Model Assessment and Selection under Temporal Distribution Shift

Elise Han [1] **Chengpiao Huang** [2] **Kaizheng Wang** [2 3]

Abstract

We investigate model assessment and selection in a changing environment, by synthesizing datasets from both the current time period and historical epochs. To tackle unknown and potentially arbitrary temporal distribution shift, we develop an adaptive rolling window approach to estimate the generalization error of a given model. This strategy also facilitates the comparison between any two candidate models by estimating the difference of their generalization errors. We further integrate pairwise comparisons into a single-elimination tournament, achieving near-optimal model selection from a collection of candidates. Theoretical analyses and empirical experiments underscore the adaptivity of our proposed methods to the non-stationarity in data.

1. Introduction

Traditionally, statistical learning theory assumes that models are trained and tested under the same data distribution. However, when practitioners train a model and deploy it into real environment, the environment often changes over time. Such temporal distribution shift may lead to serious decline in the model's quality. It is important to assess the model's performance in real time and detect possible degradation.

Moreover, one often needs to choose among multiple candidate models that result from different learning algorithms (e.g., linear regression, random forests, neural networks) and hyperparameters (e.g., penalty parameter, step size, time window for training). Temporal distribution shift poses a major challenge to model selection, as past performance may not reliably predict future outcomes. Learners usually have to work with limited data from the current time period and abundant historical data, whose distributions may vary significantly.

Main contributions. In this paper, we develop principled approaches to model assessment and model selection under temporal distribution shift.

- (Model assessment) We propose a rolling window strategy that adaptively selects historical data to estimate a model's generalization error.
- (Model selection) We then use the method above to compare any pair of models by applying it to the difference between their generalization errors. Based on this, we develop a single-elimination tournament procedure for selecting the best model from a pool of candidates.

Furthermore, we provide theoretical analyses and numericals experiments to show that our algorithms adapt to the unknown temporal distribution shift.

Related works. Model assessment and selection are classical problems in statistical learning (Hastie et al., 2009). Hold-out and cross-validation are arguably the most popular methods in practice. However, in the presence of distribution shift, the validation data may no longer accurately represent the test cases. This challenge has attracted considerable attention over the past two decades (Quinonero-Candela et al., 2022). Existing works mostly focused on the static scenario where the validation data consists of independent samples from a fixed distribution. These methods do not apply when the environment is continuously changing over time. Rolling windows offer a practical solution to this issue and have been widely adopted for learning under temporal changes (Bifet & Gavalda, 2007; Hanneke et al., 2015; Mohri & Muñoz Medina, 2012; Mazzetto & Upfal, 2023; Mazzetto et al., 2023; Huang & Wang, 2023). Our method automatically selects a window tailored to the underlying non-stationarity near a given time. The strategy is inspired by the *Goldenshluger-Lepski method* for bandwidth selection in non-parametric estimation (Goldenshluger & Lepski, 2008).

Adaptation to a non-stationary environment has recently gained popularity in online learning, where the goal is to

Author names are sorted alphabetically. [1]Department of Computer Science, Columbia University, New York, NY, United States [2]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, United States [3]Data Science Institute, Columbia University, New York, NY, United States. Correspondence to: Kaizheng Wang <kaizheng.wang@columbia.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

attain a small cumulative error over an extended time horizon (Hazan & Seshadhri, 2009; Besbes et al., 2015; Daniely et al., 2015; Jadbabaie et al., 2015; Wei & Luo, 2021; Gibbs & Candes, 2021; Bai et al., 2022). A classical problem in this area is prediction from expert advice (Littlestone & Warmuth, 1994), which aims to track the best expert in the long run. Recently, there have also been works on online model selection for bandits and reinforcement learning (Agarwal et al., 2017; Lee et al., 2021), with a similar goal of tracking the best base algorithm in the long run. In contrast, our approach focuses on evaluating a model or selecting from multiple candidates at a specific point in time, leveraging *offline* data collected from the past. Its performance is measured over the current data distribution. Consequently, our problem can be viewed as one of transfer learning, with source data from historical epochs and target data from the current distribution. We point out that our algorithms for the offline setting can be used as a sub-routine for online model selection, where the candidate models may vary across different time points.

Our problem is also related to but different from change-point detection (Niu et al., 2016; Truong et al., 2020). The latter typically assumes that distribution shifts occur only at a small number of times called *change points*. Our setting allows changes to happen in every time period with arbitrary magnitudes, covering a much broader range of shift patterns.

Outline. The rest of the paper is organized as follows. Section 2 describes the problem setup. Section 3 and Section 4 present our algorithms for model assessment and model selection, respectively. Section 5 conducts numerical experiments for the proposed algorithms on synthetic and real datasets. Finally, Section 6 concludes the paper and discusses future directions.

Notation. Let $\mathbb{Z}_+ = \{1, 2, ...\}$ be the set of positive integers. For $n \in \mathbb{Z}_+$, let $[n] = \{1, 2, ..., n\}$. For $x \in \mathbb{R}$, define $x_+ = \max\{x, 0\}$. For non-negative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = \mathcal{O}(b_n)$ if there exists $C > 0$ such that for all $n \in \mathbb{Z}_+$, $a_n \leq Cb_n$. Unless otherwise stated, $a_n \lesssim b_n$ also represents $a_n = \mathcal{O}(b_n)$. We write $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$.

2. Problem Setup

Let $\mathcal{Z}$ and $\mathcal{F}$ be spaces of samples and models, respectively. Denote by $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ a loss function. The quantity $\ell(f, z)$ measures the loss incurred by a model $f \in \mathcal{F}$ on a sample $z \in \mathcal{Z}$. At time $t \in \mathbb{Z}_+$, the quality of a model f is measured by its *generalization error* (also called *risk* or *population loss*) $L_t(f) = \mathbb{E}_{z\sim\mathcal{P}_t}\ell(f, z)$ over the current data distribution $\mathcal{P}_t$. When the environment changes over time, the distributions $\{\mathcal{P}_t\}_{t=1}^{\infty}$ can be different.

Suppose that at each time t, we receive a batch of $B_t \in \mathbb{Z}_+$ i.i.d. samples $\mathcal{B}_t = \{z_{t,i}\}_{i=1}^{B_t}$ from $\mathcal{P}_t$, independently of the history. We seek to solve the following two questions based on the data $\{\mathcal{B}_j\}_{j=1}^{t}$:

Problem 2.1 (Model assessment)**.** Given a model $f \in \mathcal{F}$, how to estimate its population loss $L_t(f)$?

Problem 2.2 (Model selection)**.** Given a collection of candidate models $\{f_r\}_{r=1}^{m} \subseteq \mathcal{F}$, how to select one with a small population loss? In other words, we want to choose $\widehat{r} \in [m]$ so that $L_t(f_{\widehat{r}}) \approx \min_{r\in[m]} L_t(f_r)$.

3. Model Assessment

In this section we study Problem 2.1, under the assumption that the loss ℓ is bounded.

Assumption 3.1 (Bounded loss)**.** The loss function ℓ takes values in a given interval $[a, b]$.

In fact, we will consider a more general problem.

Problem 3.2 (Mean estimation)**.** Let $\{\mathcal{Q}_j\}_{j=1}^{t}$ be probability distributions over $[a, b]$ and $\{\mathcal{D}_j\}_{j=1}^{t}$ be independent datasets, where $\mathcal{D}_j = \{u_{j,i}\}_{i=1}^{B_j}$ consists of $B_j \geq 1$ i.i.d. samples from $\mathcal{Q}_j$. Given $\{\mathcal{D}_j\}_{j=1}^{t}$, how to estimate the expectation μ_t of $\mathcal{Q}_t$?

Problem 2.1 is a special case of Problem 3.2 with $u_{j,i} = \ell(f, z_{j,i})$ and $\mu_t = L_t(f)$. To tackle Problem 3.2, a natural idea is to average data from the current period. More generally, we consider a *look-back window* of the k most recent periods, and approximate μ_t by their sample average:

$$\widehat{\mu}_{t,k} = \left(\sum_{j=t-k+1}^{t} B_j\right)^{-1} \sum_{j=t-k+1}^{t} \sum_{i=1}^{B_j} u_{j,i}. \qquad (1)$$

To measure the quality of $\widehat{\mu}_{t,k}$, we invoke the Bernstein inequality (Boucheron et al., 2013). See Appendix A.1 for its proof.

Lemma 3.3 (Bernstein bound)**.** *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 $1 - \delta$,*

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

Denote by μ_j and σ_j^2 the mean and the variance of $\mathcal{Q}_j$, respectively. Let $B_{t,k} = \sum_{j=t-k+1}^{t} B_j$,

$$\mu_{t,k} = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j\mu_j, \quad \sigma_{t,k}^2 = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j\sigma_j^2.$$

Using Lemma 3.3 and the triangle inequality

$$|\widehat{\mu}_{t,k} - \mu_t| \le |\widehat{\mu}_{t,k} - \mu_{t,k}| + |\mu_{t,k} - \mu_t|,$$

we obtain the following bias-variance decomposition of the approximation error of $\widehat{\mu}_{t,k}$.

Corollary 3.4. *For $t \in \mathbb{Z}_+$, $k \in [t]$ and $\delta \in (0,1)$, define $M = b - a$,*

$$\phi(t,k) = \max_{t-k+1 \le j \le t} |\mu_j - \mu_t|,$$

$$\psi(t,k,\delta) = \begin{cases} M, & \text{if } B_{t,k} = 1 \\ \sigma_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{2M\log(2/\delta)}{3B_{t,k}}, & \text{if } B_{t,k} \ge 2 \end{cases}.$$

With probability at least $1-\delta$,

$$|\widehat{\mu}_{t,k} - \mu_t| \le \phi(t,k) + \psi(t,k,\delta).$$

Here $\phi(t,k)$ upper bounds the bias induced by using data from the k most recent periods, while $\psi(t,k,\delta)$ upper bounds the statistical uncertainty associated with $\widehat{\mu}_{t,k}$. Ideally, we would like to choose a window k^* that minimizes the bias-variance decomposition $\phi(t,k) + \psi(t,k,\delta)$. This optimal window k^* depends on the pattern of distribution shift: it is large when the environment is near-stationary, and is small when the environment is heavily fluctuating. However, such structural information is generally unavailable in practice. This is reflected in the fact that both ϕ and ψ involve unknown quantities $|\mu_j - \mu_k|$ and $\sigma_{t,k}$ that depend on the unknown distribution shift. As a consequence, direct minimization of $\phi(t,k) + \psi(t,k,\delta)$ over $k \in [t]$ is infeasible. In the following, we will construct proxies for ψ and ϕ that are computable from data.

