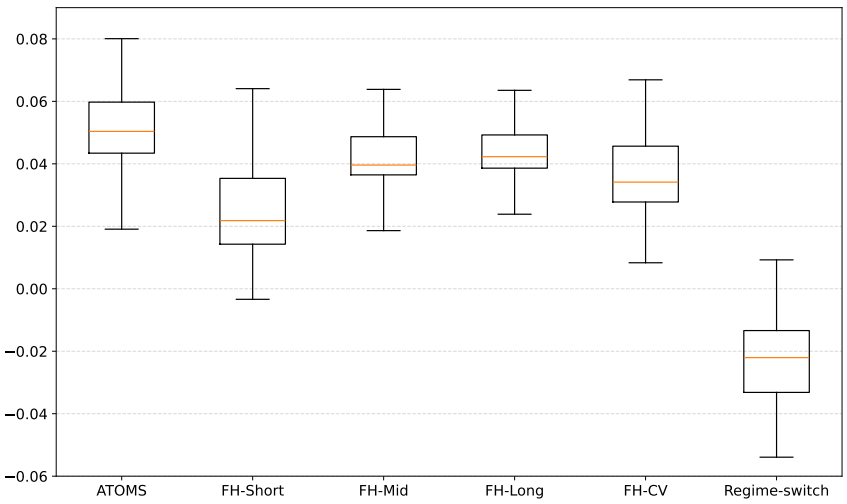
Robustness Across Industries. We conduct industry-by-industry robustness check to confirm that the recession outperformance of `ATOMS` is not driven by a subset of industries but represents a broad-based phenomenon. First, Figure 4 shows a box plot of performance across 17 industries. Each box represents one method and is constructed from 17 industry-level R^2 across all years in the OOS horizon. We observe that `ATOMS` overall has better R^2 across all years, as its median, and level position of the box is higher than those of the other methods.

In more details, Figure 5 plots the annual out-of-sample R^2 of `ATOMS` and the benchmarks in each industry. We observe that `ATOMS` maintains its advantage across diverse sectors including cyclical industries (Durbl, Cars, Trans) and defensive sectors (Food, Utils, Cnsum). This cross-sectional consistency further suggests that the observed performance reflects genuine adaptivity to changing market conditions rather than industry-specific anomalies.

Notably, `ATOMS` shows particular strength in industries most sensitive to business cycle fluctuations, such as durable goods (Durbl), consumer discretionary (Rtail), and financial services (Finan). This pattern aligns with economic intuition, as these sectors experience the most dramatic shifts in risk premia during economic transitions. Our method's ability to capture these dynamics more effectively than benchmarks suggests it successfully identifies

[9] To the best of our knowledge, we use the most established publicly available implementation for Markov regime-switching regression, with details discussed in Appendix G.1.

Figure 4: Box Plot of Out-of-Sample R^2 of `ATOMS` and Benchmarks for 17 Industry Portfolios.

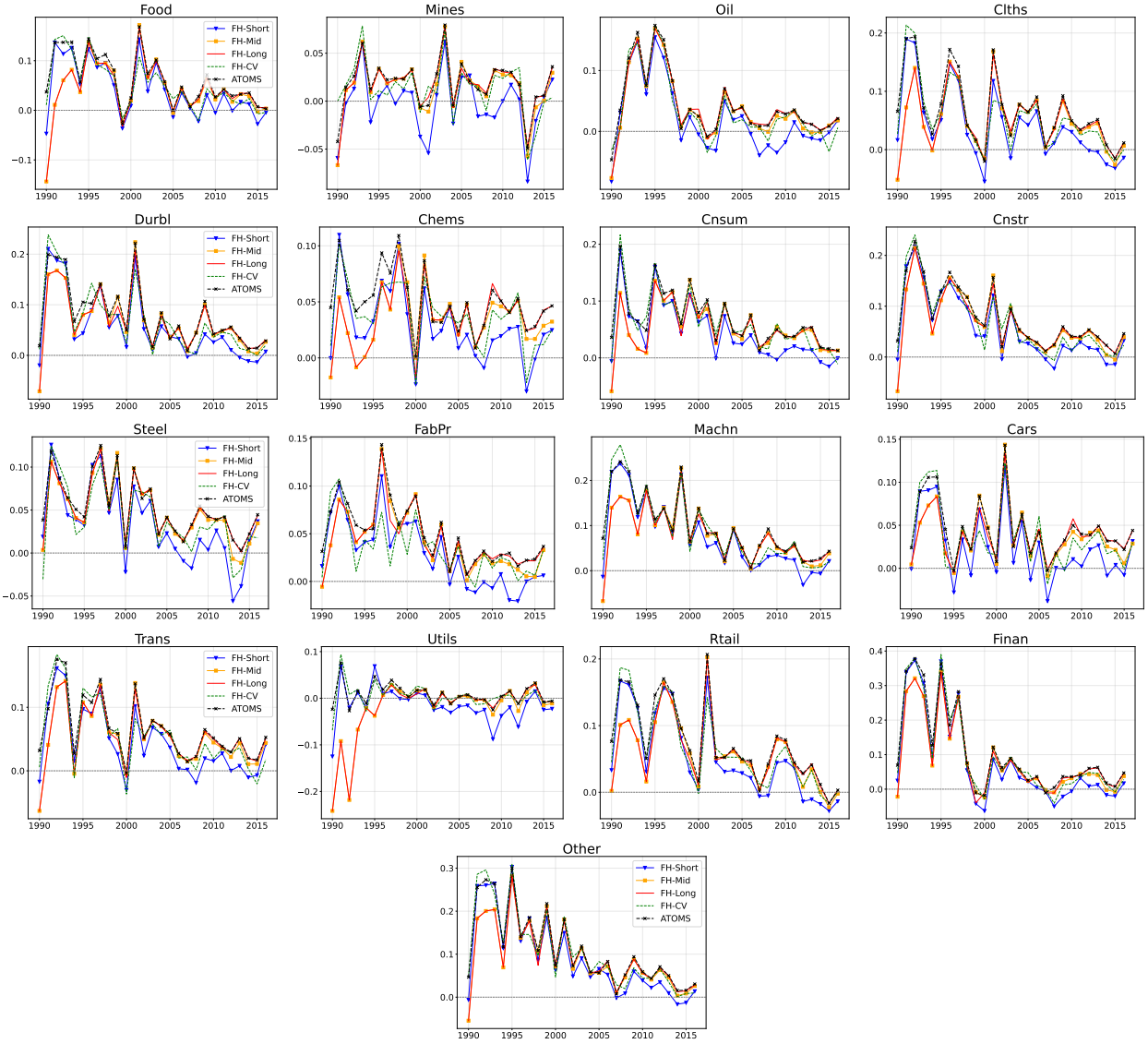


This figure describes the distribution of each method's OOS R^2. Each box represents one method and is constructed from 17 industry-level R^2 across all years in the OOS horizon. The `FH` benchmarked methods use fixed windows as described in Table 2.

the changing risk-return tradeoffs that characterize different phases of the business cycle.

Adaptation versus Best-Window Identification. Perhaps counter-intuitively, our experiments show that short validation windows may not be optimal during recession periods. In non-stationary environments, the bias-variance tradeoff involved in choosing the window size can be subtle. For example, crisis data is often highly volatile, so a longer window may stabilize model comparison. A longer window may also include data from previous crisis periods whose distribution is closer to the current crisis than samples immediately preceding it. Consequently, several window choices may achieve statistically and economically similar performance, and there need not be a universal ordering of window sizes in terms of performance, even during recessions. This is reflected in the fixed-horizon benchmarks in Table 3: among `FH-Short`, `FH-Mid`, and `FH-Long`, the best-performing fixed window varies across recession periods, with `FH-Short` performing best during the Gulf War recession, `FH-Mid` during the 2001 recession, and `FH-Long` during the 2008 Financial Crisis. Therefore, it is generally not feasible to adapt to the non-stationarity through best-window identification. We do not aim to infer a uniquely optimal validation window at each point in time, especially when several window choices may yield similar performance. Instead, `ATOMS` is designed to

Figure 5: Annual Out-of-Sample R^2 of `ATOMS` and Benchmarks for 17 Industry Portfolios.



This figure reports the annual out-of-sample R^2 of our adaptive model selection method `ATOMS` (black dashed line with $\times$'s), as well as the fixed-window baselines `FH-Short` (blue ▼'s), `FH-Mid` (orange ■'s), and `FH-Long` (red). The `FH` benchmarked methods use fixed windows as described in Table 2. The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 10.

select a window size that gives predictive performance close to the best fixed window in hindsight.

5.4 Portfolio Performance and Investment Gains

We now evaluate whether the conditional return forecasts generated by `ATOMS` translate into economic value in an implementable allocation problem. We construct two portfolio rules over the 17 industry portfolios and compare risk-adjusted performance, turnover, and cumulative wealth for `ATOMS` and the benchmarks. All portfolio results are strictly out-of-sample, with the trading period running from January 1990 (month $t = 1$) through November 2016.

Portfolio Construction and Timing. At the beginning of each month t, each method estimates $N = 17$ conditional mean functions $\widehat{f}_t^{(n)}$, $n = 1, \ldots, N$, one for each industry portfolio return, using only information available up to $t-1$. On each trading day $s \in B_t$ of month t, given predictors $\boldsymbol{x}_s$, the vector of conditional return forecasts is

$$\widehat{\boldsymbol{f}}_t(\boldsymbol{x}_s) = \big(\widehat{f}_t^{(1)}(\boldsymbol{x}_s), ..., \widehat{f}_t^{(N)}(\boldsymbol{x}_s)\big)^\top.$$

These forecasts are mapped into daily cross-sectional portfolio weights $\boldsymbol{w}_s = \big(w_s^{(1)}, ..., w_s^{(N)}\big)^\top$. We consider two allocation rules.

1. Mean-variance efficient portfolio: We form the mean-variance efficient direction

$$\widetilde{\boldsymbol{w}}_s^{\mathrm{MV}} = \widehat{\boldsymbol{\Sigma}}_s^{-1} \widehat{\boldsymbol{f}}_t(\boldsymbol{x}_s),$$

   where $\widehat{\boldsymbol{\Sigma}}_s$ is the sample covariance matrix of daily industry returns estimated with a one-year rolling window. Portfolio weights are then normalized so that their absolute values sum to 1:

$$\boldsymbol{w}_s^{\mathrm{MV}} = \frac{\widetilde{\boldsymbol{w}}_s^{\mathrm{MV}}}{\big\|\widetilde{\boldsymbol{w}}_s^{\mathrm{MV}}\big\|_1}.$$

2. Signed portfolio: We assign equal absolute weight $1/N = 1/17$ to each industry, and take long positions for industries with positive predicted returns and short positions for industries with negative predicted returns:

$$\boldsymbol{w}_s^{\mathrm{Signed}} = \frac{1}{N} \cdot \mathrm{sign}\big(\widehat{\boldsymbol{f}}_t(\boldsymbol{x}_s)\big),$$

where sign$(\cdot)$ is applied componentwise. We also compare against the equally weighted portfolio (`EW`) which longs every industry with weight $1/N$: $\boldsymbol{w}_s^{\mathrm{EW}} = (\frac{1}{N}, \ldots, \frac{1}{N})^\top$.

Performance and Implementation Metrics. For portfolio weights $\boldsymbol{w}_s$, the gross daily portfolio return is $r_s^{\mathtt{g}} = \boldsymbol{w}_s^\top \boldsymbol{y}_s$. Daily turnover is measured by the ℓ_1 rebalancing norm, $\mathrm{TO}_s = \|\boldsymbol{w}_s - \boldsymbol{w}_{s-1}\|_1$, where $s-1$ denotes the previous trading day. The net return is $r_s = r_s^{\mathtt{g}} - c \cdot \mathrm{TO}_s$, where $c = 4 \times 10^{-4}/252$[10] is the trading cost, and $c \cdot \mathrm{TO}_s$ takes into account implementation frictions.

We aggregate daily results to the monthly frequency. For each month t, we compute the monthly Sharpe ratio

$$\mathrm{SR}_t = \frac{\widehat{\mathbb{E}}_t[r]}{\sqrt{\widehat{\mathrm{var}}_t(r)}},$$

where the sample mean $\widehat{\mathbb{E}}_t[r]$ and sample variance $\widehat{\mathrm{var}}_t(r)$ are computed from daily net returns within month t:

$$\widehat{\mathbb{E}}_t[r] = \frac{1}{B_t} \sum_{s \in B_t} r_s \quad \text{and} \quad \widehat{\mathrm{var}}_t(r) = \frac{1}{B_t - 1} \sum_{s \in B_t} \Big(r_s - \widehat{\mathbb{E}}_t[r] \Big)^2.$$

We also assess the cumulative wealth across time. Starting with initial wealth $W_0 = 1$, the cumulative wealth evolves according to $W_t = W_{t-1} \prod_{s \in B_t} (1 + r_s)$ and the terminal wealth W_T at time T is then $W_T = \prod_{t=1}^{T} \prod_{s \in B_t} (1 + r_s)$.

To capture risk characteristics, we note that path-dependent downside risk is better captured by maximum drawdown (MDD) than by average return or Sharpe ratio alone. We measure the maximum drawdown of the wealth curve as the largest peak-to-trough decline over the investment horizon:

$$\mathrm{MDD} = \max_{1 \le t \le T} \left(1 - \frac{W_t}{\max\limits_{1 \le \tau \le t} W_\tau} \right).$$

For MDD, a large value indicates that the strategy's cumulative value fell sharply from its running historical peak at some point along the path—a loss event that is invisible to statistics that average across time.

[10] We consider the annualized trading cost of 4bps that are on the same scale reported by Harvey et al. (2025), and treat each year as 252 trading days.

To isolate stress-period behavior, we define the *conditional drawdown* over a subperiod $[t_0, t_1]$ with a lookback reference window of length S:

$$\text{MDD}\big|_{t_0,\,t_1} = \max_{t_0 \le t \le t_1} \left(1 - \frac{W_t}{\max\limits_{t_0 - S \le \tau \le t} W_\tau}\right).$$

This metric restricts attention to the worst trough within $[t_0, t_1]$, measured relative to the running peak computed over the immediately preceding S-length window. For example, setting $S = 3$ months and $[t_0, t_1]$ equal to the 2001 recession, $\text{MDD}|_{t_0,t_1}$ captures the deepest portfolio decline from any peak attained since the 3 months before the contraction began, through the end of the recession.

Smaller drawdowns have direct economic consequences as they reduce the probability of investor capital withdrawals, margin calls, and forced deleveraging after adverse return streaks. We therefore treat drawdown as a complementary implementability criterion because among strategies with similar risk-adjusted returns, the one with the shallower trough will remain viable in real time. Drawdown also functions as a consistency check on the cumulative wealth results where a strategy whose wealth gains concentrate in a handful of extreme observations will show sharp troughs during recessions. We would thus reveal fragility in any lack of risk diversification with the $\text{MDD}|_{t_0,t_1}$ metric.

Results: Mean-Variance Portfolio. The mean-variance portfolio provides an empirical proxy for the conditional SDF. By the Hansen and Jagannathan (1991) variance bound, for any valid pricing kernel m, $\text{std}(m)/\mathbb{E}(m) \geq \text{SR}_{\max}/(1 + R_f)$, where $\text{SR}_{\max}$ is the maximum Sharpe ratio in the traded asset span, and R_f is the risk-free rate. A higher Sharpe ratio tightens this bound, requiring the pricing kernel to vary more to remain consistent with the data.[11] The gains below therefore carry an asset-pricing interpretation beyond investor utility.

Table 4 shows a clear full-sample ordering: `ATOMS` delivers the highest out-of-sample mean-variance Sharpe ratio, 0.334, relative to 0.304 for `FH-Long`, 0.300 for `FH-CV`, 0.296 for `FH-Mid`, 0.241 for `FH-Short`, and 0.232 for `Regime-switch`. These differences correspond to gains of 9.87%, 11.33%, 38.59%, and 44.0% respectively, indicating economically meaningful

[11] The bound applies unconditionally. We interpret improvements in the realized conditional Sharpe ratio as directional evidence of better conditional SDF approximation.

improvements in excess return per unit of risk rather than marginal statistical variation.

Table 4: Average Monthly Sharpe Ratio of Mean-Variance Portfolio over Different Periods

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	0.333	0.187	0.568	0.243
`FH-Short`	0.239	0.148	0.305	0.098
`FH-Mid`	0.295	0.145	0.471	0.211
`FH-Long`	0.303	0.145	0.559	0.220
`FH-CV`	0.299	0.168	0.371	0.053
`Regime-switch`	0.231	0.113	0.399	0.138

This table reports monthly Sharpe ratio averages for mean-variance portfolios formed from conditional return forecasts for the 17 industry portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. The `FH` benchmarked methods use fixed windows as described in Table 2. Columns report Sharpe ratio averages during three recessions, as documented in NBER Business Cycle Dating:

- the 1990 Gulf War recession (06/1990∼10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001∼10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007∼06/2009).

The recession subsamples provide a sharper test because they are the episodes in which conditional means and covariance structure move most abruptly. In the Gulf War recession, `ATOMS` attains 0.188 versus 0.169 for `FH-CV`, 0.146 for `FH-Long`, and 0.115 for `Regime-switch` (11.24%, 28.77% and 63.48% improvements). In the 2001 recession, `ATOMS` reaches 0.569, close to `FH-Long` at 0.560 (1.61%), but far above `Regime-switch` at 0.400 (42.25%) and `FH-CV` at 0.371 (53.37%), consistent with substantial value from adaptive window choice under rapid repricing. In the Financial Crisis, `ATOMS` records 0.244 versus 0.221 for `FH-Long`, 0.140 for `Regime-switch`, and 0.054 for `FH-CV` (10.41%, 47.29% and 351.85% improvements), suggesting that adaptive updating remains effective even in a prolonged downturn.

For asset-pricing interpretation, these results matter beyond investor utility. Within the span of traded industry returns, the mean-variance efficient portfolio is tied to the mean-variance efficient pricing kernel. Accordingly, superior out-of-sample Sharpe ratios for this portfolio imply a tighter empirical approximation to the conditional SDF, or equivalently, smaller pricing-kernel misspecification in the traded cross section. Taken together, the evidence supports a state-dependent environment in which adaptive estimation improves both

implementable portfolio efficiency and SDF approximation.

These Sharpe ratio improvements are not generated by mechanically higher trading intensity. Table 5 shows that ATOMS has lower turnover than FH-Short, FH-Mid, FH-Long and Regime-switch in the full sample (11.94 versus 16.39, 13.87, 12.45 and 20.40), and comparable turnover in recession windows. Relative to FH-CV, ATOMS trades somewhat more but delivers materially higher Sharpe ratios, implying a more favorable return-cost trade-off.

Table 5: Average Monthly Turnover of Mean-Variance Portfolio over Different Periods

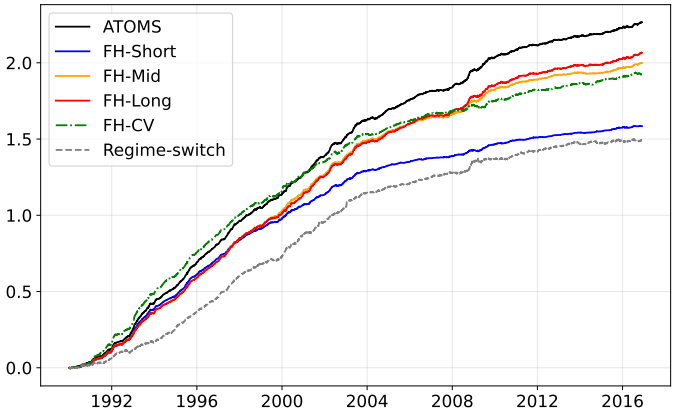
Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	11.94	16.85	13.71	13.62
FH-Short	16.39	17.43	16.91	17.86
FH-Mid	13.87	18.63	15.07	15.12
FH-Long	12.45	18.63	14.06	13.34
FH-CV	11.31	13.07	10.92	10.06
Regime-switch	20.40	22.83	19.06	20.62

This table reports monthly turnover averages for mean-variance portfolios formed from conditional return forecasts. Turnover is computed as the average number of portfolio weight changes per month across all 17 industry portfolios, i.e. the ℓ_1 norm of the difference in portfolio weights between consecutive months. The FH benchmarked methods use fixed windows as described in Table 2.

Figure 6 plots log cumulative wealth. ATOMS dominates the fixed-window alternatives and the regime-switching model over most of the out-of-sample period and separates notably from FH-CV after 2001. The wealth paths confirm that the Sharpe-ratio gains are economically meaningful after implementation costs.

In Table 6, we report the maximum drawdown of the mean-variance portfolio; for the recession periods, we compute the conditional maximum drawdown with a three-month look-back window. ATOMS has a full-sample drawdown of 3.0%, below FH-Short (4.3%) and FH-CV (4.4%), and close to FH-Long (3.1%). During recessions, ATOMS also maintains a low downside risk: it has the smallest drawdown during the Gulf War recession (0.7%), ties the best drawdown during the 2001 recession (0.6%), and matches FH-Mid during the Financial Crisis (2.4%). Although Regime-switch has a slightly smaller full-sample drawdown, it also has materially lower Sharpe ratios and much higher turnover. These results indicate that the mean-variance portfolio based on ATOMS improves risk-adjusted performance without

Figure 6: Cumulative Wealth of Mean-Variance Portfolio



This figure plots cumulative wealth (in logarithmic scale) for the mean-variance portfolio formed from conditional return forecasts under `ATOMS` and the benchmark methods. Starting with initial wealth $W_0 = 1$, the wealth in each month t evolves according to $W_t = W_{t-1} \prod_{s \in B_t} (1 + r_s^{\mathrm{g}} - c \cdot \mathrm{TO}_s)$, where B_t collects all days in month t, r_s^{g} is the return of the mean-variance portfolio on day s, $c = 4 \times 10^{-4}/252$ is the trading cost, and TO_s is the daily turnover.

generating larger peak-to-trough losses. In Appendix G.3, we report additional drawdown results using alternative look-back windows of $S = 1, 2, 6$ months.

