# When More Data Hurts: Navigating the Nonstationarity-Complexity Tradeoff in Return Prediction

**Agostino Capponi**

COLUMBIA ENGINEERING

The Center for Digital Finance and Technologies

Based on joint work with
Chengpiao Huang, Antonio Sidaoui, Kaizheng Wang, Jiacheng Zou

# Contents

# Machine Learning for Return Prediction

# Stock Return Prediction

**Objective:** Given risk characteristics $x$, predict stock return $y$.

**Standard approach:** Assume

$$y = f^*(x) + \varepsilon.$$

Build a prediction model:

$$\text{data } \{(x_i, y_i)\}_{i=1}^{t-1} \; + \; \text{model class } \mathcal{F} \; \rightsquigarrow \; \text{prediction model } \widehat{f} \in \mathcal{F}$$

Which model class $\mathcal{F}$? (e.g., linear model vs neural network)

**Challenge: the data distribution is often changing over time!**

# Learning with Non-Stationary Data

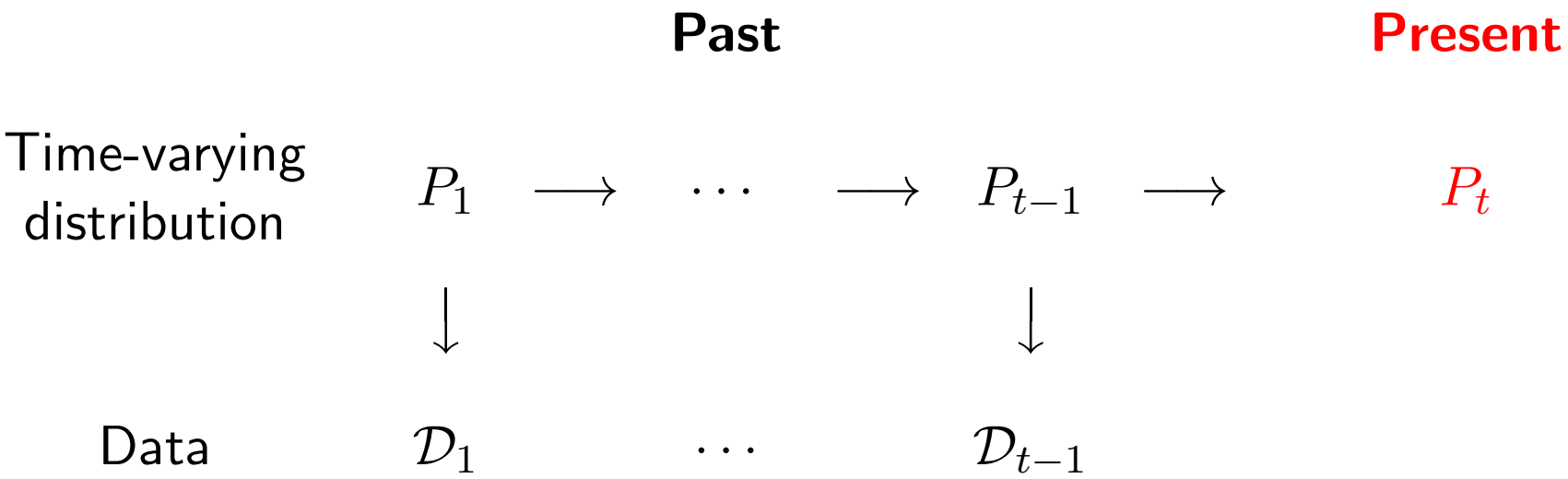
| | **Past** | | | | | **Present** |
|---|---|---|---|---|---|---|
| Time-varying distribution | $P_1$ | $\longrightarrow$ | $\cdots$ | $\longrightarrow$ | $P_{t-1}$ | $\longrightarrow$ $P_t$ |

# Learning with Non-Stationary Data

| | **Past** | | | | | **Present** |
|---|---|---|---|---|---|---|
| Time-varying distribution | $P_1$ | $\longrightarrow$ | $\cdots$ | $\longrightarrow$ | $P_{t-1}$ | $\longrightarrow$ $P_t$ |

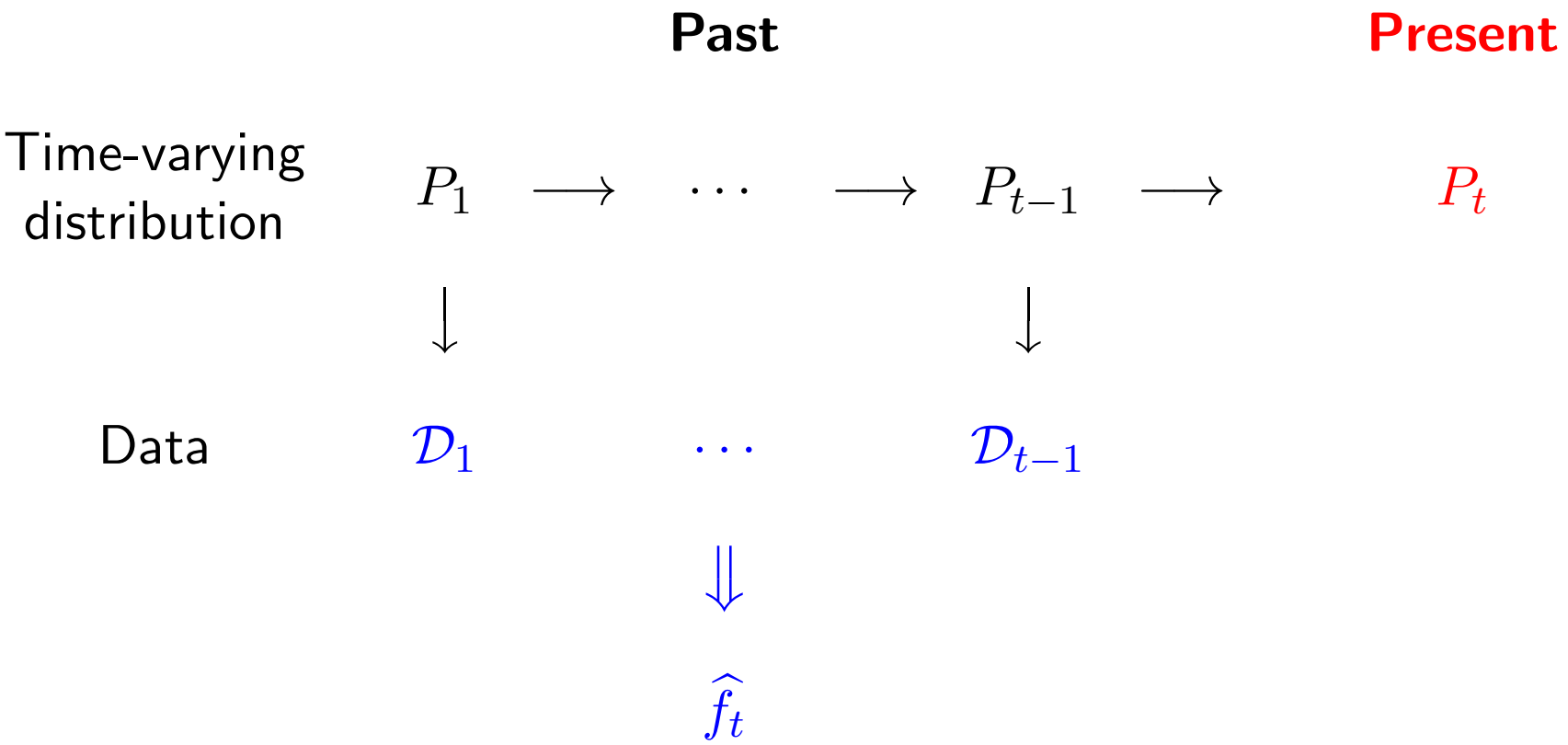
- Each sample $(x, y) \sim P_t$ satisfies $y = f_t^*(x) + \varepsilon$.

# Learning with Non-Stationary Data



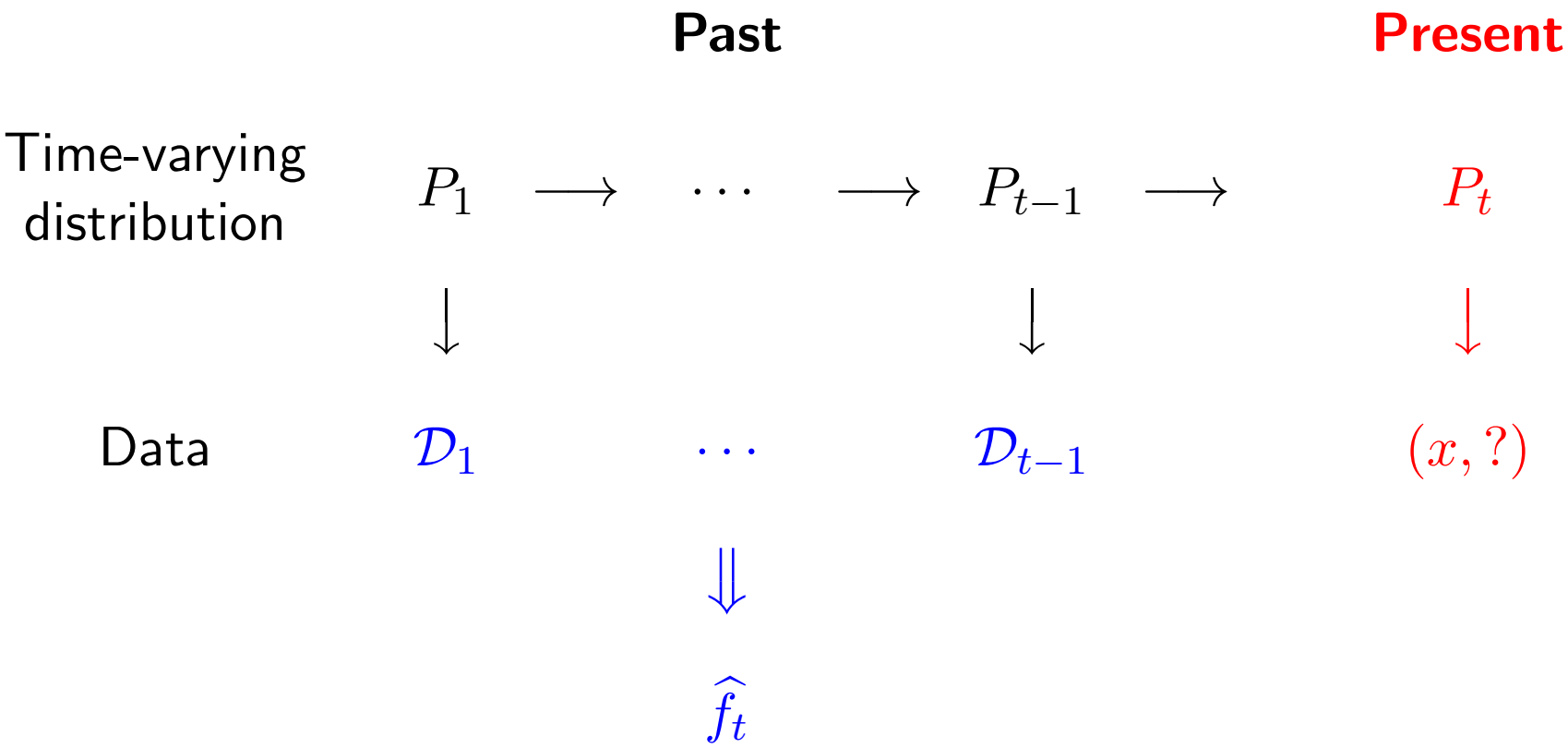

- Each sample $(x, y) \sim P_t$ satisfies $y = f_t^*(x) + \varepsilon$.

# Learning with Non-Stationary Data



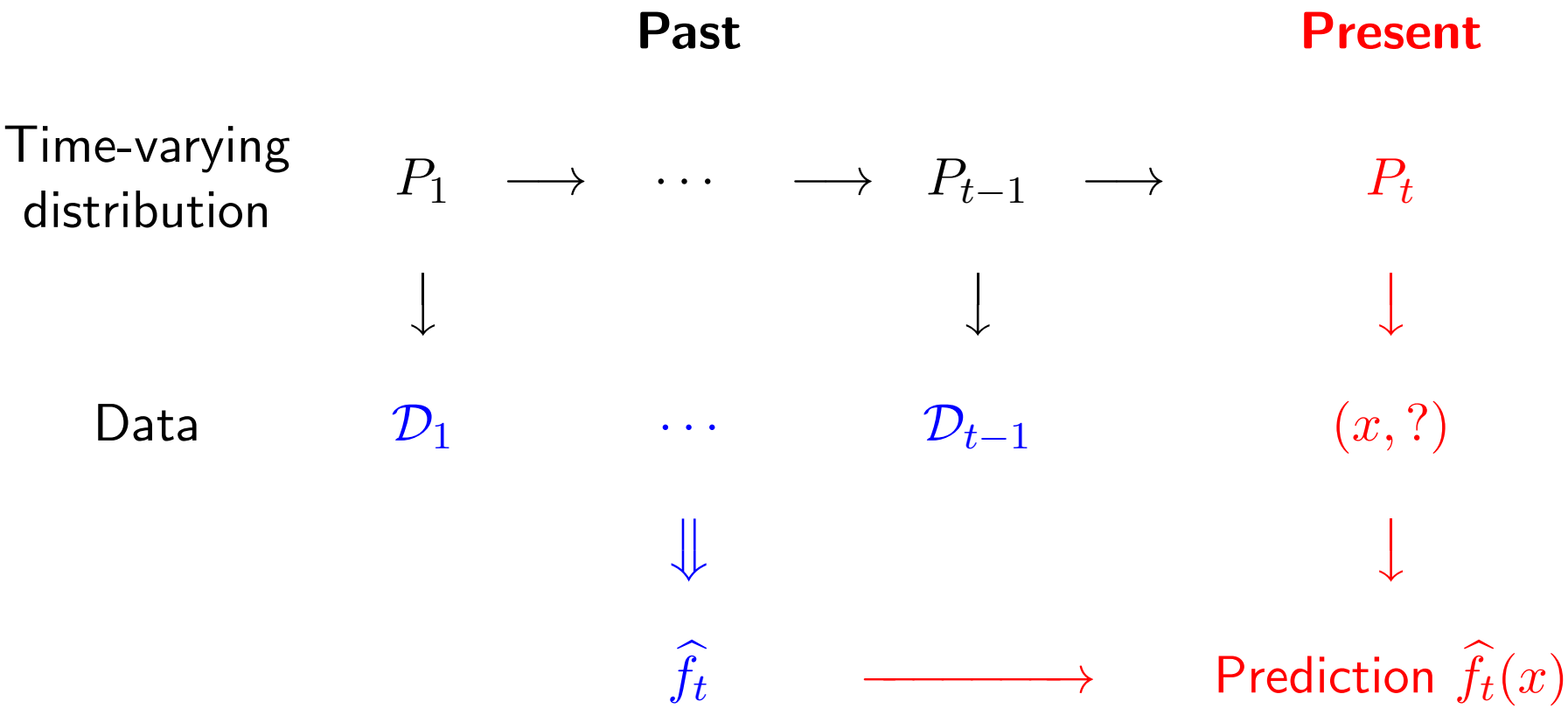

- Each sample $(x, y) \sim P_t$ satisfies $y = f_t^*(x) + \varepsilon$.

# Learning with Non-Stationary Data




- Each sample $(x, y) \sim P_t$ satisfies $y = f_t^*(x) + \varepsilon$.

# Learning with Non-Stationary Data




- Each sample $(x, y) \sim P_t$ satisfies $y = f_t^*(x) + \varepsilon$.

# Model Class and Training Window Selection

**Challenges:**

- Choice of model class $\mathcal{F}$: linear model vs neural network?
- Training window size: 3 months vs 1 year?

**Possibilities:**

- Use a simple model $\rightsquigarrow$ misspecification
- Use a complex model
  - Short training window $\rightsquigarrow$ low distributional mismatch but high variance
  - Long training window $\rightsquigarrow$ low variance but non-stationarity

# The Nonstationarity-Complexity Tradeoff

# When Does Complexity Help? Role of Non-Stationarity

**Virtue of Complexity**
(Stationary world)

All data equally valuable

↓

Use everything available

↓

**Complexity helps**

Prediction Error $\approx$

**Bias** + **Variance**

**Our Framework**
(Non-stationary world)

Data relevance varies

↓

Select training window

↓

**Non-stationarity matters**

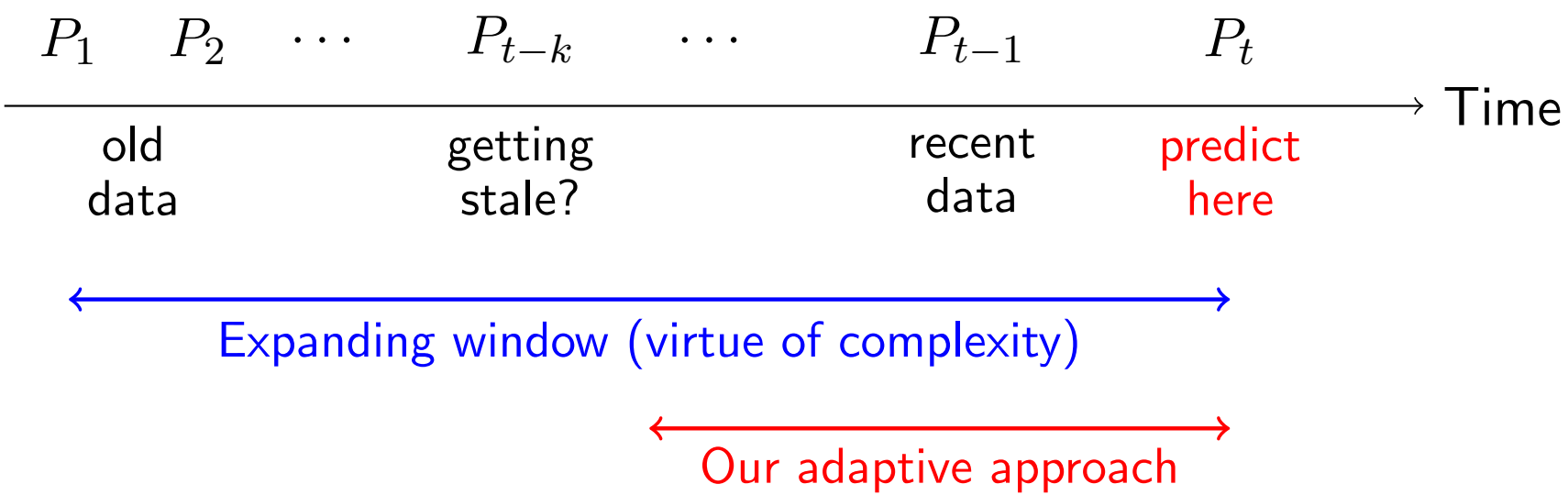
Prediction Error $\approx$

**Bias** + **Variance**

+ **Non-stationarity**

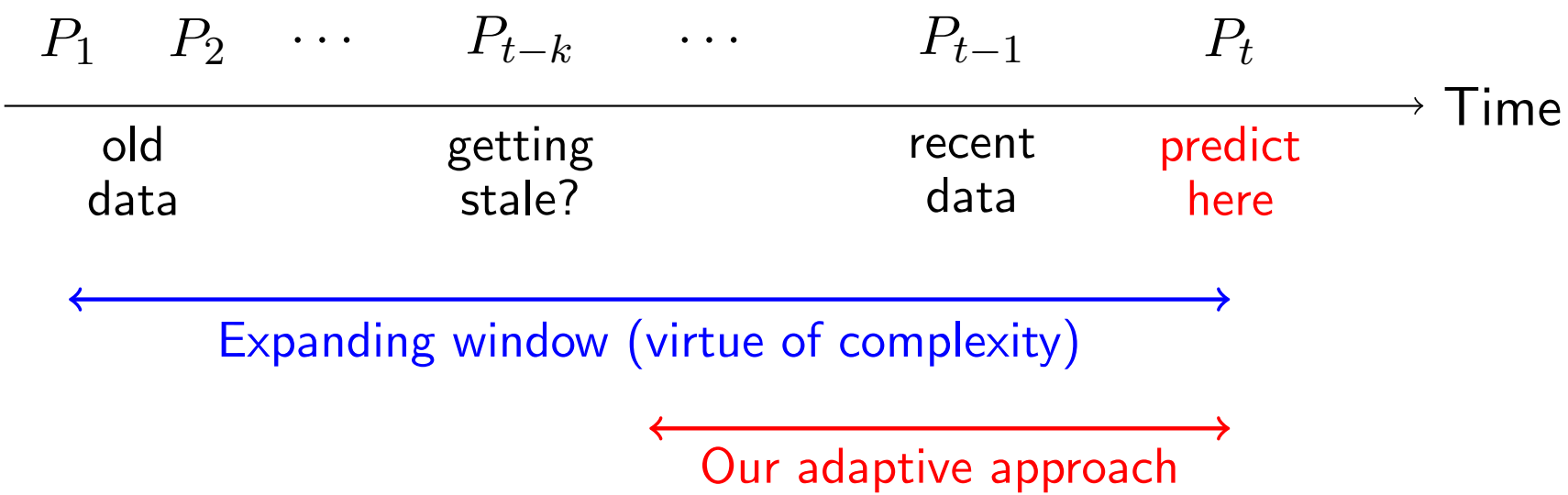
# The Nonstationarity-Complexity Tradeoff