We first construct a proxy for $\psi(t,k,\delta)$. A natural approximation of $\sigma^2_{t,k}$ is the sample variance over the k most recent periods, given by

$$\widehat{v}^2_{t,k} = \frac{1}{B_{t,k}-1} \sum_{j=t-k+1}^{t} \sum_{i=1}^{B_j} (u_{j,i} - \widehat{\mu}_{t,k})^2. \quad (2)$$

It has been used for deriving empirical versions of the Bernstein inequality (Audibert et al., 2007; Maurer & Pontil, 2009). Based on the estimate $\widehat{v}^2_{t,k}$, we define our proxy as

$$\widehat{\psi}(t,k,\delta) = \begin{cases} M, & \text{if } B_{t,k} = 1 \\ \widehat{v}_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{8M\log(2/\delta)}{3(B_{t,k}-1)}, & \text{if } B_{t,k} \ge 2 \end{cases}. \quad (3)$$

As Lemma 3.5 shows that with high probability, $\widehat{\psi}$ upper bounds ψ and their gap is not too large. Its proof is deferred to Appendix A.2.

Lemma 3.5. *For any $\delta \in (0,1)$, define $\xi(t,k,\delta) =$*

$$\sqrt{\frac{4\log(2/\delta)}{B_{t,k}}} \max_{t-k+1 \le j \le t} |\mu_j - \mu_t| + \frac{13(b-a)\log(2/\delta)}{3(B_{t,k}-1)}$$

if $B_{t,k} \ge 2$. If $B_{t,k} = 1$, define $\xi(t,k,\delta) = 0$. Then

$$\mathbb{P}\left(\psi(t,k,\delta) \le \widehat{\psi}(t,k,\delta)\right) \ge 1-\delta$$

$$\mathbb{P}\left(\widehat{\psi}(t,k,\delta) \le \psi(t,k,\delta) + \xi(t,k,\delta)\right) \ge 1-\delta.$$

Combining Corollary 3.4 with the first bound in Lemma 3.5 immediately gives the following useful corollary.

Corollary 3.6. *Let $\delta \in (0,1)$. With probability at least $1 - 2t\delta$, it holds*

$$|\widehat{\mu}_{t,k} - \mu_t| \le \phi(t,k) + \widehat{\psi}(t,k,\delta), \quad \forall k \in [t]. \quad (4)$$

To construct a proxy for $\phi(t,k)$, we borrow ideas from the Goldenshluger-Lepski method for adaptive non-parametric estimation (Goldenshluger & Lepski, 2008). Define

$$\widehat{\phi}(t,k,\delta) = \max_{i \in [k]} \left(|\widehat{\mu}_{t,k} - \widehat{\mu}_{t,i}| - \left[\widehat{\psi}(t,k,\delta) + \widehat{\psi}(t,i,\delta)\right] \right)_+. \quad (5)$$

We now give an interpretation of $\widehat{\phi}$. In light of the bias-variance decomposition in Corollary 3.4, the quantity

$$|\widehat{\mu}_{t,k} - \widehat{\mu}_{t,i}| - [\widehat{\psi}(t,k,\delta) + \widehat{\psi}(t,i,\delta)] \quad (6)$$

can be viewed as a measure of the bias between the windows k and i, where subtracting $\widehat{\psi}(t,k,\delta)$ and $\widehat{\psi}(t,i,\delta)$ eliminates the stochastic error and teases out the bias. Indeed, as the following lemma shows, $\widehat{\phi}(t,k,\delta) \le 2\phi(t,k)$ holds with high probability. Its proof is given in Appendix A.3.

Lemma 3.7. *When the event (4) happens,*

$$0 \le \widehat{\phi}(t,k,\delta) \le 2\phi(t,k).$$

We take the positive part in (5) so that when the quantity (6) is negative, we regard the bias as dominated by the stochastic error and hence negligible. Taking maximum over all windows $i \in [k]$ makes sure that we detect all possible biases between window k and the smaller windows.

Replacing ψ and ϕ with their proxies $\widehat{\psi}$ and $\widehat{\phi}$ gives Algorithm 1. We note that the quantities $\widehat{\mu}_{t,k}$ and $\widehat{v}_{t,k}$ can be conveniently computed from summary statistics of individual datasets. Define empirical first and second momends in the j-th time period, $\widehat{\mu}_j = B_j^{-1}\sum_{i=1}^{B_j} u_{j,i}$ and

Algorithm 1 Adaptive Rolling Window for Mean Estimation (Problem 3.2)

Input: Data $\{\mathcal{D}_j\}_{j=1}^t$, hyperparameters δ' and M.
for $k = 1, \cdots, t$ **do**
  Compute $\widehat{\mu}_{t,k}$, $\widehat{v}_{t,k}$, $\widehat{\psi}(t,k,\delta')$ and $\widehat{\phi}(t,k,\delta')$ according to (1), (2), (3) and (5).
end for
Choose any $\widehat{k} \in \operatorname{argmin}_{k\in[t]}\{\widehat{\phi}(t,k,\delta') + \widehat{\psi}(t,k,\delta')\}$.
Output: $\widehat{\mu}_{t,\widehat{k}}$.

$\widehat{\omega}_j = B_j^{-1}\sum_{i=1}^{B_j} u_{j,i}^2$. Then

$$\widehat{\mu}_{t,k} = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j \widehat{\mu}_j, \tag{7}$$

$$\widehat{v}_{t,k}^2 = \frac{B_{t,k}}{B_{t,k}-1} \cdot \left(\frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j \widehat{\omega}_j - \widehat{\mu}_{t,k}^2 \right). \tag{8}$$

We now present theoretical guarantees for Algorithm 1.

Theorem 3.8 (Oracle inequality)**.** *Let Assumption 3.1 hold. Choose $\delta \in (0,1)$ and take $\delta' = \delta/(3t)$ in Algorithm 1. With probability at least $1-\delta$, the output of Algorithm 1 satisfies*

$$|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \lesssim \min_{k\in[t]}\{\phi(t,k) + \psi(t,k,\delta)\}.$$

Here $\lesssim$ only hides a logarithmic factor of t and δ^{-1}.

Theorem 3.8 states that the selected window $\widehat{k}$ is near-optimal for the error bound in Corollary 3.4 derived from bias-variance decomposition. We illustrate this oracle property under the following distribution shift patterns.

Example 3.9 (Change point)**.** *Suppose that the environment remained unchanged over the last K periods but had been very different before, i.e. $\mathcal{Q}_{t-K} \neq \mathcal{Q}_{t-K+1} = \cdots = \mathcal{Q}_t$. If K were known, one could take the window size K and output $\widehat{\mu}_{t,K}$ as an estimate of μ_t. By Theorem 3.8, $\widehat{\mu}_{t,\widehat{k}}$ is at least comparable to $\widehat{\mu}_{t,K}$ in terms of the estimation error: up to a logarithmic factor,*

$$|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \lesssim \frac{\sigma_t}{\sqrt{B_{t,K}}} + \frac{1}{B_{t,K}}.$$

Therefore, Algorithm 1 automatically adapts to the local stationarity and is comparable to using $B_{t,K}$ i.i.d. samples.

Example 3.10 (Bounded drift)**.** *Suppose that the distribution shift between consecutive periods is bounded, i.e. there exists $\Delta > 0$ such that for each $j \in \mathbb{Z}_+$, $|\mu_{j+1} - \mu_j| \le \Delta$. This is a common assumption in the literature (Bartlett, 1992; Helmbold & Long, 1994; Barve & Long, 1997; Mohri & Muñoz Medina, 2012). The quantity Δ characterizes the non-stationarity: as Δ grows larger, the environment is allowed to fluctuate more wildly. For simplicity, we further assume $B_j = 1$ for all $j \in \mathbb{Z}_+$. In this case, $\phi(t,k) \le (k-1)\Delta$, so the bias-variance decomposition in Corollary 3.4 becomes*

$$|\widehat{\mu}_{t,k} - \mu| \lesssim (k-1)\Delta + \sqrt{\frac{1}{k}}$$

up to a logarithmic factor. If Δ were known, then one could pick the optimal window size $k^* \asymp \Delta^{-2/3}$, with an estimation error of $\mathcal{O}(\Delta^{1/3})$ which is known to be optimal (Barve & Long, 1997). Theorem 3.8 shows that without knowing Δ, Algorithm 1 achieves the optimal order of estimation error.

We now turn to the proof of Theorem 3.8. A key ingredient is the following lemma, which can be seen as an empirical version of Theorem 3.8 with ψ replaced by $\widehat{\psi}$.

Lemma 3.11. *When the event (4) happens,*

$$|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \le 3 \min_{k\in[t]} \left\{ \phi(t,k) + \widehat{\psi}(t,k,\delta) \right\}.$$

Proof. For any $k \in [\widehat{k}]$,

$$\begin{aligned}
|\widehat{\mu}_{t,\widehat{k}} - \mu_t| &\le |\widehat{\mu}_{t,\widehat{k}} - \widehat{\mu}_{t,k}| + |\widehat{\mu}_{t,k} - \mu_t| \\
&\le \left[\widehat{\phi}(t,\widehat{k},\delta) + \widehat{\psi}(t,\widehat{k},\delta) + \widehat{\psi}(t,k,\delta)\right] \\
&\quad + \left[\phi(t,k) + \widehat{\psi}(t,k,\delta)\right] && \text{(by (5) and (4))} \\
&= \left[\widehat{\phi}(t,\widehat{k},\delta) + \widehat{\psi}(t,\widehat{k},\delta)\right] + \left[\phi(t,k) + 2\widehat{\psi}(t,k,\delta)\right] \\
&\le \left[\widehat{\phi}(t,k,\delta) + \widehat{\psi}(t,k,\delta)\right] + \left[\phi(t,k) + 2\widehat{\psi}(t,k,\delta)\right] \\
& && \text{(by the definition of } \widehat{k}) \\
&\le \left[2\phi(t,k) + \widehat{\psi}(t,k,\delta)\right] + \left[\phi(t,k) + 2\widehat{\psi}(t,k,\delta)\right] \\
& && \text{(by Lemma 3.7)} \\
&= 3\left[\phi(t,k) + \widehat{\psi}(t,k,\delta)\right].
\end{aligned}$$

On the other hand, for any $k \in \{\widehat{k}+1, \cdots, t\}$,

$$\begin{aligned}
|\widehat{\mu}_{t,\widehat{k}} - \mu_t| &\le \phi(t,\widehat{k}) + \widehat{\psi}(t,\widehat{k},\delta) && \text{(by (4))} \\
&\le \phi(t,k) + \left[\widehat{\psi}(t,\widehat{k},\delta) + \widehat{\phi}(t,\widehat{k},\delta)\right] && \text{(by Lemma 3.7)} \\
&\le \phi(t,k) + \left[\widehat{\psi}(t,k,\delta) + \widehat{\phi}(t,k,\delta)\right] \\
& && \text{(by the definition of } \widehat{k}) \\
&\le 3\left[\phi(t,k) + \widehat{\psi}(t,k,\delta)\right]. && \text{(by Lemma 3.7)}
\end{aligned}$$

The proof is finished by taking the minimum over k. □

Combining Lemma 3.11 and the second bound in Lemma 3.5 proves Theorem 3.8. We provide a full proof in Appendix A.4, along with a more precise bound.

