https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 73, No. 6, November–December 2025, pp. 3044–3064
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

A Stability Principle for Learning Under Nonstationarity

Chengpiao Huang,[a] Kaizheng Wang[a,b,*]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027; [b] Data Science Institute, Columbia University, New York, New York 10027

*Corresponding author

Contact: chengpiao.huang@columbia.edu, https://orcid.org/0009-0008-2193-7632 (CH); kaizheng.wang@columbia.edu, https://orcid.org/0000-0002-0926-2600 (KW)

Received: January 22, 2024
Revised: October 9, 2024; March 1, 2025; April 25, 2025
Accepted: May 6, 2025
Published Online in Articles in Advance: June 10, 2025

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2024.0766

Abstract. We develop a versatile framework for statistical learning in nonstationary environments. In each time period, our approach applies a stability principle to select a look-back window that maximizes the utilization of historical data while keeping the cumulative bias within an acceptable range relative to the stochastic error. Our theory showcases the adaptivity of this approach to unknown nonstationarity. We prove regret bounds that are minimax optimal up to logarithmic factors when the population losses are strongly convex or Lipschitz only. At the heart of our analysis lie two novel components: a measure of similarity between functions and a segmentation technique for dividing the nonstationary data sequence into quasistationary pieces. We evaluate the practical performance of our approach through real-data experiments on electricity demand prediction and hospital nurse staffing.

Funding: This work was supported by the Data Science Institute, Columbia University [Seed Grant SF-181], the National Science Foundation [Grant DMS-2210907], and Fu Foundation School of Engineering and Applied Science (startup grant).

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results are available at https://doi.org/10.1287/opre.2024.0766.

Keywords: nonstationarity • online learning • distribution shift • adaptivity • look-back window

1. Introduction

It has been widely observed in economics (Clements and Hendry 2001), healthcare (Nestor et al. 2019), environmental science (Milly et al. 2008), and many other fields that the underlying environment is constantly changing over time. The pervasive nonstationarity presents formidable challenges to statistical learning and data-driven decision making, as knowledge from the past may no longer be useful for the future and the learner needs to chase a moving target. In this paper, we develop a principled approach for adapting to unknown changes in the environment.

As a motivating example, consider the problem of hospital nurse staffing, where a hospital needs to decide how many nurses to schedule every week. The number of patient visits can vary significantly across time because of the seasonality of certain diseases (e.g., influenza), the outbreak of a new disease (e.g., COVID-19), or other factors. Consequently, past data on the numbers of patient visits may not be informative or can even be misleading for making future nurse staffing decisions. Indeed, as demonstrated by our numerical studies on New York City emergency department visits (NYC Health 2024) in Section 7, blindly using data from more than six months ago leads to a cost twice as high as using data from just the previous week. In contrast, our proposed approach adaptively chooses past data to make decisions, reducing the cost by 3%–64% compared with different benchmarks. This highlights the importance and benefits of adapting to the nonstationary environment.

More generally, consider a canonical setup of online learning where, in each time period, a learner chooses a decision from a feasible set to minimize an unknown loss function and observes a noisy realization of the loss through a batch of samples. The decision is made based on historical data and incurs an excess loss, which is the difference between the learner's loss and the loss of the optimal decision. The learner's overall performance is measured by the cumulative excess loss, which is an example of the dynamic regret in online learning (Zinkevich 2003).

In the presence of nonstationarity, the historical observations gathered at different time periods are not equally informative for minimizing the present objective. Most learning algorithms are designed for stationary settings, which can lead to suboptimal outcomes if applied directly. A natural idea is to choose a

look-back window k and use the observations from the most recent k periods to compute an empirical minimizer. Selecting a good window involves a bias-variance trade-off: increasing the window size typically reduces the stochastic error but may result in a larger bias. The optimal window is smaller during fluctuating periods and larger in stable eras. Unfortunately, such structural knowledge is often lacking in practice.

We propose a *stability principle* for automatically selecting windows tailored to the unknown local variability. At each time step, our method looks for the largest look-back window in which the cumulative bias is dominated by the stochastic error. This is carried out by iteratively expanding the window and comparing it with smaller ones. Given two windows, we compare the associated solutions through their performance on the data in the smaller window. If the performance gap is too large, then the environment seems to have undergone adverse changes within the larger window, and we choose the smaller window. Otherwise, the larger window is not significantly worse than the smaller one, and we choose the larger window to promote statistical stability. This idea can be extended to the general scenario with multiple candidate windows. A window is deemed *admissible* if it passes pairwise tests against smaller ones. Our approach picks the largest admissible window to maximize the utilization of historical data while effectively managing bias. The window selection procedure can be succinctly summarized as "expansion until proven guilty."

1.1. Main Contributions

Our contributions are threefold.

1. (Flexible method) We develop a versatile framework for statistical learning in dynamic environments based on the stability principle described above. It can be easily combined with learning algorithms for stationary settings, helping them adapt to distribution shifts over time.

2. (Adaptivity guarantees in common settings) We provide sharp regret bounds for our method when the population losses are strongly convex and smooth, or Lipschitz only. We also prove matching minimax lower bounds up to logarithmic factors. Our method is shown to achieve the optimal rates while being agnostic to the nonstationarity. We further evaluate its practical performance through real-data experiments on electricity demand prediction and hospital nurse staffing.

3. (A general theory of learning under nonstationarity) We derive regret bounds based on a unified characterization of nonstationarity. We propose a novel measure of similarity between functions: two functions $f, g : \Omega \to \mathbb{R}$ are said to be (ε, δ)-*close* if, for all $\boldsymbol{\theta} \in \Omega$, it holds that

$$g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}') \le e^{\varepsilon} \left(f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}') + \delta \right),$$

$$f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}') \le e^{\varepsilon} \left(g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}') + \delta \right).$$

The closeness relation behaves nicely under common operations, providing a powerful tool for analyzing sample average approximation with non–independent and identically distributed (i.i.d.) data. We further develop a segmentation technique that partitions the whole data sequence into quasistationary pieces.

1.2. Related Works

We give a review of the most relevant works, which is by no means exhaustive. Existing approaches to learning under nonstationarity can be broadly divided into *model-based* and *model-free* ones. Model-based approaches use latent state variables to encode the underlying distributions and directly model the evolution. Examples include regime-switching and seasonality models (Hamilton 1989, Chen et al. 2025), linear dynamical systems (Kalman 1960, Mania et al. 2022), Gaussian processes (Slivkins and Upfal 2008), and autoregressive processes (Chen et al. 2023). Whereas they have nice interpretations, the prediction powers can be impaired by model misspecification (Dacco and Satchell 1999). Such an issue may mislead models to use data from past environments that are substantially different from the present one.