**Expanding Windows May Hurt When Distributions Shift**



# The Nonstationarity-Complexity Tradeoff

**Expanding Windows May Hurt When Distributions Shift**




**Complements virtue of complexity:** Complexity helps *when paired with appropriate training data*

# Setting

Consider training a model in $\mathcal{F}$ using data from the last $k$ periods:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{(x,y) \in \mathcal{D}_j^{\mathrm{tr}}} |f(x) - y|^2 \right\}.$$

where $n_{t,k}^{\mathrm{tr}}$ is the number of training samples in the last $k$ periods.

# Setting

Consider training a model in $\mathcal{F}$ using data from the last $k$ periods:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{(x,y) \in \mathcal{D}_j^{\mathrm{tr}}} |f(x) - y|^2 \right\}.$$

where $n_{t,k}^{\mathrm{tr}}$ is the number of training samples in the last $k$ periods.

# Setting

Consider training a model in $\mathcal{F}$ using data from the last $k$ periods:

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{n_{t,k}^{\mathrm{tr}}} \sum_{j=t-k}^{t-1} \sum_{(x,y) \in \mathcal{D}_j^{\mathrm{tr}}} |f(x) - y|^2 \right\}.$$

where $n_{t,k}^{\mathrm{tr}}$ is the number of training samples in the last $k$ periods.

We characterize the out-of-sample (OOS) performance of $\widehat{f}$ at time $t$:

$$L_t(\widehat{f}) = \mathbb{E}_{(x,y) \sim P_t} \left| \widehat{f}(x) - y \right|^2,$$

compared against $L_t(f^*)$, where the oracle ${f_t}^*(x) = \mathbb{E}_{(x,y) \sim P_t}[y \mid x]$.

# Model Performance Bound

## Theorem 1

*Assume the data points* $(x, y) \in \mathcal{D}_1^{\text{tr}}, ..., \mathcal{D}_{t-1}^{\text{tr}}$ *are independent.*

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \textit{misspecification} + \textit{variance} + \textit{non-stationarity}$$

# Model Performance Bound

## Theorem 1

*Assume the data points* $(x, y) \in \mathcal{D}_1^{\text{tr}}, ..., \mathcal{D}_{t-1}^{\text{tr}}$ *are independent.*

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \textit{misspecification} + \textit{variance} + \textit{non-stationarity}$$

# Model Performance Bound

## Theorem 1

*Assume the data points* $(x, y) \in \mathcal{D}_1^{\text{tr}}, ..., \mathcal{D}_{t-1}^{\text{tr}}$ *are independent.*

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \textit{ misspecification } + \textit{ variance } + \textit{ non-stationarity}$$

- Misspecification: $\min_{f \in \mathcal{F}} L_t(f) - L_t(f_t^*)$; the limitation of model class $\mathcal{F}$.

# Model Performance Bound

## Theorem 1

*Assume the data points* $(x, y) \in \mathcal{D}_1^{\text{tr}}, ..., \mathcal{D}_{t-1}^{\text{tr}}$ *are independent.*

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \textit{misspecification} + \textit{variance} + \textit{non-stationarity}$$

- Misspecification: $\min_{f \in \mathcal{F}} L_t(f) - L_t(f_t^*)$; the limitation of model class $\mathcal{F}$.
- Variance of fitting $\widehat{f} \in \mathcal{F}$ with samples from last $k$ periods.

  Bounded by local Rademacher complexity, typically $\left[\frac{\text{Comp}(\mathcal{F})}{n_{t,k}^{\text{tr}}}\right]^{\alpha}$ details
  - Finite class: $\log|\mathcal{F}|/n_{t,k}^{\text{tr}}$; linear models: $d/n_{t,k}^{\text{tr}}$;
  - Nonparametric class: $(n_{t,k}^{\text{tr}})^{-\gamma}$, $0 < \gamma < 1$.

# Model Performance Bound

## Theorem 1

*Assume the data points* $(x, y) \in \mathcal{D}_1^{\text{tr}}, ..., \mathcal{D}_{t-1}^{\text{tr}}$ *are independent.*

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \textit{misspecification} + \textit{variance} + \textit{non-stationarity}$$

- Misspecification: $\min_{f \in \mathcal{F}} L_t(f) - L_t(f_t^*)$; the limitation of model class $\mathcal{F}$.
- Variance of fitting $\widehat{f} \in \mathcal{F}$ with samples from last $k$ periods.

  Bounded by local Rademacher complexity, typically $\left[\frac{\text{Comp}(\mathcal{F})}{n_{t,k}^{\text{tr}}}\right]^{\alpha}$ details
- Non-stationarity: $\max_{t-k \le i \le t-1} \text{TV}(P_i, P_t)$ over the last $k$ periods.

# Interaction of Model Complexity and Training Window

| | Misspecification | Variance | Non-stationarity |
|---|---|---|---|
| Model complexity ↗ | ↘ | ↗ | – |
| Training window ↗ | – | ↘ | ↗ |

# Interaction of Model Complexity and Training Window

| | Misspecification | Variance | Non-stationarity |
|---|---|---|---|
| Model complexity ↗ | ↘ | ↗ | – |
| Training window ↗ | – | ↘ | ↗ |

Does more data always help?

# Interaction of Model Complexity and Training Window

| | Misspecification | Variance | Non-stationarity |
|---|---|---|---|
| Model complexity ↗ | ↘ | ↗ | – |
| Training window ↗ | – | ↘ | ↗ |

Does more data always help?

- Performance depends **jointly** on model complexity and training window.
- Greater model complexity or more training data $\not\Rightarrow$ better performance.

# Interaction of Model Complexity and Training Window

| | Misspecification | Variance | Non-stationarity |
|---|---|---|---|
| Model complexity ↗ | ↘ | ↗ | – |
| Training window ↗ | – | ↘ | ↗ |

Does more data always help?

- Performance depends **jointly** on model complexity and training window.
- Greater model complexity or more training data $\not\Rightarrow$ better performance.

**How to jointly select the optimal model class and training window?**

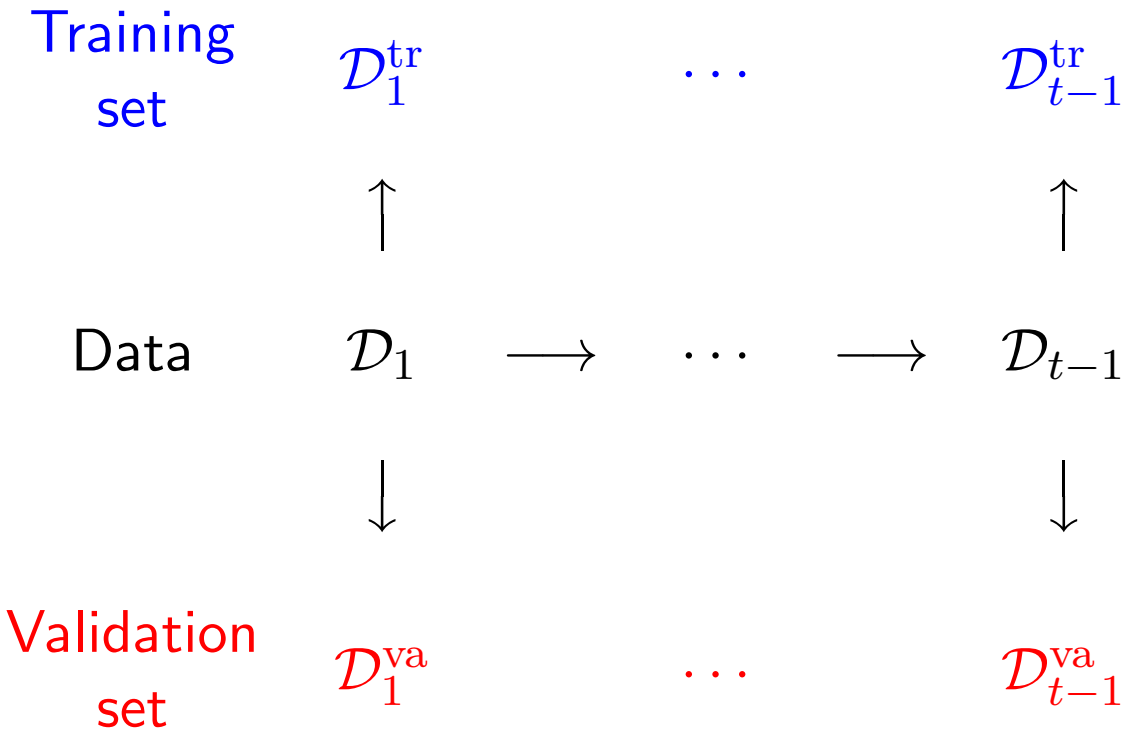
# Joint Selection of Model Class and Training Window

# Training and Validation Framework

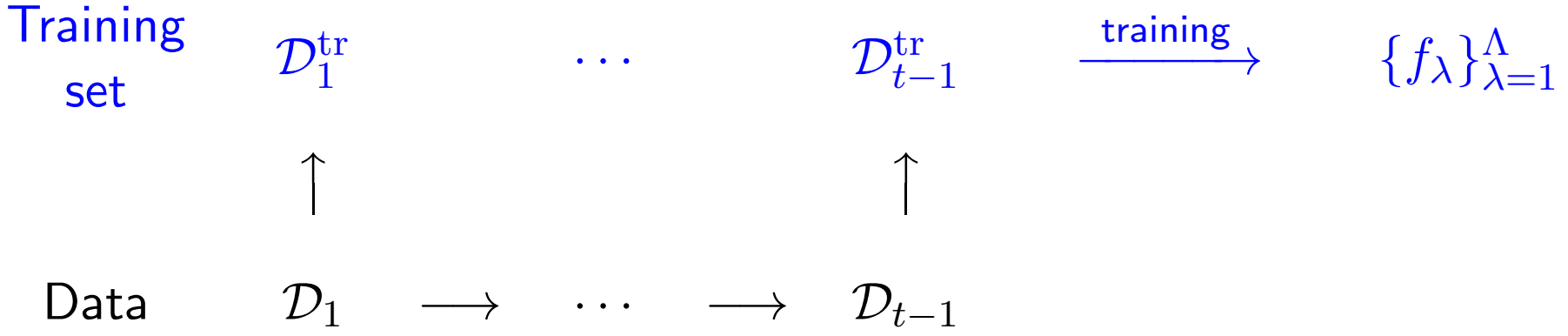
# Training and Validation Framework

Data $\mathcal{D}_1 \longrightarrow \cdots \longrightarrow \mathcal{D}_{t-1}$

# Training and Validation Framework

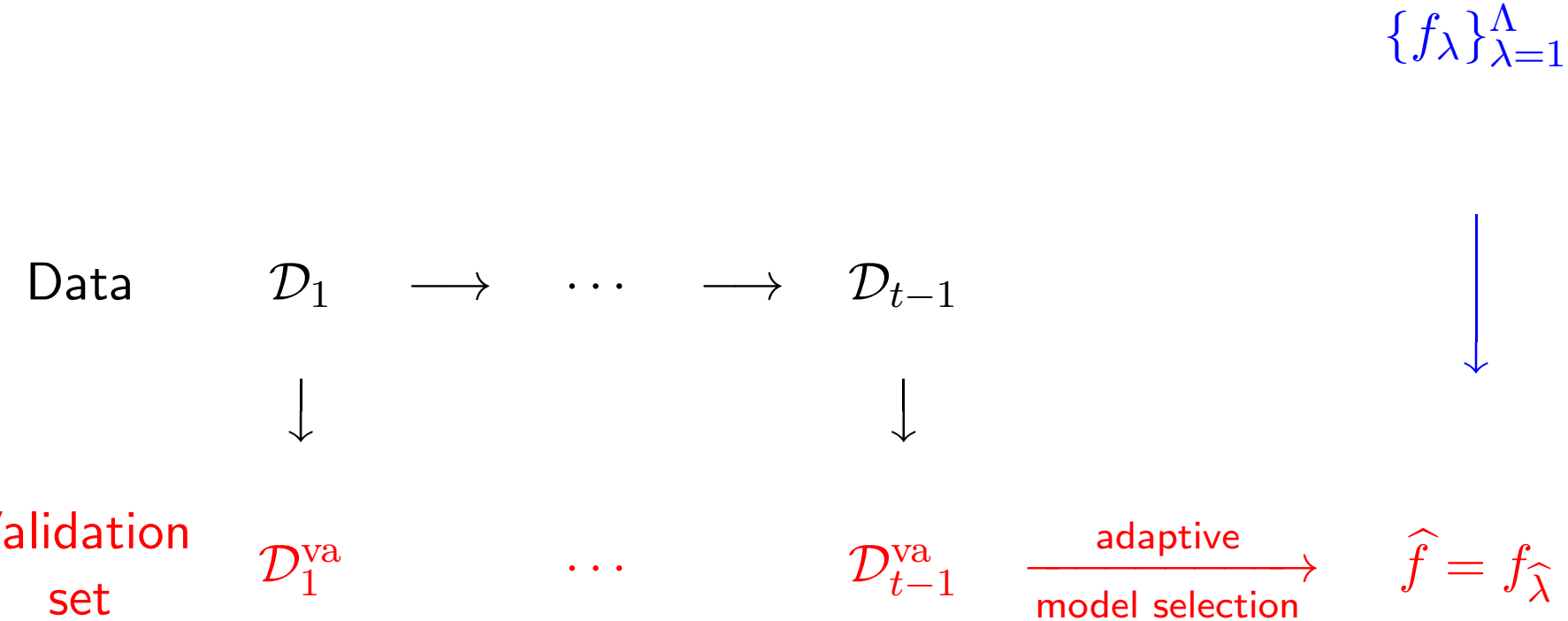


# Training and Validation Framework




**Example of $\{f_\lambda\}$:** $\{\text{linear model, random forest}\}$ trained on the most recent $\{1, 4, 16, 64, 256\}$ months of data.

# Training and Validation Framework



# Model Selection under Non-Stationarity

From now on, assume that the candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$ are given.

$$\begin{array}{ccccccc} P_1 & \longrightarrow & \cdots & \longrightarrow & P_{t-1} & \longrightarrow & P_t \\ \downarrow & & & & \downarrow & & \\ \mathcal{D}_1^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-1}^{\mathrm{va}} & & \end{array}$$

**Goal:** Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to select a good candidate model:

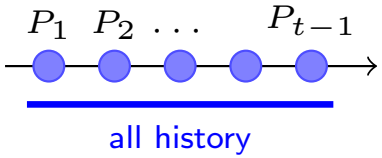
$$\widehat{\lambda} \approx \operatorname*{argmin}_{1 \le \lambda \le \Lambda} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

# How to Select the Best Candidate Model?

**Goal:** Use **past** data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$ to find the best model for **current** $P_t$.

## Approach 1

*Use all validation data*



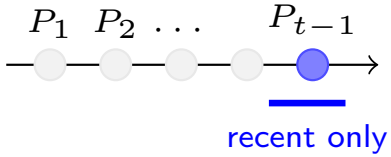

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda} \sum_{j=1}^{t-1} \sum_{\mathcal{D}_j^{\text{va}}} |f_\lambda - y|^2$$

**Problem:**

Old data from $P_1, P_2, \ldots$ may differ greatly from $P_t$ (bias)

## Approach 2

*Use only recent data*



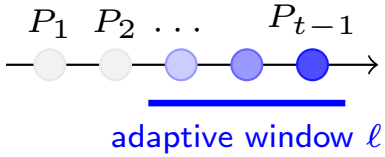

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda} \sum_{\mathcal{D}_{t-1}^{\text{va}}} |f_\lambda - y|^2$$

**Problem:**

Only one period of data too noisy, wastes available history (variance)

## ATOMS

*Adaptive selection*




Adaptively chooses $\ell$ balancing:

- **Recency:** relevance to $P_t$
- **Sample size:** statistical accuracy

Near-optimal, as if $P_t$ were known

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \underset{\lambda \in \{1,2\}}{\operatorname{argmin}} \, \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

If $\Delta_t > 0$, choose $f_2$. Otherwise choose $f_1$.

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

If $\Delta_t > 0$, choose $f_2$. Otherwise choose $f_1$.

**Challenge:** Validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ is non-stationary!

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\text{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\text{va}} & & \cdots & & \mathcal{D}_{t-1}^{\text{va}} & &
\end{array}
$$

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-1}^{\mathrm{va}} & &
\end{array}
$$

$$\Downarrow$$

Empirical performance gap

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-1}^{\mathrm{va}} & &
\end{array}
$$

$$
\Downarrow
$$

Empirical performance gap

$$
\widehat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}^{\mathrm{va}}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}_j^{\mathrm{va}}} \left( |f_1(x) - y|^2 - |f_2(x) - y|^2 \right)
$$

where $n_{t,\ell}^{\mathrm{va}}$ is the number of validation samples from the last $\ell$ periods.