Remark 3.12 (Boundedness of loss). The boundedness Assumption 3.1 can be relaxed to a light-tailed condition. For instance, we may assume that $\ell(f, z)$ is sub-exponential for $f \in \mathcal{F}$ and $z \sim \mathcal{P}_j$, and then apply a standard truncation argument. Note that for any sub-exponential random variables $\{v_i\}_{i=1}^n$, the bound $\max_{i\in[n]} |v_i| = \mathcal{O}(\log n)$ holds with high probability. Hence, we can truncate the random variables at a logarithmic level and then apply Bernstein's inequality. This will only incur an extra logarithmic factor in the oracle inequality.

Remark 3.13 (Logarithmic factor in the oracle inequality). The $\mathcal{O}(\log t)$ dependence in Theorem 3.8 comes from taking union bound on Corollary 3.4 over all $k \in [t]$. It is in fact possible to improve it to a sharp bound $\mathcal{O}(\log\log t)$, by using advanced techniques from the law of iterated logarithm and martingale concentration (Jamieson et al., 2014; Balsubramani & Ramdas, 2016).

4. Model Selection

In this section, we consider Problem 2.2. We will first study the case of selection between two models, and then extend the approach to the general case of $m \in \mathbb{Z}_+$ models.

4.1. Warmup: Model Comparison

We first consider the case $m = 2$, where the goal is to compare two models f_1 and f_2, and choose the better one. As in Section 3, using a look-back window $k \in [t]$, we can estimate $L_t(f)$ by

$$\widehat{L}_{t,k}(f) = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} \sum_{i=1}^{B_j} \ell(f, z_{j,i}).$$

We will choose $k \in [t]$ and return

$$\widehat{r}_k \in \operatorname*{argmin}_{r\in[2]} \widehat{L}_{t,k}(f_r).$$

Our approach is based on the following key observation, proved in Appendix B.1.

Lemma 4.1. *For every $k \in [t]$, the index $\widehat{r}_k$ satisfies*

$$L_t(f_{\widehat{r}_k}) - \min_{r\in[2]} L_t(f_r) \le \left| \left[\widehat{L}_{t,k}(f_1) - \widehat{L}_{t,k}(f_2)\right] - \left[L_t(f_1) - L_t(f_2)\right] \right|.$$

Define

$$\Delta_j = L_j(f_1) - L_j(f_2)$$

and $\widehat{\Delta}_{t,k} = \widehat{L}_{t,k}(f_1) - \widehat{L}_{t,k}(f_2)$. Then, Δ_j is the performance gap between f_1 and f_2 at time j, and $\widehat{\Delta}_{t,k}$ is a sample average. By Lemma 4.1, it suffices to choose k such that $|\widehat{\Delta}_{t,k} - \Delta_t|$ is small. That is, *an accurate estimate of the performance gap guarantees near-optimal selection.*

Algorithm 2 Adaptive Rolling Window for Model Comparison

Input: Models f_1 and f_2, data $\{\mathcal{B}_j\}_{j=1}^t$, hyperparameters δ' and M.
Let $u_{j,i} = \ell(f_1, z_{j,i}) - \ell(f_2, z_{j,i})$ and $\mathcal{D}_j = \{u_{j,i}\}_{i=1}^{B_j}$.
Run Algorithm 1 with inputs $\{\mathcal{D}_j\}_{j=1}^t$, δ' and M to obtain $\widehat{\Delta}_{t,\widehat{k}}$. Let $\widehat{r}_{\widehat{k}} = \begin{cases} 1, & \text{if } \widehat{\Delta}_{t,\widehat{k}} \le 0 \\ 2, & \text{if } \widehat{\Delta}_{t,\widehat{k}} > 0 \end{cases}$.
Output: $\widehat{f} = f_{\widehat{r}_{\widehat{k}}}$.

This reduces the problem to Problem 3.2, with $u_{j,i} = \ell(f_1, z_{j,i}) - \ell(f_2, z_{j,i})$ and thus $\mu_t = \Delta_t$. We can then readily apply Algorithm 1. The detailed description is given in Algorithm 2.

Theorem 3.8 and Lemma 4.1 directly yield the following guarantee of Algorithm 2.

Theorem 4.2 (Oracle inequality)**.** *Let Assumption 3.1 hold. Choose $\delta \in (0, 1)$ and take $\delta' = \delta/(3t)$ in Algorithm 2. With probability at least $1 - \delta$, Algorithm 2 outputs $\widehat{f}$ satisfying*

$$L_t(\widehat{f}) - \min_{r\in[2]} L_t(f_r) \lesssim \min_{k\in[t]} \left\{ \max_{t-k+1\le j\le t} |\Delta_j - \Delta_t| + \frac{\widetilde{\sigma}_{t,k}}{\sqrt{B_{t,k}}} + \frac{1}{B_{t,k}} \right\},$$

where

$$\widetilde{\sigma}_{t,k}^2 = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j \operatorname{var}_{z\sim\mathcal{P}_j}[\ell(f_1, z) - \ell(f_2, z)]$$

and $\lesssim$ only hides a logarithmic factor of t and δ^{-1}.

Consider again Example 3.9 where $\mathcal{P}_{t-K} \ne \mathcal{P}_{t-K+1} = \cdots = \mathcal{P}_t$ for some K. Theorem 4.2 admits a similar interpretation as Theorem 3.8: Algorithm 2 selects $\widehat{f}$ satisfying

$$L_t(\widehat{f}) - \min_{r\in[2]} L_t(f_r) \lesssim \frac{\widetilde{\sigma}_t}{\sqrt{B_{t,K}}} + \frac{1}{B_{t,K}} \lesssim \frac{1}{\sqrt{B_{t,K}}}, \quad (9)$$

where $\widetilde{\sigma}_t^2 = \operatorname{var}_{z\sim\mathcal{P}_t}[\ell(f_1, z) - \ell(f_2, z)]$.

In the setting of bounded regression without covariate shift, we may further improve the rate in (9). To state the results, we let $\mathcal{X}$ be a feature space and consider the following assumptions.

Assumption 4.3 (Bounded response)**.** For $j \in \mathbb{Z}_+$, a sample $z \sim \mathcal{P}_j$ takes the form $z = (x, y)$, where $x \in \mathcal{X}$ is the covariate and $y \in \mathbb{R}$ is the response. There exists $M_0 > 0$ such that $|y| \le M_0$ holds for $(x, y) \sim \mathcal{P}_j$ and $j \in \mathbb{Z}_+$.

Assumption 4.4 (Bounded models)**.** The loss ℓ is given by $\ell(f,(x,y)) = [f(x)-y]^2$. For all $x \in \mathcal{X}$ and $f \in \mathcal{F}$, $|f(x)| \le M_0$.

Assumption 4.5 (No covariate shift)**.** The distributions $\{\mathcal{P}_j\}_{j=1}^{\infty}$ have the same marginal distribution of the covariate, denoted by $\mathcal{P}$.

Assumptions 4.3 and 4.4 imply Assumption 3.1 with $a = 0$ and $b = 4M_0^2$. In Assumption 4.5, the distribution of the covariate $x_{j,1}$ is the same for all $j \in \mathbb{Z}_+$, while the conditional distribution of $y_{j,1}$ given $x_{j,1}$ may experience shifts. The latter is commonly known as *concept drift* (Gama et al., 2014).

Before we state the result, we introduce a few notations. Define $f_j^*(\cdot) = \mathbb{E}(y_{j,1}|x_{j,1} = \cdot)$, which minimizes the mean square error $\mathbb{E}[f(x)-y]^2$ over the class of all measurable $f: \mathcal{X} \to \mathbb{R}$. For $h_1, h_2 \in \mathcal{F}$, define an inner product

$$\langle h_1, h_2\rangle_{\mathcal{P}} = \mathbb{E}_{x\sim\mathcal{P}}\left[h_1(x)h_2(x)\right],$$

which induces a norm $\|h\|_{\mathcal{P}} = \sqrt{\langle h,h\rangle_{\mathcal{P}}}$. It can be readily checked that $L_t(f) - L_t(f_t^*) = \|f - f_t^*\|_{\mathcal{P}}^2$ for all $f \in \mathcal{F}$. Thus, we may measure the performance of a model $f \in \mathcal{F}$ by

$$\|f - f_t^*\|_{\mathcal{P}} = \sqrt{L_t(f) - L_t(f_t^*)},$$

as it admits the interpretation of being both the distance between f and f_t^* under $\|\cdot\|_{\mathcal{P}}$, and the square root of the excess risk $L_t(f) - L_t(f_t^*)$. For $j \ne t$, the quantity $\|f_j^* - f_t^*\|_{\mathcal{P}}$ serves as a measure of distribution shift between time j and time t.

We are now ready to state our result. In Appendix B.2 we provide a more precise bound and its proof.

Theorem 4.6 (Fast rate)**.** *Let Assumptions 4.3, 4.4 and 4.5 hold. Let M_0 be a constant. Choose $\delta \in (0,1)$ and take $\delta' = \delta/(3t)$ in Algorithm 2. With probability at least $1-\delta$, Algorithm 2 outputs $\widehat{f}$ satisfying*

$$\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[2]}\|f_r - f_t^*\|_{\mathcal{P}} \lesssim \min_{k\in[t]}\left\{\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{1}{\sqrt{B_{t,k}}}\right\}.$$

Here $\lesssim$ only hides a logarithmic factor of t and δ^{-1}.