In contrast, model-free approaches focus on the most recent data to ensure relevance. A popular tool is rolling window, which has seen great success in nonstationary time series (Fan and Yao 2003), probably approximately correct (PAC) learning (Mohri and Muñoz Medina 2012), classification (Hanneke et al. 2015), inventory management (Keskin et al. 2023), distribution learning (Mazzetto and Upfal 2023), and so on. Our approach belongs to this family, with wider applicability and better adaptivity to unknown changes. It draws inspiration from Lepskii's method for adaptive bandwidth selection in nonparametric estimation (Lepskii 1991). Both of them identify admissible solutions through pairwise tests. In Lepskii's method, each test compares the distance between two candidate solutions with a threshold determined by its estimated statistical uncertainties. However, it is not suitable when the empirical loss does not have a unique minimizer. Our approach, on the other hand, compares candidate solutions by their objective values. This is applicable to any loss function defined on an arbitrary domain that may not have a metric. Related ideas were also used by Spokoiny (2009) to estimate volatilities in time series, by Luo et al. (2018) to design algorithms for contextual bandits, and by Mazzetto and Upfal (2023) for window selection in distribution learning.

There has also been a great number of model-free approaches in the area of nonstationary online convex optimization (OCO) (Hazan 2016). Given access to noisy gradient information, one can modify standard first-order OCO algorithms using carefully chosen restarts (Besbes et al. 2015, Chen et al. 2019a) and learning rate schedules (Yang et al. 2016, Cutler et al. 2023, Fahrbach et al. 2023). The updating rules are much simpler than those of rolling window methods. However, they require knowledge about certain *path variation*, which is the summation of changes in loss functions or minimizers between consecutive times. Adaptation to the unknown variation is usually achieved by online ensemble methods (Hazan and Seshadhri 2009, Zhang et al. 2018, Baby and Wang 2022, Bai et al. 2022, Bilodeau et al. 2023, Zhao et al. 2024). Our measure of nonstationarity gives a more refined characterization than the path variations, especially when the changes exhibit temporal heterogeneity. Moreover, our general results imply minimax optimal regret bounds with respect to path variations. The bounds show explicit and optimal dependence on the dimension of the decision space, whereas existing works usually treat it as a constant. On the other hand, some works on nonstationary OCO studied robust utilization of side information, such as noisy forecast of the loss gradient or the data distribution before each time period (Hall and Willett 2013, Jadbabaie et al. 2015, Jiang et al. 2025). They measured the problem complexity using the sum of forecast errors, similar to the path variation. It would be interesting to extend our nonstationarity measure to that scenario.

Full observation of the noisy loss function or its gradient is not always possible. Instead, the learner may only receive a noisy realization of the function value at the decision. This motivated recent works on OCO with bandit feedback (Besbes et al. 2015, Chen et al. 2019a, Wang 2025), which reduced the problem to first-order OCO through gradient estimation. Their settings are more difficult than ours, as it is harder to detect nonstationarity from single-point observations. In contrast, our noisy observation of the whole loss function facilitates evaluation and comparison of solutions associated with different look-back windows so as to select the optimal one. Another line of research investigated dynamic pricing (Keskin and Zeevi 2017, Zhao et al. 2023) and various bandit problems (Luo et al. 2018, Auer et al. 2019, Chen et al. 2019b, Wei and Luo 2021, Cheung et al. 2022, Suk and Kpotufe 2022, Foussoul et al. 2023, Jia et al. 2023, Liu et al. 2023, Min and Russo 2023), where the learner needs to strike a balance between exploration and exploitation in the presence of nonstationarity.

1.3. Outline

The rest of the paper is organized as follows. Section 2 describes the problem setup. Section 3 introduces the stability principle and the methodology. Section 4 presents regret bounds in common settings. Section 5 develops a general theory of learning under nonstationarity. Section 6 provides minimax lower bounds to prove the adaptivity of our method. Section 7 conducts simulations and real-data experiments to test the performance of our algorithm. Finally, Section 8 concludes the paper and discusses future directions.

2. Problem Setup

In this section, we formally describe the problem of statistical learning in nonstationary environments and its main challenge.

Problem 1 (Online Statistical Learning Under Nonstationarity). Let $\mathcal{Z}$ be a sample space, Ω be a decision set, and $\ell : \Omega \times \mathcal{Z} \to \mathbb{R}$ be a known loss function. At each time $n = 1, 2, \ldots$, the environment is represented by an unknown data distribution $\mathcal{P}_n$ over $\mathcal{Z}$. A learner chooses a decision $\boldsymbol{\theta}_n \in \Omega$ based on historical information to minimize the (unknown) *population loss*

$$F_n(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}_n}[\ell(\boldsymbol{\theta}, \boldsymbol{z})], \quad \forall \boldsymbol{\theta} \in \Omega,$$

and collects a batch of $B \in \mathbb{Z}_+$ i.i.d. samples $\mathcal{D}_n = \{\boldsymbol{z}_{n,j}\}_{j=1}^B$ from $\mathcal{P}_n$. Assume that $\{\mathcal{D}_n\}_{n=1}^\infty$ are independent.

The data $\mathcal{D}_i = \{\boldsymbol{z}_{i,j}\}_{j=1}^B$ at time i defines an *empirical loss*

$$f_i(\boldsymbol{\theta}) = \frac{1}{B}\sum_{j=1}^{B} \ell(\boldsymbol{\theta}, \boldsymbol{z}_{i,j}), \quad \forall \boldsymbol{\theta} \in \Omega, \tag{1}$$

which is an unbiased estimator of F_i. At each time n, given noisy observations $\{f_i\}_{i=1}^{n-1}$, we look for $\boldsymbol{\theta}_n$ that will be good for minimizing the upcoming loss function F_n. The *excess risk* in period n is $F_n(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}'_n\in\Omega} F_n(\boldsymbol{\theta}'_n)$. Our performance measure is the cumulative excess risk, also known as the *dynamic regret*:

$$\sum_{n=1}^{N}\left[F_n(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}'_n\in\Omega} F_n(\boldsymbol{\theta}'_n)\right].$$

Here, the horizon N may not be known a priori.

To minimize F_n, it is natural to choose some *look-back window* $k \in [n-1]$ and approximate F_n by the pre-average $f_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} f_i$. Let $\widehat{\boldsymbol{\theta}}_{n,k}$ be an approximate minimizer of $f_{n,k}$. We will select some $\widehat{k} \in [n-1]$ and output $\boldsymbol{\theta}_n = \widehat{\boldsymbol{\theta}}_{n,\widehat{k}}$.