Results: Signed Portfolio. The signed portfolio isolates the directional information in the predicted returns: it holds each industry with the same absolute weight, but uses the model forecast to determine whether the position is long or short. Table 7 shows that `ATOMS` again achieves the highest full-sample Sharpe ratio, 0.359, outperforming all the benchmark methods. We observe that the best fixed-horizon benchmark among `FH-Short`, `FH-Mid` and `FH-Long` varies across recessions: `FH-Short` performs the best during the Gulf War recession, while `FH-Mid` performs the best during the 2001 recession and marginally leads during the Financial Crisis. `ATOMS` keeps up with the best fixed-horizon benchmark in hindsight, and outperforms all three fixed-horizon benchmarks in the Gulf War and 2001 recessions while remaining close to the best fixed-horizon benchmark during the Financial Crisis.

The comparison with the equally weighted benchmark `EW` is more pronounced. `EW` attains a full-sample Sharpe ratio of 0.239, so `ATOMS` improves the Sharpe ratio by 50.21%. The same pattern holds in recession subsamples: `ATOMS` delivers Sharpe ratios of 0.397, 0.430, and 0.215 during the Gulf War recession, the 2001 recession, and the Financial Crisis, compared with

Table 6: Maximum Drawdown of Mean-Variance Portfolio

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	3.0%	0.7%	0.6%	2.4%
FH-Short	4.3%	1.0%	0.9%	3.5%
FH-Mid	3.3%	1.0%	0.7%	2.4%
FH-Long	3.1%	1.0%	0.6%	2.7%
FH-CV	4.4%	0.9%	0.9%	4.2%
Regime-switch	2.7%	1.2%	0.8%	2.7%

This table reports maximum drawdowns for mean-variance portfolios formed from conditional return forecasts of ATOMS and benchmark methods. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a three-month look-back reference window ($S = 3$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

-0.123, 0.097, and 0.071 for EW. This shows that even a simple equal-weight trading strategy benefits substantially from conditioning positions on the signs of the predicted returns. The additional gains of ATOMS over the other benchmarks show that ATOMS further improves the directional information in the predicted returns.

In terms of turnover, Table 8 shows that ATOMS has lower turnover than FH-Short, FH-Mid, FH-Long, and Regime-switch in the full sample (3.22 versus 5.46, 4.09, 3.53, and 10.62), though slightly more than FH-CV (3.01). The long-only equally weighted benchmark EW has zero turnover by construction because it maintains constant equal weights. Relative to this zero-turnover benchmark, the prediction-based methods, including ATOMS and the fixed-horizon benchmarks, introduce only modest trading: ATOMS changes only 3.22 units of ℓ_1 weight per month on average, while attaining a substantially higher Sharpe ratio. During recession periods, ATOMS generally remains below or comparable to the fixed-horizon and regime-switching benchmarks, while delivering stronger risk-adjusted performance.

Figure 7 reports the log cumulative wealth for the signed portfolio. ATOMS again dominates all benchmarks for most of the sample, and substantially outperforms FH-CV and Regime-switch after 2000, indicating persistent economic gains from adaptive portfolio formation. Taken together, the mean-variance and equally weighted evidence supports a conditional asset-pricing view in which expected returns vary with the economic state and are

Table 7: Average Monthly Sharpe Ratio of Signed Portfolio over Different Periods

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	0.359	0.397	0.430	0.215
FH-Short	0.354	0.329	0.408	0.202
FH-Mid	0.352	0.293	0.425	0.223
FH-Long	0.351	0.293	0.414	0.216
FH-CV	0.352	0.283	0.426	0.172
Regime-switch	0.311	0.187	0.384	0.203
EW	0.239	−0.123	0.097	0.071

This table reports monthly Sharpe ratio averages for signed portfolios formed from conditional return forecasts for the 17 industry portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. The FH benchmarked methods use fixed windows as described in Table 2. Columns report Sharpe ratio averages during three recessions, as documented in NBER Business Cycle Dating:

- the 1990 Gulf War recession (06/1990∼10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001∼10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007∼06/2009).

better captured by adaptive training windows.

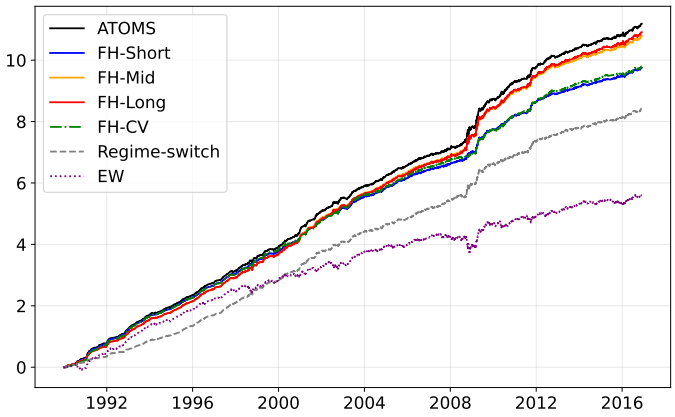
In Table 9, we report maximum drawdowns for the signed portfolio using the same three-month look-back window as the mean-variance portfolio. The prediction-based strategies have much smaller drawdowns than the long-only equal-weight benchmark: ATOMS has a full-sample drawdown of 16.3%, compared with 48.8% for EW. The difference is especially pronounced in recessions, where EW has drawdowns of 17.8%, 14.2%, and 46.9% during the Gulf War recession, the 2001 recession, and the Financial Crisis, respectively, while ATOMS has drawdowns of only 1.9%, 2.4%, and 16.3%. Although FH-Short and Regime-switch have smaller full-sample drawdowns than ATOMS, they also have lower Sharpe ratios. Overall, the adaptive prediction-based strategy delivers stronger risk-adjusted performance while avoiding the large downside exposure of the long-only equal-weight benchmark. In Appendix G.3, we report additional drawdown results using alternative look-back windows of $S = 1, 2, 6$ months.

Table 8: Average Monthly Turnover of Signed Portfolio over Different Periods

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	3.22	5.15	5.30	3.68
`FH-Short`	5.46	6.56	6.61	7.49
`FH-Mid`	4.09	8.16	5.66	4.31
`FH-Long`	3.53	8.16	5.16	3.58
`FH-CV`	3.01	4.40	2.48	3.90
`Regime-switch`	10.62	14.38	10.94	11.06
`EW`	0	0	0	0

This table reports monthly turnover averages for signed portfolios formed from conditional return forecasts. Turnover is computed as the average number of portfolio weight changes per month across all 17 industry portfolios, i.e. the ℓ_1-norm of the difference in portfolio weights between consecutive months. The `FH` benchmarked methods use fixed windows as described in Table 2.

Figure 7: Cumulative Wealth of Signed Portfolio



This figure plots cumulative wealth (in logarithmic scale) for the equally weighted portfolio formed from conditional return forecasts under `ATOMS` and the benchmark approaches. Starting with initial wealth $W_0 = 1$, the wealth in each month t evolves according to $W_t = W_{t-1} \prod_{s \in B_t} (1 + r_s^{\mathrm{g}} - c \cdot \mathrm{TO}_s)$, where B_t collects all days in month t, r_s^{g} is the return of the mean-variance portfolio on day s, $c = 4 \times 10^{-4}/252$ is the trading cost, and TO_s is the daily turnover.

6 Conclusions

Our empirical results demonstrate the practical value of our adaptive framework across multiple dimensions. Most notably, during periods of heightened economic stress, our adaptive

Table 9: Maximum Drawdown of Signed Portfolio

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	16.3%	1.9%	2.4%	16.3%
FH-Short	13.6%	2.1%	2.1%	12.8%
FH-Mid	15.2%	2.7%	2.4%	15.1%
FH-Long	16.5%	2.7%	2.3%	16.5%
FH-CV	18.4%	2.4%	1.9%	18.0%
Regime-switch	13.0%	1.2%	1.6%	13.0%
EW	48.8%	17.8%	14.2%	46.9%

This table reports maximum drawdowns for signed portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a three-month look-back reference window ($S = 3$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

method ATOMS exhibits superior performance compared to fixed-window approaches. Our approach is well motivated by the documented facts in Section 3.1 that, during recession periods including the 1990 Gulf War recession, the 2001 dot-com bubble burst and 9/11 attack, and the 2007–2009 Financial Crisis, simpler models trained on shorter windows consistently outperformed more complex models trained on longer windows. This empirical evidence validates our theory of the nonstationarity-complexity tradeoff.

The adaptive algorithm's performance during economic downturns is worth highlighting. As shown in Table 3, ATOMS achieves an out-of-sample R^2 of 0.027 during the brief but severe 1990 Gulf War recession, while the best fixed-window benchmark FH-Long produces a negative R^2 of -0.031. During the 2001 recession, ATOMS attains an impressive R^2 of 0.125, outperforming FH-Long by 6.8% (0.117). Even during the prolonged Global Financial Crisis of 2007-2009, ATOMS maintains its advantage with an R^2 of 0.041 compared to FH-Long's 0.039. Beyond statistical performance metrics, the economic significance of our approach is demonstrated through trading strategy analysis. Averaged across the industries, our model generates 31% higher cumulative wealth net of trading costs over the full out-of-sample period, as shown in Figure 6.

Several future directions are worth exploring. First, our adaptive model selection framework relies on the assumption that data is independent across time even though the distri-

bution can change arbitrarily. While numerical experiments show that our method is robust against temporal dependence in real-world financial time series, it would be interesting and important to extend the framework to handle dependent data in a principled way. Second, our framework of joint model and training window selection requires training a large number of candidate models, which can be computationally intensive. A valuable future direction is to reduce these training costs. For example, one might utilize the trained models from previous periods as warm starts for subsequent training.

References

Adrian, T. and Shin, H. S. (2014). Procyclical leverage and value-at-risk. *The Review of Financial Studies* **27** 373–403.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61** 821–856.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78.

Banerjee, A. and Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics* **129** 1–34. Modelling structural breaks.

Bartlett, P. L., Bousquet, O. and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics* **33** 1497 – 1537.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Chen, L., Pelger, M. and Zhu, J. (2024). Deep learning in asset pricing. *Management Science* **70** 714–750.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics* **86** 221–241.

Chinco, A., Clark-Joseph, A. D. and Ye, M. (2019). Sparse signals in the cross-section of returns. *The Journal of Finance* **74** 449–492.

Choi, D., Jiang, W. and Zhang, C. (2025). Alpha go everywhere: Machine learning and international stock returns. *The Review of Asset Pricing Studies* **15** 288–331.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2022). *Introduction to algorithms*. MIT press.

Didisheim, A., Ke, S. B., Kelly, B. T. and Malamud, S. (2024). APT or "AIPT"? The surprising dominance of large factor models. Tech. rep., National Bureau of Economic Research.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33** 3–56.

Fama, E. F. and French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance* **51** 55–84.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics* **116** 1–22.

Freyberger, J., Neuhierl, A. and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies* **33** 2326–2377.

Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150 – 1190.

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies* **33** 2223–2273.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57** 357–384.

Han, E., Huang, C. and Wang, K. (2024). Model assessment and selection under temporal distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*. PMLR.

Hansen, L. P. and Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy* **99** 225–262.

Harvey, C. R., Mazzoleni, M. G. and Melone, A. (2025). The unintended consequences of rebalancing. Tech. rep., National Bureau of Economic Research.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Homm, U. and Breitung, J. (2012). Testing for speculative bubbles in stock markets: A comparison of alternative methods. *Journal of Financial Econometrics* **10** 198–231.

Inoue, A., Jin, L. and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics* **196** 55–67.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems* **31**.

Jones, C. M. and Kaul, G. (1996). Oil and the stock markets. *The Journal of Finance* **51** 463–491.

Kelly, B., Kuznetsov, B., Malamud, S., Xu, T. A. and Zhang, Y. (2024a). Large and Deep Factor Models.

Kelly, B., Malamud, S. and Zhou, K. (2024b). The virtue of complexity in return prediction. *The Journal of Finance* **79** 459–503.

Kelly, B. and Xiu, D. (2023). Financial machine learning. *Foundations and Trends® in Finance* **13** 205–363.

Kelly, B. T. and Malamud, S. (2025). Understanding the virtue of complexity. Tech. Rep. 25-96, Swiss Finance Institute.

Kelly, B. T., Malamud, S. and Zhou, K. (2022). The virtue of complexity everywhere. Tech. Rep. 22-57, Swiss Finance Institute.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C. and Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*. PMLR.

Kozak, S., Nagel, S. and Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics* **135** 271–292.

Kpotufe, S. and Martinet, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics* **49** 3299 – 3323.

Liu, C. and Maheu, J. M. (2007). Are there structural breaks in realized volatility? *Journal of Financial Econometrics* **6** 326–360.

Ma, C., Pathak, R. and Wainwright, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics* **51** 738 – 761.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability* **28** 863 – 884.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance* **71** 5–32.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory*. Springer, Berlin, Heidelberg.

Pástor, Ľ. and Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy* **111** 642–685.

Patil, P., Du, J.-H. and Tibshirani, R. (2024). Optimal ridge regularization for out-of-distribution prediction. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*. PMLR.

Pesaran, M. H. and Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics* **29** 307–318.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137** 134–161.

Sugiyama, M. and Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. 1st ed. Springer Publishing Company, Incorporated.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Wang, K. (2026). Pseudo-Labeling for kernel ridge regression under covariate shift. *The Annals of Statistics* **54** 252 – 276.

Appendix

A Additional Experiment Details for Section 5

A.1 Data

We use the 17 industry portfolio returns from Kenneth French's website[12]. We first describe our rich set of covariates, which combines daily and monthly sources to construct a comprehensive time series combining macroeconomic and cross-sectional signals. The final daily-frequency sample spans the time period from September 1987 to November 2016.

Common Factors. To capture the underlying macroeconomic and systematic risk structure in our response variables, we include 15 common factors from the monthly dataset of Chen et al. (2024), who develop a deep learning framework for estimating the stochastic discount factor (SDF) consistent with the no-arbitrage condition.[13]

The Chen et al. (2024) dataset provides three distinct elements relevant to our analysis: (i) the estimated stochastic discount factor (SDF), capturing the aggregate price of risk implied by their no-arbitrage model; (ii) the returns of ten equal-weighted decile portfolios sorted by firms' exposure to the SDF (beta-sorted), which serve as cross-sectional test assets; and (iii) four macroeconomic hidden state variables, derived from the GAN model, that summarize the joint dynamics of 178 macroeconomic time series into a small set of nonlinear factors reflecting business cycle conditions and systemic risk.[14]

In addition to these 15 factors, we incorporate the daily Fama-French three factors (FF3) from Fama and French (1993), which include the excess market return (MKT), the size factor

[12] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_17_ind_port.html

[13] Their model integrates three neural network components: a feedforward network to approximate the nonlinear functional form of the SDF, a recurrent Long-Short Term Memory (LSTM) network to extract macroeconomic state variables, and a generative adversarial network (GAN) that constructs the most informative test assets for pricing. This architecture yields an empirically estimated SDF that represents the conditional mean-variance efficient portfolio of all U.S. equities.

[14] These components jointly span the main drivers of systematic variation in expected returns. The SDF and beta-sorted deciles capture the cross-sectional risk-return trade-offs, while the macroeconomic hidden states track time variation in the pricing kernel associated with expansions, recessions, and crisis periods. The dataset covers the period 1967-2016; to align frequencies with our daily response data, each monthly observation is assigned to all trading days within the corresponding month.

(SMB), and the value factor (HML). These factors provide a benchmark set of linear risk exposures that have been shown to explain broad cross-sectional patterns in stock returns. Each factor is lagged appropriately to avoid look-ahead bias.

This unified set of common factors allows us to incorporate both the traditional linear risk structure captured by FF3 and the nonlinear, macroeconomically conditioned dynamics extracted from Chen et al. (2024).

Firm-Level Returns and Characteristics. We augment the dataset with daily firm-level stock returns obtained from the CRSP database. Excess returns are computed relative to the daily risk-free rate from Kenneth French's data library.[15] To characterize the cross-sectional heterogeneity in firm fundamentals and trading behavior, we incorporate the comprehensive panel of firm-level characteristics from Gu et al. (2020). Their dataset provides 94 standardized monthly characteristics for U.S. equities, constructed to ensure comparability across firms and over time.

Specifically, Gu et al. (2020) identify three dominant groups of predictive signals: (i) price trend variables: measures of short- and long-term momentum, industry momentum, and short-term reversal that capture the persistence of returns and investor underreaction; (ii) liquidity and size variables: market capitalization, trading volume, turnover, and bid-ask spread, which reflect trading frictions and the limits to arbitrage; (iii) volatility and risk variables: total and idiosyncratic volatility, market beta, and higher-order terms such as beta squared, which proxy for systematic and residual risk exposure.[16]

A.1.1 Size Filter

We apply a size-based filter to focus the analysis on the subset of firms that are most representative of the small-cap universe. Specifically, within each day t, we retain only the

[15] To mitigate the impact of outliers that can distort cross-sectional averages, we Winsorize the daily cross-section of excess returns by removing the top and bottom 1% of firms within each day based on their excess return values. This procedure ensures that the resulting portfolio and factor constructions are not unduly influenced by extreme realizations.

[16] As the Gu et al. (2020) characteristics are available only at a monthly frequency, we assign each monthly vector of characteristics to all trading days within that month for the corresponding firm. This alignment ensures consistency between the daily CRSP return data and the lower-frequency fundamental information.

bottom 25% of firms by market equity.[17]

Formally, let $Z_{i,t}$ denote the size of firm i in month t. For all trading days in month t, we retain only those firms satisfying $Z_{i,t} \leq Q_{0.25}\big(\{Z_{j,t}\}_{j\in\mathcal{I}_t}\big)$, where $Q_{0.25}(\cdot)$ denotes the 25th percentile of the cross-sectional size distribution, and $\mathcal{I}_t$ is the set of all firms available in month t. Restricting the sample in this way allows us to study the predictive role of firm characteristics and macroeconomic factors within a homogeneous segment of the equity market, mitigating heterogeneity arising from scale effects in firm size.

A.1.2 Long-Short Decile Portfolios from Characteristics

To transform the firm-level characteristics into tradable, interpretable covariates, we construct daily long-short portfolios sorted on each of the Gu et al. (2020) characteristics. For each of the 94 characteristics, denoted by $c \in [94]$, and for each day t, we sort firms into ten equal-weighted portfolios based on the value of characteristic c.

Formally, let $X_{i,c,t}$ denote the value of characteristic c for firm i in month t, which is assigned to all trading days within that month, and let $R^{\text{ex}}_{i,s}$ denote the daily excess return of firm i on day s. We compute monthly breakpoints for characteristic c using cross-sectional quantiles $b_{c,k,t} = Q_{k/10}\big(\{X_{i,c,t}\}_{i\in\mathcal{I}_s}\big)$, $k \in [10]$, where $Q_p(\cdot)$ denotes the p-th quantile, and $\mathcal{I}_s$ is the set of all firms observed on day s.

We then compute the daily return of the k-th decile portfolio for characteristic c as:

$$D_{c,k,s} = \frac{1}{N_{c,k,s}} \sum_{i:\, b_{c,k-1,t} < X_{i,c,t} \leq b_{c,k,t}} R^{\text{ex}}_{i,s},$$

where $N_{c,k,s}$ is the number of firms assigned to the k-th decile on day s. Each decile is equal-weighted to ensure that portfolio performance reflects cross-sectional variation in firm characteristics rather than differences in market capitalization.

The corresponding long-short portfolio return for characteristic c is defined as $\text{LS}_{c,s} = D_{c,10,s} - D_{c,1,s}$, representing the daily return to a strategy that longs the highest-decile firms (with the largest $X_{i,c,t}$) and shorts the lowest-decile firms (with the smallest $X_{i,c,t}$).

This construction yields a balanced set of long-short factor returns that isolate the pricing implications of each firm characteristic. Repeating this procedure for all $c \in [94]$ produces a

[17] We measure market equity with the (monthly) "mvel1" size covariate from Gu et al. (2020). It represents the firm's market capitalization at the end of month m, standardized for comparability across firms and time.

time series matrix of 94 characteristic-sorted long-short portfolio returns at daily frequency.

A.1.3 Final Covariate Set

Our final set of predictors integrates macroeconomic, factor-based, and cross-sectional sources of variation to form a unified time series of covariates for forecasting the French 17 industry portfolio returns. The resulting dataset combines information from three complementary dimensions of the asset pricing literature:

1. **Macroeconomic and systematic factors:** the 10 equal-weighted beta-sorted decile portfolios, the 4 macroeconomic hidden states, and the estimated stochastic discount factor (SDF) from Chen et al. (2024).

2. **Benchmark risk factors:** the daily Fama-French 3 factors (FF3) from Fama and French (1993), consisting of the market excess return (MKT), the size factor (SMB), and the value factor (HML).

3. **Cross-sectional characteristic factors:** the 94 daily long-short characteristic-sorted portfolio returns, from the Gu et al. (2020) characteristics.