The oracle inequality in Theorem 4.6 shares the same bias-variance structure as that of Theorem 4.2. By squaring both sides of the bound, we see that in Example 3.9 where $\mathcal{P}_{t-K} \ne \mathcal{P}_{t-K+1} = \cdots = \mathcal{P}_t$,

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \min_{r\in[2]} L_t(f_r) - L_t(f_t^*) + \frac{1}{B_{t,K}}. \quad (10)$$

When either f_1 or f_2 has a small error so that $\min_{r\in[2]} L_t(f_r) - L_t(f_t^*) = o(1/\sqrt{B_{t,K}})$, (10) provides a much sharper guarantee on $L_t(\widehat{f})$ compared to (9). As the proof of Theorem 4.6 reveals, such an improvement relies crucially on the structure of $\mathrm{var}\left[\ell(f_1,z) - \ell(f_2,z)\right]$ in the Bernstein bound. In particular, it cannot be achieved by the naïve method of applying Algorithm 1 to f_1 and f_2 separately and choosing the one with a lower estimated generalization error. We believe that in other scenarios such as binary classification, our analysis still goes through under commonly used noise conditions in learning theory (Bartlett et al., 2005; Boucheron et al., 2005).

Algorithm 3 Single-Elimination Tournament for Model Selection

Input: Models $\{f_r\}_{r=1}^m$, data $\{\mathcal{B}_j\}_{j=1}^t$, hyperparameters δ' and M
Let $S = \lceil \log_2 m \rceil$ and $\mathcal{F}_1 = \{f_r : r \in [m]\}$.
for $s = 1, \cdots, S$ **do**
  Split $\mathcal{F}_s$ into disjoint pairs:

$$\mathcal{F}_s = \mathcal{G}_{s,1} \cup \cdots \cup \mathcal{G}_{s,m_s} \cup \mathcal{G}_{s,m_s+1},$$

  where $|\mathcal{G}_{s,i}| = 2$ for $i \in [m_s]$, and $|\mathcal{G}_{s,m_s+1}| \le 2$.
  for $i = 1, \cdots, m_s + 1$ **do**
    Run Algorithm 2 with inputs $\mathcal{G}_{s,i}$, $\{\mathcal{D}_j\}_{j=1}^t$, δ' and M to obtain $\widehat{g}_{s,i} \in \mathcal{G}_{s,i}$.
    If $|\mathcal{G}_{s,m_s+1}| = 1$, simply take $\widehat{g}_{s,m_s+1} \in \mathcal{G}_{s,m_s+1}$.
  end for
  Let $\mathcal{F}_{s+1} = \{\widehat{g}_{s,i} : i \in [m_s+1]\}$.
end for
Output: The only model $\widehat{f} \in \mathcal{F}_S$.

4.2. Selection from Multiple Candidates

We now consider the general case of selecting over $m \in \mathbb{Z}_+$ models $\{f_r\}_{r=1}^m$. We will use a straightforward single-elimination tournament procedure. In each round, we pair up the remaining models, and use Algorithm 2 to perform pairwise comparison. Within each pair, the model picked by Algorithm 2 advances to the next round. When there is only one model left, the procedure terminates and outputs the model. Algorithm 3 gives the detailed description.

Here $\mathcal{F}_{s+1}$ is the set of remaining models after round s. By design, Algorithm 3 eliminates about half of the remaining models in each round: $|\mathcal{F}_{s+1}| \le \lceil |\mathcal{F}_s|/2 \rceil$. Thus, only one model remains after $\lceil \log_2 m \rceil$ rounds. Since each call of Algorithm 2 eliminates one model, then Algorithm 3 calls Algorithm 2 exactly $m-1$ times.

We now give the theoretical guarantee of Algorithm 3 in the setting of bounded regression. We provide a more precise bound and its proof in Appendix B.3.

Theorem 4.7 (Oracle inequality)**.** *Let Assumptions 4.3, 4.4 and 4.5 hold. Let M_0 be a constant. Choose $\delta \in (0,1)$ and*

Algorithm 4 Fixed-Window Model Selection Algorithm

Input: Models $\{f_r\}_{r=1}^m$, data $\{\mathcal{B}_j\}_{j=1}^t$, window size k.
Compute $s = \max\{t, k\}$.
for $r = 1, \cdots, m$ **do**
$\widehat{L}_{t,k}(f_r) = \frac{1}{B_{t,s}} \sum_{j=t-s+1}^{t} \sum_{i=1}^{B_j} \ell(f_r, z_{j,i})$.
end for
Compute $\widehat{r} \in \operatorname{argmin}_{r\in[m]} \widehat{L}_{t,k}(f_r)$.
Output: $\widehat{f}_k = f_{\widehat{r}}$.

take $\delta' = \delta/(3m^2 t)$ in Algorithm 3. With probability at least $1-\delta$, Algorithm 3 outputs $\widehat{f}$ satisfying

$$\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[m]} \|f_r - f_t^*\|_{\mathcal{P}} \lesssim \min_{k\in[t]} \left\{ \max_{t-k+1\le j\le t} \|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{1}{\sqrt{B_{t,k}}} \right\}.$$

Here $\lesssim$ hides a polylogarithmic factor of m, t and δ^{-1}.

We remark that Theorem 4.7 takes the same form as Theorem 4.6 up to a factor of $\log_2 m$. Thus, in Example 3.9 where $\mathcal{P}_{t-K} \neq \mathcal{P}_{t-K+1} = \cdots = \mathcal{P}_t$, the model $\widehat{f}$ selected by Algorithm 3 also enjoys a fast rate similar to (10):

$$L_t(\widehat{f}) - L_t(f_t^*) \lesssim \min_{r\in[m]} L_t(f_r) - L_t(f_t^*) + \frac{1}{B_{t,K}}.$$

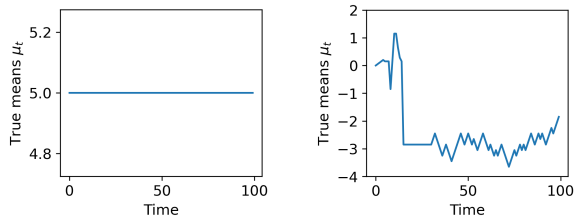
5. Numerical Experiments

We conduct simulations to verify our theories and test our algorithms on three real datasets. We will focus on Algorithm 3 for Problem 2.2. For simplicity, the hyperparameters δ' and M in our algorithms are set to be 0.1 and 0 throughout our numerical study. For notational convenience, we denote Algorithm 3 by $\mathcal{V}_{\mathrm{ARW}}$. We compare $\mathcal{V}_{\mathrm{ARW}}$ with fixed-window model selection procedures, given by Algorithm 4 and denoted by $\mathcal{V}_k$, where k is the fixed window size and takes values in a set $I = \{1, 4, 16, 64, 256\}$. Our code is available at https://github.com/eliselyhan/ARW.

5.1. Synthetic Data

We first train different models for a mean estimation task on synthetic data. Then, we deploy Algorithms 3 and 4 to select top-performing models. Finally, we assess and compare their qualities. Throughout the simulations, we consider 100 time periods. At each time period t, we generate a batch of $B_t \in \mathbb{Z}_+$ i.i.d. samples $\mathcal{B}_t = \{z_{t,i}\}_{i=1}^{B_t}$. In our setup, each sample $z_{t,i} \sim \mathcal{N}(\mu_t, \sigma^2)$, where μ_t is the given population mean of period t. The task of the models is to estimate μ_t.

At each period t, we split $\mathcal{B}_t$ into a training set $\mathcal{B}_t^{\mathrm{tr}}$ and a validation set $\mathcal{B}_t^{\mathrm{va}}$. The models are trained on the training set, and subsequent model selection is done using the validation set. The size of the validation set B_t^{va} is sampled uniformly from $\{2, 3, 4\}$. The size of the training set is $B_t^{\mathrm{tr}} = 3B_t^{\mathrm{va}}$.



(a) Example 5.1 (b) Example 5.2

Figure 1. True means $\{\mu_t\}_{t=0}^{100}$ in synthetic data.

Table 1. Mean excess risks of selection methods for Example 5.1.

$\mathcal{V}_{\mathrm{ARW}}$	$\mathcal{V}_1$	$\mathcal{V}_4$	$\mathcal{V}_{16}$	$\mathcal{V}_{64}$	$\mathcal{V}_{256}$
0.015	0.043	0.025	0.013	0.010	0.010
1.293	4.117	2.572	1.396	1.015	0.982

In each period t, we consider 5 estimates $\{\mathcal{T}(t,w)\}_{w\in I}$ of the target μ_t, where $\mathcal{T}(t,w)$ is the sample average of the data $\{\mathcal{B}_j^{\mathrm{tr}}\}_{j=t-w+1}^t$. We then select a model based on data $\{\mathcal{B}_j^{\mathrm{va}}\}_{j=1}^t$. To measure its quality, we compute the excess risk, which is the squared difference between the true mean μ_t and the selected estimate. We compare $\mathcal{V}_{\mathrm{ARW}}$ with the fixed-window benchmarks $\{\mathcal{V}_k\}_{k\in I}$. We use the following scenarios as testbeds.

Example 5.1. *Figure 1(a) illustrates the stationary case where the true mean stays constant. We explore both low-variance and high-variance regimes with $\sigma^2 = 1$ and 10, respectively. Table 1 records the average excess risks over 100 periods and 20 independent trials for $\mathcal{V}_{\mathrm{ARW}}$ and $\{\mathcal{V}_k\}_{k\in I}$. The first and second rows correspond to $\sigma^2 = 1$ and $\sigma^2 = 10$. In both regimes, $\mathcal{V}_{\mathrm{ARW}}$ leverages the underlying stationarity and yields excess risks comparable to $\mathcal{V}_k$ with large k's, whereas $\mathcal{V}_k$ with smaller k's perform worse. In Figure 2, we plot the average excess risks over 20 trials at each time t for $\mathcal{V}_{\mathrm{ARW}}, \mathcal{V}_1, \mathcal{V}_{256}$ in red, orange and blue.*

Example 5.2. *We carry out the same experiments in a scenario with sufficient non-stationarity. The true means $\{\mu_t\}_{t=1}^{100}$ in Figure 1(b) are generated using a randomized mechanism; see Appendix C.1. Similar to that of Example 5.1, Table 2 summarizes the mean excess risks. When $\sigma^2 = 1$, the non-stationarity pattern of the underlying means is largely preserved. Our algorithm $\mathcal{V}_{\mathrm{ARW}}$ outperforms all $\mathcal{V}_k$'s, demonstrating its adaptivity. When $\sigma^2 = 10$, the larger noise makes the non-stationarity less significant, so $\mathcal{V}_k$ with larger k's are be more stable. Nevertheless, $\mathcal{V}_{\mathrm{ARW}}$ is still competitive with them. We also plot the average excess risks over 20 trials at each time t in Figure 3.*

Table 2. Mean excess risks of selection methods for Example 5.2.