Choosing a good window k involves a bias-variance trade-off. Increasing the window size k improves the concentration of the empirical loss $f_{n,k}$ around its population version $F_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} F_i$ and thus reduces the stochastic error. Meanwhile, the nonstationarity can drive $F_{n,k}$ away from the target loss F_n and induce a large approximation error (bias). Achieving a low regret

requires striking a balance between the deterministic bias and the stochastic error, which is a bias-variance trade-off.

2.1. Notation

Let $\mathbb{Z}_+ = \{1,2,\dots\}$ be the set of positive integers and $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\}$ be the set of nonnegative real numbers. For $n \in \mathbb{Z}_+$, define $[n] = \{1,2,\dots,n\}$. For $a,b \in \mathbb{R}$, define $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. For $x \in \mathbb{R}$, let $x_+ = x \vee 0$. For nonnegative 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 $n \in \mathbb{Z}_+$, $a_n \leq Cb_n$. We write $a_n = \widetilde{\mathcal{O}}(b_n)$ if $a_n = \mathcal{O}(b_n)$ up to logarithmic factors $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Unless otherwise stated, $a_n \lesssim b_n$ also represents $a_n = \mathcal{O}(b_n)$. For $\boldsymbol{x} \in \mathbb{R}^d$ and $r \geq 0$, let $B(\boldsymbol{x},r) = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y}-\boldsymbol{x}\|_2 \leq r\}$ and $B_\infty(\boldsymbol{x},r) = \{\boldsymbol{y} \in \mathbb{R}^d : \|\boldsymbol{y}-\boldsymbol{x}\|_\infty \leq r\}$. Let $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. The diameter of a set $\Omega \subseteq \mathbb{R}^d$ is defined as $\operatorname{diam}(\Omega) = \sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\|\boldsymbol{x}-\boldsymbol{y}\|_2$. The sup-norm of a function $f : \Omega \to \mathbb{R}$ is defined as $\|f\|_\infty = \sup_{\boldsymbol{x}\in\Omega}|f(\boldsymbol{x})|$. For $\alpha \in \{1,2\}$ and a random variable z, define $\|z\|_{\psi_\alpha} = \sup_{p\geq 1}\{p^{-1/\alpha}\mathbb{E}^{1/p}|z|^p\}$, where $\|\cdot\|_{\psi_1}$ is the subexponential norm, and $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm. For a random vector $\boldsymbol{v}$ in $\mathbb{R}^d$, define $\|\boldsymbol{v}\|_{\psi_\alpha} = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}}\|\boldsymbol{u}^\top\boldsymbol{v}\|_{\psi_\alpha}$. The notation $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The notation $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix.

3. A Stability Principle for Adapting to Nonstationarity

In this section, we propose a stability principle for adaptive selection of the look-back window under unknown nonstationarity. We will first introduce a criterion for choosing between two windows based on the idea of hypothesis testing and then extend the approach to the general case.

3.1. Choosing Between Two Windows: To Pool or Not to Pool?

To begin with, we investigate a retrospective variant of Problem 1. Imagine that at time n, we seek to minimize the loss F_n based on noisy realizations $\{f_i\}_{i=1}^{n-1}$ and f_n of both the past losses and the present one. Suppose that $\mathcal{P}_1 = \cdots = \mathcal{P}_{n-1}$ but there is a possible distribution shift causing $\mathcal{P}_n \neq \mathcal{P}_{n-1}$. Consequently, $\{f_i\}_{i=1}^{n-1}$ are i.i.d. but possibly poor approximations of F_n. We want to decide between using the most recent observation f_n and pooling all the historical data $\{f_i\}_{i=1}^{n}$. They lead to two candidate solutions, $\widetilde{\boldsymbol{\theta}}_1 \in \arg\min_{\boldsymbol{\theta}\in\Omega} f_n(\boldsymbol{\theta})$ and $\widetilde{\boldsymbol{\theta}}_0 \in \arg\min_{\boldsymbol{\theta}\in\Omega} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\theta})$, respectively.

Our idea is to detect a harmful distribution shift between $\mathcal{P}_{n-1}$ and $\mathcal{P}_n$, get an indicator $\mathcal{I} \in \{0,1\}$, and then output $\widetilde{\boldsymbol{\theta}}_{\mathcal{I}}$. We make the following observations:

1. If $\mathcal{P}_{n-1} = \mathcal{P}_n$, then $\widetilde{\boldsymbol{\theta}}_0$ tends to be better than $\widetilde{\boldsymbol{\theta}}_1$ because of its statistical stability; that is, $F_n(\widetilde{\boldsymbol{\theta}}_0) - F_n(\widetilde{\boldsymbol{\theta}}_1) \leq 0$;
2. If there is a harmful distribution shift between $\mathcal{P}_{n-1}$ and $\mathcal{P}_n$, then $\widetilde{\boldsymbol{\theta}}_0$ will be much worse than $\widetilde{\boldsymbol{\theta}}_1$; that is, $F_n(\widetilde{\boldsymbol{\theta}}_0) - F_n(\widetilde{\boldsymbol{\theta}}_1)$ is large.

A faithful test should be likely to return $\mathcal{I} = 0$ in the first case and $\mathcal{I} = 1$ in the second case. Both cases concern the performance gap $F_n(\widetilde{\boldsymbol{\theta}}_0) - F_n(\widetilde{\boldsymbol{\theta}}_1)$. We propose to estimate it by $f_n(\widetilde{\boldsymbol{\theta}}_0) - f_n(\widetilde{\boldsymbol{\theta}}_1)$ and compare it with some threshold $\tau > 0$. The resulting test is

$$\mathcal{I} = \begin{cases} 0, & \text{if } f_n(\widetilde{\boldsymbol{\theta}}_0) - f_n(\widetilde{\boldsymbol{\theta}}_1) \leq \tau \\ 1, & \text{if } f_n(\widetilde{\boldsymbol{\theta}}_0) - f_n(\widetilde{\boldsymbol{\theta}}_1) > \tau. \end{cases} \tag{2}$$

To set the threshold τ, we need some estimates on the statistical uncertainty of the test statistic $f_n(\widetilde{\boldsymbol{\theta}}_0) - f_n(\widetilde{\boldsymbol{\theta}}_1)$ in the absence of a distribution shift. As we will demonstrate in Section 4, these are available in many common scenarios.