4. **Lagging features:** we augment the feature set with one-day lagged returns. Specifically, we use the full vector of the 17 industry portfolio lagged returns.

By combining the first three components, our dataset links the macroeconomic and cross-sectional perspectives on asset pricing. The Chen et al. (2024) factors embed the nonlinear, time-varying structure of the stochastic discount factor, while the Gu et al. (2020) characteristic-sorted long-short portfolios summarize the cross-sectional distribution of risk premia across firms. Incorporating the Fama-French 3 factors provides a benchmark for evaluating whether these modern, high-dimensional representations offer predictive power beyond the traditional linear framework.

Table 10: Name Mapping for the 17 Industries

Industry Acronym	Full Industry Name
Food	Food
Mines	Mining and Minerals
Oil	Oil and Petroleum Products
Clths	Textiles, Apparel & Footwear
Durbl	Consumer Durables
Chems	Chemicals
Cnsum	Drugs, Soap, Perfumes, Tobacco
Cnstr	Construction and Construction Materials
Steel	Steel Works Etc
FabPr	Fabricated Products
Machn	Machinery and Business Equipment
Cars	Automobiles
Trans	Transportation
Utils	Utilities
Rtail	Retail Stores
Finan	Banks, Insurance Companies, and Other Financials
Other	Other

A.2 Name Mapping for the 17 Industries

A.3 Machine Learning Model Implementation Details

A.3.1 Model Hyperparameter Grids

For each model class considered in our analysis, we systematically explore a comprehensive grid of hyperparameter values for model selection. The hyperparameter grids are designed to balance computational efficiency with thorough exploration of the model space.

Linear Models with Regularization. For the ridge regression (Ridge), LASSO (LASSO), and elastic net (E-Net) models, we consider the following hyperparameter specifications:

1. For ridge regression, we consider values of the regularization parameter α on a logarithmic scale: $\alpha \in \{10^{-3}, 10^{-1.5}, 1, 10^{1.5}, 10^{3}\}$. This range allows for both strong regularization (small α) and weak regularization (large α), accommodating different levels of multicollinearity in our high-dimensional covariate space.

2. For LASSO, we consider values of the regularization parameter α on a logarithmic scale: $\alpha \in \{10^{-5}, 10^{-3.5}, 10^{-2}, 10^{-0.5}, 10\}$. The ℓ_1 penalty in LASSO facilitates feature selection, which is particularly valuable given our large set of covariates.

3. For the elastic net, we consider the following combinations of the regularization parameter α and the mixing parameter r: $\alpha \in \{10^{-3}, 1, 10^{3}\}$ and $r \in \{0.01, 0.05, 0.1\}$. This grid explores the balance between feature selection and coefficient shrinkage.

Random Forest. For the random forest models, we consider the following combinations of the number of trees $n_{\texttt{tree}}$ and the maximum tree depth $d_{\max}$: $n_{\texttt{tree}} \in \{10, 100, 200\}$ and $d_{\max} \in \{3, 5, 10\}$. Increasing the number of trees generally improves model stability and reduces variance. Trees with shallower depths provide stronger regularization, while deeper trees can capture more complex nonlinear relationships.

A.3.2 Hyperparameters for ATOMS

In our implementation of ATOMS, we set $\delta' = 0.1$ and $M' = 10^{-3}$. Appendix F.3 provides further details on the choice of M' and reports ablation studies showing that the performance of ATOMS remains stable when M' is doubled or halved.

Internet Appendix

B Theoretical Analysis of the Nonstationarity-Complexity Tradeoff

In this section, we provide more details on Theorem 3.1 and its proof.

We first set up some mathematical notation. Recall that $f_t^*(\cdot) = \mathbb{E}_{(\boldsymbol{x},y)\sim P_t}[y \mid \boldsymbol{x} = \cdot]$ is the Bayes optimal least squares estimator, which minimizes $L_t(f)$ over all measurable $f : \mathcal{X} \to \mathbb{R}$. Let

$$\bar{f}_t = \operatorname*{argmin}_{f\in\mathcal{F}} L_t(f),$$

which minimizes $L_t(f)$ over all $f \in \mathcal{F}$. For each t, let $P_{t,\boldsymbol{x}}$ be the marginal distribution of P_t with respect to the covariates $\boldsymbol{x}$. The distribution $P_{t,\boldsymbol{x}}$ induces an inner product $\langle\cdot,\cdot\rangle_t$ and norm $\|\cdot\|_t$, given by

$$\langle f, g\rangle_t = \mathbb{E}_{\boldsymbol{x}\sim P_{t,\boldsymbol{x}}}[f(\boldsymbol{x})g(\boldsymbol{x})] \qquad \text{and} \qquad \|f\|_t = \sqrt{\langle f, f\rangle_t} = \sqrt{\mathbb{E}_{\boldsymbol{x}\sim P_{t,\boldsymbol{x}}}[f(\boldsymbol{x})^2]}.$$

It can be easily shown that $\mathcal{E}_t(f) = \|f - f_t^*\|_t^2$ for all $f : \mathcal{X} \to \mathbb{R}$.

B.1 A Non-Stationary Local Rademacher Complexity

We now formally define $r_{t,k}(\mathcal{F})$ through a non-stationary version of the *local Rademacher complexity* (Bartlett et al., 2005). We first define the *Rademacher complexity*, which reflects the richness of a function class with respect to certain data samples.

Definition B.1 (Rademacher complexity)**.** Let $\{\boldsymbol{z}_i\}_{i=1}^n$ be independent random variables in $\mathcal{X}$. Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The **Rademacher complexity** of $\mathcal{G}$ associated with $\{\boldsymbol{z}_i\}_{i=1}^n$ is defined by

$$\mathfrak{R}(\mathcal{G}; \{\boldsymbol{z}_i\}_{i=1}^n) = \mathbb{E}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(\boldsymbol{z}_i)\right],$$

where $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. random variables following the Rademacher distribution $\mathbb{P}(\varepsilon_1 =$

$1) = \mathbb{P}(\varepsilon_1 = -1) = 1/2$, and are independent of $\{z_i\}_{i=1}^n$.

The local Rademacher complexity is the Rademacher complexity of some local function class centered at $\bar{f}_t$. Let $B_j^{\mathrm{tr}} = |B_j^{\mathrm{tr}}|$.

Definition B.2 (Local function class)**.** For every $r \geq 0$, define the local function class

$$\mathcal{F}_{t,k}(r) = \left\{ f \in \mathcal{F} : \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \left\| f - \bar{f}_t \right\|_j^2 \leq r \right\}.$$

In words, $\mathcal{F}_{t,k}(r)$ consists of functions $f \in \mathcal{F}$ that are close to $\bar{f}_t$ with respect to the distributions $\{P_{j,\boldsymbol{x}}\}_{j=t-k}^{t-1}$ on average. We are now ready to define the quantity $r_{t,k}(\mathcal{F})$, which is more formally known as the *critical radius* of the function

$$\mathfrak{R}_{t,k}(r; \mathcal{F}) = \mathfrak{R}\left(\mathcal{F}_{t,k}(r); \mathcal{D}_{t,k}^{\mathrm{tr}}\right).$$

Definition B.3 (Subroot function)**.** A function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is **subroot** if it is increasing and $r \mapsto \psi(r)/\sqrt{r}$ is decreasing on $(0, \infty)$.

Definition B.4 (Critical radius)**.** The **critical radius** of $\mathfrak{R}_{t,k}(r; \mathcal{F})$ is defined by

$$r_{t,k}(\mathcal{F}) = \inf \left\{ r \geq 0 : \exists \,\text{subroot } \psi \text{ such that } \psi(r) = r, \text{ and } \psi(s) \geq \mathfrak{R}_{t,k}(s; \mathcal{F}) \ \forall s \geq r \right\}.$$

B.2 Proof of Theorem 3.1

We now prove Theorem 3.1. We will prove a more refined result that decomposes the non-stationarity into conditional shift and covariate shift, where

- the conditional shift between a past time period τ and the current time t is measured by $\|f_\tau^* - f_t^*\|_t$;
- the covariate shift between time τ and t is characterized by the following integral probability metric:

$$D_t\left(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}\right) = \sup_{h \in (\partial \mathcal{F}_t)^2} \left| \mathbb{E}_{\boldsymbol{x} \sim P_{\tau,\boldsymbol{x}}}[h(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x} \sim P_{t,\boldsymbol{x}}}[h(\boldsymbol{x})] \right|, \tag{B.1}$$

with $(\partial \mathcal{F}_t)^2 = \{(f - f_\tau^*)^2 : f \in \mathcal{F}, \tau \in [t]\}$.

We now state the more refined prediction error bound.

Theorem B.1 (Refined prediction error bound)**.** *Let Assumptions 2.1 and 3.1 hold, and fix $\delta \in (0,1)$. With probability at least $1-\delta$, the model $\widehat{f}$ defined by (3.1) satisfies*

$$\mathcal{E}_t(\widehat{f}) \lesssim \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + M^2\left(r_{t,k}(\mathcal{F}) + \frac{\log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right) + \max_{t-k\le\tau\le t-1}\left\{\|f_\tau^* - f_t^*\|_t^2 + D_t\left(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}\right)\right\}.$$

Here $\lesssim$ hides a universal constant, and $D_t\left(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}\right)$ is defined by (B.1).

From Theorem B.1, we can prove Theorem 3.1 by showing that both the conditional shift and covariate shift terms are $O\left(M^2\cdot \mathrm{TV}(P_\tau, P_t)\right)$. To bound the conditional shift term, for each τ,

$$\begin{aligned}
\|f_\tau^* - f_t^*\|_t^2 &= L_t(f_\tau^*) - L_t(f_t^*) \\
&= \left[L_t(f_\tau^*) - L_\tau(f_\tau^*)\right] + \left[L_\tau(f_\tau^*) - L_\tau(f_t^*)\right] + \left[L_\tau(f_t^*) - L_t(f_t^*)\right] \\
&\le \left[L_t(f_\tau^*) - L_\tau(f_\tau^*)\right] + 0 + \left[L_\tau(f_t^*) - L_t(f_t^*)\right].
\end{aligned}$$

For all $f : \mathcal{X} \to [-M, M]$, we have

$$|L_t(f) - L_\tau(f)| = \left|\mathbb{E}_{(\boldsymbol{x},y)\sim P_t}\left[(f(\boldsymbol{x})-y)^2\right] - \mathbb{E}_{(\boldsymbol{x},y)\sim P_\tau}\left[(f(\boldsymbol{x})-y)^2\right]\right| \le 4M^2\cdot \mathrm{TV}(P_t, P_\tau).$$

Thus,

$$\|f_\tau^* - f_t^*\|_t^2 \le 8M^2\cdot \mathrm{TV}(P_t, P_\tau). \tag{B.2}$$

To bound the covariate shift term, since every $h \in (\partial\mathcal{F}_t)^2$ satisfies $|h(\boldsymbol{x})| \le 4M^2$ for all $\boldsymbol{x}\in\mathcal{X}$, then

$$\begin{aligned}
D_t\left(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}\right) = \sup_{h\in(\partial\mathcal{F}_t)^2}\left|\mathbb{E}_{\boldsymbol{x}\sim P_{\tau,\boldsymbol{x}}}[h(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim P_{t,\boldsymbol{x}}}[h(\boldsymbol{x})]\right| &\le 4M^2\cdot \mathrm{TV}(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}) \\
&\le 4M^2\cdot \mathrm{TV}(P_\tau, P_t). \qquad \text{(B.3)}
\end{aligned}$$

Combining (B.2) and (B.3), we obtain

$$\max_{t-k\le\tau\le t-1}\left\{\|f_\tau^* - f_t^*\|_t^2 + D_t\left(P_{\tau,\boldsymbol{x}}, P_{t,\boldsymbol{x}}\right)\right\} \le 12M^2\cdot \max_{t-k\le\tau\le t-1}\mathrm{TV}(P_\tau, P_t).$$

In the next section, we prove Theorem B.1.

B.3 Proof of Theorem B.1

In the proof, we will write $\widehat{f} = \widehat{f}_{t,k}$ to emphasize its dependence on the time t and the training window size k. The key of the proof is the following lemma, which is a non-stationary version of Theorem 3.3 in Bartlett et al. (2005).

Lemma B.1 (Localized uniform concentration)**.** *Let $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$ be independent random variables taking values in a space $\mathcal{Z} \subseteq \mathbb{R}^{d+1}$. Let $\mathcal{G}$ be a collection of functions from $\mathcal{Z}$ to $[a, b]$. Suppose that there exist $T : \mathcal{G} \to \mathbb{R}_+$ and $C, \eta_1, ..., \eta_n \geq 0$ such that the following noise condition holds:*

$$\frac{1}{n}\sum_{i=1}^{n} \operatorname{var}\left[g(\boldsymbol{z}_i)\right] \leq T(g) \leq C \cdot \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}\left[g(\boldsymbol{z}_i)\right] + \eta_i\right), \qquad \forall g \in \mathcal{G}. \tag{B.4}$$

Let ψ be a sub-root function with a fixed point r^* satisfying

$$\psi(r) \geq C \cdot \mathfrak{R}\big(\{g \in \mathcal{G} : T(g) \leq r\}; \{\boldsymbol{z}_i\}_{i=1}^n\big), \qquad \forall r \geq r^*.$$

Take $\delta \in (0, 1)$. With probability at least $1 - \delta$, for all $g \in \mathcal{G}$ and $K > 1$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[g(\boldsymbol{z}_i)\right] \\
&\qquad \leq \frac{K}{K-1} \cdot \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{z}_i) + c\left[\frac{K}{C} r^* + ((b-a) + CK)\frac{\log(1/\delta)}{n}\right] + \frac{1}{K-1} \cdot \frac{1}{n}\sum_{i=1}^{n} \eta_i.
\end{aligned}$$

Here $c > 0$ is a universal constant.

Proof of Lemma B.1. See Appendix B.4. □

For every $f \in \mathcal{F}$, define $\ell_f : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$ by $\ell_f((\boldsymbol{x}, y)) = [f(\boldsymbol{x}) - y]^2$. We also denote $(\boldsymbol{x}, y)$ by $\boldsymbol{z}$. In Lemma B.1, take $\{\boldsymbol{z}_i\}_{i=1}^n = \mathcal{B}_{t,k}^{\text{tr}}$ and $\mathcal{G} = \{\ell_f - \ell_{\bar{f}_t} : f \in \mathcal{F}\}$. The following lemma suggests a choice of g_i and T for which (B.4) holds.

Lemma B.2 (Noise condition). *Let Assumption 3.1 hold. For all $f, \bar{f} \in \mathcal{F}$,*

$$\begin{aligned}
&\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \operatorname{var}_{\boldsymbol{z}\sim P_j}\left[\ell_f(\boldsymbol{z}) - \ell_{\bar{f}}(\boldsymbol{z})\right] \\
&\quad \le 16M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \|f - \bar{f}\|_j^2 \\
&\quad \le 32M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \left\{\mathbb{E}_{\boldsymbol{z}\sim P_j}\left[\ell_f(\boldsymbol{z}) - \ell_{\bar{f}}(\boldsymbol{z})\right] + 2\mathcal{E}_j(\bar{f})\right\}.
\end{aligned}$$

Proof of Lemma B.2. See Appendix B.5. □

Define

$$T(\ell_f - \ell_{\bar{f}_t}) = 16M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \left\|f - \bar{f}_t\right\|_j^2, \qquad \forall f \in \mathcal{F}.$$

By Lemma B.2, the noise condition (B.4) holds with $C = 32M^2$, and Lemma B.1 is applicable. Moreover, for all $r \ge 0$,

$$\begin{aligned}
&\mathfrak{R}\left(\left\{\ell_f - \ell_{\bar{f}_t} : f \in \mathcal{F},\, T(\ell_f - \ell_{\bar{f}_t}) \le r\right\}; \mathcal{D}_{t,k}^{\mathrm{tr}}\right) \\
&\quad = \mathbb{E}\sup\left\{\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{s\in B_j^{\mathrm{tr}}} \varepsilon_s \left[f(\boldsymbol{x}_s) - y_s\right]^2 : f \in \mathcal{F},\, T(\ell_f - \ell_{\bar{f}_t}) \le r\right\} \\
&\quad \le 4M \cdot \mathbb{E}\sup\left\{\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{s\in B_j^{\mathrm{tr}}} \varepsilon_s \left[f(\boldsymbol{x}_s) - y_s\right] : f \in \mathcal{F},\, T(\ell_f - \ell_{\bar{f}_t}) \le r\right\} \\
&\quad = 4M \cdot \mathbb{E}\sup\left\{\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{s\in B_j^{\mathrm{tr}}} \varepsilon_s f(\boldsymbol{x}_s) : f \in \mathcal{F},\, T(\ell_f - \ell_{\bar{f}_t}) \le r\right\} \\
&\quad = 4M \cdot \mathfrak{R}\left(\left\{f \in \mathcal{F} : T(\ell_f - \ell_{\bar{f}_t}) \le r\right\}; \mathcal{D}_{t,k}^{\mathrm{tr}}\right),
\end{aligned}$$

where $\{\varepsilon_s\}$ are i.i.d. Rademacher random variables, and the inequality is due to Theorem A.6 in Bartlett et al. (2005).

Define

$$L_{t,k}^{\mathrm{tr}}(f) = \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} L_j(f).$$

Applying Lemma B.1 with $g = \widehat{f}_{t,k} - \bar{f}_t$, we obtain that if $\widetilde{\psi} : \mathbb{R}_+ \to \mathbb{R}_+$ is a sub-root function with fixed point $\widetilde{r}$ satisfying

$$\widetilde{\psi}(r) \geq 128M^3 \cdot \mathfrak{R}\left(\left\{f \in \mathcal{F} : T(\ell_f - \ell_{\bar{f}_t}) \leq r\right\}; \mathcal{D}_{t,k}^{\mathrm{tr}}\right), \qquad \forall r \geq \widetilde{r}, \tag{B.5}$$

then with probability at least $1-\delta$,

$$\begin{aligned} L_{t,k}^{\mathrm{tr}}(\widehat{f}_{t,k}) - L_{t,k}^{\mathrm{tr}}(\bar{f}_t) &\lesssim \left(\widehat{L}_{t,k}^{\mathrm{tr}}(\widehat{f}_{t,k}) - \widehat{L}_{t,k}^{\mathrm{tr}}(\bar{f}_t)\right) + \left[\frac{\widetilde{r}}{M^2} + \frac{M^2 \log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t) \\ &\lesssim \left[\frac{\widetilde{r}}{M^2} + \frac{M^2 \log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t), \end{aligned} \tag{B.6}$$

where $\lesssim$ hides a universal constant, and the second inequality is due to $\widehat{L}_{t,k}^{\mathrm{tr}}(\widehat{f}_{t,k}) \leq \widehat{L}_{t,k}^{\mathrm{tr}}(\bar{f}_t)$.