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$P_1 \to \cdots \to P_{t-\ell} \to \cdots \to P_{t-1} \to P_t$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$\mathcal{D}_1^{\mathrm{va}} \quad \cdots \quad \mathcal{D}_{t-\ell}^{\mathrm{va}} \quad \cdots \quad \mathcal{D}_{t-1}^{\mathrm{va}}$$

$$\Downarrow$$

Empirical performance gap

$$\widehat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}^{\mathrm{va}}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}_j^{\mathrm{va}}} \left( |f_1(x) - y|^2 - |f_2(x) - y|^2 \right)$$

How to choose validation window $\ell$?

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell} \sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

- $B(\ell)$: bias due to non-stationarity ($\uparrow$)
- $V(\ell)$: variance due to finite samples ($\downarrow$)

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

- Construct $\widehat{V}(\ell) \approx V(\ell)$ using empirical variance. details

# Adaptive Validation Window Selection

## Lemma 2 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

- Construct $\widehat{V}(\ell) \approx V(\ell)$ using empirical variance. details
- How to estimate the non-stationarity bias $B(\ell)$?

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

Choose validation window $\widehat{\ell} = \operatorname{argmin}_\ell \left\{ \widehat{B}(\ell) + \widehat{V}(\ell) \right\}$. details

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

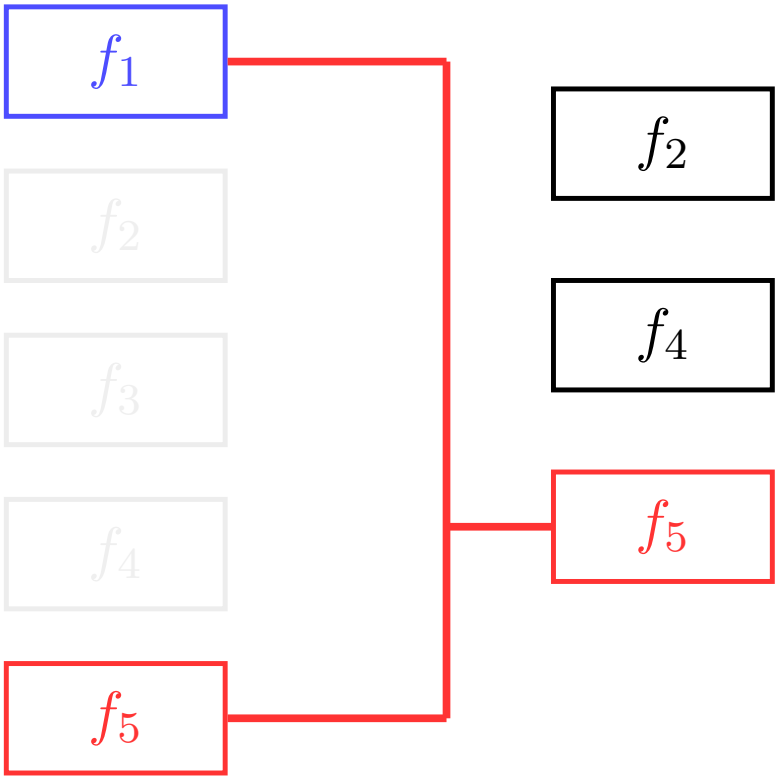
Choose validation window $\widehat{\ell} = \operatorname{argmin}_\ell \left\{ \widehat{B}(\ell) + \widehat{V}(\ell) \right\}$. details

## Theorem 3 (Near-Optimal Validation Window Selection)

*With high probablity,* $\left| \widehat{\Delta}_{t,\widehat{\ell}} - \Delta_t \right| \lesssim \min_\ell \left\{ B(\ell) + V(\ell) \right\}$.

# Adaptive Tournament Model Selection (ATOMS)

**Sequential elimination tournament** based on **pairwise comparison**:




For each pairwise comparison,

- Choose validation window $\widehat{\ell}$ to estimate performance gap $\widehat{\Delta}_{t,\widehat{\ell}}$.
- Compare based on sign of $\widehat{\Delta}_{t,\widehat{\ell}}$.

# Sequential Elimination Tournament

**Example:** We start with $5$ candidate models.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

# Sequential Elimination Tournament

**Round 1:** Pivot model: $f_1$.

$f_1$

$f_2$

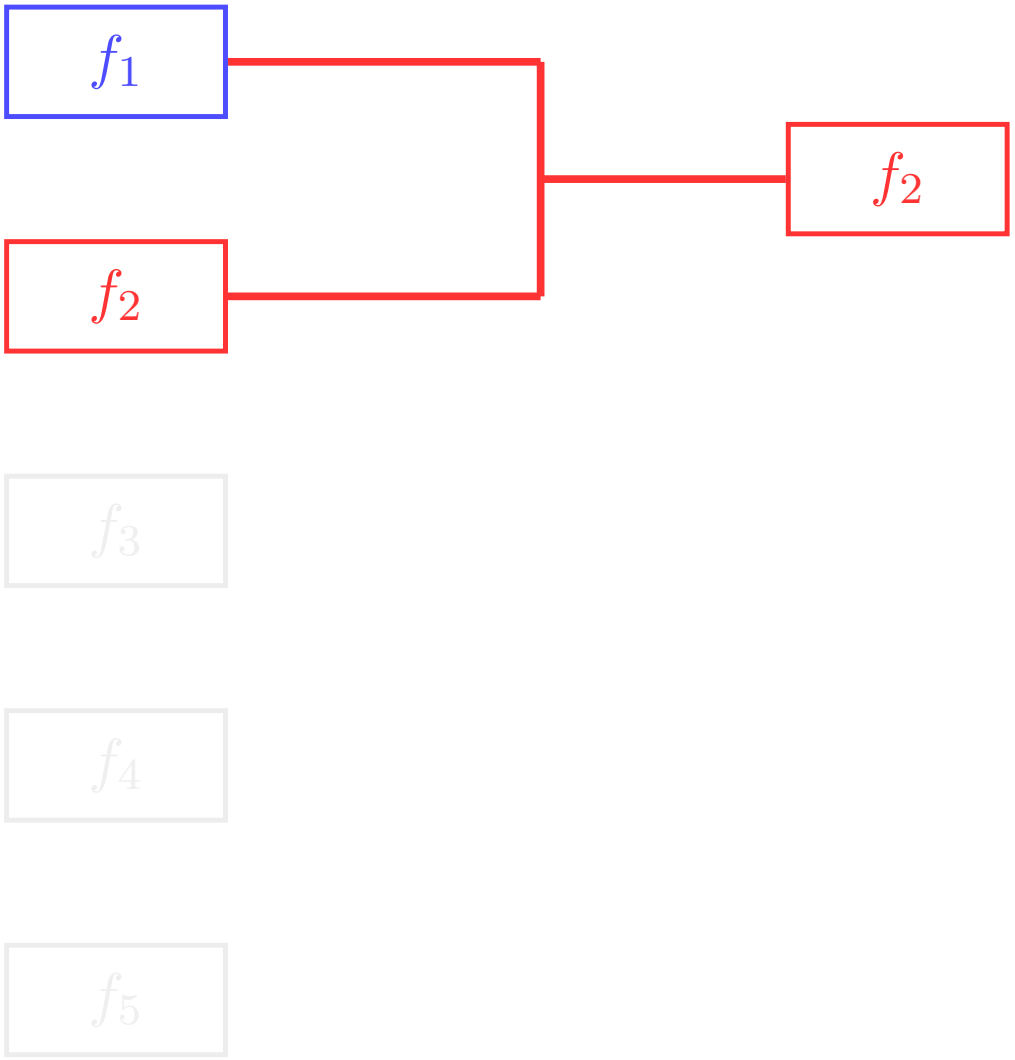
$f_3$

$f_4$

$f_5$

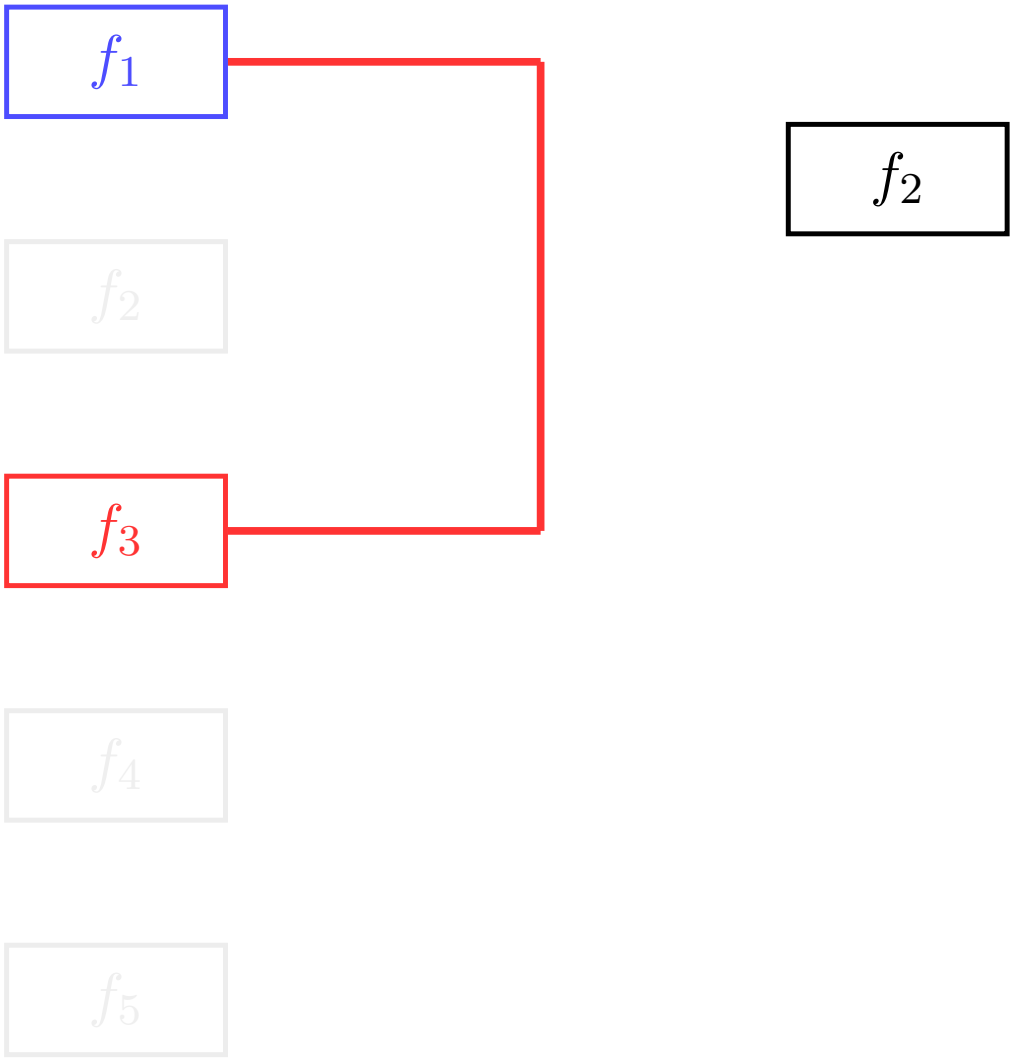
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_2$. Winner is $f_2$.



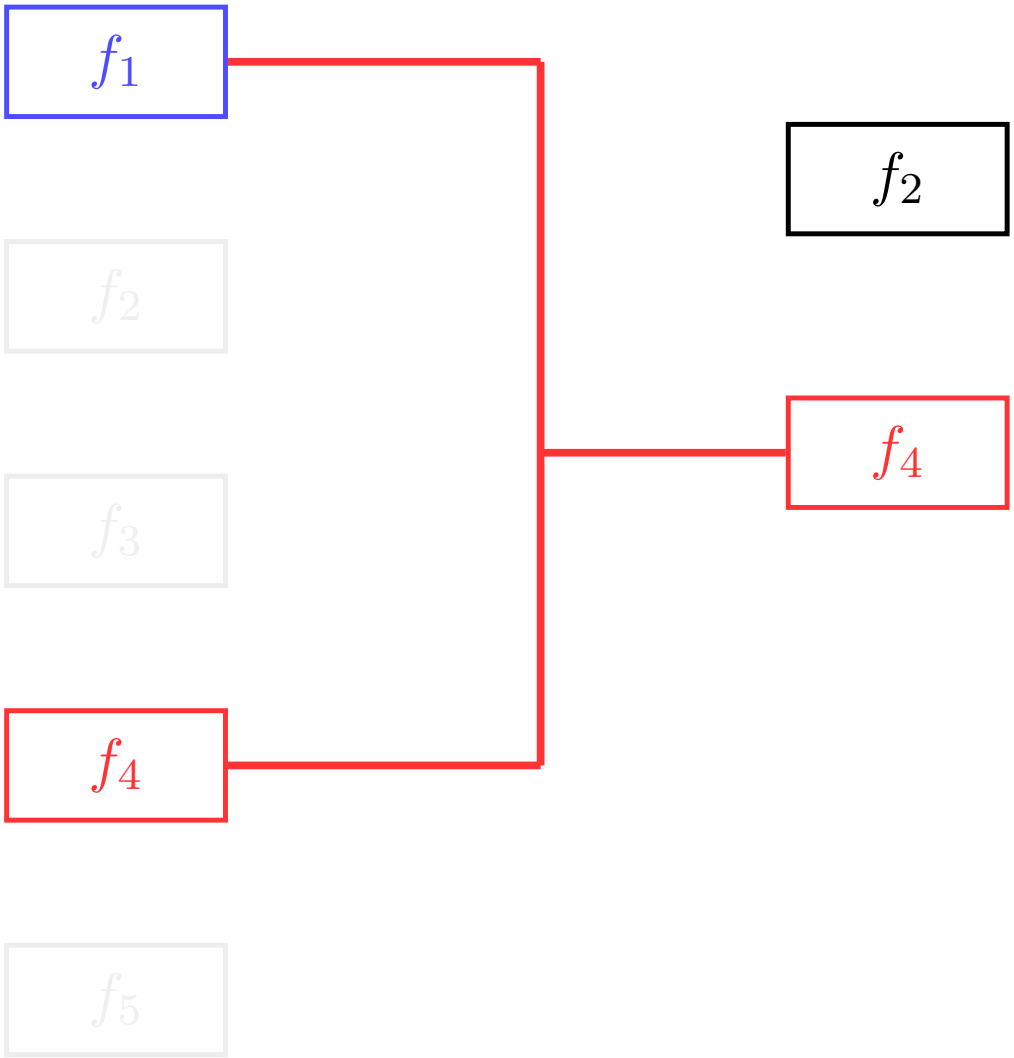
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_3$. Winner is $f_1$.



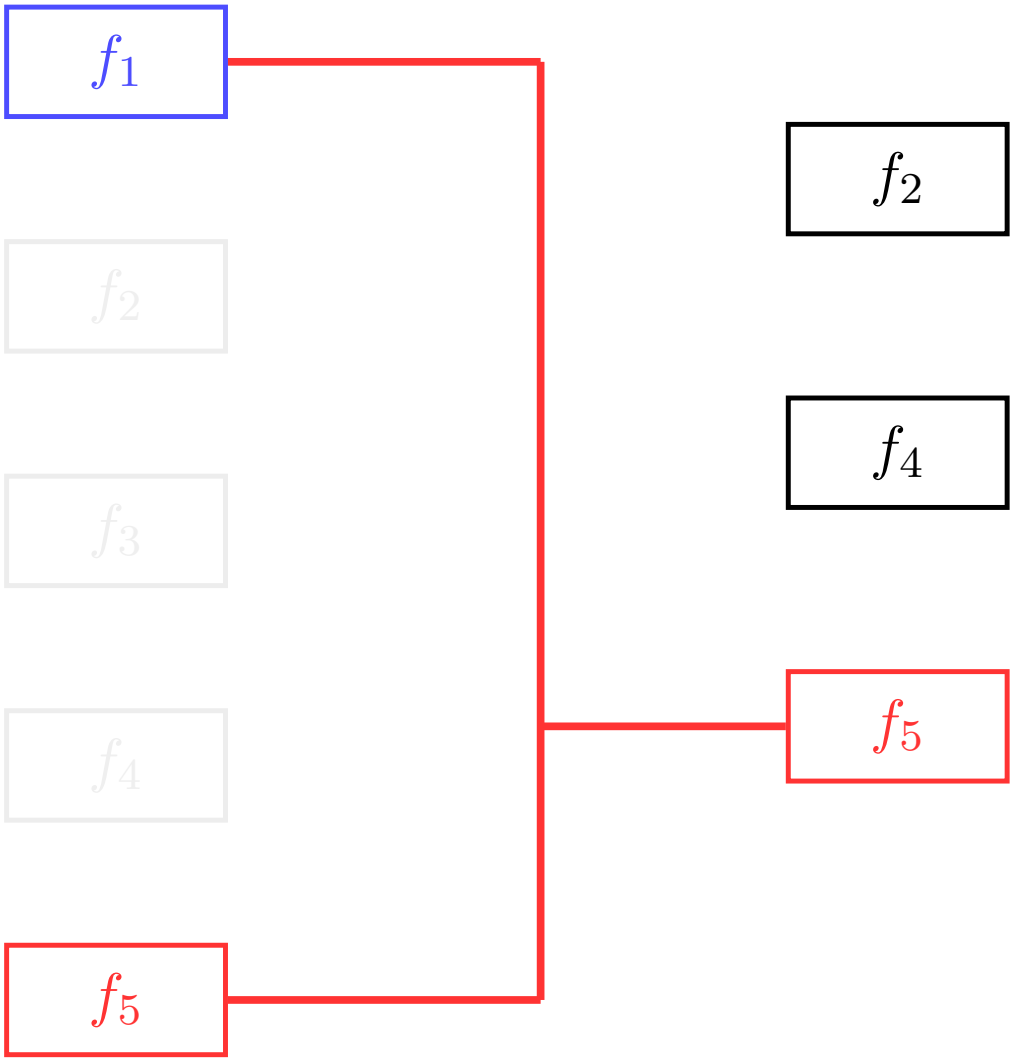
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_4$. Winner is $f_4$.



# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_5$. Winner is $f_5$.