In words, our principle can be summarized as follows: *we prefer a statistically more stable solution unless it appears significantly worse.*

3.2. Choosing from Multiple Windows

We now develop a general framework for window selection. Recall that for any $n \geq 2$, each look-back window $k \in [n-1]$ is associated with a loss function $f_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} f_i$ and its minimizer $\widehat{\boldsymbol{\theta}}_{n,k}$. Following the idea in (2), we choose positive thresholds $\{\tau(n,i)\}_{i=1}^{n-1}$ and construct a test

$$\mathcal{I}_{i,k} = \begin{cases} 0, & \text{if } f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,k}) - f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,i}) \leq \tau(n,i) \\ 1, & \text{if } f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,k}) - f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,i}) > \tau(n,i) \end{cases} \tag{3}$$

for every pair of windows $i \leq k$. If $\{\mathcal{P}_i\}_{i=n-k}^{n-1}$ are close and the thresholds are suitably chosen, then $\mathcal{I}_{1,k} = \cdots = \mathcal{I}_{k,k} = 0$ with high probability. Such test results give us the green light to pool $\{\mathcal{D}_i\}_{i=n-k}^{n-1}$. When $\mathcal{I}_{i,k} = 1$ for some $i < k$, a harmful distribution shift seems to have occurred in the last k time periods, and the positive test result raises a red flag.

The pairwise tests lead to a notion of admissibility: a window size $k \in [n-1]$ is said to be *admissible* if $\mathcal{I}_{i,k} = 0$ for all $i \in [k]$. Our stability principle suggests choosing the largest admissible window. In doing so, we maximize the utilization of historical data while keeping the cumulative bias within an acceptable range relative to the stochastic error. We name the procedure stability-based adaptive window selection, or SAWS for short. A formal description is given by Algorithm 1.

Algorithm 1 (Stability-Based Adaptive Window Selection (Subroutine))

Input: Samples $\{\mathcal{D}_i\}_{i=1}^{n-1}$, nonincreasing sequence of thresholds $\{\tau(n,k)\}_{k=1}^{n-1} \subseteq [0,\infty)$.

For $k = 1,\dots,n-1$**:**

Compute a minimizer $\widehat{\boldsymbol{\theta}}_{n,k}$ of $f_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} f_i$, where f_i is defined in (1).

Let $\mathcal{I}_k = 0$ if $f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,k}) - f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,i}) \le \tau(n,i)$ holds for all $i \in [k]$, and $\mathcal{I}_k = 1$ otherwise.

Let $\widehat{k} = \max\{k \in [m] : \mathcal{I}_k = 0\}$.

Output: $\boldsymbol{\theta}_n = \widehat{\boldsymbol{\theta}}_{n,\widehat{k}}$.

Algorithm 2 (Stability-Based Adaptive Window Selection (Online Version))

Input: Thresholds $\{\tau(n,k)\}_{n\in\mathbb{Z}_+,k\in[n-1]} \subseteq [0,\infty)$.

Choose any $\boldsymbol{\theta}_1 \in \Omega$.

For $n = 2,\dots,N$**:**

Run Algorithm 1 with samples $\{\mathcal{D}_i\}_{i=1}^{n-1}$ and thresholds $\{\tau(n,k)\}_{k=1}^{n-1}$ to obtain $\boldsymbol{\theta}_n$.

Output: $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$.

To tackle online learning under nonstationarity (Problem 1), we apply Algorithm 2, which runs Algorithm 1 as a subroutine in each period $n \in \mathbb{Z}_+$. In Section 4, we will design $\tau(n,k)$ to get sharp theoretical guarantees simultaneously for all horizons $N \in \mathbb{Z}_+$. Roughly speaking, when the population losses $\{F_n\}_{n=1}^{N}$ are strongly convex, we choose $\tau(n,k) \asymp \frac{d\log n}{Bk}$, with d being the dimension of the decision space Ω; when the population losses $\{F_n\}_{n=1}^{N}$ are only Lipschitz, we choose $\tau(n,k) \asymp \sqrt{\frac{d\log n}{Bk}}$.

3.3. Efficiency Improvements

Algorithm 3 (Stability-Based Adaptive Window Selection (General Subroutine))

Input: Samples $\{\mathcal{D}_i\}_{i=n_0}^{n-1}$, nonincreasing sequence of thresholds $\{\tau(n,k)\}_{k=1}^{n-1} \subseteq [0,\infty)$, window sizes $\{k_s\}_{s=1}^{m} \subseteq [n-1]$ that satisfy $1 = k_1 < \cdots < k_m = n - n_0$.

For $s = 1,\dots,m$**:**

Compute a minimizer $\widehat{\boldsymbol{\theta}}_{n,k_s}$ of $f_{n,k_s} = \frac{1}{k_s}\sum_{i=n-k_s}^{n-1} f_i$, where f_i is defined in (1).

Let $\mathcal{I}_s = 0$ if $f_{n,k_i}(\widehat{\boldsymbol{\theta}}_{n,k_s}) - f_{n,k_i}(\widehat{\boldsymbol{\theta}}_{n,k_i}) \le \tau(n,k_i)$ holds for all $i \in [s]$, and $\mathcal{I}_s = 1$ otherwise.

Let $\widehat{s} = \max\{s \in [m] : \mathcal{I}_s = 0\}$.

Output: $\boldsymbol{\theta}_n = \widehat{\boldsymbol{\theta}}_{n,k_{\widehat{s}}}$ and $\widehat{k} = k_{\widehat{s}}$.

Algorithm 4 (Stability-Based Adaptive Window Selection (Online Version with Improved Efficiency))

Input: Thresholds $\{\tau(n,k)\}_{n\in\mathbb{Z}_+,k\in[n-1]} \subseteq [0,\infty)$.

Let $K_1 = 0$ and choose any $\boldsymbol{\theta}_1 \in \Omega$.

For $n = 2,\dots,N$**:**

Let $m = \lceil \log_2(K_{n-1}+1) \rceil + 1$, $k_s = 2^{s-1}$ for $s \in [m-1]$, and $k_m = K_{n-1}+1$.

Run Algorithm 3 with inputs $\{\mathcal{D}_i\}_{i=n-k_m}^{n-1}$, $\{\tau(n,k)\}_{k=1}^{n-1}$ and $\{k_s\}_{s=1}^{m}$ to obtain $\boldsymbol{\theta}_n$ and $\widehat{k}$.

Let $K_n = \widehat{k}$.

Output: $\{\boldsymbol{\theta}_n\}_{n=1}^{N}$.