It remains to express $\widetilde{r}$ in terms of $r_{t,k}$, and convert (B.6) into a bound for $L_t(\widehat{f}_{t,k})$. We work on $\widetilde{r}$ first. Take a subroot function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with fixed point $r_{t,k}$ such that

$$\psi(r) \geq \mathfrak{R}\left(\left\{f \in \mathcal{F} : \frac{1}{n_{t,k}} \sum_{j=t-k}^{t-1} n_j \left\|f - \bar{f}_t\right\|_j^2 \leq r\right\}\right), \quad \forall r \geq r_{t,k}.$$

We now show that $\widetilde{\psi}(r) = 128M^3 \psi\left(\frac{r}{16M^2}\right)$ satisfies (B.5). By Lemma D.1, the fixed point $\widetilde{r}$ of $\widetilde{\psi}$ satisfies

$$\min\{1, 8M\}^2 16M^2 r_{t,k} \leq \widetilde{r} \leq \max\{1, 8M\}^2 16M^2 r_{t,k}. \tag{B.7}$$

For all $r \geq \widetilde{r}$, since $r \geq r_{t,k}$, then

$$\widetilde{\psi}(r) \geq 128M^3 \cdot \mathfrak{R}\left(\left\{f \in \mathcal{F} : 16M^2 \cdot \frac{1}{n_{t,k}} \sum_{j=t-k}^{t-1} n_j \left\|f - \bar{f}_t\right\|_j^2 \leq r\right\}\right),$$

so (B.5) holds for this choice of $\widetilde{\psi}$, and (B.6) becomes

$$L^{\mathrm{tr}}_{t,k}(\widehat{f}_{t,k}) - L^{\mathrm{tr}}_{t,k}(\bar{f}_t) \leq M^2 \left[r_{t,k} + \frac{\log(1/\delta)}{B^{\mathrm{tr}}_{t,k}} \right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t). \tag{B.8}$$

To convert (B.8) into a bound for $\mathcal{E}_t(\widehat{f}_{t,k})$, we note that

$$\begin{aligned}
\mathcal{E}_t(\widehat{f}_{t,k}) - \mathcal{E}_t(\bar{f}_t) &\leq \frac{1}{B^{\mathrm{tr}}_{t,k}} \sum_{j=t-k}^{t-1} B^{\mathrm{tr}}_j \mathcal{E}_j(\widehat{f}_{t,k}) - \frac{1}{B^{\mathrm{tr}}_{t,k}} \sum_{j=t-k}^{t-1} B^{\mathrm{tr}}_j \mathcal{E}_j(\bar{f}_t) \\
&\quad + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\widehat{f}_{t,k}) - \mathcal{E}_t(\widehat{f}_{t,k})| + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\bar{f}_t) - \mathcal{E}_t(\bar{f}_t)| \\
&= L^{\mathrm{tr}}_{t,k}(\widehat{f}_{t,k}) - L^{\mathrm{tr}}_{t,k}(\bar{f}_t) \\
&\quad + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\widehat{f}_{t,k}) - \mathcal{E}_t(\widehat{f}_{t,k})| + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\bar{f}_t) - \mathcal{E}_t(\bar{f}_t)| \\
&\leq C \left\{ M^2 \left[r_{t,k} + \frac{\log(1/\delta)}{B^{\mathrm{tr}}_{t,k}} \right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t) \right\} \\
&\quad + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\widehat{f}_{t,k}) - \mathcal{E}_t(\widehat{f}_{t,k})| + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\bar{f}_t) - \mathcal{E}_t(\bar{f}_t)| \\
&\leq C \left\{ M^2 \left[r_{t,k} + \frac{\log(1/\delta)}{B^{\mathrm{tr}}_{t,k}} \right] + \mathcal{E}_t(\bar{f}_t) \right\} \\
&\quad + \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\widehat{f}_{t,k}) - \mathcal{E}_t(\widehat{f}_{t,k})| + (C+1) \max_{t-k \leq j \leq t-1} |\mathcal{E}_j(\bar{f}_t) - \mathcal{E}_t(\bar{f}_t)|
\end{aligned} \tag{B.9}$$

for some universal constant $C > 0$. We now control the last two terms in (B.9). For all $j \in \mathbb{Z}_+$ and f, we have

$$\begin{aligned}
|\mathcal{E}_j(f) - \mathcal{E}_t(f)| &= \left| \|f - f^*_j\|^2_j - \|f - f^*_t\|^2_t \right| \\
&\leq \left| \|f - f^*_j\|^2_j - \|f - f^*_j\|^2_t \right| + \left| \|f - f^*_j\|^2_t - \|f - f^*_t\|^2_t \right|,
\end{aligned} \tag{B.10}$$

where the first term on the right hand side of (B.10) involves covariate shift and can be

bounded by the integral probability metric (B.1) via

$$\|f - f_j^*\|_j^2 - \|f - f_j^*\|_t^2 \le D_t(P_{j,\boldsymbol{x}}, P_{t,\boldsymbol{x}}), \tag{B.11}$$

and the second term involves conditional shift and can be bounded via

$$\begin{aligned}\left|\|f - f_j^*\|_t^2 - \|f - f_t^*\|_t^2\right| &= \left|\langle f_j^* - f_t^*,\, 2f - f_j^* - f_t^*\rangle_t\right| \\ &= \left|\langle f_j^* - f_t^*,\, 2(f - f_t^*) - (f_j^* - f_t^*)\rangle_t\right| \\ &\le 2\|f_j^* - f_t^*\|_t\|f - f_t^*\|_t + \|f_j^* - f_t^*\|_t^2. \end{aligned} \tag{B.12}$$

Substituting (B.10), (B.11), and (B.12) into (B.9) yields

$$\begin{aligned}\|\widehat{f}_{t,k} - f_t^*\|_t^2 - \|\bar{f}_t - f_t^*\|_t^2 \le C\left\{M^2\left[r_{t,k} + \frac{\log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right] + \|\bar{f}_t - f_t^*\|_t^2\right\} \\ + C'\Big[2\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t\|\widehat{f}_{t,k} - f_t^*\|_t \\ + 2\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t\|\bar{f}_t - f_t^*\|_t \\ + 2\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t^2 + 2\max_{t-k\le j\le t-1} D_t(P_{j,\boldsymbol{x}}, P_{t,\boldsymbol{x}})\Big]\end{aligned}$$

for some universal constant $C' > 0$. Completing the squares for $\|\widehat{f}_{t,k} - f_t^*\|_t$ and $\|\bar{f}_t - f_t^*\|_t$ yields

$$\begin{aligned}&\left(\|\widehat{f}_{t,k} - f_t^*\|_t - C'\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t\right)^2 \\ &\quad\le \left(\|\bar{f}_t - f_t^*\|_t + C'\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t\right)^2 \\ &\qquad + C\left\{M^2\left[r_{t,k} + \frac{\log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right] + \|\bar{f}_t - f_t^*\|_t^2\right\} \\ &\qquad + C'\left[2\max_{t-k\le j\le t-1}\|f_j^* - f_t^*\|_t^2 + 2\max_{t-k\le j\le t-1} D_t(P_{j,\boldsymbol{x}}, P_{t,\boldsymbol{x}})\right].\end{aligned}$$

This implies

$$\|\widehat{f}_{t,k} - f_t^*\|_t \lesssim \|\bar{f}_t - f_t^*\|_t + M\left[r_{t,k} + \frac{\log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right]^{1/2} + \max_{t-k\le j\le t-1}\left\{\|f_j^* - f_t^*\|_t + D_t(P_{j,\boldsymbol{x}}, P_{t,\boldsymbol{x}})^{1/2}\right\}.$$

Squaring both sides yields

$$\mathcal{E}_t(\widehat{f}_{t,k}) \lesssim \mathcal{E}_t(\bar{f}_t) + M^2\left[r_{t,k} + \frac{\log(1/\delta)}{B_{t,k}^{\mathrm{tr}}}\right] + \max_{t-k\le j\le t-1}\left\{\|f_j^* - f_t^*\|_t^2 + D_t(P_{j,\boldsymbol{x}}, P_{t,\boldsymbol{x}})\right\}.$$

This finishes the proof.

B.4 Proof of Lemma B.1

The core techniques are the same as those of Theorem 3.3 in Bartlett et al. (2005), with small changes in the quantities to bound. For $r, \lambda > 0$, let

$$w(g) = \min\{r\lambda^k : k \in \{0\} \cup \mathbb{Z}_+,\, r\lambda^k \ge T(g)\}, \qquad \mathcal{G}_r = \left\{\frac{r}{w(g)} g : g \in \mathcal{G}\right\},$$

and

$$V_r^+ = \sup_{g\in\mathcal{G}}\left\{\frac{r}{w(g)} \cdot \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}[g(\boldsymbol{z}_i)] - g(\boldsymbol{z}_i)\right)\right\}.$$

Similar to Lemma 3.8 of Bartlett et al. (2005), for every $K > 1$ and $g \in \mathcal{G}$,

$$V_r^+ \le \frac{r}{\lambda CK} \quad \text{implies} \quad \frac{1}{n}\sum_{i=1}^n \mathbb{E}[g(\boldsymbol{z}_i)] \le \frac{K}{K-1}\cdot\frac{1}{n}\sum_{i=1}^n g(\boldsymbol{z}_i) + \frac{r}{\lambda CK} + \frac{1}{K-1}\cdot\frac{1}{n}\sum_{i=1}^n \eta_i.$$

We now invoke a uniform convergence result to give a bound for V_r^+. It is a non-stationary version of Theorem 2.1 in Bartlett et al. (2005).

Lemma B.3 (Uniform concentration)**.** *Consider the setting of Lemma B.1. Define*

$$v = \frac{1}{n}\sup_{g\in\mathcal{G}}\sum_{i=1}^n \mathrm{var}[g(\boldsymbol{z}_i)].$$

Let $\delta \in (0,1)$. With probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{E}[g(\boldsymbol{z}_i)] - g(\boldsymbol{z}_i)\big) \lesssim \inf_{\alpha>0}\left\{(1+\alpha)\mathfrak{R}(\mathcal{G};\{\boldsymbol{z}_i\}_{i=1}^n) + \sqrt{\frac{v\log(1/\delta)}{n}} + (1+\alpha^{-1})\frac{(b-a)\log(1/\delta)}{n}\right\},$$

where $\lesssim$ hides a universal constant.

Proof of Lemma B.3. Let $Z = \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}[g(\boldsymbol{z}_i) - \mathbb{E}g(\boldsymbol{z}_i)]$. Adapting the proof of Theorem 4 in Massart (2000), with probability at least $1-\delta$,

$$Z - \mathbb{E}Z \lesssim \sqrt{\frac{\log(1/\delta)}{n}(v + (b-a)\mathbb{E}Z)} + \frac{(b-a)\log(1/\delta)}{n},$$

where $\lesssim$ hides a universal constant. By Young's inequality, for every $\alpha > 0$,

$$\sqrt{\frac{\log(1/\delta)}{n}(b-a)\mathbb{E}Z} \le \alpha\mathbb{E}Z + \alpha^{-1}\frac{(b-a)\log(1/\delta)}{n}.$$

By Rademacher symmetrization (e.g., Lemma A.5 in Bartlett et al. (2005)), for i.i.d. Rademacher random variables $\varepsilon_1, ..., \varepsilon_n$ independent of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$, it holds that

$$\mathbb{E}Z \le 2\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(\boldsymbol{z}_i)\right] = 2\mathfrak{R}(\mathcal{G};\{\boldsymbol{z}_i\}_{i=1}^n).$$

This completes the proof. □

By Lemma B.3, each of the following inequalities holds with probability at least $1-\delta$:

$$V_r^+ \lesssim \inf_{\alpha>0}\left\{(1+\alpha)\mathfrak{R}(\mathcal{G}_r;\{\boldsymbol{z}_i\}_{i=1}^n) + \sqrt{\frac{v\log(1/\delta)}{n}} + (1+\alpha^{-1})\frac{(b-a)\log(1/\delta)}{n}\right\},$$

The rest of the proof follows that in Section 3.2 of Bartlett et al. (2005).

B.5 Proof of Lemma B.2

Recall that for $\boldsymbol{z} = (\boldsymbol{x}, y)$, we define $\ell_f(\boldsymbol{z}) = [f(\boldsymbol{x}) - y]^2$. Then

$$
\begin{aligned}
&\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \operatorname{var}_{\boldsymbol{z}\sim P_j}\left[\ell_f(\boldsymbol{z}) - \ell_{\bar{f}}(\boldsymbol{z})\right] \\
&\quad= \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \operatorname{var}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f(\boldsymbol{x}) - y\big)^2 - \big(\bar{f}(\boldsymbol{x}) - y\big)^2\right] \\
&\quad\le \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left\{\left[\big(f(\boldsymbol{x}) - y\big)^2 - \big(\bar{f}(\boldsymbol{x}) - y\big)^2\right]^2\right\} \\
&\quad= \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\big)^2\big(f(\boldsymbol{x}) + \bar{f}(\boldsymbol{x}) - 2y\big)^2\right] \\
&\quad\le 16M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f(\boldsymbol{x}) - \bar{f}(\boldsymbol{x})\big)^2\right] = 16M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \|f - \bar{f}\|_j^2 \\
&\quad\le 32M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \left(\|f - f_j^*\|_j^2 + \|\bar{f} - f_j^*\|_j^2\right) \\
&\quad= 32M^2 \cdot \frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} B_j^{\mathrm{tr}} \left\{\mathbb{E}_{\boldsymbol{z}\sim P_j}\left[\ell_f(\boldsymbol{z}) - \ell_{\bar{f}}(\boldsymbol{z})\right] + 2\mathcal{E}_j(\bar{f})\right\}.
\end{aligned}
$$

This finishes the proof.

B.6 Unavoidable Cost of Non-stationarity

Section 3.2 argues informally that the non-stationarity term in Theorem 3.1 cannot be removed by a cleverer estimator. We state and prove this claim formally here.

Proposition B.1 (Unavoidable cost of non-stationarity)**.** *Fix $t \ge 2$ and a window $k \in [t-1]$. Let $|\mathcal{X}| \ge 2$. Suppose for simplicity that $M = 1$ and $|\mathcal{D}_\tau^{\mathrm{tr}}| = B$ for all $\tau \in [t-1]$. For $V \in [0, 1]$, define a class of non-stationary data distributions over $\mathcal{X} \times [-1, 1]$:*

$$\mathscr{Q}(V) = \left\{(P_1, ..., P_t) : \text{supp}(P_\tau) \subseteq \mathcal{X} \times [-1, 1] \ \forall \tau \in [t], \ \max_{t-k \leq \tau \leq t-1} \text{TV}(P_\tau, P_t) \leq V \right\}.$$

Then any estimator $\widehat{f}_t \in \mathcal{F}$ *based on the data* $\{\mathcal{D}_\tau^{\text{tr}}\}_{\tau=t-k}^{t-1}$ *must have a worst-case out-of-sample prediction error of at least*

$$\sup_{(P_1,...,P_t) \in \mathscr{Q}(V)} \mathbb{E}\left[\mathcal{E}_t(\widehat{f}_t)\right] \gtrsim \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + \frac{1}{Bk} + V.$$

Proposition B.1 considers all non-stationary data-generating processes for which the current distribution P_t is at most V away, in total variation, from each historical data distribution P_τ used for training. When V is small, the last k periods are close to the current environment; when V is large, the training window may contain observations from highly different regimes.

The lower bound decomposes the unavoidable error into the same three components as Theorem 3.1, and by cross-checking, establishes that our upper bound in Theorem 3.1 is indeed tight. The term $\min_{f \in \mathcal{F}} \mathcal{E}_t(f)$ is the approximation error from using the chosen model class at the current distribution. The term $1/(Bk)$ is the statistical uncertainty due to having only Bk observations in the training window. The term V is the irreducible cost of training on historical distributions that differ from the current distribution. Thus, even if the model class is correctly chosen and the estimator is optimal, non-stationarity alone can generate an error of order V. This establishes that the upper bound in Theorem 3.1 has the correct dependence on the non-stationarity.

B.7 Proof of Proposition B.1

Let

$$\mathcal{L} = \sup_{(P_1,...,P_t) \in \mathscr{Q}(V)} \mathbb{E}\left[\mathcal{E}_t(\widehat{f}_t)\right]$$

It suffices to prove the following three lower bounds separately:

$$\mathcal{L} \geq \min_{f \in \mathcal{F}} \mathcal{E}_t(f), \qquad \mathcal{L} \gtrsim \frac{1}{Bk}, \qquad \mathcal{L} \gtrsim V.$$

The first one is immediate since $\widehat{f}_t \in \mathcal{F}$ always. The second one holds regardless of the non-stationarity, and follows from standard minimax lower bound results (Tsybakov, 2008). We now prove the third bound.

Let $p = \min\{V, 1/4\}$. Take two distinct elements $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. We define three distributions Q, Q^+ and Q^- as follows.

- A sample $(\boldsymbol{x}, y) \sim Q$ satisfies $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_1, 0)\big) = 1 - p$ and $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_2, 0)\big) = p$.
- A sample $(\boldsymbol{x}, y) \sim Q^+$ satisfies $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_1, 0)\big) = 1 - p$ and $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_2, 1)\big) = p$. For Q^+, the Bayes least squares optimal estimator $f_+^*(\cdot) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q^+}[y \mid \boldsymbol{x} = \cdot]$ is given by $f_+^*(\boldsymbol{x}_1) = 0$ and $f_-^*(\boldsymbol{x}_2) = 1$.
- A sample $(\boldsymbol{x}, y) \sim Q^-$ satisfies $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_1, 0)\big) = 1 - p$ and $\mathbb{P}\big((\boldsymbol{x}, y) = (\boldsymbol{x}_2, -1)\big) = p$. For Q^-, the Bayes least squares optimal estimator $f_-^*(\cdot) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q^-}[y \mid \boldsymbol{x} = \cdot]$ is given by $f_+^*(\boldsymbol{x}_1) = 0$ and $f_-^*(\boldsymbol{x}_2) = -1$.

We have $\mathrm{TV}(Q, Q^+) = \mathrm{TV}(Q, Q^-) = p \in [V/4, V]$.

Define two non-stationary environments $\mathcal{P}^+ = \big(P_1^+, ..., P_t^+\big)$ and $\mathcal{P}^- = \big(P_1^-, ..., P_t^-\big)$ by

- $P_1^+ = \cdots = P_{t-1}^+ = Q$ and $P_t^+ = Q^+$,
- $P_1^- = \cdots = P_{t-1}^- = Q$ and $P_t^- = Q^-$.

Clearly $\mathcal{P}^+, \mathcal{P}^- \in \mathscr{Q}(V)$. Consider any estimator $\widehat{f}_t$. Under both environments, the training data have exactly the same distribution, because all training sample are drawn from Q. Thus, $\widehat{f}_t$ has the same distribution under $\mathcal{P}^+$ and $\mathcal{P}^-$.

Under $\mathcal{P}^+$,

$$\mathcal{E}_t(\widehat{f}_t) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q^+}\left[\big(\widehat{f}_t(\boldsymbol{x}) - f_+^*(\boldsymbol{x})\big)^2\right] \geq p\big(\widehat{f}_t(\boldsymbol{x}_2) - f_+^*(\boldsymbol{x}_2)\big)^2 = p\big(\widehat{f}_t(\boldsymbol{x}_2) - 1\big)^2.$$

Under $\mathcal{P}^-$,

$$\mathcal{E}_t(\widehat{f}_t) = \mathbb{E}_{(\boldsymbol{x},y)\sim Q^+}\left[\big(\widehat{f}_t(\boldsymbol{x}) - f_-^*(\boldsymbol{x})\big)^2\right] \geq p\big(\widehat{f}_t(\boldsymbol{x}_2) - f_-^*(\boldsymbol{x}_2)\big)^2 = p\big(\widehat{f}_t(\boldsymbol{x}_2) + 1\big)^2.$$

Therefore,

$$\sup_{(P_1,...,P_t)\in\mathscr{Q}(V)} \mathbb{E}\left[\mathcal{E}_t(\widehat{f}_t)\right] \geq \mathbb{E}\left[\max\left\{p\big(\widehat{f}_t(\boldsymbol{x}_2) - 1\big)^2,\ p\big(\widehat{f}_t(\boldsymbol{x}_2) + 1\big)^2\right\}\right] \geq p \geq \frac{V}{4}.$$

This completes the proof.

B.8 Proofs for Example 3.1 and Example 3.2

To obtain the results in Example 3.1 and Example 3.2, we invoke a useful lemma. It is an extension of Theorem 41 in Mendelson (2002) to the non-i.i.d. case. The proof is omitted.

Lemma B.4. *Take $\bar{\boldsymbol{\theta}} \in B(\mathbf{0}, \sqrt{M}) \subseteq \mathbb{R}^d$. Suppose $\boldsymbol{x}_i \sim Q_i$, $i \in [n]$ are independent random vectors in $\mathbb{R}^d$. Let $\varepsilon_1, ..., \varepsilon_n$ be i.i.d. Rademacher random variables independent of $\{\boldsymbol{x}_i\}_{i=1}^n$. For $r \geq 0$, define*

$$\mathcal{R}(r) = \mathbb{E} \sup \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i \boldsymbol{x}_i^\top \boldsymbol{\theta} : \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\left(\boldsymbol{x}_i^\top (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \right)^2 \right] \leq r, \ \|\boldsymbol{\theta}\|_2^2 \leq M \right\}.$$

Let $\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i^\top)$, and denote by $\{\lambda_i\}_{i=1}^d$ its eigenvalues sorted in descending order. Then

$$\mathcal{R}(r) \leq \sqrt{\frac{c}{n} \sum_{i=1}^d \min\{r, M\lambda_i\}}.$$

The above lemma leads to $\mathcal{R}(r) \leq \sqrt{crd/n}$. The right-hand side is subroot and has a unique fixed point $r = cd/n$. This verifies Example 3.1.

To get Example 3.2, we apply Lemma B.4 to the transformed features $\phi(\boldsymbol{x}_{j,i})$ rather than the raw ones. Correspondingly, the feature space becomes $\mathbb{H}$, and the matrix $\boldsymbol{\Sigma}$ in Example 3.2 becomes $\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{\tau=t-k}^{t-1} \sum_{s \in B_\tau} \mathbb{E}[\phi(\boldsymbol{x}_s) \otimes \phi(\boldsymbol{x}_s)]$. Here $\otimes$ denotes the tensor product.

In Example 3.2, the assumption regarding the trace-class operator $\boldsymbol{S}$ forces $\boldsymbol{\Sigma} \preceq \boldsymbol{S}$. Then, Lemma B.4 yields

$$\mathcal{R}(r) \lesssim \sqrt{\frac{1}{B_{t,k}^{\mathrm{tr}}} \sum_{i=1}^\infty \min\{r, \mu_i\}} \leq \frac{1}{\sqrt{B_{t,k}^{\mathrm{tr}}}} \min_{j \geq 1} \sqrt{jr + \sum_{i=j+1}^\infty \mu_i},$$

where $\lesssim$ hides a constant factor.

- If there are constants $c_1, c_2 > 0$ such that $\mu_j \leq c_1 e^{-c_2 j}$ holds for all j, then $\sum_{i=j+1}^\infty \mu_i \lesssim$

$e^{-c_2 j}$. Taking $j = \lceil c_2^{-1} \log(1/r) \rceil$, we get

$$\mathcal{R}(r) \lesssim \frac{1}{\sqrt{B_{t,k}^{\mathrm{tr}}}} \sqrt{r \lceil c_2^{-1} \log(1/r) + 1 \rceil}.$$

The right-hand side is sub-root. Elementary calculation yields $r_{t,k}(\mathcal{F}) \lesssim (\log B_{t,k}^{\mathrm{tr}})/B_{t,k}^{\mathrm{tr}}$.

- If there are constants $c > 0$ and $\alpha \geq 1$ such that $\mu_j \leq c j^{-2\alpha}$ for all j, then $\sum_{i=j+1}^{\infty} \mu_i \lesssim j^{1-2\alpha}$. Taking $j = \lceil r^{-1/(2\alpha)} \rceil$, we get

$$\mathcal{R}(r) \lesssim \sqrt{\frac{r^{1-1/(2\alpha)}}{B_{t,k}^{\mathrm{tr}}}}.$$

The right-hand side is sub-root. Then, we can easily get $r_{t,k}(\mathcal{F}) \lesssim \left(B_{t,k}^{\mathrm{tr}}\right)^{-\frac{2\alpha}{2\alpha+1}}$.