$\mathcal{V}_{\mathrm{ARW}}$	$\mathcal{V}_1$	$\mathcal{V}_4$	$\mathcal{V}_{16}$	$\mathcal{V}_{64}$	$\mathcal{V}_{256}$
0.139	0.157	0.171	0.539	1.034	1.067
2.052	4.425	2.934	1.920	1.771	1.784

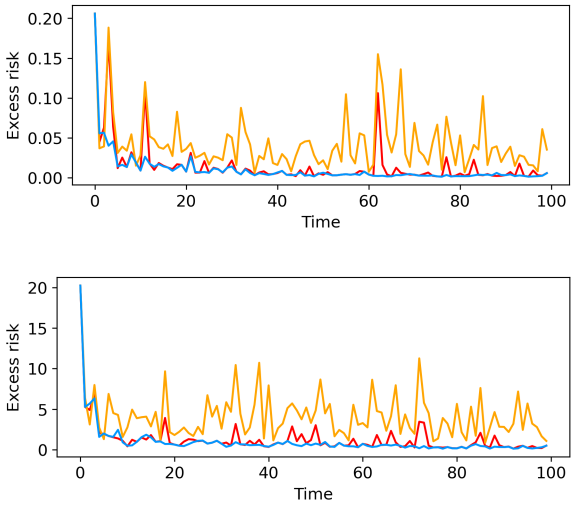


Figure 2. Excess risks in Example 5.1: $\sigma^2 = 1$ (top) and $\sigma^2 = 10$ (bottom). Red: $\mathcal{V}_{\mathrm{ARW}}$. Orange: $\mathcal{V}_1$. Blue: $\mathcal{V}_{256}$.

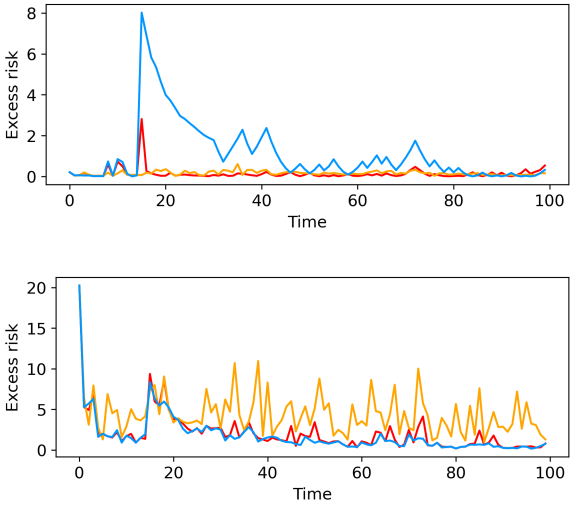


Figure 3. Excess risks in Example 5.2: $\sigma^2 = 1$ (top) and $\sigma^2 = 10$ (bottom). Red: $\mathcal{V}_{\mathrm{ARW}}$. Orange: $\mathcal{V}_1$. Blue: $\mathcal{V}_{256}$.

Table 3. Overall average excess risks ($\times 10^{-3}$) of selection methods on the arXiv data.

$\mathcal{V}_{\mathrm{ARW}}$	$\mathcal{V}_1$	$\mathcal{V}_4$	$\mathcal{V}_{16}$	$\mathcal{V}_{64}$	$\mathcal{V}_{256}$
2.4	6.7	4.5	2.4	1.7	1.9

5.2. Real Data: Topic Frequency Estimation

The first real dataset we use is the arXiv dataset[1]. It consists of basic information of papers submitted to `arXiv.org`, such as title, abstract and categories. We study the topics of the papers in the category `cs.LG` from Match 1st, 2020 to December 31st, 2023. There are $118,883$ papers in total, and we would like to estimate the proportion of deep learning papers in this category. Each week is a time period, and there are 200 weeks in total. We regard a paper as a deep learning paper if its abstract contains at least one of the words "deep", "neural", "dnn" and "dnns". The data in each period is randomly split into training, validation and test sets. The training set $\mathcal{B}_j^{\mathrm{tr}}$ has 15 samples, the validation set $\mathcal{B}_j^{\mathrm{va}}$ has 5 samples and the rest of the samples $\mathcal{B}_j^{\mathrm{te}}$ are used for testing. Typically $|\mathcal{B}_j^{\mathrm{te}}| \in [300, 1200]$. In Figure 6(a), we plot the frequencies over 200 weeks estimated from $\{\mathcal{B}_j^{\mathrm{te}}\}_{j=1}^{200}$, which exhibits a slowly drifting pattern.

In each period t, we consider 5 models $\{\mathcal{T}(t,w)\}_{w\in I}$, where $\mathcal{T}(t,w)$ computes the average frequency from the data $\{\mathcal{B}_j^{\mathrm{tr}}\}_{j=t-w+1}^{t}$. The selection and evaluation procedures are similar to those in the synthetic data experiment. We compare $\mathcal{V}_{\mathrm{ARW}}$ with the fixed-window benchmarks $\{\mathcal{V}_k\}_{k\in I}$. The average excess risks over 200 weeks and 20 independent runs are listed in Table 3. In Figure 4, we also plot the excess risks in every period. We observe that the performance of $\mathcal{V}_{\mathrm{ARW}}$ is comparable to that of the large-window benchmark $\mathcal{V}_{256}$, while the small-window model selection algorithm $\mathcal{V}_1$ performs poorly.

5.3. Real Data: House Price Prediction

Finally, we test our method using a real-world dataset maintained by the Dubai Land Department[2]. We study sales of apartments during the past 16 years (from January 1st, 2008 to December 31st, 2023). There are 211,432 samples (sales) in total, and we want to predict the final price given characteristics of an apartment (e.g., number of rooms, size, location). Each month is treated as a time period, and there are 192 of them in total. Our goal is to build a prediction model for each period using historical data. In Figure 6(b), we plot the monthly average prices. Compared with the arXiv dataset, the distribution shift in this case is more abrupt.

The data in each period is randomly split into training, validation and test sets with proportions 60%, 20% and 20%, respectively. We follow the standard practice to apply a logarithmic transform to the price and target that in our prediction. See Appendix C.3 for details of preprocessing.

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv

[2] https://www.dubaipulse.gov.ae/data/dld-transactions/dld_transactions-open

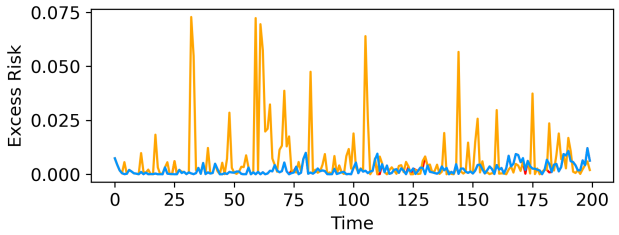


Figure 4. Error curves of different model selection methods on the arXiv data. Red: $\mathcal{V}_{\mathrm{ARW}}$. Orange: $\mathcal{V}_1$. Blue: $\mathcal{V}_{256}$.

Table 4. Overall MSE of different methods on the housing data.

$\mathcal{V}_{\mathrm{ARW}}$	$\mathcal{V}_1$	$\mathcal{V}_4$	$\mathcal{V}_{16}$	$\mathcal{V}_{64}$	$\mathcal{V}_{256}$
0.071	0.071	0.069	0.071	0.091	0.095

For the t-th period, we use each of the 5 training windows $w \in I$ to select training data $\{\mathcal{B}_j^{\mathrm{tr}}\}_{j=t-w+1}^{t}$ for regression with 2 algorithms: random forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016). This results in 10 candidate models. One of them is selected using validation data and finally evaluated on the test data $\mathcal{B}_t^{\mathrm{te}}$ by the mean squared error (MSE). For each model selection method, we compute the average MSE over all of the 192 time periods and 20 independent runs. We compare our proposed approach with fixed-window benchmarks $\{\mathcal{V}_k\}_{k\in I}$. The mean values are reported in Table 4. In addition, we also plot the test MSEs of our method, $\mathcal{V}_1$ and $\mathcal{V}_{256}$ in all individual time periods, see Figure 5. Due to the strong non-stationarity, model selection based on large windows does not work well. Our method still nicely adapts to the changes.

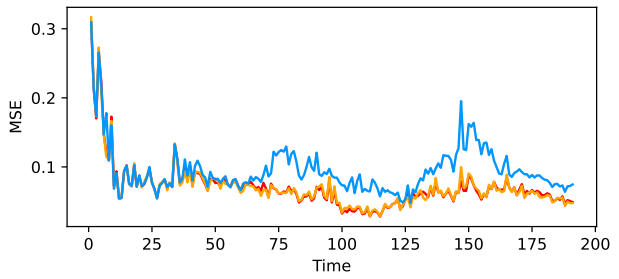


Figure 5. Error curves of different model selection methods on the housing data. Red: $\mathcal{V}_{\mathrm{ARW}}$. Orange: $\mathcal{V}_1$. Blue: $\mathcal{V}_{256}$.

5.4. Summary of Experiments

The synthetic and real data experiments above show that for different patterns of non-stationarity, the best window can be different. It is large for stationary environments, medium for the arXiv data, and small for the Dubai housing data. In practice, the non-stationarity pattern is generally unknown, so it is not clear *a priori* what the best window should be, or even what candidate windows to choose from. Our experiments show that our algorithm *adaptively* selects a window that is comparable to the best one *in hindsight*.

6. Discussions

We developed adaptive rolling window approaches to model assessment and selection in a changing environment. Theoretical analyses and numerical experiments demonstrate their adaptivity to unknown temporal distribution shift. Several future directions are worth exploring. First, our rolling window methodology does not assume any structures on the non-stationarity. In practice, temporal distribution shift often exhibits regularities such as seasonalities and trends. It is important to develop methods that automatically detect patterns of distribution shift and make use of data from the more distant past. Second, our model selection algorithm only applies to finitely many models, and it would be interesting to extend it to infinite model classes. Finally, another direction is to use our model selection algorithm to perform hyperparameter tuning for online learning algorithms, so as to further boost their performance.

Acknowledgement

Elise Han, Chengpiao Huang, and Kaizheng Wang's research is supported by an NSF grant DMS-2210907 and a startup grant at Columbia University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Luo, H., Neyshabur, B., and Schapire, R. E. Corralling a band of bandit algorithms. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 12–38. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/agarwal17b.html.