In the worst case, Algorithm 1 solves $\mathcal{O}(n)$ empirical risk minimization problems at time $n \in \mathbb{Z}_+$. Running Algorithm 2 up to time N requires solving $\mathcal{O}(N^2)$ empirical risk minimization problems and storing $\mathcal{O}(NB)$ samples. To improve computational and memory efficiency, we further develop more efficient versions of Algorithms 1 and 2, given by Algorithms 3 and 4, respectively. They incorporate the following two efficiency improvements.

First, Algorithm 3 allows for a general collection of candidate windows $\{k_s\}_{s=1}^{m}$ that is not necessarily the whole set $\{1,2,\dots,n-1\}$. In particular, we will use the geometric sequence $k_s = 2^{s-1}$ so that at most $\mathcal{O}(\log n)$ empirical risk minimization problems are solved at each time $n \in \mathbb{Z}_+$. Improving the efficiency of a search procedure by adopting a geometric candidate sequence is a standard technique in learning under nonstationarity (Hazan and Seshadhri 2009) and beyond.

Second, Algorithm 4, which runs Algorithm 3 as a subroutine, employs a caching mechanism that discards irrelevant past data upon detection of nonstationarity. Specifically, at each time $n \in \mathbb{Z}_+$, Algorithm 4 applies Algorithm 3 to obtain a window $K_n = \widehat{k}$. As the window K_n indicates that a significant distribution shift has been detected at time $n - K_n$, Algorithm 4 discards all past data before time $n - K_n$. Thus, in the next period $n+1$, it suffices to consider look-back windows with lengths at most $K_n + 1$, which leads to the candidate window sequence $\{k_s\}_{s=1}^{m}$ with $k_s = 2^{s-1}$ $\forall$ $s \in [m-1]$ and $k_m = K_n + 1$. Similar ideas are also used in multiple change-point detection (Niu et al. 2016).

Finally, we emphasize that our algorithms do not require any prior information on the nonstationarity of the underlying environment.

4. Regret Analysis in Common Settings

In this section, we will provide theoretical guarantees for SAWS (Algorithm 4) in two scenarios where the population losses are strongly convex and smooth, or

Lipschitz only. Throughout this section, we make the following standard assumption.

Assumption 1 (Regularity of Domain). *The decision set Ω is a closed convex subset of $\mathbb{R}^d$, and $\operatorname{diam}(\Omega) = M < \infty$ is a constant.*

4.1. Strongly Convex Population Losses

Our first study concerns the case where each population loss F_n is strongly convex and thus has a unique minimizer. To set the stage, we make the following standard assumptions.

Assumption 2 (Strong Convexity and Smoothness). *The loss function $\ell\colon \Omega \times \mathcal{Z} \to \mathbb{R}$ is convex and continuously differentiable with respect to its first argument. There exist constants $0 < \rho \le L < \infty$ such that for every $n \in \mathbb{Z}_+$, F_n is ρ-strongly convex and L-smooth:*

$$F_n(\boldsymbol{\theta}') \ge F_n(\boldsymbol{\theta}) + \langle \nabla F_n(\boldsymbol{\theta}),\ \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \frac{\rho}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2.$$

$$\|\nabla F_n(\boldsymbol{\theta}) - \nabla F_n(\boldsymbol{\theta}')\|_2 \le L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \quad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Omega$$

Moreover, for each $n \in \mathbb{Z}_+$, F_n attains its minimum at an interior point $\boldsymbol{\theta}_n^*$ of Ω.

Assumption 3 (Concentration). *There exist constants $\sigma, \lambda > 0$ such that for all $n \in \mathbb{Z}_+$ and $\boldsymbol{z}_n \sim \mathcal{P}_n$,*

$$\sup_{\boldsymbol{\theta}\in\Omega} \|\nabla \ell(\boldsymbol{\theta}, \boldsymbol{z}_n) - \nabla F_n(\boldsymbol{\theta})\|_{\psi_1} \le \sigma,$$

$$\mathbb{E}\left[\sup_{\substack{\boldsymbol{\theta},\boldsymbol{\theta}'\in\Omega \\ \boldsymbol{\theta}\neq\boldsymbol{\theta}'}} \frac{\|\nabla \ell(\boldsymbol{\theta}, \boldsymbol{z}_n) - \nabla \ell(\boldsymbol{\theta}', \boldsymbol{z}_n)\|_2}{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2}\right] \le \lambda^2 d.$$

Here, the gradient of ℓ is taken with respect to the first argument $\boldsymbol{\theta}$.

Assumption 2 states that F_n is strongly convex and smooth and attains its minimum at some interior point of the domain. The interior minimizer assumption is common in the literature of nonstationary stochastic optimization (Besbes et al. 2015, Wang 2025). Assumption 3 states that the empirical losses have subexponential tails and Lipschitz continuous gradients. Below, we present canonical examples that satisfy Assumptions 2 and 3. In these examples, $\Omega = B(\mathbf{0}, M/2)$ is a ball with diameter M, and $\sigma_0 > 0$ is a constant. We defer their verifications to Section EC.2.1 in the electronic companion.

Example 1 (Gaussian Mean Estimation). Suppose $\mathcal{Z} = \mathbb{R}^d$, $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{z}\|_2^2$, and $\mathcal{P}_n = N(\boldsymbol{\theta}_n^*, \sigma_0^2 \boldsymbol{I}_d)$ for some $\boldsymbol{\theta}_n^*$ with $\|\boldsymbol{\theta}_n^*\|_2 < M/2$. Then, Assumptions 2 and 3 hold with $\rho = L = \lambda = 1$ and $\sigma = c\sigma_0$ for some universal constant $c \ge 1/2$.

Example 2 (Linear Regression). Each sample $\boldsymbol{z}_n \sim \mathcal{P}_n$ takes the form $\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, where the covariate vector $\boldsymbol{x}_n$ and the response y_n satisfy $\mathbb{E}(y_n|\boldsymbol{x}_n) = \boldsymbol{x}_n^\top \boldsymbol{\theta}_n^*$. Define the squared loss $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \frac{1}{2}(y - \boldsymbol{x}^\top \boldsymbol{\theta})^2$ and the error term $\varepsilon_n = y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta}_n^*$. Suppose that $\|\boldsymbol{\theta}_n^*\|_2 < M/2$, $\|\boldsymbol{x}_n\|_{\psi_2} \le \sigma_0$, $\|\varepsilon_n\|_{\psi_2} \le \sigma_0$, and $\mathbb{E}(\boldsymbol{x}_n \boldsymbol{x}_n^\top) \ge \gamma\sigma_0^2 \boldsymbol{I}_d$ for some constant $\gamma \in (0, 1]$. Then, Assumptions 2 and 3 hold with $\sigma \asymp (M+1)\sigma_0^2$, $\lambda \asymp \sigma_0$, $\rho \asymp \gamma\sigma_0^2$, and $L \asymp \sigma_0^2$.