C Proofs for Section 4 and Appendix E

C.1 Algorithmic Details

We record here the full pseudocode for the tournament procedure and the pairwise comparison subroutine described in prose in Section 4. The proofs below refer directly to objects defined in this pseudocode, such as the pivot model and the surviving set S'.

C.2 Proof of Lemma 4.1

Given two models $f, f' \in \mathcal{F}$, denote the output of $\mathcal{A}$ by $\mathcal{A}(f, f') \in \{f, f'\}$. Let $T(\Lambda)$ be the maximum expected number of times Algorithm 2 can call $\mathcal{A}$ after there are Λ remaining models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$ at the end of a while loop, where the maximum is taken over all possible choices of $\{f_\lambda\}_{\lambda=1}^{\Lambda}$. Then $T(\Lambda)$ is increasing in Λ, and $T(\Lambda) \leq \Lambda^2/2$. Let N denote the number of remaining models at the end of the next while loop. Since that while loop calls $\mathcal{A}$ at most $\Lambda - 1$ times, then

$$T(\Lambda) \leq (\Lambda - 1) + \mathbb{E}\left[T(N)\right]. \tag{C.2}$$

Algorithm 2 Adaptive Tournament Model Selection (`ATOMS`)

Input: Candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$, validation data $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$, pairwise model comparison subroutine $\mathcal{A}$.
Initialize $S = \{f_\lambda\}_{\lambda=1}^{\Lambda}$. `// collection of remaining models`
while $|S| > 1$
  Choose a pivot model $f \in S$ uniformly at random.
  Initialize $S' \leftarrow \emptyset$. `// collection of models in` S `that outperform` f
  for $f' \in S \backslash \{f\}$
    Run $\mathcal{A}$ to compare $\{f, f'\}$ to obtain $\mathcal{A}(f, f')$.
    If $\mathcal{A}(f, f') = f'$, set $S' \leftarrow S' \cup \{f'\}$.
  if $S' = \emptyset$
    return $\widehat{f} = f$. `// if no model outperforms` f`, output` f
  else
    Set $S \leftarrow S'$.
return the only model $\widehat{f} \in S$.

Algorithm 3 Adaptive Rolling Window for Model Comparison

Input: Candidate models $\{f_1, f_2\}$, validation data $\{\mathcal{D}_\tau^{\text{va}}\}_{\tau=1}^{t-1}$, hyperparameters δ' and M'.
for $\ell = 1, \cdots, t-1$
  Compute $\widehat{\Delta}_{t,\ell}$, $\widehat{\psi}(t, \ell, \delta')$ and $\widehat{\phi}(t, \ell, \delta')$ according to (4.1), (4.4) and (4.5).
Choose window size

$$\widehat{\ell} \in \operatorname*{argmin}_{\ell \in [t-1]} \left\{ \widehat{\phi}(t, \ell, \delta') + \widehat{\psi}(t, \ell, \delta') \right\}. \tag{C.1}$$

Select $\widehat{\lambda} = 1$ if $\widehat{\Delta}_{t,\widehat{\ell}} \leq 0$, and $\widehat{\lambda} = 2$ otherwise.
return $\widehat{f} = f_{\widehat{\lambda}}$.

For each $\lambda \in [\Lambda]$, let $n_\lambda = |\{\lambda' \in [\Lambda] \backslash \{\lambda\} : \mathcal{A}(f_\lambda, f_{\lambda'}) = f_{\lambda'}\}|$ be the number of remaining models that would beat f_λ if they were paired. Since Algorithm 2 chooses each f_λ as the pivot model uniformly at random, then

$$\mathbb{E}\left[T(N)\right] = \mathbb{E}\left[\frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} T(n_\lambda)\right]. \tag{C.3}$$

Since $\sum_{\lambda=1}^{\Lambda} n_\lambda$ counts exactly one of $(f_\lambda, f_{\lambda'})$ and $(f_{\lambda'}, f_\lambda)$ for all $\lambda \neq \lambda'$, then $\sum_{\lambda=1}^{\Lambda} n_\lambda = \Lambda(\Lambda - 1)/2$. Let $n_{(1)} \leq \cdots \leq n_{(\Lambda)}$ be the order statistics of $n_1, ..., n_\Lambda$. Then for all $i =$

$1, ..., \lceil \Lambda/3 \rceil$,

$$n_{(i)} \le n_{(\lceil \Lambda/3 \rceil)} \le \frac{1}{\Lambda - \lceil \Lambda/3 \rceil + 1} \sum_{i=\lceil \Lambda/3 \rceil}^{\Lambda} n_i \le \frac{3}{2\Lambda} \cdot \frac{\Lambda(\Lambda-1)}{2} = \frac{3(\Lambda-1)}{4}.$$

By the monotonicity of T,

$$\begin{aligned}
\frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda} T(n_\lambda) = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda} T(n_{(i)}) &\le \frac{1}{\Lambda}\Big(\lfloor \Lambda/3 \rfloor T(\lceil 3\Lambda/4 \rceil) + (\Lambda - \lfloor \Lambda/3 \rfloor)\, T(\Lambda - 1) \Big] \\
&\le \frac{1}{3} T(\lceil 3\Lambda/4 \rceil) + \left(\frac{2}{3} + \frac{1}{\Lambda} \right) T(\Lambda) \\
&\le \frac{1}{3} T(\lceil 3\Lambda/4 \rceil) + \frac{2}{3} T(\Lambda) + \Lambda.
\end{aligned}$$

Substituting this into (C.3) and (C.2) gives

$$T(\Lambda) \le 6\Lambda + T(\lceil 3\Lambda/4 \rceil).$$

By the Master Theorem (Theorem 4.1) in Cormen et al. (2022), we conclude that $T(\Lambda) = \Theta(\Lambda)$.

C.3 Proof of Theorem 4.1

We will prove the following stronger result: with probability at least $1 - \delta$,

$$\sqrt{\mathcal{E}_t(\widehat{f})} - \sqrt{\mathcal{E}_t(f_2)} \lesssim M \sqrt{\log(t/\delta)} \cdot \left(\max_{t-\ell \le j \le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{va}}_{t,\ell}} \right)^{1/2}. \tag{C.4}$$

The bound in Theorem 4.1 is obtained by squaring both sides of (C.4).

Without loss of generality, assume $L_t(f_1) \ge L_t(f_2)$, so $\min_{\lambda \in [2]} L_t(f_\lambda) = L_t(f_2)$. By Theorem 4.2 in Han et al. (2024), with probability at least $1 - \delta$, Algorithm 3 outputs a model $\widehat{f}$ satisfying

$$L_t(\widehat{f}) - L_t(f_2) \le |\widehat{\Delta}_{t,\widehat{\ell}} - \Delta_t|$$

$$\lesssim \min_{\ell\in[t-1]}\left\{\sqrt{\log(t/\delta)}\cdot\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\sigma_{t,\ell}\sqrt{\frac{\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}}+\frac{M^2\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}\right\},$$

where $\lesssim$ hides a universal constant. When this event happens, it holds for all $\ell\in[t-1]$ that

$$L_t(\widehat{f})-L_t(f_2)\lesssim\sqrt{\log(t/\delta)}\cdot\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\sigma_{t,\ell}\sqrt{\frac{\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}}+\frac{M^2\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}.$$

When $\widehat{f}=f_2$, (C.4) automatically holds and there is nothing to prove.

Now consider the case when $\widehat{f}=f_1$. Fix $\ell\in[t-1]$. Define

$$\mathcal{E}^{\mathrm{va}}_{t,\ell}(f)=\frac{1}{B^{\mathrm{va}}_{t,\ell}}\sum_{j=t-\ell}^{t-1}B^{\mathrm{va}}_j\big[L_j(f)-L_j(f^*_j)\big].$$

Since

$$\big|\big[\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_1)-\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_2)\big]-\big[L_t(f_1)-L_t(f_2)\big]\big|\le\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|,$$

then

$$\begin{aligned}\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_1)-\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_2)&\le L_t(f_1)-L_t(f_2)+\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|\\&\le C\left[\sqrt{\log(t/\delta)}\cdot\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\sigma_{t,\ell}\sqrt{\frac{\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}}+\frac{M^2\log(t/\delta)}{B^{\mathrm{va}}_{t,\ell}}\right]\end{aligned}\tag{C.5}$$

for some universal constant $C\ge1$. The variance term $\sigma^2_{t,\ell}$ can be bounded by

$$\begin{aligned}\sigma^2_{t,\ell}&=\frac{1}{B^{\mathrm{va}}_{t,\ell}}\sum_{j=t-\ell}^{t-1}B^{\mathrm{va}}_j\operatorname{var}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f_1(\boldsymbol{x})-y\big)^2-\big(f_2(\boldsymbol{x})-y\big)^2\right]\\&\le2\sum_{\lambda=1}^{2}\left[\frac{1}{B^{\mathrm{va}}_{t,\ell}}\sum_{j=t-\ell}^{t-1}B^{\mathrm{va}}_j\operatorname{var}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f_\lambda(\boldsymbol{x})-y\big)^2-\big(f^*_j(\boldsymbol{x})-y\big)^2\right]\right].\end{aligned}$$

For each $\lambda \in [2]$ and $j \in \mathbb{Z}_+$,

$$\begin{aligned}
\operatorname{var}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f_\lambda(\boldsymbol{x})-y\big)^2-\big(f_j^*(\boldsymbol{x})-y\big)^2\right] &\le \mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left\{\left[\big(f_\lambda(\boldsymbol{x})-y\big)^2-\big(f_j^*(\boldsymbol{x})-y\big)^2\right]^2\right\} \\
&= \mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f_\lambda(\boldsymbol{x})-f_j^*(\boldsymbol{x})\big)^2\big(f_\lambda+f_j^*(\boldsymbol{x})-2y\big)^2\right] \\
&\le 16M^2\mathbb{E}_{(\boldsymbol{x},y)\sim P_j}\left[\big(f_\lambda(\boldsymbol{x})-f_j^*(\boldsymbol{x})\big)^2\right] \\
&= 16M^2\mathcal{E}_j(f_\lambda).
\end{aligned}$$

Thus,

$$\sigma_{t,\ell}^2 \le 32M^2\cdot\sum_{\lambda=1}^{2}\left[\frac{1}{B_{t,\ell}^{\mathrm{va}}}\sum_{j=t-\ell}^{t-1}B_j^{\mathrm{va}}\mathcal{E}_j(f_\lambda)\right] = 32M^2\big[\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_1)+\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_2)\big]. \tag{C.6}$$

Substituting (C.6) into (C.5) yields

$$\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_1)-\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_2) \le 2A\left(\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_1)}+\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_2)}\right)+D,$$

where

$$A = 2\sqrt{2}CM\sqrt{\frac{\log(t/\delta)}{B_{t,\ell}^{\mathrm{va}}}} \quad\text{and}\quad D = C\left[\sqrt{\log(t/\delta)}\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\frac{M^2\log(t/\delta)}{B_{t,\ell}^{\mathrm{va}}}\right]$$

Completing the squares gives

$$\left(\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_1)}-A\right)^2 \le \left(\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_2)}+A\right)^2+D,$$

which implies

$$\begin{aligned}
\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_1)}-\sqrt{\mathcal{E}_{t,\ell}^{\mathrm{va}}(f_2)} &\le 2A+\sqrt{D} \\
&\lesssim M\sqrt{\frac{\log(t/\delta)}{B_{t,\ell}^{\mathrm{va}}}}+\left[\sqrt{\log(t/\delta)}\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\frac{M^2\log(t/\delta)}{B_{t,\ell}^{\mathrm{va}}}\right]^{1/2} \\
&\lesssim \sqrt{\log(t/\delta)}\cdot\left(\max_{t-\ell\le j\le t-1}|\Delta_j-\Delta_t|+\frac{M^2}{B_{t,\ell}^{\mathrm{va}}}\right)^{1/2}
\end{aligned}$$

$$\lesssim M\sqrt{\log(t/\delta)}\cdot\left(\max_{t-\ell\le j\le t-1}\mathrm{TV}(P_j,P_t)+\frac{1}{B^{\mathrm{va}}_{t,\ell}}\right)^{1/2},$$

where the last inequality is due to

$$|\Delta_j-\Delta_t|\lesssim M^2\cdot\mathrm{TV}(P_j,P_t).$$

Finally, for every $f\in\{f_1,f_2\}$,

$$\begin{aligned}\left|\sqrt{\mathcal{E}^{\mathrm{va}}_{t,\ell}(f)}-\sqrt{\mathcal{E}_t(f)}\right|\le\sqrt{\left|\mathcal{E}^{\mathrm{va}}_{t,\ell}(f)-\mathcal{E}_t(f)\right|}&\le\max_{t-\ell\le j\le t-1}\sqrt{\left|\mathcal{E}_j(f)-\mathcal{E}_t(f)\right|}\\&\le 2M\max_{t-\ell\le j\le t-1}\sqrt{\mathrm{TV}(P_j,P_t)},\end{aligned}$$

so

$$\begin{aligned}\sqrt{\mathcal{E}_t(f_1)}-\sqrt{\mathcal{E}_t(f_2)}&\le\sqrt{\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_1)}-\sqrt{\mathcal{E}^{\mathrm{va}}_{t,\ell}(f_2)}+4M\max_{t-\ell\le j\le t-1}\sqrt{\mathrm{TV}(P_j,P_t)}\\&\lesssim M\sqrt{\log(t/\delta)}\cdot\left(\max_{t-\ell\le j\le t-1}\mathrm{TV}(P_j,P_t)+\frac{1}{B^{\mathrm{va}}_{t,\ell}}\right)^{1/2}.\end{aligned}$$

As $\widehat{f}=f_1$, this finishes the proof.

C.4 Proof of Theorem 4.2

We first prove the following lemma, which converts any performance guarantee of the subroutine $\mathcal{A}$ to that of Algorithm 2.

Lemma C.1 (From comparison to selection)**.** *Take a performance metric $\mathcal{L}:\{f_\lambda\}_{\lambda=1}^{\Lambda}\to\mathbb{R}$, and $U:(0,1)\to\mathbb{R}_+$. Fix $\delta\in(0,1)$. Suppose that the model comparison subroutine $\mathcal{A}$ in Algorithm 2 satisfies the following property: given two models $h_1,h_2\in\{f_\lambda\}_{\lambda=1}^{\Lambda}$, it outputs a model $\widehat{h}\in\{h_1,h_2\}$ satisfying*

$$\mathbb{P}\left(\mathcal{L}(\widehat{h})-\min\{\mathcal{L}(h_1),\mathcal{L}(h_2)\}\le U(\delta)\right)\ge 1-\delta.$$

Then the output $\widehat{f}$ *of Algorithm 2 satisfies*

$$\mathbb{P}\left(\mathcal{L}(\widehat{f}) - \min_{\lambda\in[\Lambda]} \mathcal{L}(f_\lambda) \le 2U(\delta)\right) \ge 1 - \Lambda^2\delta.$$

Proof of Lemma C.1. Given two models $f, f' \in \mathcal{F}$, denote the output of $\mathcal{A}$ by $\mathcal{A}(f, f') \in \{f, f'\}$. For notational convenience we also set $\mathcal{A}(f, f) = f$ for every $f \in \mathcal{F}$. By a union bound, with probability at least $1 - \Lambda^2\delta$,

$$\mathcal{L}\left(\mathcal{A}\big(f_{\lambda'}, f_{\lambda''}\big)\right) - \min_{\lambda\in\{\lambda',\lambda''\}} \mathcal{L}(f_\lambda) \le U(\delta), \qquad \forall \lambda', \lambda'' \in [\Lambda]. \tag{C.7}$$

From now on suppose that (C.7) holds. Take $\bar{f} \in \{f_\lambda\}_{\lambda=1}^{\Lambda}$ such that $\mathcal{L}(\bar{f}) = \min_{\lambda\in[\Lambda]} \mathcal{L}(f_\lambda)$.

If Algorithm 2 outputs $\widehat{f} = \bar{f}$, then there is nothing to prove. Now assume that $\widehat{f} \neq \bar{f}$. Then, there exists $\ell \in \mathbb{Z}_+$ such that at the end of the K-th while loop, $\bar{f}$ is not in S'. Take the smallest such K. Let g_K denote the pivot model f during the K-th while loop, and let S'_K denote the set S' at the end of the K-th while loop. There are two cases.

1. If at the end of the K-th while loop, $S'_K = \emptyset$ and Algorithm 2 outputs $\widehat{f}$, then during this while loop, a call of Algorithm 3 has yielded $\mathcal{A}(\{\widehat{f}, \bar{f}\}) = \widehat{f}$, so by (C.7), $\mathcal{L}(\widehat{f}) - \mathcal{L}(\bar{f}) \le U(\delta)$.

2. Otherwise, at the end of the K-th while loop, $S'_K \neq \emptyset$. There are two cases.

   (a) If $g_K = \bar{f}$, then every $f \in S'_K$, a call of $\mathcal{A}$ has yielded $\mathcal{A}(\{\bar{f}, f\}) = f$, so by (C.7), $\mathcal{L}(f) - \mathcal{L}(\bar{f}) \le U(\delta)$. Since the output $\widehat{f}$ must come from S'_K, then automatically $\mathcal{L}(\widehat{f}) - \mathcal{L}(\bar{f}) \le U(\delta)$.

   (b) If $g_K \neq \bar{f}$, then for every $f \in S'_K$, a call of $\mathcal{A}$ has yielded $\mathcal{A}(\{g_K, f\}) = f$, so by (C.7),

   $$\mathcal{L}(f) - \mathcal{L}(g_K) \le U(\delta). \tag{C.8}$$

   Since $\bar{f} \notin S'_K$, then a call of Algorithm 3 has yielded $\mathcal{A}(\{g_K, \bar{f}\}) = g_K$, so by (C.7),

   $$\mathcal{L}(g_K) - \mathcal{L}(\bar{f}) \le U(\delta). \tag{C.9}$$

Putting together (C.8) and (C.9) yields that for all $f \in S'_K$,

$$\mathcal{L}(f) \le \mathcal{L}(g_K) + U(\delta) \le \mathcal{L}(\bar{f}) + 2U(\delta).$$

Since the output $\widehat{f}$ must come from S'_K, then automatically $\mathcal{L}(\widehat{f}) - \mathcal{L}(\bar{f}) \le 2U(\delta)$.

In all the cases above, we have $\mathcal{L}(\widehat{f}) - \mathcal{L}(\bar{f}) \le 2U(\delta)$. □

We now prove Theorem 4.2. By the stronger bound (C.4) in the proof of Theorem 4.1, we can set in Lemma C.1 $\mathcal{L}(f) = \sqrt{\mathcal{E}_t(f)}$ and

$$U(\delta) = CM\sqrt{\log(t/\delta)} \cdot \min_{\ell \in [t-1]} \left(\max_{t-\ell \le j \le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{va}}_{t,\ell}} \right)^{1/2},$$

for some universal constant $C > 0$. Then, by Lemma C.1 and the choice of δ', with probability $1 - \delta$, the output $\widehat{f}$ of Algorithm 2 satisfies

$$\sqrt{\mathcal{E}_t(\widehat{f})} - \min_{\lambda \in [\Lambda]} \sqrt{\mathcal{E}_t(f_\lambda)} \lesssim U(\delta/\Lambda^2) \lesssim M\sqrt{\log(\Lambda t/\delta)} \cdot \min_{\ell \in [t-1]} \left(\max_{t-\ell \le j \le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{va}}_{t,\ell}} \right)^{1/2}.$$

Here $\lesssim$ only hides universal constants. This finishes the proof.