Audibert, J.-Y., Munos, R., and Szepesvári, C. Tuning bandit algorithms in stochastic environments. In Hutter, M., Servedio, R. A., and Takimoto, E. (eds.), *Algorithmic Learning Theory*, pp. 150–165, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75225-7.

Bai, Y., Zhang, Y.-J., Zhao, P., Sugiyama, M., and Zhou, Z.-H. Adapting to online label shift with provable guarantees.

Advances in Neural Information Processing Systems, 35: 29960–29974, 2022.

Balsubramani, A. and Ramdas, A. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 42 – 51, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

Bartlett, P. L. Learning with a slowly changing distribution. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 243–252, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130412. URL https://doi.org/10.1145/130385.130412.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282.

Barve, R. D. and Long, P. M. On the complexity of learning from drifting distributions. *Information and Computation*, 138(2):170–193, 1997. ISSN 0890-5401. doi: https://doi.org/10.1006/inco.1997.2656. URL https://www.sciencedirect.com/science/article/pii/S0890540197926560.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/24540443.

Bifet, A. and Gavalda, R. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining*, pp. 443–448. SIAM, 2007.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005. doi: 10.1051/ps:2005018.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

Breiman, L. Random forests. *Machine learning*, 45:5–32, 2001.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1405–1411, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/daniely15.html.

Gama, J. a., Žliobaitundefined, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), mar 2014. ISSN 0360-0300. doi: 10.1145/2523813. URL https://doi.org/10.1145/2523813.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Goldenshluger, A. and Lepski, O. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008. ISSN 13507265. URL http://www.jstor.org/stable/20680135.

Hanneke, S., Kanade, V., and Yang, L. Learning with a drifting target concept. In Chaudhuri, K., GENTILE, C., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 149–164, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24486-0.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hazan, E. and Seshadhri, C. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pp. 393–400, 2009.

Helmbold, D. P. and Long, P. M. Tracking drifting concepts by minimizing disagreements. *Machine learning*, 14: 27–45, 1994.

Huang, C. and Wang, K. A stability principle for learning under non-stationarity. *arXiv preprint arXiv:2310.18304*, 2023.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online Optimization : Competing with Dynamic Comparators. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 398–406, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/jadbabaie15.html.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' UCB : An optimal exploration algorithm for multi-armed bandits. In Balcan, M. F., Feldman, V., and SzepesvÃ¡ri, C. (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 423–439, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/jamieson14.html.

Lee, J., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3340–3348. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lee21d.html.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Maurer, A. and Pontil, M. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

Mazzetto, A. and Upfal, E. An adaptive algorithm for learning with unknown distribution drift. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=exiXmAfuDK.

Mazzetto, A., Esfandiarpoor, R., Upfal, E., and Bach, S. H. An adaptive method for weak supervision with drifting data. *arXiv preprint arXiv:2306.01658*, 2023.

Mohri, M. and Muñoz Medina, A. New analysis and algorithm for learning with drifting distributions. In Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T. (eds.), *Algorithmic Learning Theory*, pp. 124–138, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34106-9.

Niu, Y. S., Hao, N., and Zhang, H. Multiple change-point detection: A selective overview. *Statistical Science*, 31 (4):611–623, 2016. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/26408091.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. MIT Press, 2022.

Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2019.107299. URL https://www.sciencedirect.com/science/article/pii/S0165168419303494.

Wei, C.-Y. and Luo, H. Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. In Belkin, M. and Kpotufe, S. (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4300–4354. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/wei21b.html.

A. Proofs for Section 3

A.1. Proof of Lemma 3.3

Inequality (2.10) in (Boucheron et al., 2013) implies that for any $t \geq 0$,

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

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

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

Hence,

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

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

A.2. Proof of Lemma 3.5

The result is trivial when $B_{t,k} = 1$ (i.e. $k = B_t = 1$), as $\widehat{\psi}(t,k,\delta) = \psi(t,k,\delta)$. From now on, we assume that $B_{t,k} \geq 2$. We first present a useful lemma.

Lemma A.1. *Let $\{x_i\}_{i=1}^n$ be independent, $[0,1]$-valued random variables. Define the sample mean $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$ and the sample variance $\widehat{v}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \widehat{\mu})^2$. Let $\mu = \frac{1}{n}\sum_{i=1}^n \mathbb{E}x_i$ and $v^2 = \mathbb{E}\widehat{v}^2$. We have*

$$v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{E}x_i - \mu)^2 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(x_i)$$

and for any $\delta \in (0,1)$,

$$\mathbb{P}\left(v \leq \widehat{v} + \sqrt{\frac{2\log(1/\delta)}{n-1}}\right) \geq 1-\delta \qquad \textit{and} \qquad \mathbb{P}\left(v \geq \widehat{v} - \sqrt{\frac{2\log(1/\delta)}{n-1}}\right) \geq 1-\delta.$$

Proof of Lemma A.1. Define $\boldsymbol{x} = (x_1, \cdots, x_n)^\top$, $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{x}$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\mathrm{var}(x_1), \cdots, \mathrm{var}(x_n))$. Let $\mathbf{1}_n$ be the n-dimensional all-one vector. We have $\widehat{\mu} = \mathbf{1}_n^\top \boldsymbol{x}/n$ and

$$\sum_{i=1}^{n}(x_i - \widehat{\mu})^2 = \|(\boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^\top/n)\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top(\boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^\top/n)\boldsymbol{x}.$$

Then,

$$\begin{aligned}
(n-1)\mathbb{E}\widehat{v}^2 &= \langle \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}, \boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^\top/n\rangle = \boldsymbol{\mu}^\top(\boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^\top/n)\boldsymbol{\mu} + \langle \boldsymbol{\Sigma}, \boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^\top/n\rangle \\
&= \sum_{i=1}^{n}(\mathbb{E}x_i - \mu)^2 + \left(1 - \frac{1}{n}\right)\sum_{i=1}^{n}\mathrm{var}(x_i).
\end{aligned}$$

This verifies the expression of v^2. The concentration bounds come from Theorem 10 in (Maurer & Pontil, 2009). □

We now come back to Lemma 3.5. It suffices to consider the special case $b - a = 1$. From Lemma A.1 we immediately get $\mathbb{E}\widehat{v}_{t,k}^2 = v_{t,k}^2$, where

$$v_{t,k}^2 = \sigma_{t,k}^2 + \frac{1}{B_{t,k} - 1} \sum_{j=t-k+1}^{t} B_j(\mu_j - \mu_{t,k})^2. \tag{11}$$

In addition, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(v_{t,k} \le \widehat{v}_{t,k} + \sqrt{\frac{2\log(1/\delta)}{B_{t,k} - 1}}\right) \ge 1 - \delta \qquad \text{and} \qquad \mathbb{P}\left(v_{t,k} \ge \widehat{v}_{t,k} - \sqrt{\frac{2\log(1/\delta)}{B_{t,k} - 1}}\right) \ge 1 - \delta.$$

With probability at least $1 - \delta$, we have $\sigma_{t,k} \le v_{t,k} \le \widehat{v}_{t,k} + \sqrt{\frac{2\log(1/\delta)}{B_{t,k}-1}}$ and thus

$$\begin{aligned}
\psi(t, k, \delta) &\le \left(\widehat{v}_{t,k} + \sqrt{\frac{2\log(1/\delta)}{B_{t,k} - 1}}\right)\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{2\log(2/\delta)}{3(B_{t,k} - 1)} \\
&\le \widehat{v}_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{8\log(2/\delta)}{3(B_{t,k} - 1)} = \widehat{\psi}(t, k, \delta).
\end{aligned}$$

To prove the second bound, note that

$$\begin{aligned}
v_{t,k} &= \left(\sigma_{t,k}^2 + \frac{1}{B_{t,k} - 1} \sum_{j=t-k+1}^{t} B_j(\mu_j - \mu_{t,k})^2\right)^{1/2} \\
&\le \sigma_{t,k} + \left(\frac{1}{B_{t,k} - 1} \sum_{j=t-k+1}^{t} B_j(\mu_j - \mu_{t,k})^2\right)^{1/2} \\
&= \sigma_{t,k} + \sqrt{\frac{B_{t,k}}{B_{t,k} - 1}} \cdot \left(\frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j(\mu_j - \mu_t)^2\right)^{1/2} \\
&\le \sigma_{t,k} + \sqrt{\frac{B_{t,k}}{B_{t,k} - 1}} \cdot \max_{t-k+1 \le j \le t} |\mu_j - \mu_t|.
\end{aligned}$$

By Lemma A.1, with probability at least $1 - \delta$, we have $v_{t,k} \ge \widehat{v}_{t,k} - \sqrt{\frac{2\log(1/\delta)}{B_{t,k}-1}}$ and thus

$$\widehat{v}_{t,k} \le v_{t,k} + \sqrt{\frac{2\log(1/\delta)}{B_{t,k} - 1}} \le \sigma_{t,k} + \sqrt{\frac{B_{t,k}}{B_{t,k} - 1}} \cdot \max_{t-k+1 \le j \le t} |\mu_j - \mu_t| + \sqrt{\frac{2\log(1/\delta)}{B_{t,k} - 1}}.$$

As a result,

$$\begin{aligned}
\psi(t,k,\delta) &\geq \sigma_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{\log(2/\delta)}{3(B_{t,k}-1)} \\
&\geq \left(\widehat{v}_{t,k} - \sqrt{\frac{B_{t,k}}{B_{t,k}-1}} \cdot \max_{t-k+1\leq j\leq t} |\mu_j - \mu_t| - \sqrt{\frac{2\log(1/\delta)}{B_{t,k}-1}}\right)\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{\log(2/\delta)}{3(B_{t,k}-1)} \\
&\geq \widehat{v}_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} - \sqrt{\frac{2\log(2/\delta)}{B_{t,k}-1}} \cdot \max_{t-k+1\leq j\leq t} |\mu_j - \mu_t| - \frac{5\log(2/\delta)}{3(B_{t,k}-1)} \\
&= \left(\widehat{v}_{t,k}\sqrt{\frac{2\log(2/\delta)}{B_{t,k}}} + \frac{8\log(2/\delta)}{3(B_{t,k}-1)}\right) - \sqrt{\frac{2\log(2/\delta)}{B_{t,k}-1}} \cdot \max_{t-k+1\leq j\leq t} |\mu_j - \mu_t| - \frac{13\log(2/\delta)}{3(B_{t,k}-1)} \\
&= \widehat{\psi}(t,k,\delta) - \sqrt{\frac{2\log(2/\delta)}{B_{t,k}-1}} \cdot \max_{t-k+1\leq j\leq t} |\mu_j - \mu_t| - \frac{13\log(2/\delta)}{3(B_{t,k}-1)} \\
&\geq \widehat{\psi}(t,k,\delta) - \xi(t,k,\delta).
\end{aligned}$$

The first and last inequalities follow from the fact that $B_{t,k} \geq 2$.