Example 3 (Logistic Regression). Each sample $\boldsymbol{z}_n \sim \mathcal{P}_n$ takes the form $\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{0, 1\}$, where the covariate vector $\boldsymbol{x}_n$ and the binary label y_n satisfy $\mathbb{P}(y_n = 1|\boldsymbol{x}_n) = 1/[1 + \exp(-\boldsymbol{x}_n^\top \boldsymbol{\theta}_n^*)]$. Define the logistic loss $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \log[1 + \exp(\boldsymbol{x}^\top \boldsymbol{\theta})] - y\boldsymbol{x}^\top \boldsymbol{\theta}$. Suppose that $\|\boldsymbol{\theta}_n^*\|_2 < M/2$, $\|\boldsymbol{x}_n\|_{\psi_1} \le \sigma_0$, and $\mathbb{E}(\boldsymbol{x}_n \boldsymbol{x}_n^\top) \ge \gamma\sigma_0^2 \boldsymbol{I}_d$ for some constant $\gamma \in (0, 1]$. Then, Assumptions 2 and 3 hold with $\sigma \asymp \sigma_0$, $\lambda \asymp \sigma_0$, $L \asymp \sigma_0^2$, and $\rho = c\gamma\sigma_0^2$ for some $c > 0$ determined by M, γ, and σ_0.

Example 4 (Robust Linear Regression). Each sample $\boldsymbol{z}_n \sim \mathcal{P}_n$ takes the form $\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, where the covariate vector $\boldsymbol{x}_n$ and the response y_n satisfy $y_n = \boldsymbol{x}_n^\top \boldsymbol{\beta}_n^* + \varepsilon_n$. Suppose $M \ge 1$, $\|\boldsymbol{\beta}_n^*\|_2 \le M/4$, $\|\boldsymbol{x}_n\|_{\psi_2} \le \sigma_0$, and $\mathbb{E}(\boldsymbol{x}_n \boldsymbol{x}_n^\top) \ge \gamma\sigma_0^2 \boldsymbol{I}_d$ for some constant $\gamma \in (0, 1]$. Assume that the noise ε_n follows the Huber contamination model (Huber 1964): $\varepsilon_n \sim (1-p)\mathcal{Q}_n^* + p\mathcal{Q}_n$ for some $p \in (0, 1)$, where $\mathcal{Q}_n^*$ is symmetric with respect to zero and has a sub-Gaussian norm bounded by σ_0, whereas $\mathcal{Q}_n$ can be arbitrary and may have a nonzero mean and a heavy tail. For $u > 0$, define the Huber loss

$$h_u(t) = \begin{cases} \dfrac{1}{2}t^2, & \text{if } |t| \le u \\ u\left(|t| - \dfrac{1}{2}u\right), & \text{otherwise.} \end{cases}$$

Choose $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = h_u(y - \boldsymbol{x}^\top \boldsymbol{\theta})$ with $u = cM\sigma_0$ for some constant $c > 0$. Then, for $(p^{-1} - 1)\gamma$ sufficiently large, Assumptions 2 and 3 hold with $\sigma \asymp M\sigma_0^2$, $\lambda \asymp \sigma_0$, $\rho \asymp \gamma\sigma_0^2$, and $L \asymp \sigma_0^2$.

We emphasize that only the population loss F_n, but not the empirical loss f_n, is assumed to be strongly convex. This is much weaker than assuming that f_n is strongly convex or exp-concave, as is commonly done in the literature (Hazan and Seshadhri 2009, Mokhtari et al. 2016, Baby and Wang 2022). For example, f_n is not strongly convex in Examples 2 and 3 when the batch size B in each time period is smaller than the dimension d. In Example 4, f_n is neither strongly convex nor exp-concave because of the linearity of h_u in $(-\infty, -u) \cup (u, \infty)$.

The regret bound of our algorithm will depend on the nonstationarity of the environment. We propose to measure it by decomposing the minimizer sequence $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ into *quasistationary segments*. Within each segment, the environment has small variations and can be treated as stationary. In this way, the nonstationarity is reflected by the number of such segments: a stationary environment is just one segment itself,

whereas a heavily fluctuating environment needs to be divided into a large number of short segments. Figure 1 provides a visualization of segmentation.

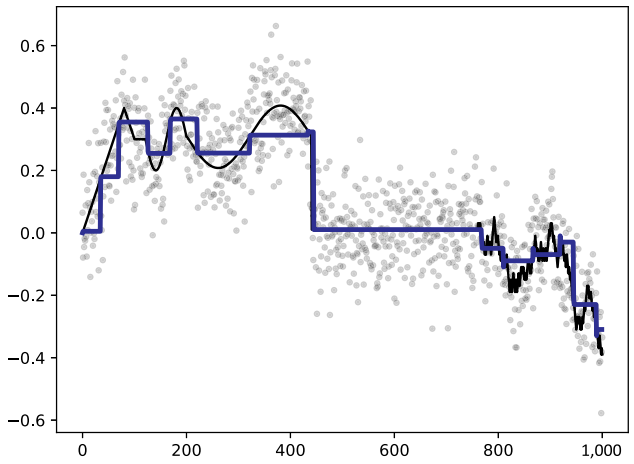
To motivate our segmentation criterion, consider the mean estimation problem in Example 1 with $d = 1$ and $\sigma_0 = 1$, and let $\Omega = \mathbb{R}$ for simplicity. For any time $n \in [N-1]$ and look-back window $k \in [n-1]$, the empirical minimizer $\widehat{\boldsymbol{\theta}}_{n,k} = \arg\min_{\boldsymbol{\theta}\in\Omega} f_{n,k}(\boldsymbol{\theta})$ has distribution $N\left(\frac{1}{k}\sum_{i=n-k}^{n-1} \boldsymbol{\theta}_i^*, \frac{1}{Bk}\right)$. If $\{\boldsymbol{\theta}_i^*\}_{i=n-k}^{n-1}$ differ by at most $\mathcal{O}(1/\sqrt{Bk})$, then the bias of $\widehat{\boldsymbol{\theta}}_{n,k}$ in estimating $\boldsymbol{\theta}_{n-1}^*$ is at most comparable to its stochastic error, so the distribution shift over the past k periods can be ignored. In general, we treat a length-k segment of $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ as stationary if its variation does not exceed $\mathcal{O}(\sqrt{\frac{d}{Bk}})$, which leads to the following Definition 1.