C.5 Proof of Theorem 4.3

Recall that $|\mathcal{D}^{\mathrm{tr}}_j| = B^{\mathrm{tr}}_j$, $|\mathcal{D}^{\mathrm{va}}_j| = B^{\mathrm{va}}_j$, $|\mathcal{D}_j| = B^{\mathrm{tr}}_j + B^{\mathrm{va}}_j$, $B^{\mathrm{tr}}_{t,k} = \sum_{j=t-k}^{t-1} B^{\mathrm{tr}}_j$, $B^{\mathrm{va}}_{t,\ell} = \sum_{j=t-\ell}^{t-1} B^{\mathrm{va}}_j$, and $B_{t,k} = B^{\mathrm{tr}}_{t,k} + B^{\mathrm{va}}_{t,k}$. Since there are $(t-1)|\mathscr{F}|$ candidate models, then by Theorem 4.2, with probability at least $1 - \delta/2$, the output $\widehat{f}$ of Algorithm 2 satisfies

$$\mathcal{E}_t(\widehat{f}) \lesssim \min_{\mathcal{F} \in \mathscr{F},\, k \in [t-1]} \mathcal{E}_t\big(\widehat{h}(\mathcal{F}, k)\big) + M^2 \log\left(t^2|\mathscr{F}|/\delta\right) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \le j \le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{va}}_{t,\ell}} \right\}. \tag{C.10}$$

By Theorem 3.1 and a union bound, with probability at least $1 - \delta/2$,

$$\mathcal{E}_t\big(\widehat{h}(\mathcal{F}, k)\big) \lesssim \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathscr{F}|/\delta)}{B^{\mathrm{tr}}_{t,k}} \right) + M^2 \max_{t-k \le j \le t-1} \mathrm{TV}\left(P_j, P_t\right). \tag{C.11}$$

Combining (C.10) and (C.11) yields that, with probability at least $1-\delta$,

$$\begin{aligned}
\mathcal{E}_t(\widehat{f}) &\lesssim \min_{\mathcal{F}\in\mathscr{F},\, k\in[t-1]} \left\{ \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathscr{F}|/\delta)}{B^{\mathrm{tr}}_{t,k}} \right) + M^2 \max_{t-k\le j\le t-1} \mathrm{TV}\left(P_j, P_t\right) \right\} \\
&\quad + M^2 \log\left(t^2|\mathscr{F}|/\delta\right) \cdot \min_{\ell\in[t-1]} \left\{ \max_{t-\ell\le j\le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{va}}_{t,\ell}} \right\} \\
&\lesssim \min_{\mathcal{F}\in\mathscr{F},\, k\in[t-1]} \left\{ \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathscr{F}|/\delta)}{B^{\mathrm{tr}}_{t,k}} \right) + M^2 \max_{t-k\le j\le t-1} \mathrm{TV}\left(P_j, P_t\right) \right\} \\
&\quad + M^2 \log\left(t^2|\mathscr{F}|/\delta\right) \cdot \min_{k\in[t-1]} \left\{ \max_{t-k\le j\le t-1} \mathrm{TV}(P_j, P_t) + \frac{1}{B^{\mathrm{tr}}_{t,k}} \right\} \quad \text{(by Assumption 4.1)} \\
&\lesssim \log\left(t|\mathscr{F}|/\delta\right) \cdot \min_{\mathcal{F}\in\mathscr{F},\, k\in[t-1]} \left\{ \min_{f\in\mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{1}{B^{\mathrm{tr}}_{t,k}} \right) + M^2 \max_{t-k\le j\le t-1} \mathrm{TV}\left(P_j, P_t\right) \right\}.
\end{aligned}$$

Here the second inequality hides the constant c in Assumption 4.1. This completes the proof.

C.6 Proof of Lemma E.1

By the triangle inequality,

$$\left|\widehat{\Delta}^R_{t,\ell} - \Delta^R_t\right| \le \left|\widehat{\Delta}^R_{t,\ell} - \Delta^R_{t,\ell}\right| + \left|\Delta^R_{t,\ell} - \Delta^R_t\right|, \tag{C.12}$$

where

$$\Delta^R_{t,\ell} = \mathbb{E}\big[\widehat{\Delta}^R_{t,\ell}\big] = \frac{\Delta_{t,\ell}}{V_{t,\ell}}.$$

By Lemma D.3,

$$\left|\Delta^R_{t,\ell} - \Delta^R_t\right| = \left|\frac{\Delta_{t,\ell}}{V_{t,\ell}} - \frac{\Delta_t}{V_t}\right| \le \max_{t-\ell\le j\le t-1} |\Delta^R_j - \Delta^R_t|. \tag{C.13}$$

We have

$$\widehat{\Delta}^R_{t,\ell} = \frac{1}{B^{\mathrm{va}}_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{i=1}^{B^{\mathrm{va}}_j} \frac{u_{j,i}}{V_{t,\ell}},$$

where $u_{j,i} = \big[f_1(\boldsymbol{x}^{\mathrm{va}}_{j,i}) - y^{\mathrm{va}}_{j,i}\big]^2 - \big[f_2(\boldsymbol{x}^{\mathrm{va}}_{j,i}) - y^{\mathrm{va}}_{j,i}\big]^2$. By Assumptions 3.1 and E.1, $|u_{j,i}/V_{t,\ell}| \le 8M^2/v$ for all j and i. By Bernstein's concentration inequality (Lemma D.2), with probability

at least $1-\delta$,

$$\left|\widehat{\Delta}^R_{t,\ell} - \Delta^R_{t,\ell}\right| \le \sigma^R_{t,\ell}\sqrt{\frac{2\log(2/\delta)}{B^{\mathrm{va}}_{t,\ell}}} + \frac{16(M^2/v)\log(2/\delta)}{3B^{\mathrm{va}}_{t,\ell}}. \tag{C.14}$$

Substituting (C.13) and (C.14) into (C.12) completes the proof.

C.7 Proof of Theorem E.1

We first prove the following theoretical guarantee for the R^2-based comparison subroutine Algorithm 4.

Theorem C.1 (Near-optimal model comparison with R^2)**.** *Let Assumptions 3.1 and E.1 hold. Choose $\delta \in (0,1)$ and set $\delta' = 1/(3t)$ in Algorithm 2. With probability at least $1-\delta$, the output $\widehat{f}$ of Algorithm 4 satisfies*

$$\max_{\lambda\in[2]} \widetilde{R}^2_t(f_\lambda) - \widetilde{R}^2_t(\widehat{f}) \lesssim \log(t/\delta)\cdot \min_{\ell\in[t-1]} \left\{ \max_{t-\ell\le j\le t-1} \max_{\lambda\in[2]} \left|\widetilde{R}^2_j(f_\lambda) - \widetilde{R}^2_t(f_\lambda)\right| + \frac{M^2/v}{\sqrt{B^{\mathrm{va}}_{t,\ell}}} \right\}. \tag{C.15}$$

Here $\lesssim$ hides a universal constant.

Proof of Theorem C.1. Following the same argument as Theorem 4.2 in Han et al. (2024), we can show that with probability at least $1-\delta$,

$$\begin{aligned}\max_{\lambda\in[2]} \widetilde{R}^2_t(f_\lambda) - \widetilde{R}^2_t(\widehat{f}) &\le \left|\widehat{\Delta}^R_{t,\widehat{\ell}} - \Delta^R_t\right| \\ &\lesssim \log(t/\delta)\cdot \min_{\ell\in[t-1]} \left\{ \max_{t-\ell\le j\le t-1} \left|\Delta^R_j - \Delta^R_t\right| + \frac{\widehat{v}^R_{t,\ell}}{\sqrt{B^{\mathrm{va}}_{t,\ell}}} + \frac{M^2/v}{B^{\mathrm{va}}_{t,\ell}} \right\}.\end{aligned}$$

We finish the proof by noting that $\sigma^R_{t,\ell} \lesssim M^2/v$ and

$$\left|\Delta^R_j - \Delta^R_t\right| = \left|\left[\widetilde{R}^2_j(f_1) - \widetilde{R}^2_j(f_2)\right] - \left[\widetilde{R}^2_t(f_1) - \widetilde{R}^2_t(f_2)\right]\right| \le 2\max_{\lambda\in[2]} \left|\widetilde{R}^2_j(f_\lambda) - \widetilde{R}^2_j(f_\lambda)\right|.$$

□

We can now use Lemma C.1 to translate Theorem 4.1 to a theoretical guarantee for

general model selection. Set $\mathcal{L}(f) = 1 - \widetilde{R}_t^2(f)$ and

$$U(\delta) = C \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \le j \le t-1} \max_{\lambda \in [\Lambda]} \left| \widetilde{R}_j^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda) \right| + \frac{M^2/v}{\sqrt{B_{t,\ell}^{\mathrm{va}}}} \right\},$$

for a sufficiently large universal constant $C > 0$. Then for any f,

$$\mathcal{L}(f) - \min_{\lambda \in [\Lambda]} \mathcal{L}(f_\lambda) = \max_{\lambda \in [\Lambda]} \widetilde{R}_t^2(f_\lambda) - \widetilde{R}_t^2(f).$$

By Lemma C.1 and the choice of δ', the output $\widehat{f}$ of Algorithm 2 satisfies that with probability at least $1 - \delta$,

$$\begin{aligned} \max_{\lambda \in [\Lambda]} \widetilde{R}_t^2(f_\lambda) - \widetilde{R}_t^2(\widehat{f}) &\lesssim U(\delta/\Lambda^2) \\ &\lesssim \log(\Lambda t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \le j \le t-1} \max_{\lambda \in [\Lambda]} \left| \widetilde{R}_j^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda) \right| + \frac{M^2/v}{\sqrt{B_{t,\ell}^{\mathrm{va}}}} \right\}. \end{aligned}$$

Here $\lesssim$ only hides universal constants.

D Technical Lemmas

Lemma D.1. *Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a sub-root function with fixed point $r^* > 0$. For all $A, a > 0$, the function $\widetilde{\psi}(r) = A\psi(ar)$ is sub-root and its fixed point $\widetilde{r}$ satisfies*

$$\frac{\min\{1, Aa\}^2}{a} r^* \le \widetilde{r} \le \frac{\max\{1, Aa\}^2}{a} r^*.$$

Proof of Lemma D.1. It is easy to verify that $\widetilde{\psi}$ is subroot. We now study $\widetilde{r}$. First consider the case $a = 1$. Since $A\psi(\widetilde{r}) = \widetilde{r}$, then $\psi(\widetilde{r})/\sqrt{\widetilde{r}} = \sqrt{\widetilde{r}}/A$. There are two cases.

- If $\widetilde{r} \ge r^*$, then $\psi(\widetilde{r})/\sqrt{\widetilde{r}} \le \psi(r^*)/\sqrt{r^*} = \sqrt{r^*}$, so $\widetilde{r} \le A^2 r^*$.
- If $\widetilde{r} \le r^*$, then $\psi(\widetilde{r})/\sqrt{\widetilde{r}} \ge \psi(r^*)/\sqrt{r^*} = \sqrt{r^*}$, so $\widetilde{r} \ge A^2 r^*$.

Therefore, if $A < 1$, then $A^2 r^* \le \widetilde{r} < r^*$. If $A > 1$, then $r^* < \widetilde{r} \le A^2 r^*$. This shows that

$$\min\{1, A\}^2 r^* \le \widetilde{r} \le \max\{1, A\}^2 r^*.$$

In the general case of $a > 0$, the function $r \mapsto a^{-1}\psi(ar)$ is sub-root and has fixed point $a^{-1}r^*$. The proof is finished by noting $\widetilde{\psi}(r) = (Aa) \cdot a^{-1}\psi(ar)$. □

Lemma D.2 (Bernstein's concentration inequality)**.** *Let $\{x_i\}_{i=1}^n$ be independent random variables taking values in $[a, b]$ almost surely. Define the average variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{var}(x_i)$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n (x_i - \mathbb{E}x_i)\right| \le \sigma\sqrt{\frac{2\log(2/\delta)}{n}} + \frac{2(b-a)\log(2/\delta)}{3n}.$$

Proof of Lemma D.2. Inequality (2.10) in Boucheron et al. (2013) implies that for any $t \ge 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (x_i - \mathbb{E}x_i) > t\right) \le \exp\left(-\frac{nt^2/2}{\sigma^2 + (b-a)t/3}\right).$$

Fix $\delta \in (0, 1)$. Then,

$$\begin{aligned}
& \exp\left(-\frac{nt^2/2}{\sigma^2 + (b-a)t/3}\right) \le \delta \\
\Leftrightarrow\quad & \frac{nt^2}{2} \ge \sigma^2 \log(1/\delta) + \frac{t(b-a)\log(1/\delta)}{3} \\
\Leftrightarrow\quad & \frac{n}{2}\left(t - \frac{(b-a)\log(1/\delta)}{3n}\right)^2 \ge \sigma^2 \log(1/\delta) + \frac{n}{2}\left(\frac{(b-a)\log(1/\delta)}{3n}\right)^2 \\
\Leftarrow\quad & \left(t - \frac{(b-a)\log(1/\delta)}{3n}\right)^2 \ge \left(\sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{(b-a)\log(1/\delta)}{3n}\right)^2 \\
\Leftarrow\quad & t \ge \sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{2(b-a)\log(1/\delta)}{3n}.
\end{aligned}$$

Hence,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (x_i - \mathbb{E}x_i) > \sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{2(b-a)\log(1/\delta)}{3n}\right) \le \delta.$$

Replacing each x_i by $-x_i$ gives bounds on the lower tail and the absolute deviation. □

Lemma D.3. *For all $a, a_1, ..., a_n \geq 0$ and $b, b_1, ..., b_n > 0$, it holds that*

$$\left| \frac{a}{b} - \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right| \leq \max_{i \in [n]} \left| \frac{a}{b} - \frac{a_i}{b_i} \right|.$$

Proof of Lemma D.3. This is due to

$$\begin{aligned} \left| \frac{a}{b} - \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right| = \left| \frac{a}{b} - \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \frac{a_i}{b_i} \right| &= \left| \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \left(\frac{a}{b} - \frac{a_i}{b_i} \right) \right| \\ &\leq \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \left| \frac{a}{b} - \frac{a_i}{b_i} \right| \leq \max_{i \in [n]} \left| \frac{a}{b} - \frac{a_i}{b_i} \right|. \end{aligned}$$

This finishes the proof. □

E Extension: Model Selection with the R^2 Metric

E.1 Algorithm Design

In this section, we propose a variant of our model selection method that is tailored to the R^2 metric. We consider the following population form:

$$\widetilde{R}_t^2(f) = 1 - \frac{\mathbb{E}_{(\boldsymbol{x},y)\sim P_t}\left[(f(\boldsymbol{x}) - y)^2\right]}{\mathbb{E}_{(\boldsymbol{x},y)\sim P_t}[y^2]}, \tag{E.1}$$

Define $V_t = \mathbb{E}_{(\boldsymbol{x},y)\sim P_t}[y^2]$, then $\widetilde{R}_t^2(f) = 1 - L_t(f)/V_t$. For simplicity, we assume that at the beginning of each period $t \in \mathbb{Z}_+$, we have access to $\{V_\tau\}_{\tau=1}^{t-1}$. In our numerical experiments, we will approximate V_τ by its empirical counterpart computed from the validation data $\mathcal{D}_\tau^{\text{va}}$. For the population R^2 metric to be well defined, we assume for simplicity that $\{V_t\}_{t=1}^\infty$ are bounded away from zero.

Assumption E.1 (Uniformly lower bounded second moments)**.** There exists $v > 0$ such that $V_t \geq v$ for all $t \in \mathbb{Z}_+$.

We first define the model comparison subroutine, which aims to output the better of two given candidate models f_1 and f_2. Define the R^2 performance gap

$$\Delta_t^R = \widetilde{R}_t^2(f_2) - \widetilde{R}_t^2(f_1) = \frac{L_t(f_1) - L_t(f_2)}{V_t}.$$

Then f_2 outperforms f_1 if and only if $\Delta_t^R > 0$. For each window size $\ell \in [t-1]$, we can form a rolling-window estimator of Δ_t^R, given by

$$\widehat{\Delta}_{t,\ell}^R = \frac{\widehat{\Delta}_{t,\ell}}{V_{t,\ell}} \tag{E.2}$$

where $\widehat{\Delta}_{t,\ell}$ is defined by (4.1), and $V_{t,\ell} = \frac{1}{B_{t,\ell}^{\mathrm{va}}} \sum_{\tau=t-k}^{t-1} B_\tau^{\mathrm{va}} V_\tau$. We establish a bias-variance decomposition for the estimation error of $\widehat{\Delta}_{t,\ell}^R$.

Lemma E.1 (Bias-variance decomposition)**.** *Let Assumptions 2.1, 3.1 and E.1 hold. Let $\sigma_{t,\ell}^R = \sigma_{t,\ell}/V_{t,\ell}$, where $\sigma_{t,\ell}$ is defined by (4.4). For $\delta \in (0,1)$, define $M' = 8M^2$,*

$$\phi_R(t,\ell) = \max_{t-\ell \le \tau \le t-1} \left|\Delta_\tau^R - \Delta_t^R\right|,$$

$$\psi_R(t,\ell,\delta) = \begin{cases} M'/v, & \text{if } B_{t,\ell}^{\mathrm{va}} = 1 \\ \sigma_{t,\ell}^R \sqrt{\dfrac{2\log(2/\delta)}{B_{t,\ell}^{\mathrm{va}}}} + \dfrac{2(M'/v)\log(2/\delta)}{3B_{t,\ell}^{\mathrm{va}}}, & \text{if } B_{t,\ell}^{\mathrm{va}} \ge 2 \end{cases}.$$

With probability at least $1-\delta$,

$$\left|\widehat{\Delta}_{t,\ell}^R - \Delta_t^R\right| \le \phi_R(t,\ell) + \psi_R(t,\ell,\delta).$$

Proof. See Appendix C.6. □

Based on Lemma E.1 and following the same idea as (4.4), we form a data-driven proxy

for $\psi_R(t,\ell,\delta)$, given by

$$\widehat{\psi}_R(t,\ell,\delta) = \begin{cases} M'/v, & \text{if } B^{\mathrm{va}}_{t,\ell} = 1 \\ \widehat{v}^R_{t,\ell}\sqrt{\dfrac{2\log(2/\delta)}{B^{\mathrm{va}}_{t,\ell}}} + \dfrac{8(M'/v)\log(2/\delta)}{3(B^{\mathrm{va}}_{t,\ell}-1)}, & \text{if } B^{\mathrm{va}}_{t,\ell} \geq 2 \end{cases}, \tag{E.3}$$

where

$$\widehat{v}^R_{t,\ell} = \frac{\widehat{v}_{t,\ell}}{V_{t,\ell}}, \tag{E.4}$$

and $\widehat{v}_{t,\ell}$ is defined in (4.3). Following the same idea as (4.5), we also form a data-driven proxy for $\phi_R(t,\ell,\delta)$:

$$\widehat{\phi}_R(t,\ell,\delta) = \max_{i\in[\ell]} \left(\left|\widehat{\Delta}^R_{t,\ell} - \widehat{\Delta}^R_{t,i}\right| - \left[\widehat{\psi}_R\left(t,\ell,\delta\right) + \widehat{\psi}_R\left(t,i,\delta\right)\right] \right)_+ . \tag{E.5}$$

This yields Algorithm 4 as the model comparison routine.

Algorithm 4 Adaptive Rolling Window for Model Comparison (R^2 Metric)

Input: Candidate models $\{f_1, f_2\}$, validation data $\{\mathcal{D}^{\mathrm{va}}_\tau\}_{\tau=1}^{t-1}$, variances $\{V_\tau\}_{\tau=1}^{t-1}$, hyperparameters δ', M' and v.

for $\ell = 1, \cdots, t-1$

Compute $\widehat{\Delta}^R_{t,\ell}$, $\widehat{v}^R_{t,\ell}$, $\widehat{\psi}_R(t,\ell,\delta')$ and $\widehat{\phi}_R(t,\ell,\delta')$ according to (E.2), (E.4), (E.3) and (E.5). Choose window size

$$\widehat{\ell} \in \operatorname*{argmin}_{\ell\in[t-1]} \left\{ \widehat{\phi}_R(t,\ell,\delta') + \widehat{\psi}_R(t,\ell,\delta') \right\}. \tag{E.6}$$

Select $\widehat{\lambda} = 1$ if $\widehat{\Delta}^R_{t,\widehat{\ell}} \leq 0$, and $\widehat{\lambda} = 2$ otherwise.

return $\widehat{f} = f_{\widehat{\lambda}}$.

By using Algorithm 4 as the pairwise comparison subroutine $\mathcal{A}$ in Algorithm 2, we obtain an R^2-based model selection algorithm, which we call `ATOMS-R2`. We establish the following guarantee in terms of the R^2 metric.

Theorem E.1 (Near-optimal model selection with R^2)**.** *Let Assumptions 2.1 and 3.1 hold. Choose $\delta \in (0,1)$ and set $\delta' = 1/(3\Lambda^2 t)$ in Algorithm 2. With probability at least $1-\delta$, the*

output $\widehat{f}$ *of* `ATOMS-R2` *satisfies*

$$\max_{\lambda\in[\Lambda]} \widetilde{R}_t^2(f_\lambda) - \widetilde{R}_t^2(\widehat{f}) \lesssim \log(\Lambda t/\delta) \cdot \min_{\ell\in[t-1]} \left\{ \max_{t-\ell\le\tau\le t-1} \max_{\lambda\in[\Lambda]} \left|\widetilde{R}_\tau^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda)\right| + \frac{M^2/v}{\sqrt{B_{t,\ell}^{\mathrm{va}}}} \right\}.$$

Here $\lesssim$ *hides a universal constant.*

Proof of Theorem E.1. See Appendix C.7. □

Theorem E.1 shares a similar interpretation as Theorem 4.2. The term

$$\max_{t-\ell\le\tau\le t-1} \max_{\lambda\in[\Lambda]} \left|\widetilde{R}_\tau^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda)\right|$$

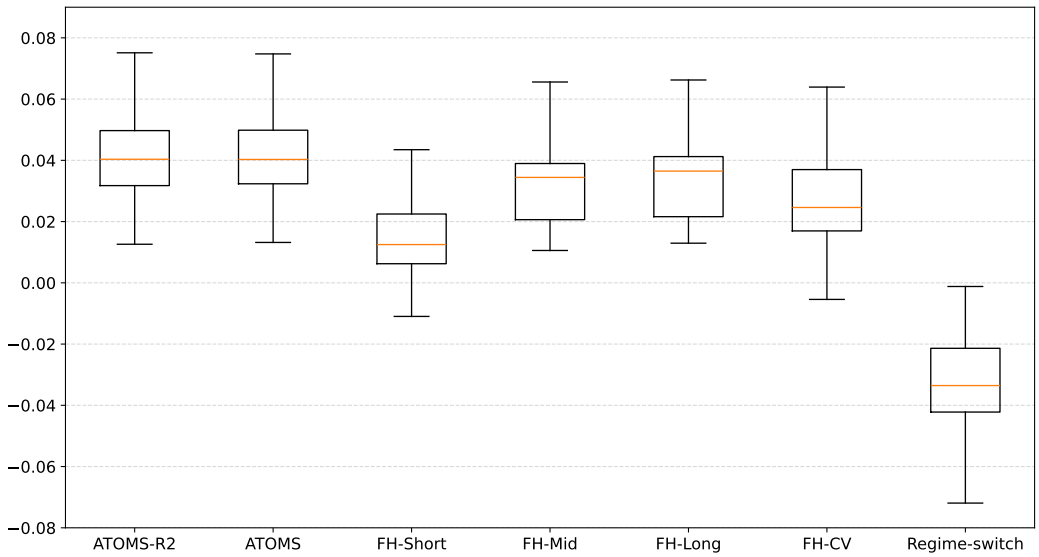
quantifies the shift in the R^2 metric of the models within the last ℓ periods, and $(M^2/v)/\sqrt{B_{t,\ell}^{\mathrm{va}}}$ represents the statistical uncertainty associated with the $B_{t,\ell}^{\mathrm{va}}$ validation data points. Together, they represent the errors that arise when using a fixed validation window ℓ to select models. Theorem E.1 shows that our R^2-based model selection algorithm `ATOMS-R2` is comparable to an oracle that uses the optimal validation window in hindsight to attain the highest R^2.