# Sequential Elimination Tournament

**Round 2:** Remaining models are $f_2$, $f_4$ and $f_5$.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

$f_2$

$f_4$

$f_5$

# Sequential Elimination Tournament

**Round 2:** Pivot model: $f_2$.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

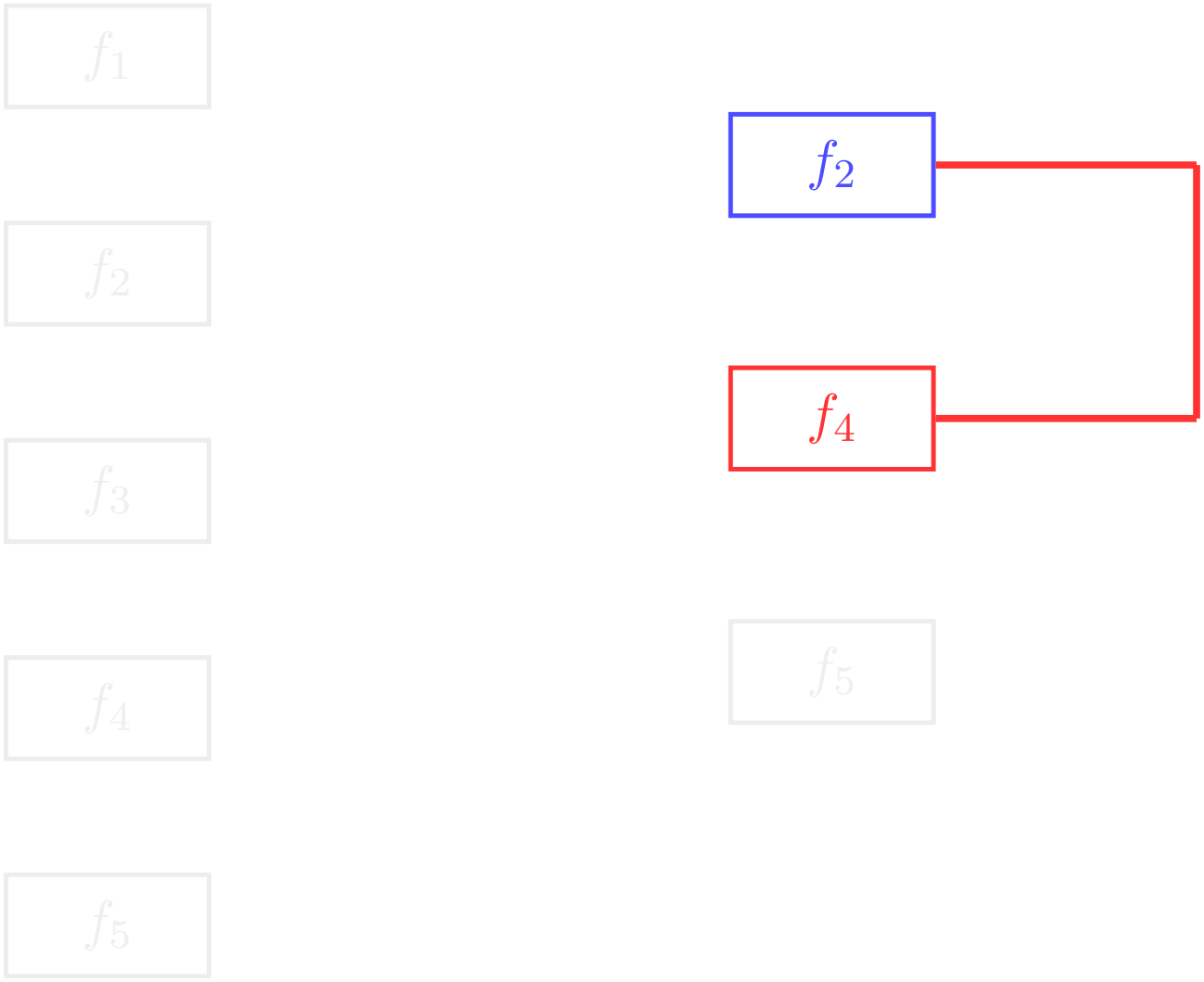
$f_2$

$f_4$

$f_5$

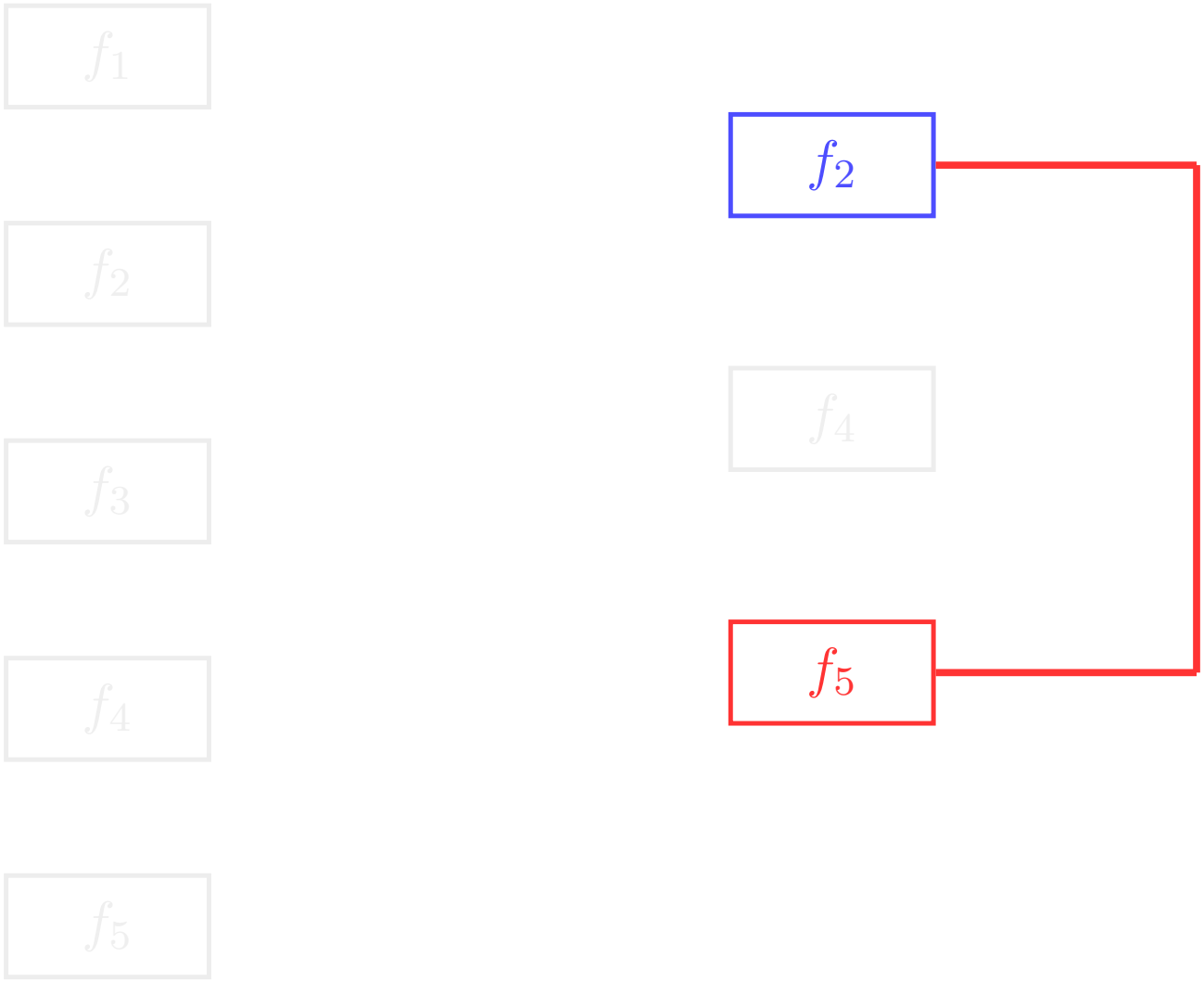
# Sequential Elimination Tournament

**Round 2:** Compare $f_2$ and $f_4$. Winner is $f_2$.



# Sequential Elimination Tournament

**Round 2:** Compare $f_2$ and $f_5$. Winner is $f_2$.



# Sequential Elimination Tournament

**Round 2:** No model outperforms $f_2$. Final winner is $f_2$.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

$f_2$ **Winner!**

$f_4$

$f_5$

# Empirical Analysis

# Empirical Setup: 17 Industry Portfolios

**Target:** daily excess returns of Kenneth French's 17 industry portfolios.

**Covariates:** combine macro + factor + characteristic information:

- Chen et al. (2024) macro/SDF-related factors (incl. beta-sorted deciles and hidden states)
- Fama–French 3 factors (MKT, SMB, HML)
- 94 characteristic-sorted long–short portfolios (based on Gu et al. 2020)
- Include the full vector of 17 industry lagged returns

# Candidate Specifications and Baselines

**Candidate model classes:**

- Linear + regularization: **Ridge**, **LASSO**, **Elastic Net**
- Nonlinear: **Random Forest**

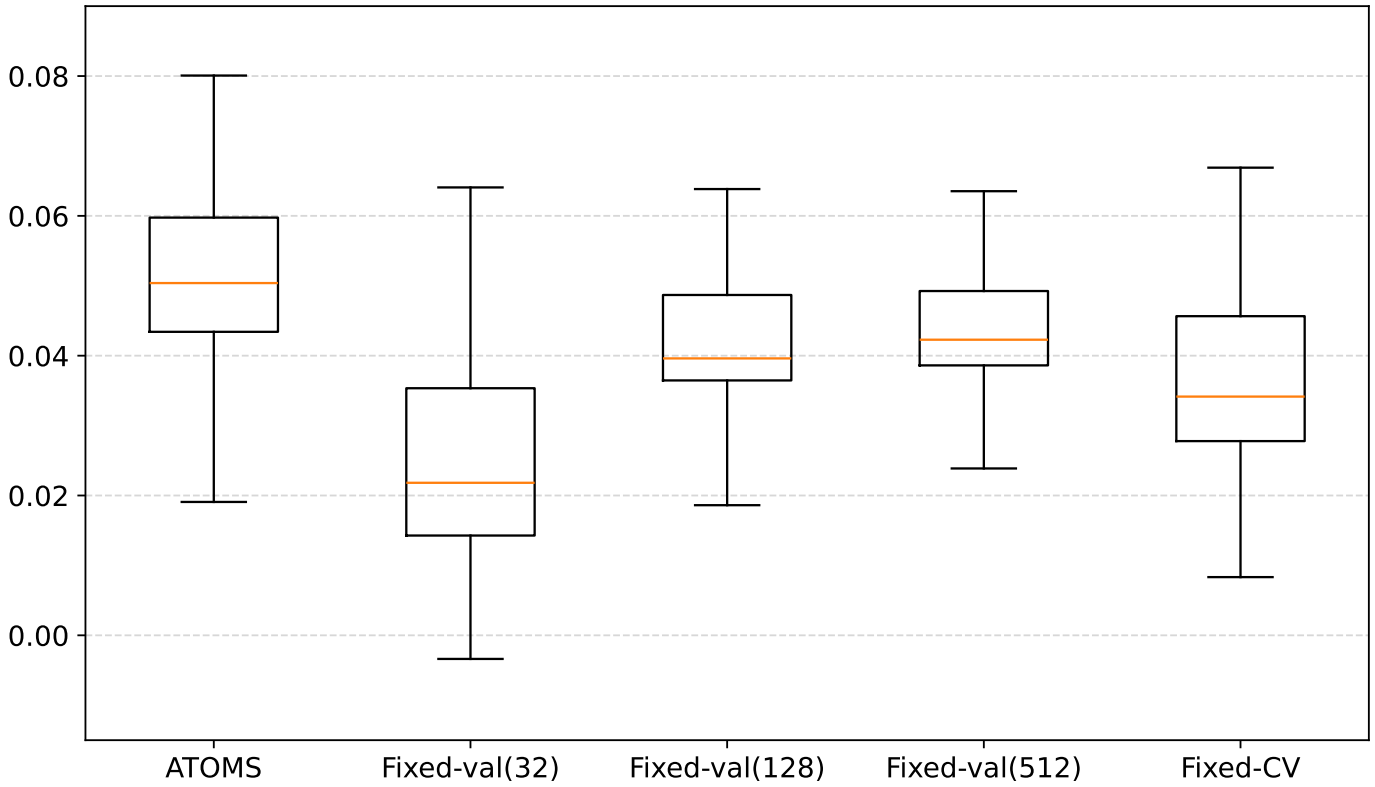
**Training window grid:** in month $t$, trained on windows of length $4^k \wedge (t-1)$ months, $k \in \{0, 1, 2, 3, 4, 5\}$.

**Benchmarks (non-adaptive):**

- **Fixed-val**$(\ell)$: fixed validation window $\ell \in \{32, 128, 512\}$ months; the non-adaptive counterpart of ATOMS
- **Fixed-CV**: 5-fold CV on a fixed 36-month rolling window

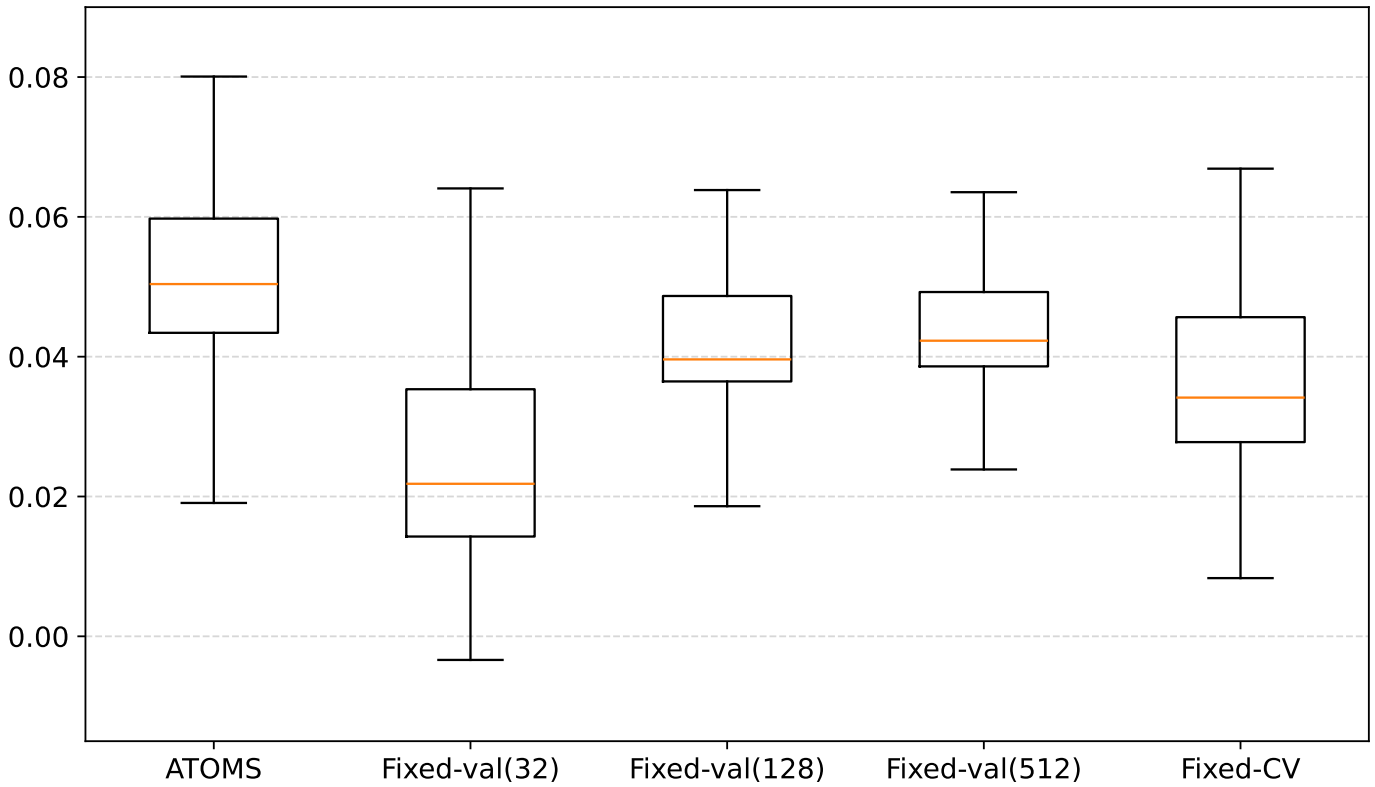
# Overall Performance

Box plot of overall OOS $R^2$ across industries



# Overall Performance

Box plot of overall OOS $R^2$ across industries




ATOMS outperforms fixed-window baselines across industries

# Performance over Recession Periods

Adaptivity matters most in recessions (strong non-stationarity).

**OOS $R^2$ averaged across industries**

| Method | Full Period | Gulf War | 2001 | Fin. Crisis |
|---|---|---|---|---|
| ATOMS | $\mathbf{0.049}$ | $\mathbf{0.027}$ | $\mathbf{0.125}$ | $\mathbf{0.041}$ |
| Fixed-val$(32)$ | $0.022$ | $0.009$ | $0.096$ | $-0.001$ |
| Fixed-val$(512)$ | $0.043$ | $-0.031$ | $0.117$ | $0.039$ |
| Fixed-CV | $0.035$ | $-0.007$ | $0.071$ | $0.014$ |

ATOMS adapts to non-stationarity even in significant recession periods.

# Performance over Recession Periods

Adaptivity matters most in recessions (strong non-stationarity).

**OOS $R^2$ averaged across industries**

| Method | Full Period | Gulf War | 2001 | Fin. Crisis |
|---|---|---|---|---|
| ATOMS | $\mathbf{0.049}$ | $\mathbf{0.027}$ | $\mathbf{0.125}$ | $\mathbf{0.041}$ |
| Fixed-val$(32)$ | $0.022$ | $0.009$ | $0.096$ | $-0.001$ |
| Fixed-val$(512)$ | $0.043$ | $-0.031$ | $0.117$ | $0.039$ |
| Fixed-CV | $0.035$ | $-0.007$ | $0.071$ | $0.014$ |

ATOMS adapts to non-stationarity even in significant recession periods.

**Remark:** While ATOMS is developed under the independent data assumption, it performs well on real-world dependent time series data.

# Portfolio Trading Experiments

**Setup:** In every month $t$,

- each algorithm yields prediction models $\widehat{f}_t^{(1)}, ..., \widehat{f}_t^{(17)}$ for 17 industries
- on each day $s$ of month $t$, use predicted returns $\widehat{f}_t^{(1)}(x_s), ..., \widehat{f}_t^{(17)}(x_s)$ to construct portfolio weights $w_s^{(1)}, ..., w_s^{(17)}$
- gross daily return $r_s = w_s^{(1)} y_s^{(1)} + \cdots + w_s^{(17)} y_s^{(17)}$, where $y_s^{(n)}$ is the *realized excess return* of industry $n$ on day $s$

# Mean-Variance Efficient Portfolio

Compute $\widetilde{\boldsymbol{w}}_s = (\widetilde{w}_s^{(1)}, ..., \widetilde{w}_s^{(17)})^T$ by

$$\widetilde{\boldsymbol{w}}_s = \widehat{\boldsymbol{\Sigma}}_s^{-1} \begin{pmatrix} \widehat{f}_t^{(1)}(x_s) \\ \vdots \\ \widehat{f}_t^{(17)}(x_s) \end{pmatrix},$$

where $\widehat{\boldsymbol{\Sigma}}_s \in \mathbb{R}^{17\times 17}$ is the sample covariance matrix of daily industry returns estimated using a one-year look-back window.