A.3. Proof of Lemma 3.7

When the event (4) happens, for every $i \in [k]$,

$$|\widehat{\mu}_{t,k} - \widehat{\mu}_{t,i}| \leq |\widehat{\mu}_{t,k} - \mu_t| + |\widehat{\mu}_{t,i} - \mu_t| \leq \left[\phi(t,k) + \widehat{\psi}(t,k,\delta)\right] + \left[\phi(t,i) + \widehat{\psi}(t,i,\delta)\right],$$

so

$$|\widehat{\mu}_{t,k} - \widehat{\mu}_{t,i}| - \left[\widehat{\psi}(t,k,\delta) + \widehat{\psi}(t,i,\delta)\right] \leq \phi(t,k) + \phi(t,i) \leq 2\phi(t,k),$$

where we used $\phi(t,i) \leq \phi(t,k)$. Taking maximum over all $i \in [k]$ gives $\widehat{\phi}(t,k,\delta) \leq 2\phi(t,k)$.

A.4. Proof of Theorem 3.8

We will prove that with probability at least $1-\delta$,

$$|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \leq 3\min_{k\in[t]}\left\{3\sqrt{\log(6t/\delta)} \cdot \max_{t-k+1\leq j\leq t} |\mu_j - \mu_t| + \sigma_{t,k}\sqrt{\frac{2\log(6t/\delta)}{B_{t,k}}} + \frac{10(b-a)\log(6t/\delta)}{B_{t,k}}\right\}. \tag{12}$$

By Lemma 3.11,

$$\mathbb{P}\left(|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \leq 3\min_{k\in[t]}\left\{\phi(t,k) + \widehat{\psi}(t,k,\delta')\right\}\right) \geq 1 - 2t\delta'. \tag{13}$$

By the second bound in Lemma 3.5,

$$\mathbb{P}\left(\widehat{\psi}(t,k,\delta') \leq \psi(t,k,\delta') + \xi(t,k,\delta'),\ \forall k \in [t]\right) \geq 1 - t\delta'. \tag{14}$$

By (13), (14) and the union bound, we obtain that

$$\mathbb{P}\left(|\widehat{\mu}_{t,\widehat{k}} - \mu_t| \leq 3\min_{k\in[t]}\left\{\phi(t,k) + \psi(t,k,\delta') + \xi(t,k,\delta')\right\}\right) \geq 1 - 3t\delta' = 1 - \delta.$$

When $B_{t,k} \geq 2$,

$$
\begin{aligned}
&\phi(t,k) + \psi(t,k,\delta') + \xi(t,k,\delta') \\
&\leq \max_{t-k+1 \leq j \leq t} |\mu_j - \mu_t| + \sigma_{t,k}\sqrt{\frac{2\log(2/\delta')}{B_{t,k}}} + \frac{2(b-a)\log(2/\delta')}{3(B_{t,k}-1)} \\
&\quad + \sqrt{\frac{4\log(2/\delta')}{B_{t,k}}} \cdot \max_{t-k+1 \leq j \leq t} |\mu_j - \mu_t| + \frac{13(b-a)\log(2/\delta')}{3(B_{t,k}-1)} \\
&\leq 3\sqrt{\log(2/\delta')} \cdot \max_{t-k+1 \leq j \leq t} |\mu_j - \mu_t| + \sigma_{t,k}\sqrt{\frac{2\log(2/\delta')}{B_{t,k}}} + \frac{5(b-a)\log(2/\delta')}{B_{t,k}-1} \\
&\leq 3\sqrt{\log(2/\delta')} \cdot \max_{t-k+1 \leq j \leq t} |\mu_j - \mu_t| + \sigma_{t,k}\sqrt{\frac{2\log(2/\delta')}{B_{t,k}}} + \frac{10(b-a)\log(2/\delta')}{B_{t,k}} \\
&= 3\sqrt{\log(6t/\delta)} \cdot \max_{t-k+1 \leq j \leq t} |\mu_j - \mu_t| + \sigma_{t,k}\sqrt{\frac{2\log(6t/\delta)}{B_{t,k}}} + \frac{10(b-a)\log(6t/\delta)}{B_{t,k}}.
\end{aligned}
$$

When $B_{t,k} = 1$, we have $\phi(t,k) + \psi(t,k,\delta') = b - a$, which is dominated by the bound above.

B. Proofs for Section 4

B.1. Proof of Lemma 4.1

Fix $k \in [t]$. For every $f \in \{f_1, f_2\}$,

$$
\begin{aligned}
L_t(f_{\widehat{r}}) - L_t(f) &= \left[L_t(f_{\widehat{r}_k}) - \widehat{L}_{t,k}(f_{\widehat{r}_k})\right] + \left[\widehat{L}_{t,k}(f_{\widehat{r}_k}) - L_t(f)\right] \\
&\leq \left[L_t(f_{\widehat{r}_k}) - \widehat{L}_{t,k}(f_{\widehat{r}_k})\right] + \left[\widehat{L}_{t,k}(f) - L_t(f)\right] \\
&= \left[\widehat{L}_{t,k}(f) - \widehat{L}_{t,k}(f_{\widehat{r}_k})\right] - \left[L_t(f) - L_t(f_{\widehat{r}_k})\right] \\
&\leq \left|\left[\widehat{L}_{t,k}(f_1) - \widehat{L}_{t,k}(f_2)\right] - \left[L_t(f_1) - L_t(f_2)\right]\right|.
\end{aligned}
$$

Taking minimum over $f \in \{f_1, f_2\}$ yields

$$
L_t(f_{\widehat{r}}) - \min_{r \in [2]} L_t(f_r) \leq \left|\left[\widehat{L}_{t,k}(f_1) - \widehat{L}_{t,k}(f_2)\right] - \left[L_t(f_1) - L_t(f_2)\right]\right|.
$$

B.2. Proof of Theorem 4.6

More precisely, we will prove that with probability at least $1 - \delta$, Algorithm 2 outputs $\widehat{f}$ satisfying

$$
\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r \in [2]} \|f_r - f_t^*\|_{\mathcal{P}} \leq C'\sqrt{\log(6t/\delta)} \min_{k \in [t]} \left\{ \max_{t-k+1 \leq j \leq t} \|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M}{\sqrt{B_{t,k}}} \right\},
$$

where $C' > 0$ is a universal constant.

Following the notation in Problem 3.2 with $u_{j,i} = \ell(f_1, z_{j,i}) - \ell(f_2, z_{j,i})$, we define

$$
\sigma_j^2 = \mathrm{var}(u_{j,1}) = \mathrm{var}\left(\ell(f_1, z_{j,1}) - \ell(f_2, z_{j,1})\right) \quad \text{and} \quad \sigma_{t,k}^2 = \frac{1}{B_{t,k}} \sum_{j=t-k+1}^{t} B_j \sigma_j^2.
$$

By Lemma 4.1 and Theorem 3.8, with probability at least $1-\delta$,

$$L_t(\widehat{f}) - \min_{r\in[2]} L_t(f_r) \le |\widehat{\Delta}_{t,\widehat{k}} - \Delta_t|$$
$$\le 3\min_{k\in[t]}\left\{3\sqrt{\log(6t/\delta)}\cdot \max_{t-k+1\le j\le t}|\Delta_j-\Delta_t| + \sigma_{t,k}\sqrt{\frac{2\log(6t/\delta)}{B_{t,k}}} + \frac{40M_0^2\log(6t/\delta)}{B_{t,k}}\right\}. \quad (15)$$

From now on suppose that this event happens.

We first derive a bound for $|\Delta_j - \Delta_t|$. Since

$$\Delta_j = \|f_1 - f_j^*\|_{\mathcal{P}}^2 - \|f_2 - f_j^*\|_{\mathcal{P}}^2 = \langle f_1 - f_2, f_1 + f_2 - 2f_j^*\rangle_{\mathcal{P}},$$

then for every $j \in \mathbb{Z}_+$,

$$|\Delta_j - \Delta_t| = 2|\langle f_1 - f_2, f_j^* - f_t^*\rangle_{\mathcal{P}}| \le 2\|f_1 - f_2\|_{\mathcal{P}}\|f_j^* - f_t^*\|_{\mathcal{P}}.$$

Hence

$$\max_{t-k+1\le j\le t}|\Delta_j - \Delta_t| \le 2\|f_1 - f_2\|_{\mathcal{P}} \max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}}. \quad (16)$$

Next, we bound $\sigma_{t,k}$. We have

$$\sigma_j^2 = \mathrm{var}\left([f_1(x_{j,1}) - y_{j,1}]^2 - [f_2(x_{j,1}) - y_{j,1}]^2\right) \le \mathbb{E}\left[\left([f_1(x_{j,1}) - y_{j,1}]^2 - [f_2(x_{j,1}) - y_{j,1}]^2\right)^2\right]$$
$$= \mathbb{E}\left[\left(f_1(x_{j,1}) - f_2(x_{j,1})\right)^2\left(f_1(x_{j,1}) + f_2(x_{j,1}) - 2y_{j,1}\right)^2\right]$$
$$\lesssim M_0^2\|f_1 - f_2\|_{\mathcal{P}}^2,$$

which implies

$$\sigma_{t,k} \lesssim M_0\|f_1 - f_2\|_{\mathcal{P}} \le M_0\big(\|f_1 - f_t^*\|_{\mathcal{P}} + \|f_2 - f_t^*\|_{\mathcal{P}}\big). \quad (17)$$

Substituting (16) and (17) into (15), there exist a universal constant $C > 0$ such that for every $k \in [t]$,

$$L_t(\widehat{f}) - \min_{r\in[2]} L_t(f_r) \le C\Bigg[\sqrt{\log(6t/\delta)}\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}}\|f_1 - f_2\|_{\mathcal{P}}$$
$$+ M_0\sqrt{\frac{\log(6t/\delta)}{B_{t,k}}}\left(\|f_1 - f_t^*\|_{\mathcal{P}} + \|f_2 - f_t^*\|_{\mathcal{P}}\right) + \frac{M_0^2\log(6t/\delta)}{B_{t,k}}\Bigg]$$
$$= C\Bigg[2\Phi(t,k)\left(\|f_1 - f_t^*\|_{\mathcal{P}} + \|f_2 - f_t^*\|_{\mathcal{P}}\right) + \Psi(t,k)\Bigg], \quad (18)$$

where

$$\Phi(t,k) = \frac{1}{2}\sqrt{\log(6t/\delta)}\left(\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right) \quad \text{and} \quad \Psi(t,k) = \frac{M_0^2\log(6t/\delta)}{B_{t,k}},$$

and we used the triangle inequality $\|f_1 - f_2\|_{\mathcal{P}} \le \|f_1 - f_t^*\|_{\mathcal{P}} + \|f_2 - f_t^*\|_{\mathcal{P}}$.