E.2 Experiment Results for `ATOMS-R2`

We now present experiment results for `ATOMS-R2`. We set its hyperparameters $\delta' = 0.1$ and $M' = 10$. Its performance is similar to the MSE-based approach `ATOMS`. We will report results for the R^2 metric that benchmarks against a zero forecast.

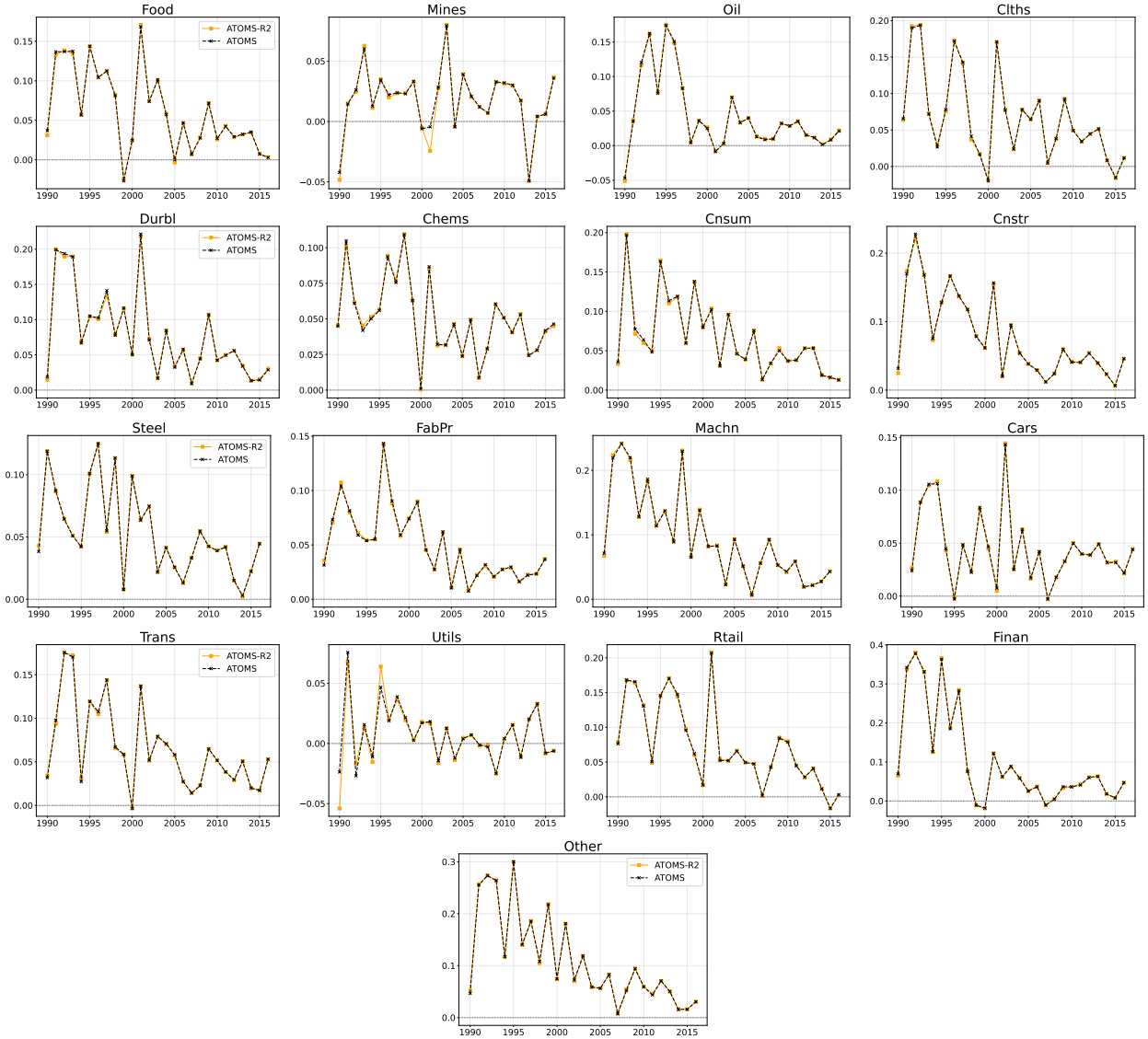
In Figure 8, we give a box plot of the overall out-of-sample R^2 of `ATOMS-R2` along with `ATOMS` and the fixed-window benchmarks across the 17 industries. In Figure 9, we compare the annual out-of-sample R^2 of `ATOMS-R2` and `ATOMS` for the 17 industries. Again, `ATOMS-R2` and `ATOMS` have similar performance.

Figure 8: Box Plot of Out-of-Sample R^2 of `ATOMS-R2`, `ATOMS` and Benchmarks for 17 Industry Portfolios.



This figure describes the distribution of each method's OOS R^2. Each box corresponds to all industries and all years in our OOS horizon.

Figure 9: Annual Out-of-Sample R^2 of `ATOMS-R2` and `ATOMS` for 17 Industry Portfolios.



This figure reports the annual OOS R^2 of our adaptive model selection algorithms `ATOMS-R2` (orange line with ■'s) and `ATOMS` (black dashed line with ×'s). The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 10.

F Additional Experiment Details

F.1 Experiment Details for Section 3.1

In this section, we given more details on our empirical investigations of the nonstationarity-complexity tradeoff in Section 3.1.

The three prediction models, along with their hyperparameters, are: (1) a linear model trained by Ridge regression using the most recent 64 months of data, with $\alpha = 1$, (2) a random forest trained on the most recent 64 months of data, with $n_{\texttt{tree}} = 200$ and $d_{\max} = 5$, and (3) a random forest trained on all historical data, with $n_{\texttt{tree}} = 200$ and $d_{\max} = 5$.

In each month t, we construct training data by randomly subsampling 4/5 of the observations $\mathcal{D}_j$ in each previous month $j \in [t-1]$. The process is repeated 20 times with independent random seeds. We then average the out-of-sample R^2 over the 20 random seeds, which are then used to produce the figures in Section 3.1.

F.2 Training Window Configurations in Section 5

To assess the performance of our adaptive model selection algorithm across different data regimes, we train models on windows of different lengths. For each month t, we consider training windows of $4^k \wedge (t-1)$ months with $0 \leq k \leq 5$. In particular, since $t \in \{1, ..., 350\}$, then $k = 5$ corresponds to a full training window of $(t-1)$ months. This yields the following window lengths:

Table 11: Training Window Lengths by Value of k

k	Window Length (months)	Approximate Years
0	1	0.08
1	4	0.33
2	16	1.33
3	64	5.33
4	256	21.33
5	$t-1$	$(t-1)/12$

This exponential scaling allows us to examine how model performance varies with the amount of historical data available for training. Shorter windows capture recent market dy-

namics but may be susceptible to noise, while longer windows provide more stable parameter estimates but may miss structural changes in the data-generating process.

F.3 Hyperparameters for ATOMS

In our implementation of ATOMS, we set $\delta' = 0.1$ and $M' = 10^{-3}$. The theoretical prescription is $M' = 8M^2$, where M is an upper bound on the magnitude of returns y. In finite samples, however, using the literal maximum return can make this bound overly conservative because a small number of extreme observations dominate the scale. We therefore calibrate M' to cover typical values of $8y^2$ in the pre-OOS period. From September 1987 to December 1989, before the OOS period begins in January 1990, the 90th percentile of $8y^2$ is 9×10^{-4}, while the maximum is 0.1. We therefore set $M' = 10^{-3}$.

To assess the sensitivity of ATOMS to the choice of M', Table 12 reports its OOS performance after doubling and halving the baseline choice $M' = 10^{-3}$, corresponding to $M' = 2 \times 10^{-3}$ and $M' = 5 \times 10^{-4}$, respectively. The results are stable across these perturbations: the full-period OOS R^2 remains between 0.049 and 0.050, and the results during the recession periods change only modestly. For all these three values of M', ATOMS continues to outperform the benchmark methods reported in Table 3.

Table 12: Ablation Study of Hyperparameter M' in `ATOMS`: Out-of-Sample R^2 Averages Across Industries by Time Period

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS` ($M' = 2 \times 10^{-3}$)	0.050	0.024	0.124	0.041
`ATOMS` ($M' = 1 \times 10^{-3}$)	0.049	0.027	0.125	0.041
`ATOMS` ($M' = 5 \times 10^{-4}$)	0.049	0.028	0.124	0.040

This table reports the sensitivity of `ATOMS`'s performance to the choice of M', in terms of the out-of-sample (OOS) R^2 averages for return prediction models across all 17 industry portfolios. The baseline choice is $M' = 10^{-3}$, which corresponds to the result reported in Table 3. The other two rows double and halve this value, setting $M' = 2 \times 10^{-3}$ and $M' = 5 \times 10^{-4}$, respectively. Full OOS Period refers our largest available OOS period covering 01/1990∼11/2016. Columns report OOS R^2 averages across all industries and highlight this metric during three recessions, as documented in NBER Business Cycle Dating:

- the 1990 Gulf War recession (06/1990∼10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001∼10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007∼06/2009).

That is, the OOS performance in Gulf War column focuses on model performance comparisons exclusively in the out-of-sample period of 06/1990∼10/1990. All values are calculated using monthly return data.

F.4 Computational Implementation

All models are implemented using Python 3.9 with the following libraries:

- Linear models: `scikit-learn` version 1.0.2, specifically `Ridge`, `Lasso`, and `ElasticNet`.
- Random forest: `scikit-learn`'s `RandomForestRegressor` with `random_state=0`.
- Regime-switching model: `statsmodels.tsa.regime_switching.markov_regression`.
- Data manipulation: `pandas` version 1.4.2 and `numpy` version 1.21.5.

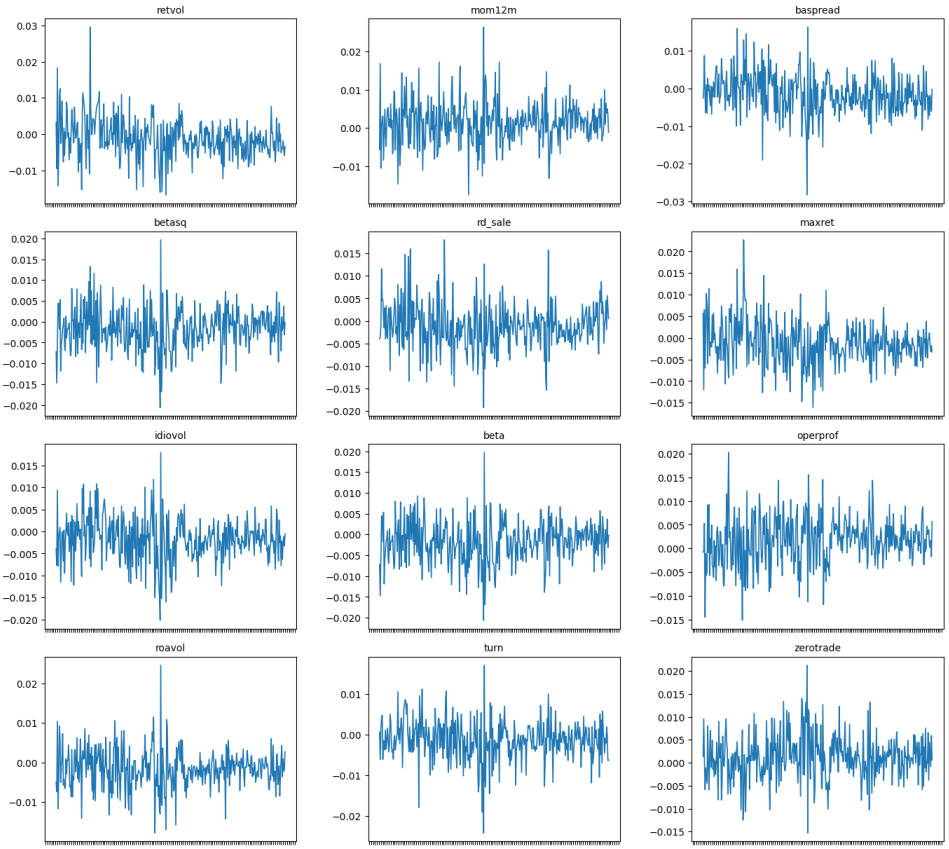
F.5 Summary Statistics of the Dataset

We provide an overview of the long–short firm characteristic covariates used in the analysis, their time-series behavior, and cross-sectional dependence as well as a brief summary of

the stochastic discount factor (SDF) and decile portfolios from Chen et al. (2024). Recall that all of the long-short characteristic portfolios are computed at the daily frequency, for the subsequent summary plots, we have aggregated them into the monthly frequency using within-month averages, in line with the standard practice of aligning signals with monthly returns. The monthly aggregation smooths out day-to-day noise and highlights the economically relevant medium-horizon variations.

Monthly Evolution of Covariates. Figure 10 displays the time series of monthly mean values for the twelve most volatile covariates, ranked by their total-sample standard deviation. The figure highlights that variables such as `retvol`, `mom12m`, and `baspread` exhibit pronounced month-to-month fluctuations, while others such as `turn` and `operprof` remain relatively stable. These series reveal persistent heteroskedasticity and regime shifts over time, particularly during market dislocations such as the early 2000s and 2008 crises.

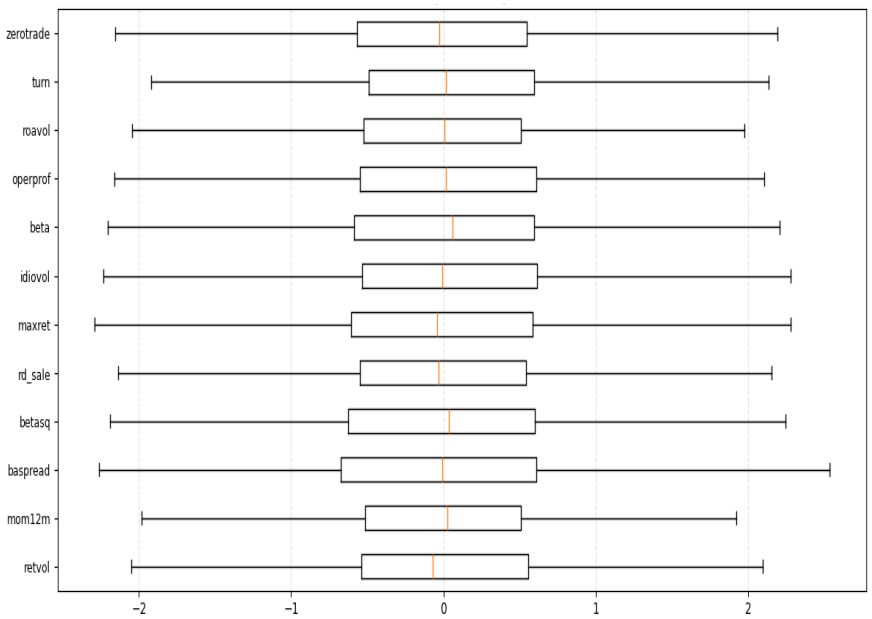
Figure 10: Monthly Means of the 12 Most Volatile Covariates.



Each panel shows the monthly mean of a long–short firm characteristic. Covariates are ranked by total-sample volatility. The time series reveal which characteristics exhibit the greatest month-to-month variation and long-run persistence.

Distributional and Correlation Structure. Figure 11 summarizes the time-series distributions of the same twelve covariates using standardized (z-scored) monthly values. The median, interquartile range, and whiskers capture the magnitude and symmetry of fluctuations across time. Most variables display near-zero median values but differ in dispersion and tail behavior, consistent with heterogeneous economic mechanisms underlying each characteristic.

Figure 11: Distributions (Boxplots) of Standardized Monthly Covariates.



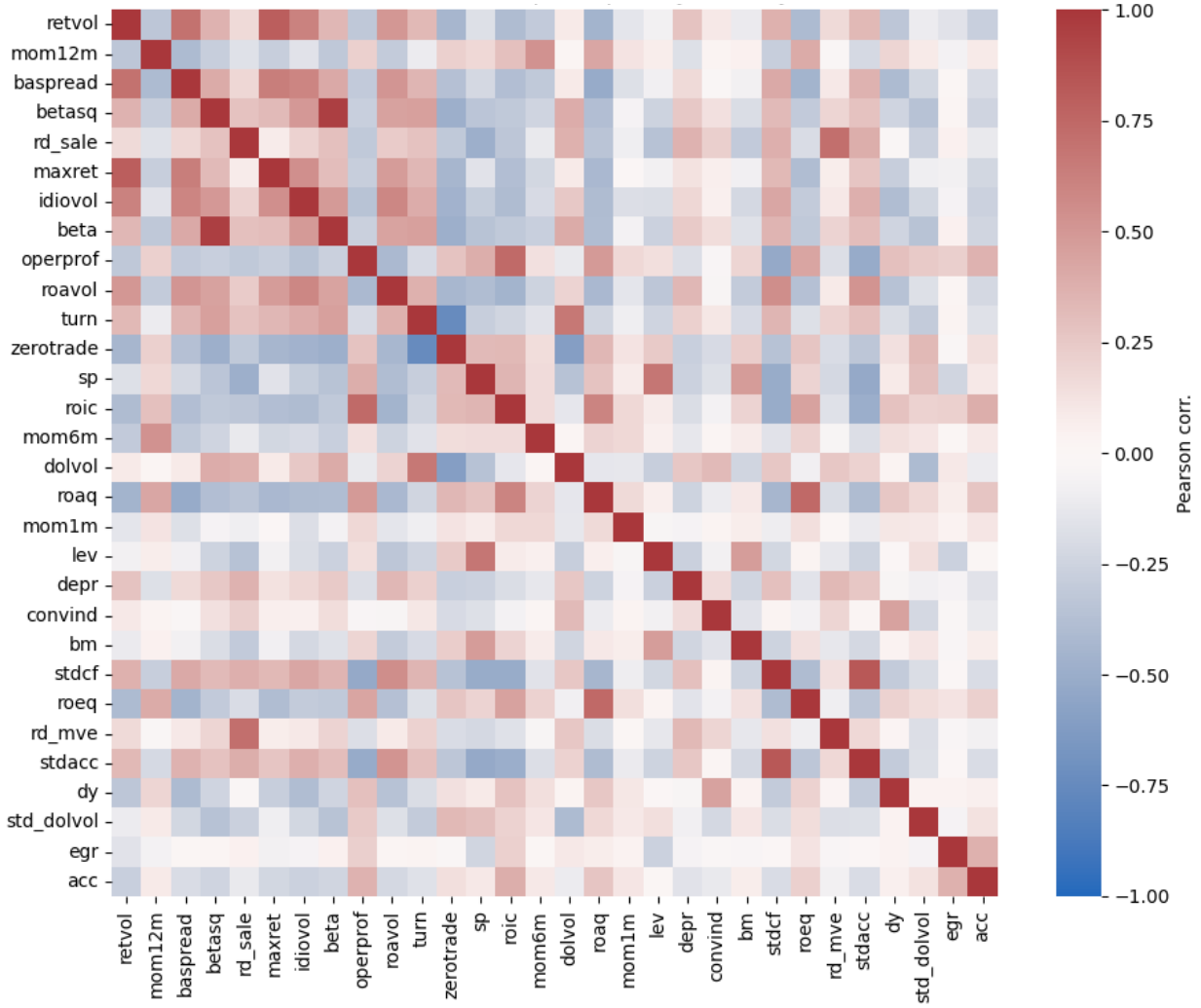
Z-scored monthly series for the twelve most volatile covariates. The figure compares dispersion and tail behavior across characteristics, highlighting differences in amplitude and symmetry.

For example, variables such as `retvol`, `baspread`, and `mom12m` exhibit wide interquartile ranges and thick tails, suggesting that these signals experience substantial time variation and occasional extreme realizations. In contrast, variables such as `turn` and `operprof` have narrower boxes, implying greater stability through time.

Each distribution is constructed by pooling monthly observations over the entire sample period for that specific covariate. This provides a concise view of the temporal heterogeneity and persistence of each characteristic after accounting for scale differences. The figure thus complements the time-series plots in Figure 10 by providing a scale-free summary of long-run variability and skewness in the underlying long–short characteristics.

The pairwise dependence structure among the top thirty covariates (in terms of volatility) is visualized in Figure 12. The heatmap reveals clusters of strongly correlated signals, such as volatility-related measures (`retvol`, `idiovol`, `roavol`) and liquidity-related variables (`baspread`, `zerotrade`, `turn`). The presence of such correlation blocks indicates there could be shared economic channels.