Set portfolio weights

$$w_s^{(n)} = \frac{\widetilde{w}_s^{(n)}}{\sum_{i=1}^{17} \widetilde{w}_s^{(i)}}.$$

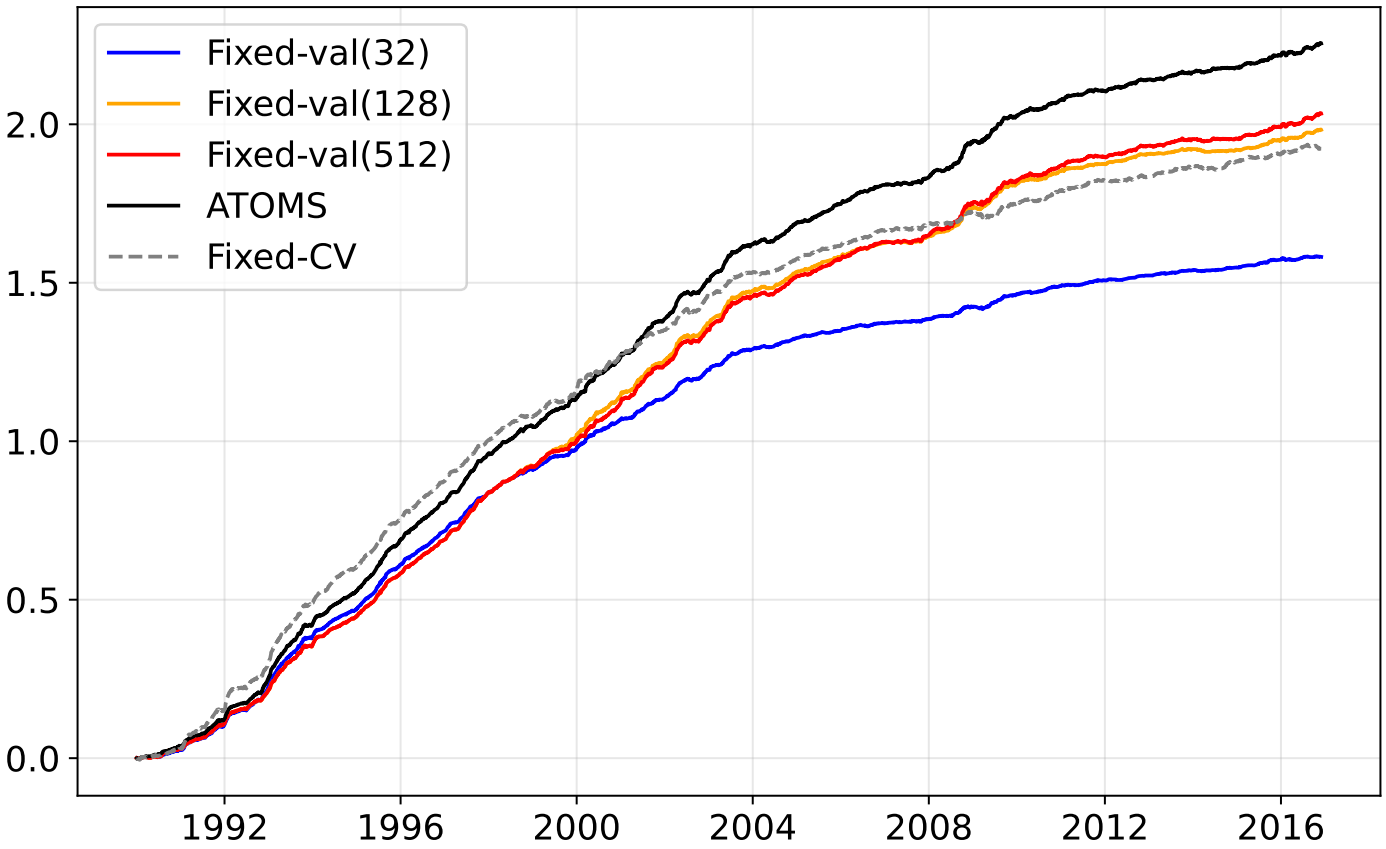
# Results: Average Monthly Sharpe Ratio

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | $\mathbf{0.334}$ | $\mathbf{0.188}$ | $\mathbf{0.569}$ | $\mathbf{0.244}$ |
| Fixed-val$(32)$ | $0.241$ | $0.150$ | $0.306$ | $0.099$ |
| Fixed-val$(512)$ | $0.304$ | $0.146$ | $0.560$ | $0.221$ |
| Fixed-CV | $0.300$ | $0.169$ | $0.371$ | $0.054$ |

ATOMS consistently achieves the highest Sharpe ratios.

# Results: Cumulative Wealth

- Initial wealth $W_0 = 1$. Trading cost $c = 4 \times 10^{-4}/252$.
- Cumulative wealth in month $t$: $W_t = W_{t-1} \prod_{s=1}^{B_t} (1 + r_s - c \cdot \mathrm{TO}_s)$.




Cumulative wealth is plotted in logarithmic scale.

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | **11.94** | **16.85** | **13.71** | **13.62** |
| Fixed-val(32) | 16.39 | 17.43 | 16.91 | 17.86 |
| Fixed-val(512) | 12.45 | 18.63 | 14.06 | 13.34 |
| Fixed-CV | 11.31 | 13.07 | 10.92 | 10.06 |

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | **11.94** | **16.85** | **13.71** | **13.62** |
| Fixed-val(32) | 16.39 | 17.43 | 16.91 | 17.86 |
| Fixed-val(512) | 12.45 | 18.63 | 14.06 | 13.34 |
| Fixed-CV | 11.31 | 13.07 | 10.92 | 10.06 |

ATOMS has turnover lower than or comparable to all baselines.

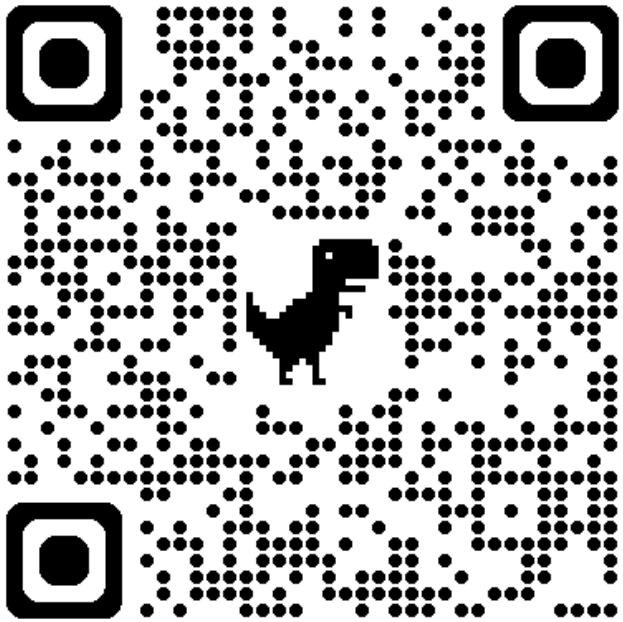
# Conclusions

# Summary

- New framework capturing nonstationarity-complexity tradeoff
- Adaptive algorithm: joint selection of model class and window size
- Theoretical analysis: near-optimal model selection
- Data Analysis: superior performance to standard benchmarks, especially during recessions.

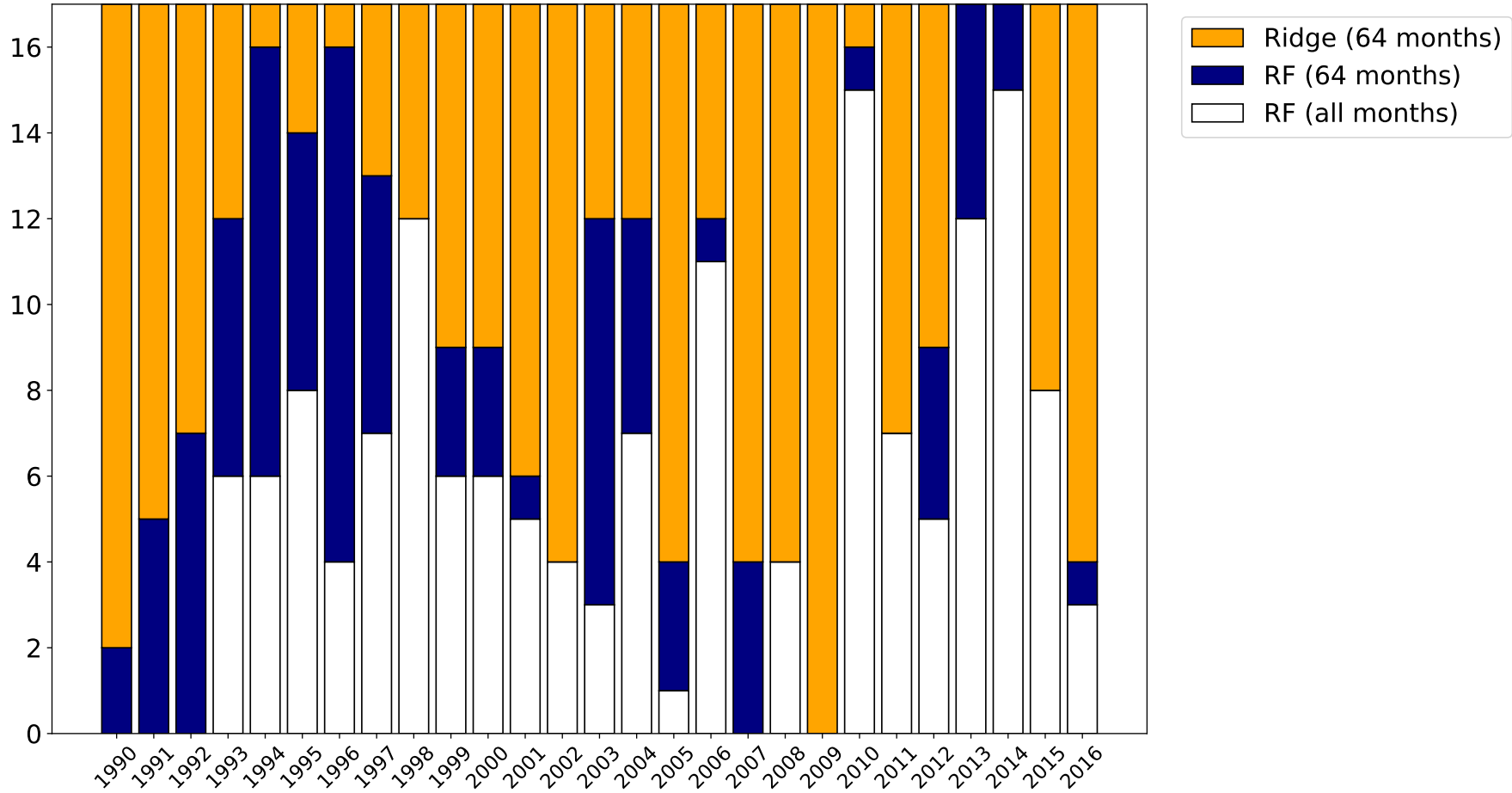
# Thank you!

A. Capponi, C. Huang, J. A. Sidaoui, K. Wang, and J. Zou (2025). "The Nonstationarity-Complexity Tradeoff in Return Prediction". *Available at SSRN 5980654*



# Results on 17 Industry Portfolios

**# Industries where each model attains the highest annual OOS $R^2$:**




Simple model + short window may outperform complex model + long window, and vice versa.

# Sequential Elimination Tournament

**Main idea:** In each round, choose a pivot model $f$ from remaining models.

Use the pairwise comparison subroutine to compare other models with $f$.

- Models that outperform $f$ advance to the next round.
- If no models outperform $f$, then output $f$ as the final winner.

# Sequential Elimination Tournament

**Example:** We start with $5$ candidate models.

$f_1$

$f_2$

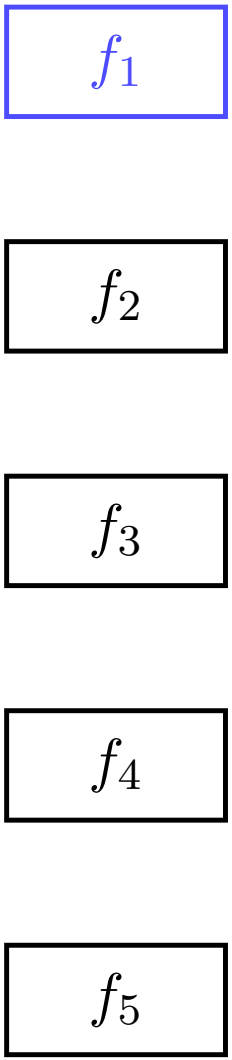
$f_3$

$f_4$

$f_5$

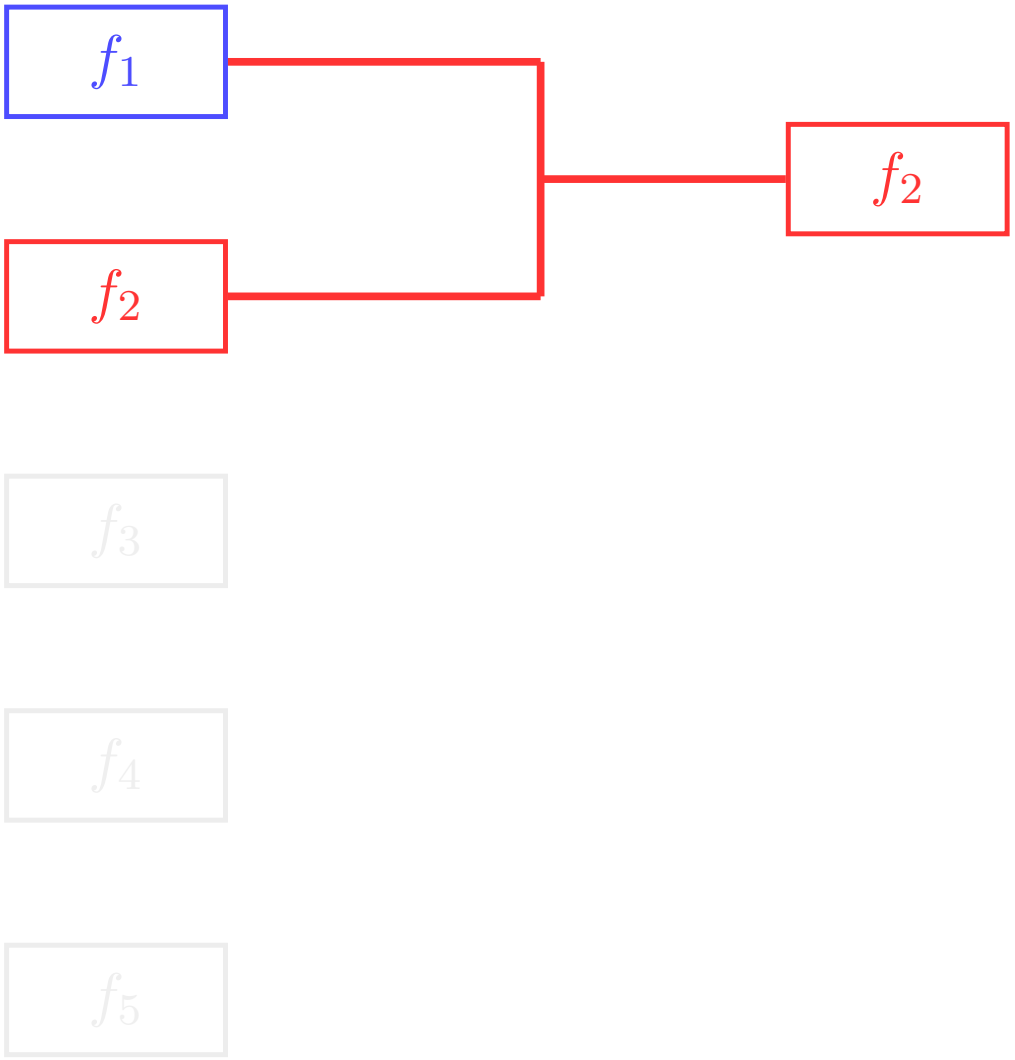
# Sequential Elimination Tournament

**Round 1:** Pivot model: $f_1$.



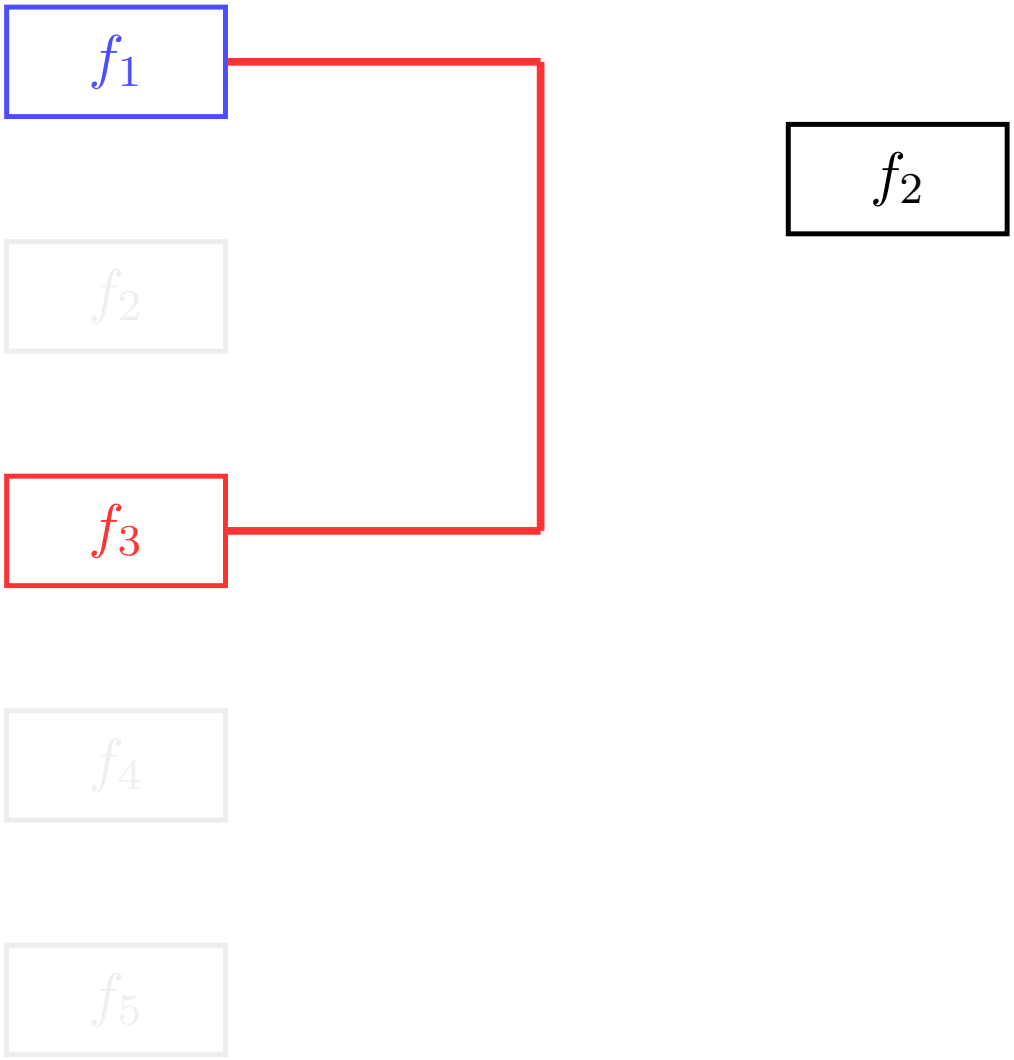
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_2$. Winner is $f_2$.