Without loss of generality, assume $L_t(f_1) \ge L_t(f_2)$. When $\widehat{r}_{\widehat{k}} = 1$, we have $\widehat{f} = f_1$. Then (18) yields

$$\|f_1 - f_t^*\|_{\mathcal{P}}^2 - \|f_2 - f_t^*\|_{\mathcal{P}}^2 = L_t(\widehat{f}) - \min_{r\in[2]} L_t(f_r) \le C\Bigg[2\Phi(t,k)\left(\|f_1 - f_t^*\|_{\mathcal{P}} + \|f_2 - f_t^*\|_{\mathcal{P}}\right) + \Psi(t,k)\Bigg].$$

Completing the squares gives

$$\Bigg[\|f_1 - f_t^*\|_{\mathcal{P}} - C\Phi(t,k)\Bigg]^2 \le \Bigg[\|f_2 - f_t^*\|_{\mathcal{P}} + C\Phi(t,k)\Bigg]^2 + C\Psi(t,k).$$

Hence

$$\begin{aligned}\|f_1 - f_t^*\|_{\mathcal{P}} &\le \|f_2 - f_t^*\|_{\mathcal{P}} + 2C\Phi(t,k) + \sqrt{C\Psi(t,k)} \\ &\le \|f_2 - f_t^*\|_{\mathcal{P}} + C'\sqrt{\log(6t/\delta)}\left(\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right) \end{aligned} \tag{19}$$

for some universal constant $C' > 0$. Since this holds for all $k \in [t]$, we get

$$\begin{aligned}&\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[2]}\|f_r - f_t^*\|_{\mathcal{P}} \\ &= \|f_1 - f_t^*\|_{\mathcal{P}} - \|f_2 - f_t^*\|_{\mathcal{P}} \\ &\le C'\sqrt{\log(6t/\delta)}\min_{k\in[t]}\left\{\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right\}. \end{aligned} \tag{20}$$

We have proved this bound for the case $\widehat{r}_{\widehat{k}} = 1$. When $\widehat{r}_{\widehat{k}} = 2$, we have $\widehat{f} = f_2$ and hence $\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[2]}\|f_r - f_t^*\|_{\mathcal{P}} = 0$, so the bound (20) continues to hold.

B.3. Proof of Theorem 4.7

More precisely, we will prove that with probability at least $1-\delta$, Algorithm 3 outputs $\widehat{f}$ satisfying

$$\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[m]}\|f_r - f_t^*\|_{\mathcal{P}} \le C'\sqrt{(\log m)\log(mt/\delta)}\min_{k\in[t]}\left\{\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right\}$$

for some universal constant $C' > 0$.

Denote Algorithm 2 with data $\{\mathcal{B}_j\}_{j=1}^t$ by $\mathcal{A}$, which takes as input two models $f', f'' \in \mathcal{F}$ and outputs the selected model $\mathcal{A}(\{f', f''\}) \in \{f', f''\}$. For notational convenience, we set $\mathcal{A}(\{f\}) = f$ for every $f \in \mathcal{F}$. By Theorem 4.6 and the union bound, with probability at least $1-\delta$, the following holds for all $f', f'' \in \{f_r\}_{r=1}^m$:

$$\|\mathcal{A}(\{f', f''\}) - f_t^*\|_{\mathcal{P}} - \min_{f\in\{f',f''\}}\|f - f_t^*\|_{\mathcal{P}} \le C\sqrt{\log(mt/\delta)}\min_{k\in[t]}\left\{\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right\}, \tag{21}$$

where $C > 0$ is a universal constant. From now on, suppose that this event happens.

For each $s \in [S+1]$, let

$$g_s \in \operatorname*{argmin}_{g\in\mathcal{F}_s} L_t(g) = \operatorname*{argmin}_{g\in\mathcal{F}_s}\|g - f_t^*\|_{\mathcal{P}}.$$

Since $\mathcal{F} = \mathcal{F}_1 \supseteq \cdots \supseteq \mathcal{F}_S \supseteq \mathcal{F}_{S+1} = \{\widehat{f}\}$, then

$$\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[m]}\|f_r - f_t^*\|_{\mathcal{P}} = \sum_{s=1}^{S}\left(\|g_{s+1} - f_t^*\|_{\mathcal{P}} - \|g_s - f_t^*\|_{\mathcal{P}}\right).$$

For each $s \in [S]$, there exists $i_s \in [m_s+1]$ such that $g_s \in \mathcal{G}_{s,i_s}$. Let $\widehat{g}_s = \mathcal{A}(\mathcal{G}_{s,i_s})$ be the model selected from $\mathcal{G}_{s,i_s}$ in Algorithm 3. By (21) and the definition of g_s,

$$\|\widehat{g}_s - f_t^*\|_{\mathcal{P}} - \|g_s - f_t^*\|_{\mathcal{P}} \le C\sqrt{\log(mt/\delta)}\min_{k\in[t]}\left\{\max_{t-k+1\le j\le t}\|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}}\right\}.$$

Moreover, since $\widehat{g}_s \in \mathcal{F}_{s+1}$, then $\|g_{s+1} - f_t^*\|_{\mathcal{P}} \leq \|\widehat{g}_s - f_t^*\|_{\mathcal{P}}$. Therefore,

$$\begin{aligned}
\|\widehat{f} - f_t^*\|_{\mathcal{P}} - \min_{r\in[m]} \|f_r - f_t^*\|_{\mathcal{P}} &= \sum_{s=1}^{S} \left(\|g_{s+1} - f_t^*\|_{\mathcal{P}} - \|g_s - f_t^*\|_{\mathcal{P}}\right) \\
&\leq \sum_{s=1}^{S} \left(\|\widehat{g}_s - f_t^*\|_{\mathcal{P}} - \|g_s - f_t^*\|_{\mathcal{P}}\right) \\
&\leq \sum_{s=1}^{S} C\sqrt{\log(mt/\delta)} \min_{k\in[t]} \left\{ \max_{t-k+1\leq j\leq t} \|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}} \right\} \\
&\leq C'\sqrt{(\log m)\log(mt/\delta)} \min_{k\in[t]} \left\{ \max_{t-k+1\leq j\leq t} \|f_j^* - f_t^*\|_{\mathcal{P}} + \frac{M_0}{\sqrt{B_{t,k}}} \right\}
\end{aligned}$$

for some universal constant $C' > 0$.

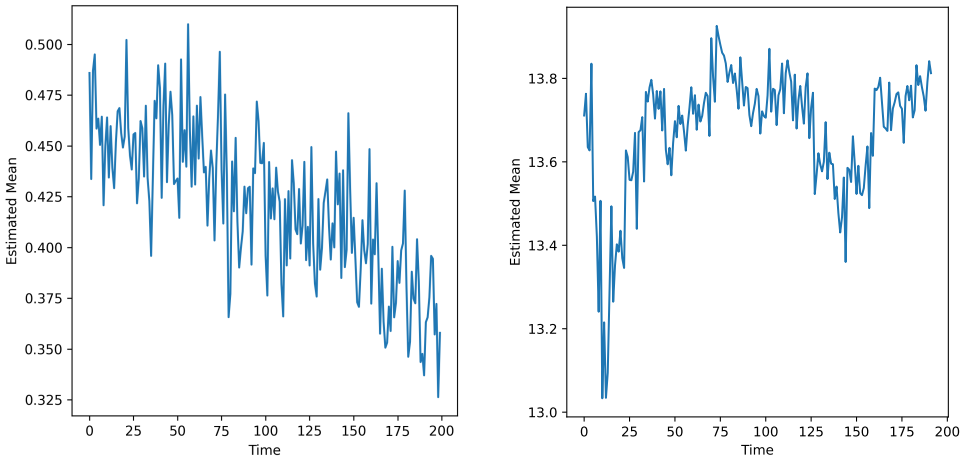
C. Numerical Experiments: Additional Details

C.1. Example 2 of the Synthetic Data

We give an outline of how the true mean sequence $\{\mu_t\}$ in Example 5.2 is generated. The sequence is constructed by combining 4 parts, each representing a distribution shift pattern. In the first part, the sequence experiences large shifts. Then, it switches to a sinusoidal pattern. Following that, the environment stays stationary for some time. Finally, the sequence drifts randomly at every period, where the drift sizes are independently sampled from $\{1, -1\}$ with equal probability and scaled with a constant.

The function for generating the sequence takes in 3 parameters `N`, `n` and `seed`, where `N` is the total number of periods, `n` is the parameter determining the splitting points of the 4 parts, and `seed` is the random seed used for code reproducibility. In our experiment, we set `N = 100`, `n = 2` and `seed = 2024`. The exact function can be found in our code at https://github.com/eliselyhan/ARW.

C.2. Non-Stationarity in arXiv and Housing Data



(a) arXiv data (b) Dubai housing data

Figure 6. Two patterns of temporal distribution shift: slow drift and abrupt changes.

C.3. Experiments on the Housing Data

We focus on transactions of studios and apartments with 1 to 4 bedrooms, between January 1st, 2008 and December 31st, 2023. We import variables `instance_date` (transaction date), `area_name_en` (English name of the area where the apartment is located in), `rooms_en` (number of bedrooms), `has_parking` (whether or not the apartment has a parking spot), `procedure_area` (area in the apartment), `actual_worth` (final price) from the data.

We use `instance_date` (transaction date) to construct monthly datasets. The target for prediction is the logarithmic of `actual_worth`. The predictors are `area_name_en`, `rooms_en`, `has_parking` and `procedure_area`. `area_name_en` has 58 possible values and encoded as an integer variable. We remove a sample when its `actual_worth` or `procedure_area` is among the largest or smallest 2.5% of the population, whichever is true. After the trimming, 91.6% of the data remains.

For random forest regression, we use the function `RandomForestRegressor` in the Python library `scikit-learn`. For XGBoost regression, we use the function `XGBRegressor` in the Python library `xgboost`. In both cases, we set `random_state = 0` and do not change any other default parameters. Implementation details can be found in our code at https://github.com/eliselyhan/ARW.