Definition 1 (Segmentation). The minimizer sequence $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ of $\{F_n\}_{n=1}^N$ is said to consist of J ***quasistationary segments*** if there exist $0 = N_0 < N_1 < \cdots < N_J = N-1$ such that for each $j \in [J]$,

$$\max_{N_{j-1}<i,k\le N_j} \|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_k^*\|_2 \le \sqrt{\frac{2M\sigma}{\rho}\max\left\{\frac{\sigma}{\rho M}, 1\right\}\cdot\frac{d}{B(N_j - N_{j-1})}}.$$

We can always decompose any $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ into $N-1$ quasistationary segments by setting $N_j = j$ for each $j \in [J]$, where each segment only contains a single time period. In what follows, we will always take a segmentation of $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ that results in the smallest J so that a larger J indicates greater nonstationarity. When the environment is stationary, that is, $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_N^*$, we have $J = 1$. The lemma below bounds J in terms of the *path variation* (*PV*) or *path length* $\sum_{n=1}^{N-1}\|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2$, which is a popular measure of nonstationarity (Zinkevich 2003, Jadbabaie et al. 2015, Zhang et al. 2018). The proof is deferred to Section EC.2.2 in the electronic companion.

Figure 1. (Color online) Visualization of Segmentation for Example 1



Notes. Horizontal axis: time n. Vertical axis: values of $\theta_n^* \in \mathbb{R}$. Black curve: trajectory of $\{\theta_n^*\}_{n=1}^N$. Gray dots: samples from $N(\theta_n^*, 0.01)$. Blue curve: quasistationary segments of $\{\theta_n^*\}_{n=1}^N$. The sequence $\{\theta_n^*\}_{n=1}^N$ is approximated by multiple constant segments, and within each segment, θ_n^* only has small variations.

Lemma 1 (From Path Variation to Segmentation). *Suppose* $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ *consists of* J *quasistationary segments, and define* $V = \sum_{n=1}^{N-1}\|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2$. *Then,* $J \le 1 + C(BN/d)^{1/3}V^{2/3}$, *where* $C > 0$ *is a constant depending on* M, ρ, *and* σ.

On the other hand, Example 5 below shows that sequences with the same path variation may have very different numbers of segments. An important reason is that whereas the path variation tracks all the distribution shifts, our segmentation aims to capture only those that lead to significant changes in the optimal solution. As a consequence, our measure of nonstationarity is often more optimistic and refined than the path variation. Indeed, we will later see that the former yields a tighter regret bound than the latter.

Example 5. We consider several patterns of nonstationarity in the setting of Example 1. For simplicity, we assume that $B = 1$, $d = 1$, $\Omega = [0,1]$ and N is large, and we omit rounding a number to its nearest integer. The following sequences $\{\theta_n^*\}_{n=1}^N$ all have path variation $V \asymp N^{1/2}$, so Lemma 1 implies $J \lesssim N^{2/3}$.

1. Large zigzags (Figure 2(a)): For every $n \in [N]$, $|\theta_{n+1}^* - \theta_n^*| = N^{-1/2}$. Moreover, for each $k \in [N^{2/3}]$, θ_n^* is monotone on $(k-1)N^{1/3} < n \le kN^{1/3}$. Then, we can take $N_j \asymp jN^{1/3}$ and $J \asymp N^{2/3}$.

2. Small zigzags (Figure 2(b)): For every $n \in [N]$, $\theta_{n+1}^* = \theta_n^* - (-1)^n N^{-1/2}$. Then, we can take $J = 1$.

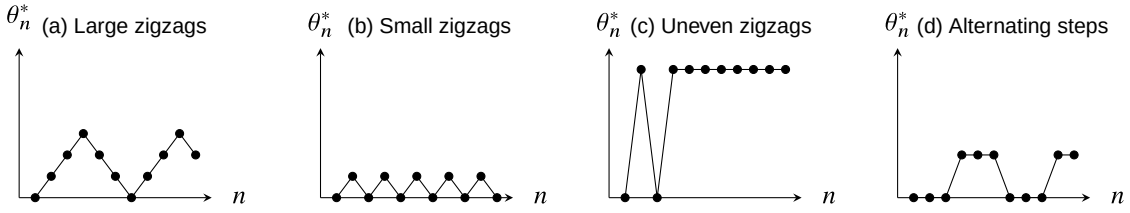
3. Uneven zigzags (Figure 2(c)): For every $n \in [N^{1/2}]$, $|\theta_{n+1}^* - \theta_n^*| = 1$. Moreover, θ_n^* is constant on $N^{1/2} < n \le N$. Then, we can take $J \asymp N^{1/2}$, with $N_j = j$ for $j \in [J-1]$ and $N_J = N-1$.

4. Alternating steps (Figure 2(d)): Choose any $u \in [N^{-1/2}, N^{-1/6}]$. For $k \in [N^{1/2}u^{-1}]$, the sequence θ_n^* is constant on $kN^{1/2}u < n \le (k+1)N^{1/2}u$, and $\theta_{kN^{1/2}u+1}^* = \theta_{kN^{1/2}u}^* - (-1)^k u$. Then, each constant piece has length $N^{1/2}u$; each segment contains $N^{-1/2}u^{-3}$ constant pieces and thus has length u^{-2}. We can take $N_j \asymp ju^{-2}$ and $J \asymp Nu^2 \in [1, N^{2/3}]$.

We are now ready to present the dynamic regret of Algorithm 4. See Appendix C for a sketch of the proof and Section EC.2.3 for a full proof. In both proofs, we present a more refined bound.

Theorem 1 (Regret Bound). *Let Assumptions 1, 2, and 3 hold. Let* J_N *be the number of quasistationary segments in* $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$. *Choose any* $\alpha \in (0,1]$. *There exists a constant* $\bar{C}_\tau > 0$ *such that if we choose* $C_\tau \ge \bar{C}_\tau$ *and run Algorithm 4*

Figure 2. Several Nonstationarity Patterns in Example 5



with

$$\tau(n,k) = C_\tau \frac{d}{Bk} \log(\alpha^{-1} + B + n),$$

then with probability at least $1-\alpha$, *the output* $\{\boldsymbol{\theta}_n\}_{n=1}^N$ *of Algorithm 4 satisfies*

$$\sum_{n=1}^{N} [F_n(\boldsymbol{\theta}_n) - F_n(\boldsymbol{\theta}_n^*)] \lesssim \min\left\{ J_N\left(\frac{d}{B}+1\right), N \right\}, \quad \forall N \in \mathbb{Z}_+. \tag{4}$$

Here, $\lesssim$ *only hides a polylogarithmic factor of* B, N, *and* α^{-1}.