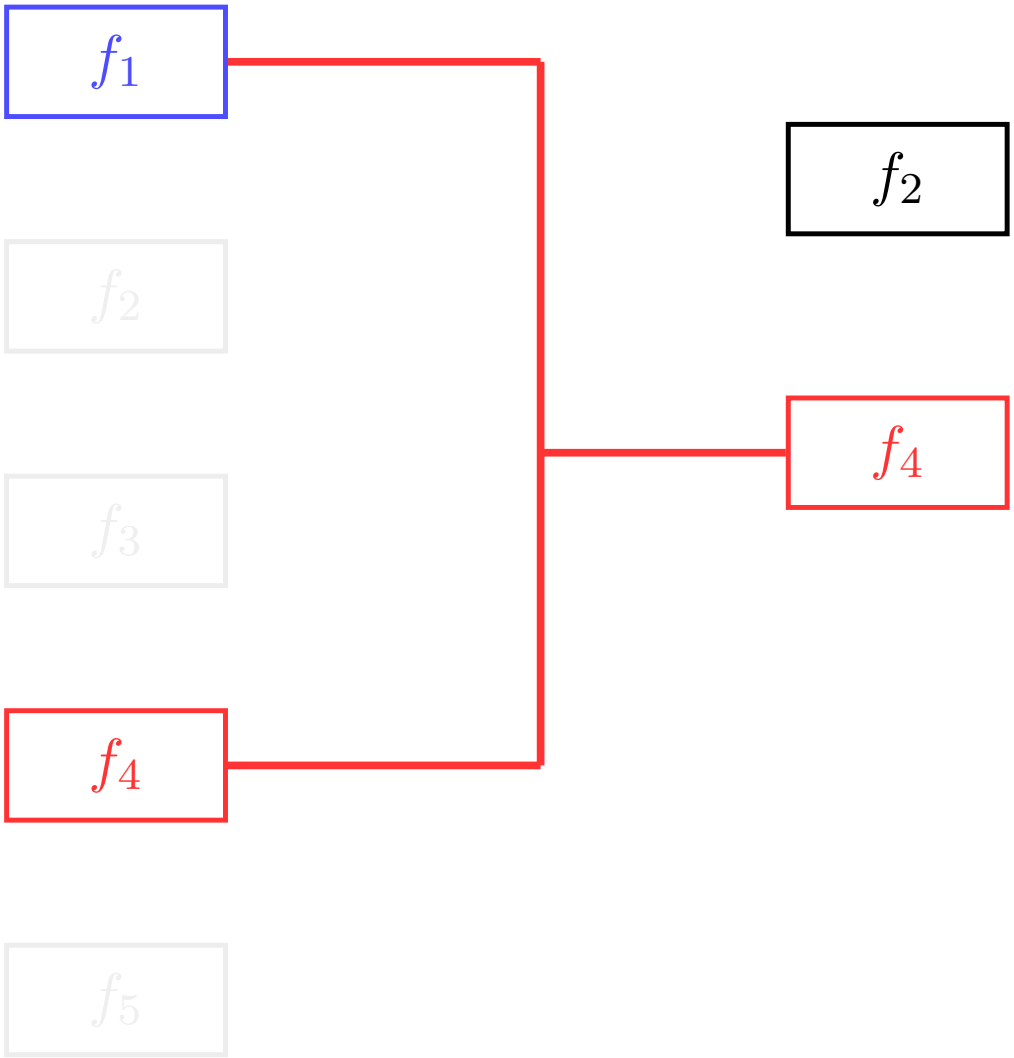
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_3$. Winner is $f_1$.



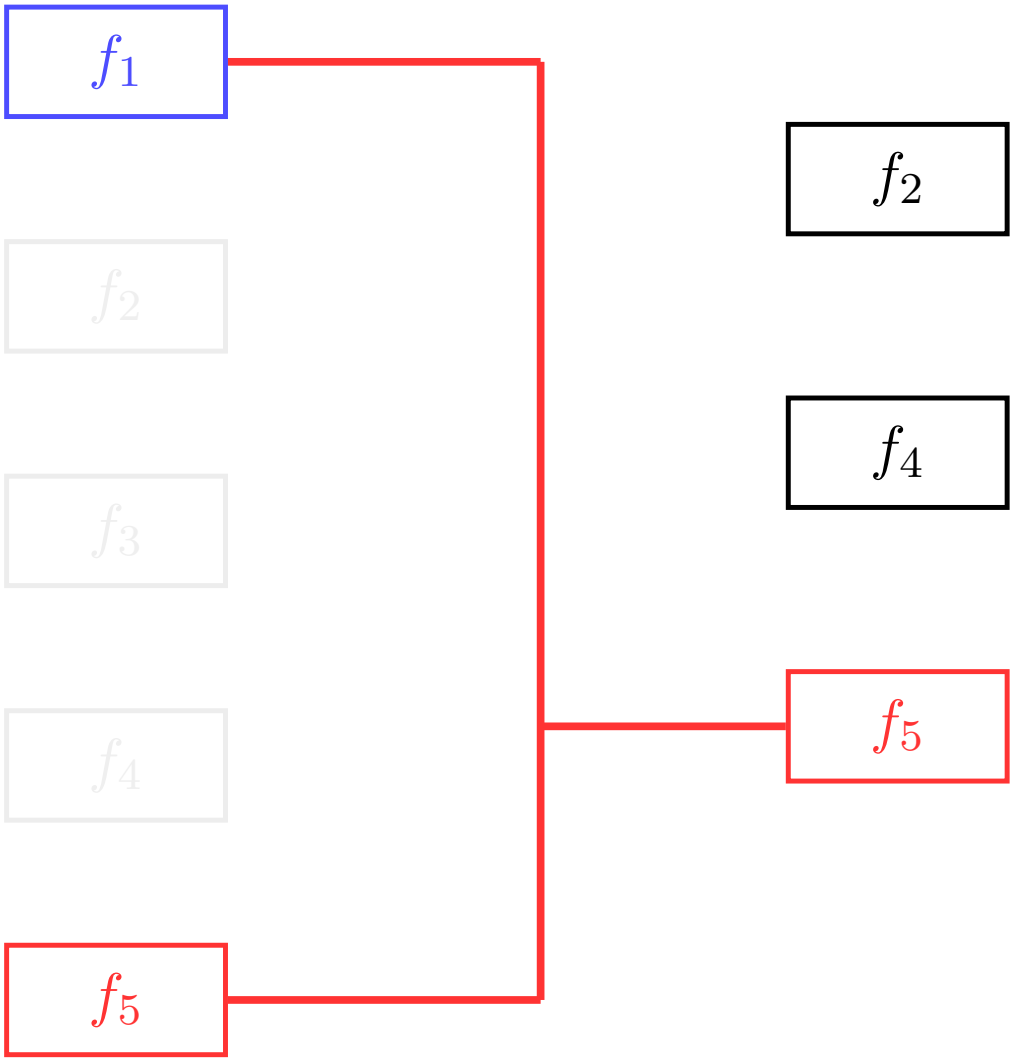
# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_4$. Winner is $f_4$.



# Sequential Elimination Tournament

**Round 1:** Compare $f_1$ and $f_5$. Winner is $f_5$.



# Sequential Elimination Tournament

**Round 2:** Remaining models are $f_2$, $f_4$ and $f_5$.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

$f_2$

$f_4$

$f_5$

# Sequential Elimination Tournament

**Round 2:** Pivot model: $f_2$.

$f_1$

$f_2$

$f_3$

$f_4$

$f_5$

$f_2$

$f_4$

$f_5$

# Sequential Elimination Tournament

**Round 2:** Compare $f_2$ and $f_4$. Winner is $f_2$.

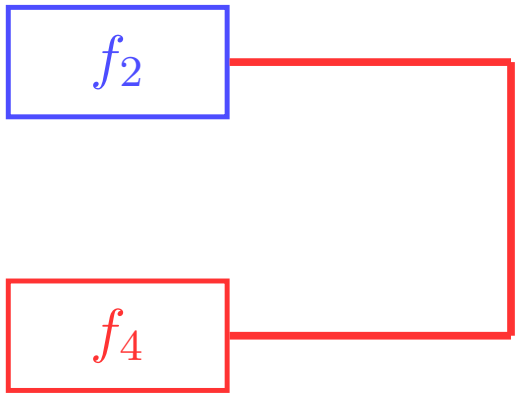
$f_1$

$f_2$

$f_3$

$f_4$

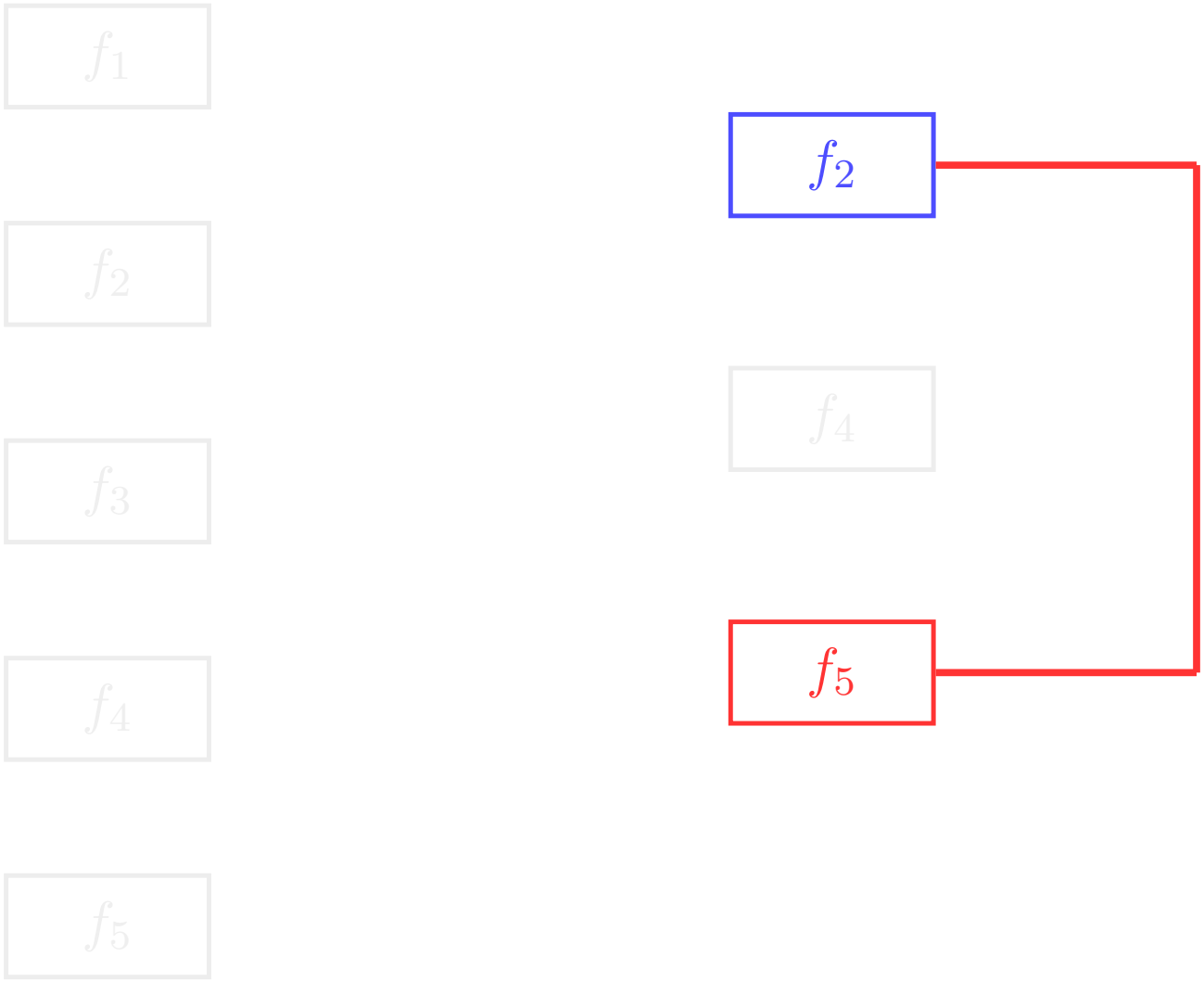
$f_5$




$f_5$

# Sequential Elimination Tournament

**Round 2:** Compare $f_2$ and $f_5$. Winner is $f_2$.



# Sequential Elimination Tournament

**Round 2:** No model outperforms $f_2$. Final winner is $f_2$.

$f_1$

$f_2$ **Winner!**

$f_2$

$f_3$

$f_4$

$f_4$

$f_5$

$f_5$

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | 11.94 | 16.85 | 13.71 | 13.62 |
| Fixed-val(32) | 16.39 | 17.43 | 16.91 | 17.86 |
| Fixed-val(512) | 12.45 | 18.63 | 14.06 | 13.34 |
| Fixed-CV | 11.31 | 13.07 | 10.92 | 10.06 |

# Results: Average Monthly Turnover

- Turnover on day $s$: $\mathrm{TO}_s = \sum_{n=1}^{17} \left| w_s^{(n)} - w_{s-1}^{(n)} \right|$.
- Turnover in month $t$ (with $B_t$ days): $\sum_{s=2}^{B_t} \mathrm{TO}_s$.

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | 11.94 | 16.85 | 13.71 | 13.62 |
| Fixed-val(32) | 16.39 | 17.43 | 16.91 | 17.86 |
| Fixed-val(512) | 12.45 | 18.63 | 14.06 | 13.34 |
| Fixed-CV | 11.31 | 13.07 | 10.92 | 10.06 |

ATOMS has turnover lower than or comparable to all baselines.

# Evaluation Criterion

Out-of-sample $R^2$ (benchmarking against a zero forecast), both *overall* and *annual*.

$$R^2_{[t_1,t_2]}(\texttt{alg}) = 1 - \frac{\sum_{t=t_1}^{t_2} \sum_{(x,y)\in\mathcal{D}_t} \left( f_t^{\texttt{alg}}(x) - y \right)^2}{\sum_{t=t_1}^{t_2} \sum_{(x,y)\in\mathcal{D}_t} y^2}.$$

- Squared prediction errors divided by sum of squared returns (zero forecast benchmark)
- Historical mean is noisy and often worse than zero for excess returns (Gu et al. 2020)
- Annual $R^2$ shows how performance varies over time

# Theoretical Analysis

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \underset{\lambda \in \{1,2\}}{\operatorname{argmin}} \, \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

If $\Delta_t > 0$, choose $f_2$. Otherwise choose $f_1$.

# Pairwise Comparison Subroutine

Given $f_1$ and $f_2$, we want to select the best model:

$$\widehat{\lambda} = \operatorname*{argmin}_{\lambda \in \{1,2\}} \mathbb{E}_{(x,y)\sim P_t} |f_\lambda(x) - y|^2.$$

But $P_t$ is unknown! Use validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ to estimate best model

Equivalently, it suffices to estimate the sign of *performance gap*

$$\Delta_t = \mathbb{E}_{(x,y)\sim P_t} |f_1(x) - y|^2 - \mathbb{E}_{(x,y)\sim P_t} |f_2(x) - y|^2.$$

If $\Delta_t > 0$, choose $f_2$. Otherwise choose $f_1$.

**Challenge:** Validation data $\{\mathcal{D}_j^{\mathrm{va}}\}_{j=1}^{t-1}$ is non-stationary!

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\text{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\text{va}} & & \cdots & & \mathcal{D}_{t-1}^{\text{va}} & &
\end{array}
$$

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-1}^{\mathrm{va}} & &
\end{array}
$$

$$\Downarrow$$

Empirical performance gap

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$
\begin{array}{ccccccccccc}
P_1 & \to & \cdots & \to & P_{t-\ell} & \to & \cdots & \to & P_{t-1} & \to & P_t \\
\downarrow & & & & \downarrow & & & & \downarrow & & \\
\mathcal{D}_1^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-\ell}^{\mathrm{va}} & & \cdots & & \mathcal{D}_{t-1}^{\mathrm{va}} & &
\end{array}
$$

$$\Downarrow$$

Empirical performance gap

$$\widehat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}^{\mathrm{va}}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}_j^{\mathrm{va}}} \left( |f_1(x) - y|^2 - |f_2(x) - y|^2 \right)$$

where $n_{t,\ell}^{\mathrm{va}}$ is the number of validation samples from the last $\ell$ periods.

# Rolling Window for Estimation under Non-stationarity

Use a rolling validation window $\ell$:

$$P_1 \rightarrow \cdots \rightarrow P_{t-\ell} \rightarrow \cdots \rightarrow P_{t-1} \rightarrow P_t$$

$$\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow$$

$$\mathcal{D}_1^{\mathrm{va}} \quad \cdots \quad \mathcal{D}_{t-\ell}^{\mathrm{va}} \quad \cdots \quad \mathcal{D}_{t-1}^{\mathrm{va}}$$

$$\Downarrow$$

Empirical performance gap

$$\widehat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}^{\mathrm{va}}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}_j^{\mathrm{va}}} \left( |f_1(x) - y|^2 - |f_2(x) - y|^2 \right)$$

How to choose validation window $\ell$?

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

- $B(\ell)$: bias due to non-stationarity ($\uparrow$)
- $V(\ell)$: variance due to finite samples ($\downarrow$)

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

- Construct $\widehat{V}(\ell) \approx V(\ell)$ using empirical variance. details

# Adaptive Validation Window Selection

## Lemma 4 (Bias-Variance Decompostion)

*For every validation window size $\ell$,*

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \quad \lesssim \quad \underbrace{\max_{t-\ell \le j \le t-1} |\Delta_j - \Delta_t|}_{B(\ell)} \quad + \quad \underbrace{\sigma_{t,\ell}\sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}_{V(\ell)},$$

*where $\sigma^2_{t,\ell}$ = average variance of performance gaps over data in window $\ell$.*

**Choose $\ell$ to minimize $B(\ell) + V(\ell)$?**

- Construct $\widehat{V}(\ell) \approx V(\ell)$ using empirical variance. details
- How to estimate the non-stationarity bias $B(\ell)$?