Figure 12: Correlation Heatmap of the 30 Most Volatile Covariates.

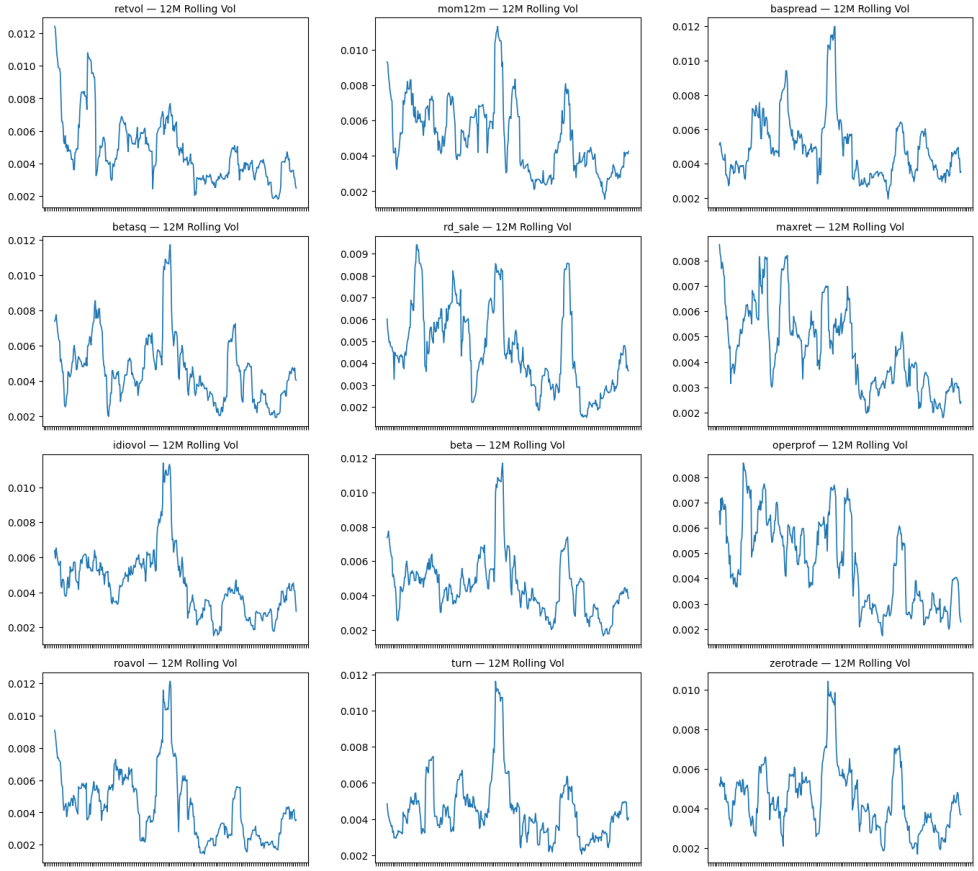


The matrix shows pairwise Pearson correlations between the thirty most volatile monthly covariates. Red indicates positive correlation and blue indicates negative correlation. Distinct blocks suggest clusters of related characteristics.

Time-Varying Volatility of Covariates. Figure 13 plots the 12-month rolling standard deviation of the twelve most volatile covariates. Unlike Figure 10, which ranks variables by overall volatility, the rolling volatility tracks how the variability of each covariate evolves through time. Periods such as the dot-com bubble and the global financial crisis correspond to distinct spikes in volatility across multiple signals, indicating that the informational

strength and instability of certain factors are regime-dependent.

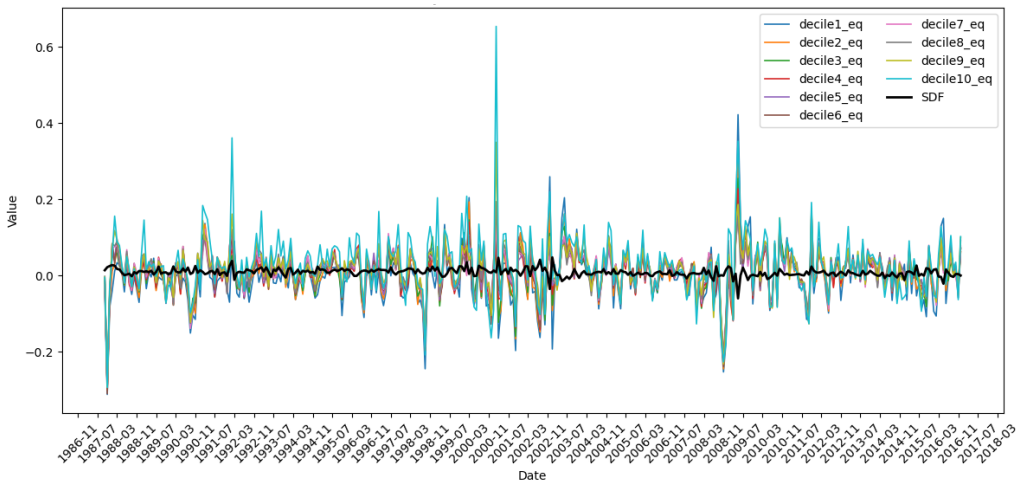
Figure 13: Twelve-Month Rolling Volatility of the Most Volatile Covariates.



The panels show rolling standard deviations computed using a 12-month moving window for each of the twelve most volatile covariates. This highlights temporal variation in the stability and amplitude of the long-short signals.

SDF and Decile Portfolios. Finally, Figure 14 plots the monthly time series of the stochastic discount factor (SDF) alongside the ten equal-weighted decile portfolios sorted by the underlying characteristic. The decile portfolios exhibit substantial comovement, with the SDF (shown in black) fluctuating more smoothly. This figure provides a benchmark for comparing the magnitude and temporal alignment of the SDF with characteristic-sorted portfolio returns, and serves as a diagnostic for whether the constructed SDF captures systematic components of asset pricing variation.

Figure 14: Monthly SDF and Decile Portfolios.



The black line represents the stochastic discount factor (SDF), while the colored lines correspond to ten equal-weighted decile portfolios sorted on firm characteristics. The comovement between the SDF and the characteristic-sorted portfolios provides an initial indication of factor relevance.

G Additional Experiment Results

G.1 Additional Results for Regime-Switching Model

In this section, we show that the regimes learned by the regime-switching model capture meaningful market variations. To do this, we will examine the correlation between the learned regimes and CBOE Volatility Index (VIX)[18].

Specifically, for each industry, for each month t, the regime-switching model estimates model parameters, including the regime transition matrix, using data available up to month $(t-1)$. Then, for each day s, we recursively compute the daily *filtered regime probability*, which is the estimated regime distribution given the data up to day s:

$$\widehat{p}_s(z) = \widehat{\mathbb{P}}(z_s = z \mid \mathcal{G}_s), \quad \forall z \in \{1, 2\},$$

where $\mathcal{G}_s$ denote all data points up to and including day s. Then, we compute the monthly

[18] https://fred.stlouisfed.org/series/VIXCLS

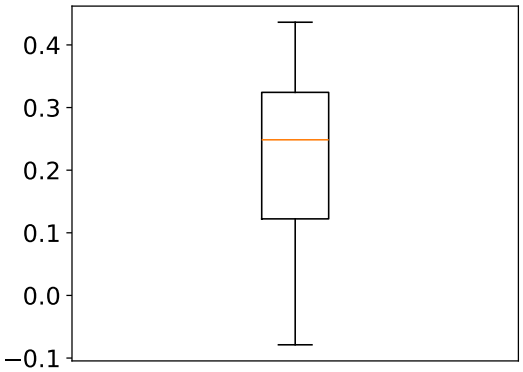
prevalence of state 2 in month t:

$$S_t = \frac{1}{B_t} \sum_{s \in B_t} \mathbf{1} \left\{ \widehat{p}_s(2) > \widehat{p}_s(1) \right\}.$$

The variation of S_t over time reflects changes in the learned regimes.

We compute the Pearson correlation between $\{S_t\}$ with the monthly averages of CBOE Volatility Index (VIX), from January 1990 to November 2016. Figure 15 shows a box plot across the 17 industries. In particular, the 25th and 75th percentiles of the 17 correlations are 0.12 and 0.32, respectively. Moreover, across the 17 industries, the raw p-values for the correlation between the estimated regimes and the VIX have 70th percentile equal to 0.0015. Since 17 correlations are tested, Bonferroni correction at the 0.05 level requires $p < 0.05/17 \approx 0.0029$. Thus, 70% of the correlations remain statistically significant. This shows that across most industries, there are positive correlations between the learned regimes and the VIX, and thus the regime-switching model captures meaningful market variations.

Figure 15: Box Plot of Pearson Correlation between Learned Industry Regimes and VIX.



This figure describes the distribution of the Pearson correlation between each industry's learned regimes and the VIX. For each of the 17 industries, we compute the prevalence of state 2 in each month, and calculate its correlation with the VIX. The box plot is over the 17 correlations, each from one industry.

G.2 Experiment Results with the Standard R^2 Metric

In Section 5, we evaluated predictive performance using the zero-benchmark R^2 to avoid the noise inherent in historical mean estimation. For completeness, this section reports the corresponding results using the standard out-of-sample R^2 metric, which benchmarks model performance against the historical sample mean. Qualitatively, the relative performance among models remains consistent with the the observations in Section 5: our adaptive algorithm `ATOMS` continues to outperform fixed-window benchmarks. Quantitatively, we observe that the standard R^2 values are generally lower than their zero-benchmark counterparts.

Table 13 presents out-of-sample standard R^2 values of `ATOMS` and baselines across distinct economic regimes, serving as the counterpart to Table 3 in the main text.

Table 13: OOS Standard R^2 Averages Across Industries by Time Period.

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	0.041	−0.019	0.115	0.039
`Fixed-val`(32)	0.013	−0.038	0.085	−0.003
`Fixed-val`(128)	0.032	−0.080	0.113	0.034
`Fixed-val`(512)	0.034	−0.080	0.107	0.037
`FH-CV`	0.026	−0.056	0.060	0.012
`Regime-switch`	−0.035	−0.508	0.070	−0.010

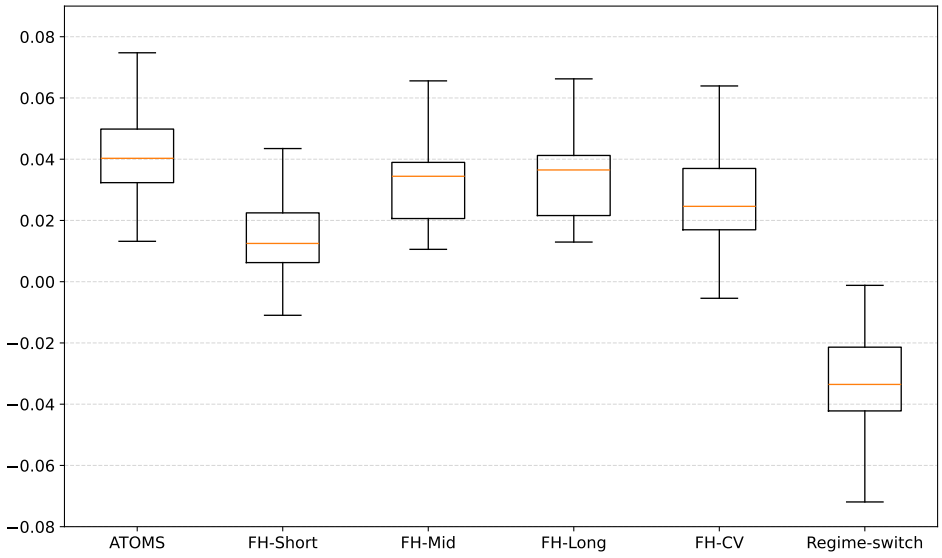
This table reports OOS standard R^2 averages for return prediction models across all 17 industry portfolios. Full OOS Period refers to OOS period covering 01/1990∼11/2016. Columns report OOS R^2 averages across all industries and highlight this metric during three recessions, as documented in NBER Business Cycle Dating:

- the 1990 Gulf War recession (06/1990∼10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001∼10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007∼06/2009).

That is, the OOS performance in Gulf War column focuses on model performance comparisons exclusively in the out-of-sample period of 06/1990∼10/1990. All values are calculated using monthly return data.

Figure 16 gives a box plot of the OOS standard R^2 of `ATOMS` and the fixed-window baselines over the 17 industry portfolios, mirroring Figure 4.

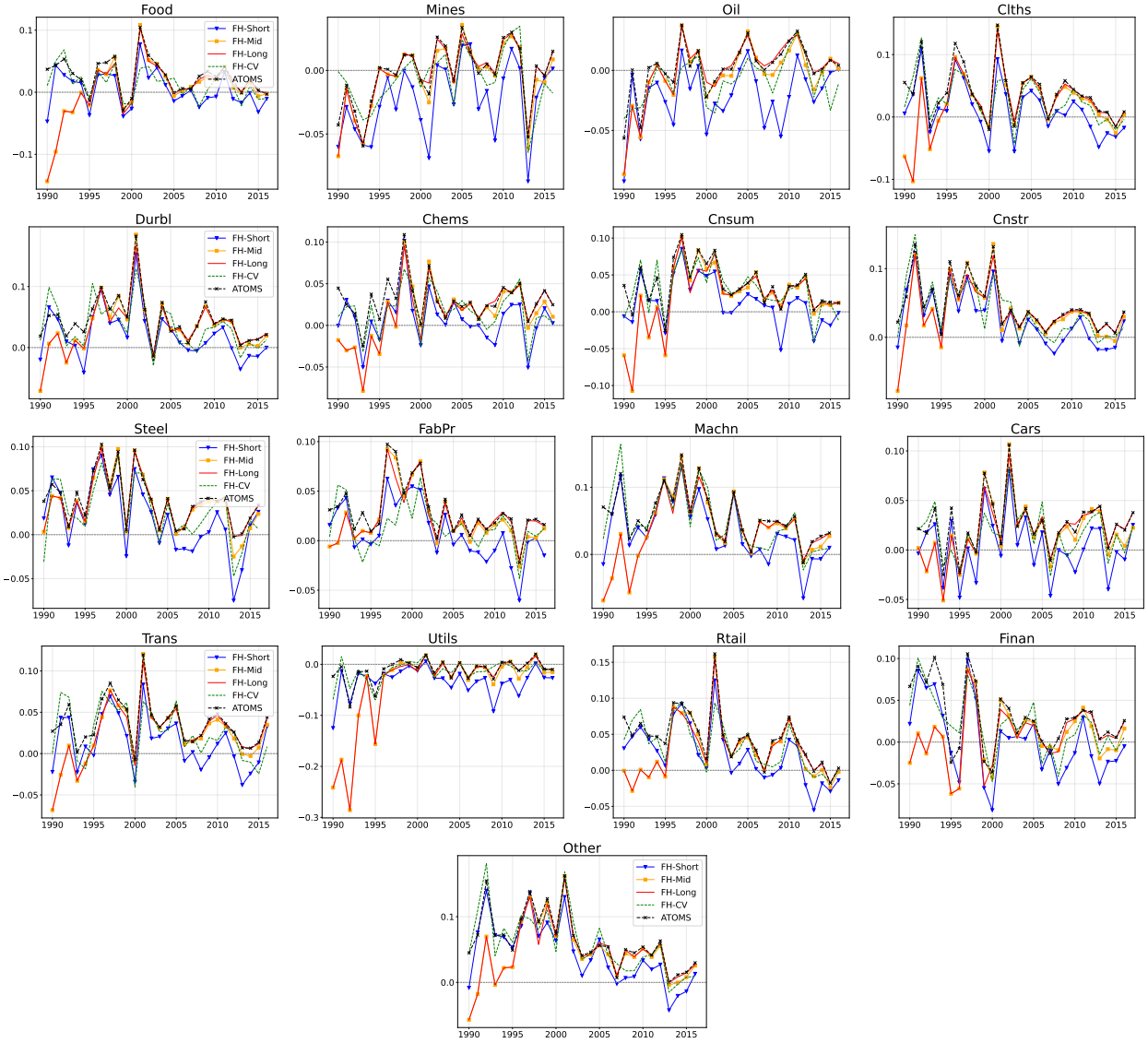
Figure 16: Box Plot of OOS Standard R^2 of `ATOMS` and Benchmarks for 17 Industry Portfolios.



This figure describes the distribution of each method's OOS R^2. Each box corresponds to all industries and all years in our OOS horizon.

Finally, Figure 17 plots the annual out-of-sample R^2 for the 17 industry portfolios, paralleling Figure 5.

Figure 17: Annual OOS Standard R^2 of Different Approaches for 17 Industry Portfolios.



This figure reports the annual OOS standard R^2 of our adaptive model selection algorithm `ATOMS` (black dashed line with ×'s), as well as the fixed-window baselines `Fixed-val`(32) (blue ▼'s), `Fixed-val`(128) (orange ■'s), and `Fixed-val`(512) (red), which use the last 32, last 128 and all months of validation data. The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 10.

G.3 Additional Maximum Drawdown Results

This section reports maximum drawdown results for alternative look-back reference windows of $S = 1, 2, 6$ months.

Results: Mean-Variance Portfolio. Tables 14 to 16 report the maximum drawdown results for the mean-variance portfolio, with reference windows of $S = 1, 2, 6$ months, respectively. The drawdowns are similar to those computed with a reference window of $S = 3$ months, presented in Table 6 of the main text.

Table 14: Maximum Drawdown of Mean-Variance Portfolio with One-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	3.0%	0.7%	0.6%	2.4%
`FH-Short`	4.3%	0.8%	0.9%	3.5%
`FH-Mid`	3.3%	0.9%	0.7%	2.4%
`FH-Long`	3.1%	0.9%	0.6%	2.7%
`FH-CV`	4.4%	0.8%	0.9%	4.2%
`Regime-switch`	2.7%	1.2%	0.7%	2.7%

This table reports maximum drawdowns for mean-variance portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a one-month look-back reference window ($S = 1$ months). The `FH` benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

Results: Signed Portfolio. Tables 17 to 19 report the maximum drawdown results for the signed portfolio, with reference windows of $S = 1, 2, 6$ months, respectively. The drawdowns are similar to those computed with a reference window of $S = 3$ months, presented in Table 9 of the main text.

Table 15: Maximum Drawdown of Mean-Variance Portfolio with Two-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	3.0%	0.7%	0.6%	2.4%
FH-Short	4.3%	0.9%	0.9%	3.5%
FH-Mid	3.3%	0.9%	0.7%	2.4%
FH-Long	3.1%	0.9%	0.6%	2.7%
FH-CV	4.4%	0.8%	0.9%	4.2%
Regime-switch	2.7%	1.2%	0.8%	2.7%

This table reports maximum drawdowns for mean-variance portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a two-month look-back reference window ($S = 2$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

Table 16: Maximum Drawdown of Mean-Variance Portfolio with Six-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	3.0%	0.7%	0.6%	2.4%
FH-Short	4.3%	1.0%	1.0%	3.5%
FH-Mid	3.3%	1.0%	0.7%	2.4%
FH-Long	3.1%	1.0%	0.6%	2.7%
FH-CV	4.4%	0.9%	0.9%	4.2%
Regime-switch	2.7%	1.2%	0.8%	2.7%

This table reports maximum drawdowns for mean-variance portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a six-month look-back reference window ($S = 6$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

Table 17: Maximum Drawdown of Signed Portfolio with One-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	16.3%	1.9%	2.4%	16.3%
FH-Short	13.6%	2.1%	2.1%	12.8%
FH-Mid	15.2%	2.7%	2.4%	15.1%
FH-Long	16.5%	2.7%	2.3%	16.5%
FH-CV	18.4%	2.4%	1.9%	18.0%
Regime-switch	13.0%	1.2%	1.6%	13.0%
EW	48.8%	17.8%	14.2%	46.9%

This table reports maximum drawdowns for signed portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a one-month look-back reference window ($S = 1$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

Table 18: Maximum Drawdown of Signed Portfolio with Two-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	16.3%	1.9%	2.4%	16.3%
FH-Short	13.6%	2.1%	2.1%	12.8%
FH-Mid	15.2%	2.7%	2.4%	15.1%
FH-Long	16.5%	2.7%	2.3%	16.5%
FH-CV	18.4%	2.4%	1.9%	18.0%
Regime-switch	13.0%	1.2%	1.6%	13.0%
EW	48.8%	17.8%	14.2%	46.9%

This table reports maximum drawdowns for signed portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a two-month look-back reference window ($S = 2$ months). The FH benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.

Table 19: Maximum Drawdown of Signed Portfolio with Six-Month Reference Window

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
`ATOMS`	16.3%	1.9%	2.4%	16.3%
`FH-Short`	13.6%	2.1%	2.1%	12.8%
`FH-Mid`	15.2%	2.7%	2.4%	15.1%
`FH-Long`	16.5%	2.7%	2.3%	16.5%
`FH-CV`	18.4%	2.4%	1.9%	18.0%
`Regime-switch`	13.0%	1.2%	1.6%	13.0%
EW	48.8%	17.8%	14.2%	48.8%

This table reports maximum drawdowns for signed portfolios. Full OOS Period refers to the full out-of-sample period, 01/1990∼11/2016. Recession columns report conditional drawdowns computed within each recession window using a six-month look-back reference window ($S = 6$ months). The `FH` benchmarked methods use fixed windows as described in Table 2. Smaller values indicate smaller peak-to-trough losses.