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

Choose validation window $\widehat{\ell} = \operatorname{argmin}_\ell \left\{ \widehat{B}(\ell) + \widehat{V}(\ell) \right\}$. details

# Adaptive Validation Window Selection

$$|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell) \text{ where } B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|.$$

Inspired by Goldenshluger and Lepski (2008) in nonparametric estimation:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \left( \left| \widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i} \right| - \widehat{V}(\ell) - \widehat{V}(i) \right)_+ .$$

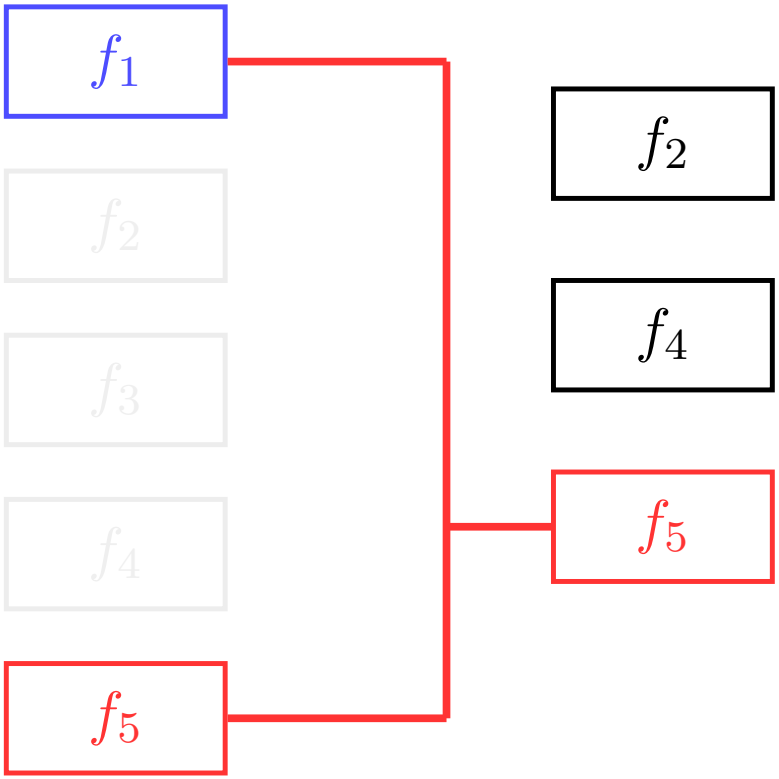
Choose validation window $\widehat{\ell} = \operatorname{argmin}_\ell \left\{ \widehat{B}(\ell) + \widehat{V}(\ell) \right\}$. details

## Theorem 5 (Near-Optimal Validation Window Selection)

*With high probablity,* $\left| \widehat{\Delta}_{t,\widehat{\ell}} - \Delta_t \right| \lesssim \min_\ell \left\{ B(\ell) + V(\ell) \right\}$.

# Adaptive Tournament Model Selection (ATOMS)

**Sequential elimination tournament** based on **pairwise comparison**:




For each pairwise comparison,

- Choose validation window $\widehat{\ell}$ to estimate performance gap $\widehat{\Delta}_{t,\widehat{\ell}}$.
- Compare based on sign of $\widehat{\Delta}_{t,\widehat{\ell}}$.

# Theoretical Guarantee for Model Selection

# Theoretical Guarantee for Model Selection

Suppose the candidate models $\{f_\lambda\}$ come from training model classes $\mathcal{F} \in \{\mathcal{F}_1, ..., \mathcal{F}_R\}$ on training windows $k \in \{1, 2, ..., t-1\}$.

# Theoretical Guarantee for Model Selection

Suppose the candidate models $\{f_\lambda\}$ come from training model classes $\mathcal{F} \in \{\mathcal{F}_1, ..., \mathcal{F}_R\}$ on training windows $k \in \{1, 2, ..., t-1\}$.

## Theorem 6 (Near-Optimal Model-and-Data Selection)

*With high probability, ATOMS selects a model $\widehat{f}$ satisfying*

$$L_t(\widehat{f}) - L_t(f_t^*)$$

$$\lesssim \min_{\mathcal{F}, k} \Big\{ \textit{misspecification}\,(\mathcal{F}) + \textit{variance}\,(\mathcal{F}, k) + \textit{non-stationarity}\,(k) \Big\}.$$

# Theoretical Guarantee for Model Selection

Suppose the candidate models $\{f_\lambda\}$ come from training model classes $\mathcal{F} \in \{\mathcal{F}_1, ..., \mathcal{F}_R\}$ on training windows $k \in \{1, 2, ..., t-1\}$.

## Theorem 6 (Near-Optimal Model-and-Data Selection)

*With high probability, ATOMS selects a model $\widehat{f}$ satisfying*

$$L_t(\widehat{f}) - L_t(f_t^*)$$

$$\lesssim \min_{\mathcal{F}, k} \Big\{ \textit{misspecification}\,(\mathcal{F}) + \textit{variance}\,(\mathcal{F}, k) + \textit{non-stationarity}\,(k) \Big\}.$$

- Right-hand side: an **oracle** analyst that knows the optimal model class $\mathcal{F}$ and window $k$ in advance.
- ATOMS nearly matches this oracle using only past data.

**Near-optimal selection of model class $\mathcal{F}$ and training window $k$**

# Our Method vs Virtue of Complexity

| **Aspect** | **Virtue of Complexity** | **Our Work** |
|---|---|---|
| **Environment** | Stationary $(P_1 = \cdots = P_t)$ | Non-stationary $(P_j \neq P_t)$ |
| **Training data** | Expanding window | Adaptive window (select $k$) |
| **Model selection** | Choose $\mathcal{F}$ | Choose $\mathcal{F}$ **and** $k$ jointly |
| **Tradeoff** | Bias vs Variance | Bias vs Variance vs **Drift** |
| **Key finding** | Complex model helps | Complexity should be paired **with right data** |

**Theoretical guarantee:**

$$\text{Performance} \lesssim \min_{\mathcal{F},k} \left\{ \underbrace{\text{Bias}(\mathcal{F})}_{\text{complexity}} + \underbrace{\text{Variance}(\mathcal{F},k)}_{\text{variance}} + \underbrace{\max_{j\in[t-k,t-1]} \text{TV}(P_j,P_t)}_{\text{non-stationarity}} \right\}$$

# References I


- Goldenshluger, A. and O. Lepski (2008). "Universal pointwise selection rule in multivariate function estimation". *Bernoulli* 14.4, pp. 1150–1190.
- Chen, L. et al. (2024). "Deep Learning in Asset Pricing". *Management Science* 70.2, pp. 714–750.
- Gu, S. et al. (Feb. 2020). "Empirical Asset Pricing via Machine Learning". *The Review of Financial Studies* 33.5, pp. 2223–2273.
- Capponi, A. et al. (2025). "The Nonstationarity-Complexity Tradeoff in Return Prediction". *Available at SSRN 5980654*.
- Bartlett, P. L. et al. (2005). "Local Rademacher complexities". *The Annals of Statistics* 33.4, pp. 1497–1537.

# Details of Nonstationarity-Complexity Tradeoff

# Non-Stationary Local Rademacher Complexity

A non-stationary local Rademacher complexity (Bartlett et al. 2005).

# Non-Stationary Local Rademacher Complexity

A non-stationary local Rademacher complexity (Bartlett et al. 2005).

**Rademacher complexity** of model class $\mathcal{G} \subseteq \mathcal{F}$ w.r.t. data $\mathcal{D} = \{x_i\}_{i=1}^n$:

$$\mathfrak{R}(\mathcal{G}; \mathcal{D}) = \mathbb{E}\left[\sup_{f \in \mathcal{G}} \sum_{i=1}^{n} \varepsilon_i f(x_i)\right],$$

where $\varepsilon_i$ are i.i.d. with $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$.

Measures the ability of $\mathcal{G}$ to fit random noise using $\mathcal{D}$.

# Non-Stationary Local Rademacher Complexity

A non-stationary local Rademacher complexity (Bartlett et al. 2005).

**Rademacher complexity** of model class $\mathcal{G} \subseteq \mathcal{F}$ w.r.t. data $\mathcal{D} = \{x_i\}_{i=1}^n$:

$$\mathfrak{R}(\mathcal{G}; \mathcal{D}) = \mathbb{E}\left[\sup_{f \in \mathcal{G}} \sum_{i=1}^{n} \varepsilon_i f(x_i)\right],$$

where $\varepsilon_i$ are i.i.d. with $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$.

Measures the ability of $\mathcal{G}$ to fit random noise using $\mathcal{D}$.

**Local class:** near-optimal models in $\mathcal{F}$ within window $[t-k, t-1]$:

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

where $n_{t,k}^{\mathrm{tr}} = \sum_{j=t-k}^{t-1} n_j^{\mathrm{tr}}$ and $\bar{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} L_t(f)$.

# Non-Stationary Local Rademacher Complexity

**Local Rademacher complexity** of $\mathcal{F}$ w.r.t. data within last $k$ periods $\mathcal{D}^{\mathrm{tr}}_{t,k} = \{\mathcal{D}^{\mathrm{tr}}_j\}_{j=t-k}^{t-1}$:

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

Ability of near-optimal models in $\mathcal{F}$ to fit noise using data in window $k$.

# Non-Stationary Local Rademacher Complexity

**Local Rademacher complexity** of $\mathcal{F}$ w.r.t. data within last $k$ periods $\mathcal{D}^{\mathrm{tr}}_{t,k} = \{\mathcal{D}^{\mathrm{tr}}_j\}_{j=t-k}^{t-1}$:

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

Ability of near-optimal models in $\mathcal{F}$ to fit noise using data in window $k$.

**Estimation variance** of $\mathcal{F}$ w.r.t. window $k$:

$$r_{t,k}(\mathcal{F}) = \text{ fixed point of the function } r \mapsto \mathfrak{R}\Big(\mathcal{F}_{t,k}(r);\, \mathcal{D}^{\mathrm{tr}}_{t,k}\Big).$$

Also known as the *critical radius* of local Rademacher complexity.

# Examples of Critical Radius $r_{t,k}(\mathcal{F})$

- If $\mathcal{F}$ is finite:

$$r_{t,k}(\mathcal{F}) \asymp \frac{\log|\mathcal{F}|}{n_{t,k}^{\mathrm{tr}}}.$$

- If $\mathcal{F}$ is $d$-dimensional linear models:

$$r_{t,k}(\mathcal{F}) \asymp \frac{d}{n_{t,k}^{\mathrm{tr}}}.$$

- If $\mathcal{F}$ is a kernel class with exponential eigenvalue decay:

$$r_{t,k}(\mathcal{F}) \asymp \frac{\log n_k}{n_{t,k}^{\mathrm{tr}}}.$$

- If $\mathcal{F}$ is a kernel class with polynomial eigenvalue decay:

$$r_{t,k}(\mathcal{F}) \asymp \left(n_{t,k}^{\mathrm{tr}}\right)^{-\frac{2\alpha}{2\alpha+1}}.$$

back

# Details on Adaptive Model Selection

# Benefits of Pairwise Comparison

**Pairwise comparison:** choose a window $\ell$ to minimize bias + variance in performance gap estimation $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$, where bias is

$$B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| \quad \text{with} \quad \Delta_t = L_t(f_1) - L_t(f_2).$$

**If compare all models simultaneously:** choosing a window $\ell$ to evaluate all models yields a bias-variance trade-off, with much larger bias

$$\widetilde{B}(\ell) = \max_{t-\ell \leq j \leq t-1} \max_{f \in \{f_\lambda\}_{\lambda=1}^{\Lambda}} |L_t(f) - L_j(f)|.$$

$\Rightarrow$ smaller selected window $\ell$ $\Rightarrow$ less accurate comparison

# Benefits of Performance Gap Estimation

Instead of estimating the individual model performance

$$\mathbb{E}_{(x,y)\sim P_t}|f_1(x)-y|^2 \qquad \text{and} \qquad \mathbb{E}_{(x,y)\sim P_t}|f_2(x)-y|^2,$$

we estimate the performance gap

$$\mathbb{E}_{(x,y)\sim P_t}|f_1(x)-y|^2 - \mathbb{E}_{(x,y)\sim P_t}|f_2(x)-y|^2.$$

**Benefits:** Taking the difference improves prediction accuracy by

- canceling out common errors,
- reducing random noise.

# Explicit Formulas for Variance Terms $V(\ell)$ and $\widehat{V}(\ell)$

# Explicit Formulas for Variance Terms $V(\ell)$ and $\widehat{V}(\ell)$

Performance gap between $f_1$ and $f_2$ w.r.t. sample $(x, y)$:

$$u(x, y) = |f_1(x) - y|^2 - |f_2(x) - y|^2.$$

# Explicit Formulas for Variance Terms $V(\ell)$ and $\widehat{V}(\ell)$

Performance gap between $f_1$ and $f_2$ w.r.t. sample $(x, y)$:

$$u(x, y) = |f_1(x) - y|^2 - |f_2(x) - y|^2.$$

In $V(\ell) \asymp \sigma_{t,\ell} \sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}$, the variance $\sigma^2_{t,\ell}$ is defined by

$$\sigma^2_{t,\ell} = \frac{1}{n^{\mathrm{va}}_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}^{\mathrm{va}}_j} \mathrm{var}\left(u(x, y)\right).$$

# Explicit Formulas for Variance Terms $V(\ell)$ and $\widehat{V}(\ell)$

Performance gap between $f_1$ and $f_2$ w.r.t. sample $(x, y)$:

$$u(x,y) = |f_1(x) - y|^2 - |f_2(x) - y|^2.$$

In $V(\ell) \asymp \sigma_{t,\ell} \sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}$, the variance $\sigma^2_{t,\ell}$ is defined by

$$\sigma^2_{t,\ell} = \frac{1}{n^{\mathrm{va}}_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{(x,y)\in\mathcal{D}^{\mathrm{va}}_j} \mathrm{var}\left(u(x,y)\right).$$

We form $\widehat{V}(\ell) \asymp \widehat{\sigma}_{t,\ell} \sqrt{\frac{1}{n^{\mathrm{va}}_{t,\ell}} + \frac{1}{n^{\mathrm{va}}_{t,\ell}}}$, where $\widehat{\sigma}^2_{t,\ell}$ is the sample variance of

$$\left\{u(x,y) : \ (x,y) \in \mathcal{D}^{\mathrm{va}}_{t-\ell}, ..., \mathcal{D}^{\mathrm{va}}_{t-1}\right\}.$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008):

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}|$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| \leq |\widehat{\Delta}_{t,\ell} - \Delta_t| + |\widehat{\Delta}_{t,i} - \Delta_t|$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$
\begin{aligned}
|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| &\leq |\widehat{\Delta}_{t,\ell} - \Delta_t| + |\widehat{\Delta}_{t,i} - \Delta_t| \\
&\leq \big[B(\ell) + V(\ell)\big] + \big[B(i) + V(i)\big]
\end{aligned}
$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$
\begin{aligned}
|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| &\leq |\widehat{\Delta}_{t,\ell} - \Delta_t| + |\widehat{\Delta}_{t,i} - \Delta_t| \\
&\leq \big[B(\ell) + V(\ell)\big] + \big[B(i) + V(i)\big] \\
&\leq 2B(\ell) + V(\ell) + V(i).
\end{aligned}
$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$
\begin{aligned}
|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| &\leq |\widehat{\Delta}_{t,\ell} - \Delta_t| + |\widehat{\Delta}_{t,i} - \Delta_t| \\
&\leq \big[B(\ell) + V(\ell)\big] + \big[B(i) + V(i)\big] \\
&\leq 2B(\ell) + V(\ell) + V(i).
\end{aligned}
$$

Subtract the variance terms $V(\ell) + V(i)$ to tease out the bias:

$$
\widehat{B}(\ell) = \qquad |\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| - \widehat{V}(\ell) - \widehat{V}(i)
$$

back

# Intuition for Bias Estimator $\widehat{B}(\ell)$

Recall $|\widehat{\Delta}_{t,\ell} - \Delta_t| \leq B(\ell) + V(\ell)$ where $B(\ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$.

Inspired by Goldenshluger and Lepski (2008): Replace $\Delta_t$ by $\widehat{\Delta}_{t,i}$ for $i \leq \ell$

$$\begin{aligned} |\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| &\leq |\widehat{\Delta}_{t,\ell} - \Delta_t| + |\widehat{\Delta}_{t,i} - \Delta_t| \\ &\leq \big[B(\ell) + V(\ell)\big] + \big[B(i) + V(i)\big] \\ &\leq 2B(\ell) + V(\ell) + V(i). \end{aligned}$$

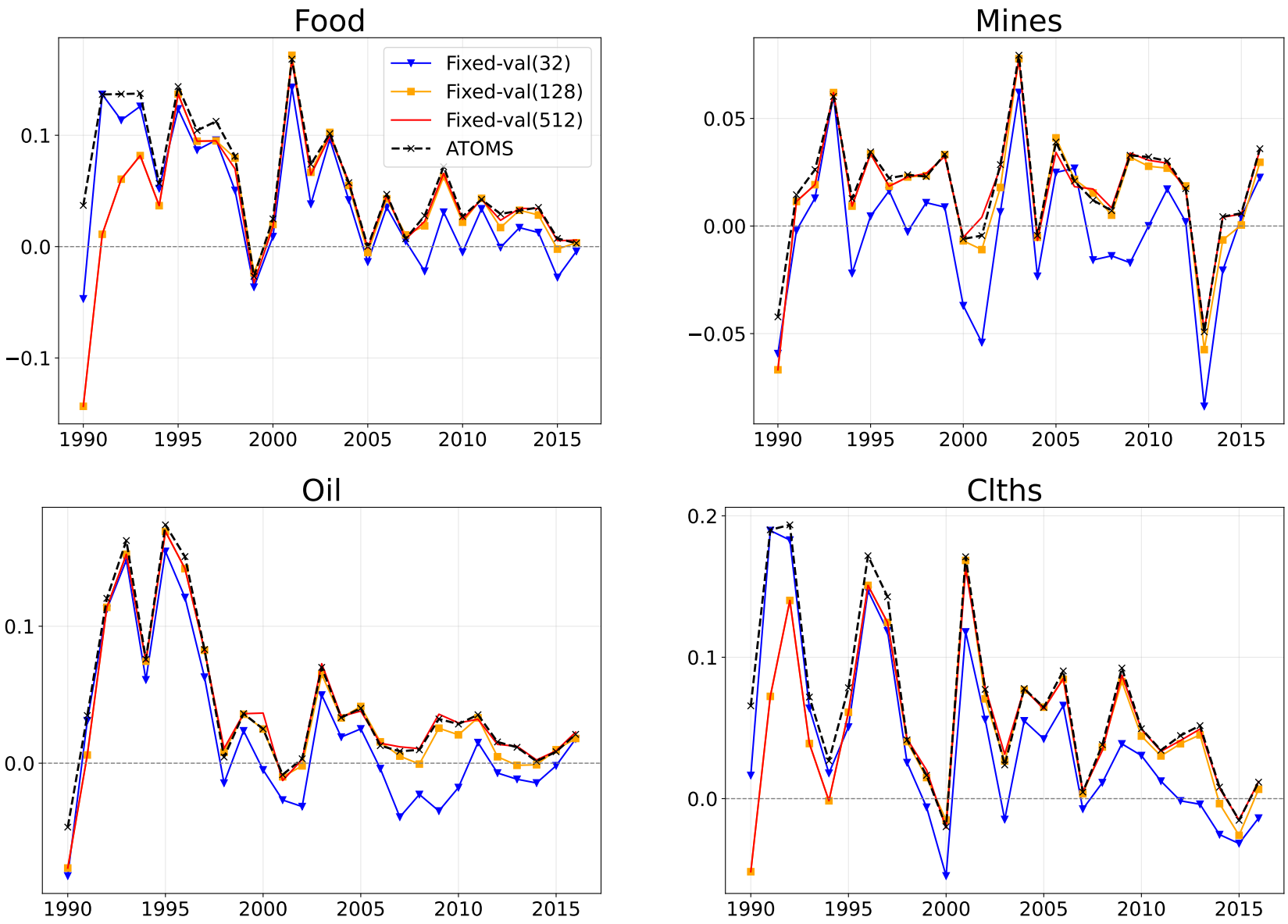
Subtract the variance terms $V(\ell) + V(i)$ to tease out the bias:

$$\widehat{B}(\ell) = \max_{1 \leq i \leq \ell} \Big( |\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| - \widehat{V}(\ell) - \widehat{V}(i) \Big)_+ .$$

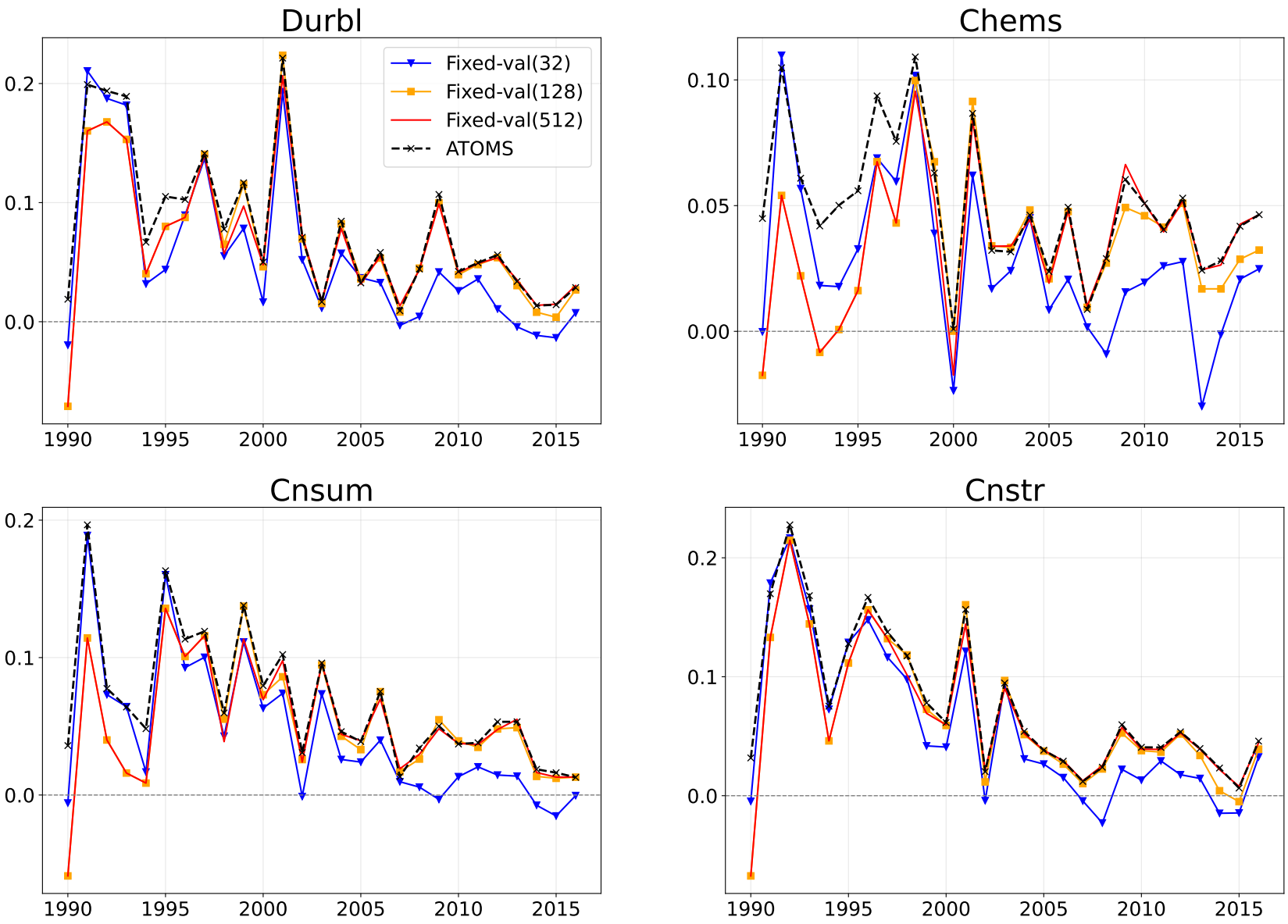
back

# Annual $R^2$ on $17$ Industries

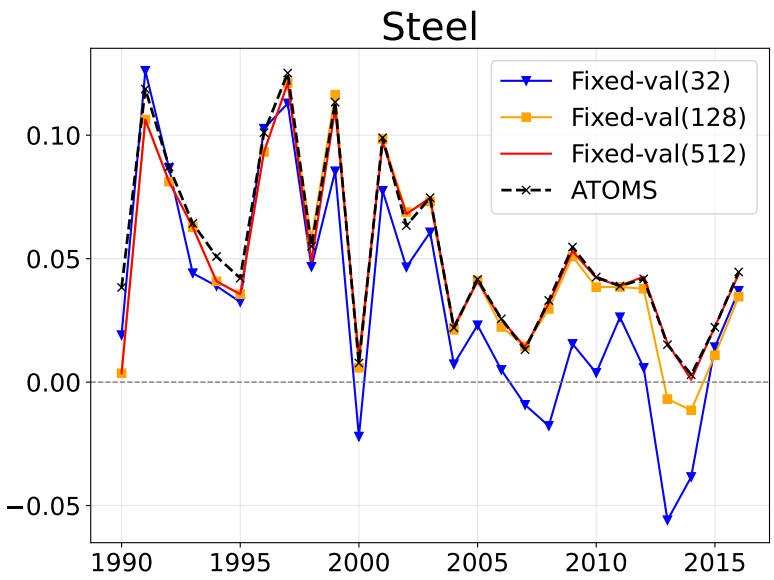
# Annual $R^2$ on 17 Industries (I)

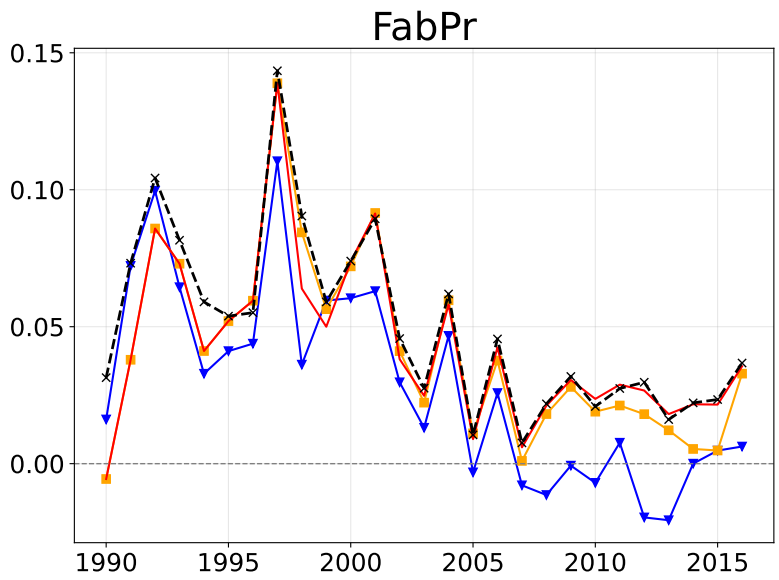


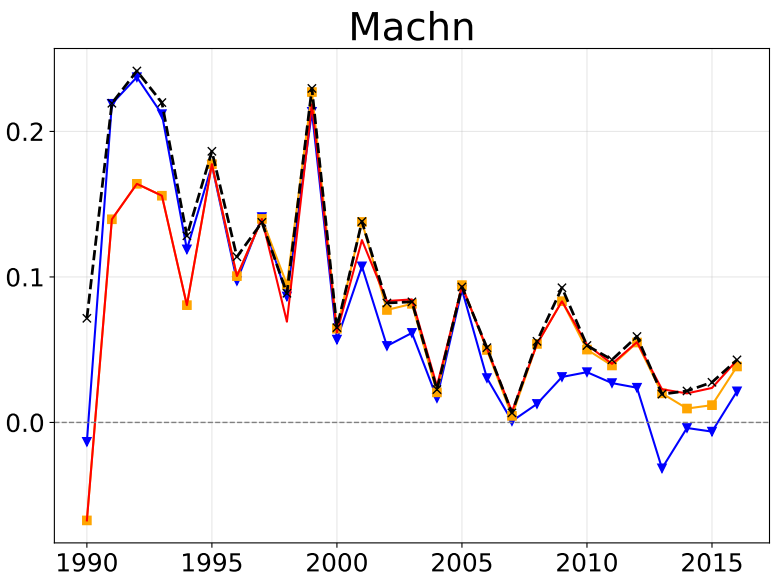
# Annual $R^2$ on 17 Industries (II)

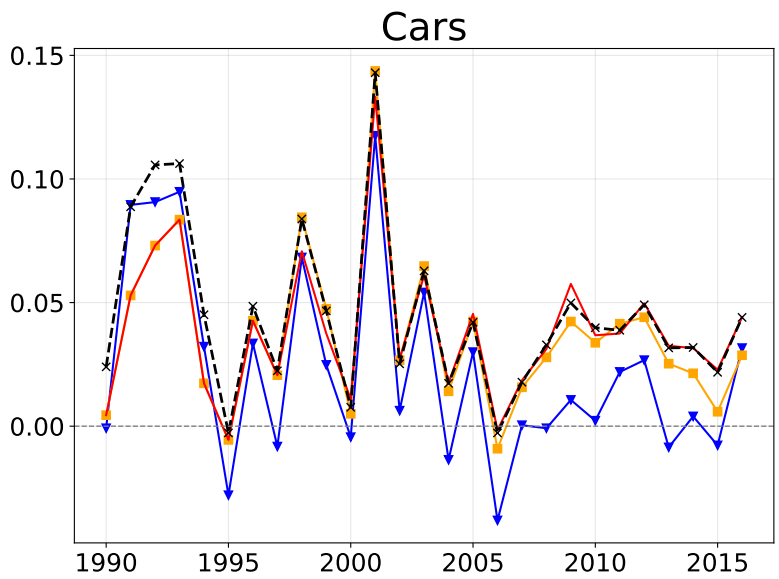


# Annual $R^2$ on 17 Industries (III)

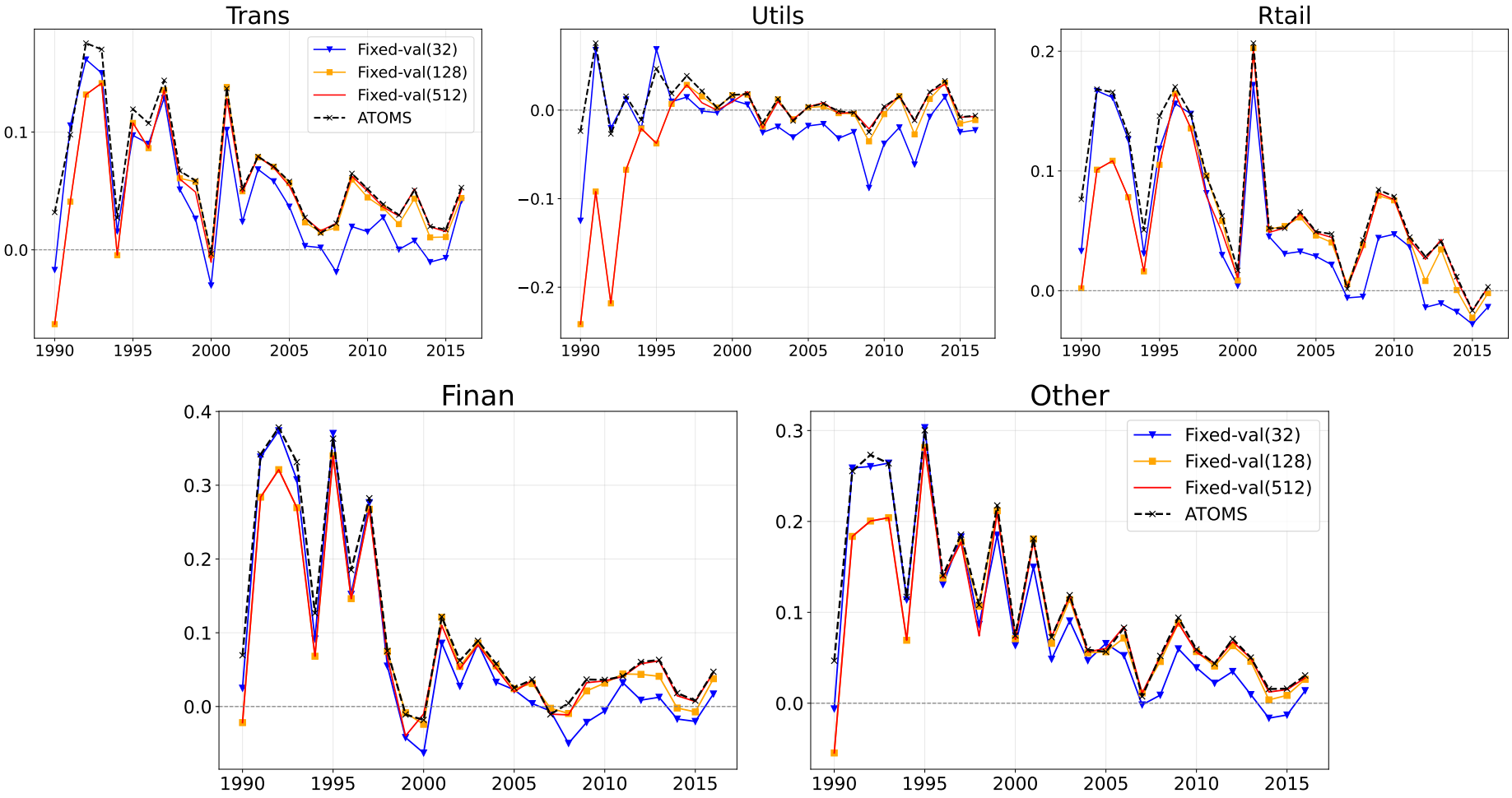


# Annual $R^2$ on 17 Industries (IV)



# Additional Portfolio Trading Experiments

# Long-Short Portfolio

Set weights

$$w_s^{(n)} = \begin{cases} 1/2, & \text{if } n = \text{argmax}_{1 \le i \le 17} \, \widehat{f}_t^{(i)}(x_s) \\ -1/2, & \text{if } n = \text{argmin}_{1 \le i \le 17} \, \widehat{f}_t^{(i)}(x_s) \\ 0, & \text{otherwise} \end{cases} .$$

# Average Monthly Sharpe Ratio for Long-Short Portfolio

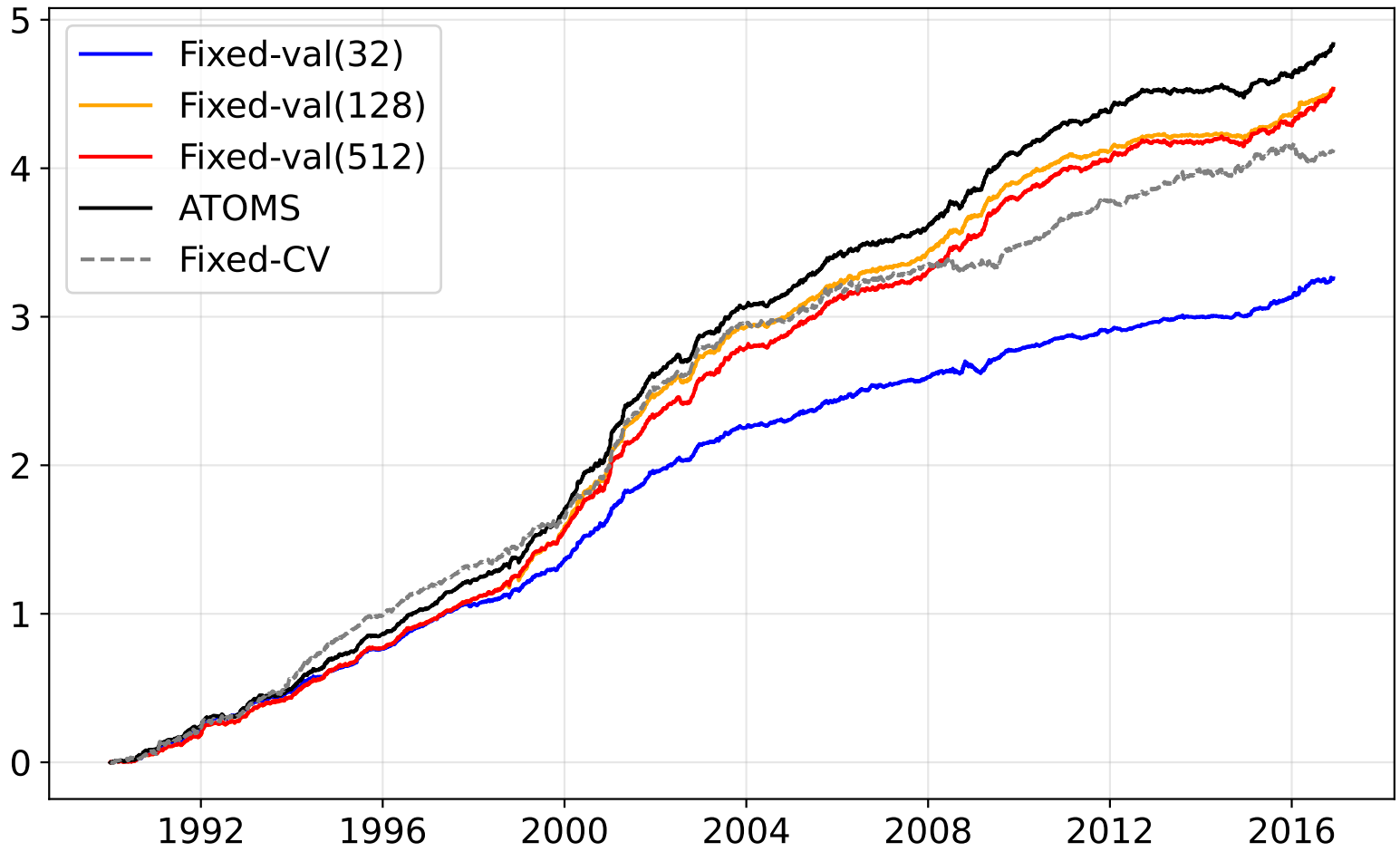
| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | **0.190** | **0.191** | **0.351** | 0.178 |
| Fixed-val(32) | 0.127 | 0.162 | 0.203 | 0.057 |
| Fixed-val(512) | 0.180 | 0.146 | 0.327 | **0.195** |
| Fixed-CV | 0.168 | 0.144 | 0.337 | 0.031 |

ATOMS has Sharpe ratios higher than or comparable to the best baseline.

# Average Monthly Turnover for Long-Short Portfolio

| **Method** | **Full OOS** | **Gulf War** | **2001** | **Fin. Crisis** |
|---|---|---|---|---|
| ATOMS | 17.65 | 22.40 | 21.90 | 18.76 |
| Fixed-val(32) | 23.53 | 23.68 | 25.33 | 24.64 |
| Fixed-val(512) | 17.96 | 24.46 | 22.90 | 18.06 |
| Fixed-CV | 15.41 | 17.68 | 15.42 | 12.69 |

# Cumulative Wealth for Long-Short Portfolio




Cumulative wealth is plotted in logarithmic scale.