Theorem 1 states that the dynamic regret of Algorithm 4 scales linearly with the number of quasistationary segments J_N, so a less variable sequence $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ leads to a smaller regret bound. We emphasize that Algorithm 4 attains this bound without any knowledge of the nonstationarity. In Section 6.1, we provide a minimax lower bound that matches the Regret Bound (4) up to logarithmic factors, showing the adaptivity of our algorithm to the unknown nonstationarity. In the stationary case where $\boldsymbol{\theta}_1^* = \cdots = \boldsymbol{\theta}_N^*$, we have $J_N = 1$, and Algorithm 4 attains a logarithmic regret. We also mention that Theorem 1 continues to hold when $\widehat{\boldsymbol{\theta}}_{n,k}$ is only an approximate minimizer of $f_{n,k}$ satisfying $f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) - \min_{\boldsymbol{\theta}\in\Omega} f_{n,k}(\boldsymbol{\theta}) = \mathcal{O}(\frac{d}{Bk})$.

As a corollary of a more refined version of Theorem 1 in Section EC.2.3 in the electronic companion, we derive a near-optimal regret bound for Algorithm 4 in terms of the path variation $\sum_{n=1}^{N-1} \|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2$. We prove Corollary 1 in Section EC.2.6 and show its near optimality in Section 6 by providing a minimax lower bound that matches it up to logarithmic factors.

Corollary 1 (PV-Based Regret Bound). *Consider the setting of Theorem 1 and define* $V_N = \sum_{n=1}^{N-1} \|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2$. *With probability at least* $1-\alpha$, *the output* $\{\boldsymbol{\theta}_n\}_{n=1}^N$ *of Algorithm 4 satisfies*

$$\sum_{n=1}^{N} [F_n(\boldsymbol{\theta}_n) - F_n(\boldsymbol{\theta}_n^*)] \lesssim 1 + \frac{d}{B} + N^{1/3}\left(\frac{V_N d}{B}\right)^{2/3} + V_N, \quad \forall N \in \mathbb{Z}_+.$$

Here, $\lesssim$ only hides a polylogarithmic factor of B, N, and α^{-1}.

We now revisit Example 5 to illustrate that the segmentation-based bound in Theorem 1 can be much tighter than the PV-based bound in Corollary 1. For the sequences in Example 5, Theorem 1 gives a regret bound of $\widetilde{\mathcal{O}}(J_N)$, which is often much smaller than $N^{2/3}$. In contrast, because $V_N \asymp N^{1/2}$, then Corollary 1 always gives a regret bound $\widetilde{\mathcal{O}}(N^{2/3})$, failing to capture refined structures of nonstationarity.

Remark 1 (Other Variation Metrics). The nonstationarity of the environment can also be quantified through other variation metrics. In the noiseless case where $f_n = F_n$ is assumed to be strongly convex, Zhao and Zhang (2021) studied the squared path length $S_N = \sum_{n=1}^{N-1} \|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2^2$ and the functional variation $W_N = \sum_{n=1}^{N-1} \|F_{n+1} - F_n\|_\infty$ and derived an $\mathcal{O}(\min\{S_N, V_N, W_N\})$ regret bound (ignoring the dependence on the dimension); Baby and Wang (2022) defined $C_N = \sum_{n=1}^{N-1} \|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_1$ as the path length and derived an $\widetilde{\mathcal{O}}(d^{1/3} C_N^{2/3} N^{1/3})$ regret bound. Our results hold for the more challenging setting where f_n is a random realization of F_n and is not necessarily strongly convex. Besbes et al. (2015) considered the functional variation $W_N^* = \sum_{n=1}^{N-1} \sup_{\boldsymbol{\theta}\in\Omega^*} |F_{n+1}(\boldsymbol{\theta}) - F_n(\boldsymbol{\theta})|$, where Ω^* is the convex hull of the minimizers $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$. For learning with noisy first-order feedback in constant dimension, they showed that the minimax optimal regret is $\widetilde{\mathcal{O}}(\sqrt{W_N^* N})$. In Section EC.2.7 in the electronic companion, we recover the same regret bound from Theorem 1 by showing that the number of quasistationary segments J_N is bounded by $1 + \mathcal{O}(\sqrt{N W_N^* B/d})$.

Remark 2 (Segmentation). The idea of quasistationary segments has appeared in various forms. Baby and Wang (2019) and Chen et al. (2019b) performed segmentation by comparing a certain path variation within a time interval against the stochastic error in the settings of one-dimensional mean estimation and contextual bandits, respectively. In contrast, our segmentation uses the maximum variation between any two time periods within a segment, which can be substantially smaller than the path variation, enabling detection of more refined nonstationarity. In the noiseless

($f_n = F_n$) and exp-concave setting, Baby and Wang (2021) performed segmentation on a dynamic comparator sequence for regret analysis, which is not intrinsic to the environment. Moreover, it is not clear how their analysis can be extended to the noisy setting, where the empirical loss f_n may not be strongly convex or exp-concave even if the population loss F_n is. Finally, Suk and Kpotufe (2022) proposed a similar concept for bandits named "significant phases" by comparing the dynamic regret under nonstationarity against the regret in the stationary case.

4.2. Lipschitz Population Losses

Our second study concerns a less regular case where each F_n is only assumed to be Lipschitz. The presentation parallels that of the strongly convex case in Section 4.1. We make the following assumption, which states that the empirical losses have sub-Gaussian tails and that the empirical and population losses are Lipschitz. In particular, the loss functions ℓ and F_n need not be convex.

Assumption 4 (Concentration and Smoothness). *There exist constants $\sigma, \lambda > 0$ such that for all $n \in \mathbb{Z}_+$ and $\boldsymbol{z}_n \sim \mathcal{P}_n$,*

$$\|\ell(\boldsymbol{\theta}_1, \boldsymbol{z}_n) - \ell(\boldsymbol{\theta}_2, \boldsymbol{z}_n) - [F_n(\boldsymbol{\theta}_1) - F_n(\boldsymbol{\theta}_2)]\|_{\psi_2} \le \sigma,$$
$$|F_n(\boldsymbol{\theta}_1) - F_n(\boldsymbol{\theta}_2)| \le \lambda \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Omega,$$

and

$$\mathbb{E}\left(\sup_{\substack{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Omega \\ \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2}} \frac{|\ell(\boldsymbol{\theta}_1, \boldsymbol{z}_n) - \ell(\boldsymbol{\theta}_2, \boldsymbol{z}_n)|}{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2} \right) \le \lambda \sqrt{d}.$$

Below, we give several classical examples satisfying Assumption 4, where $\mathcal{Z} = \mathbb{R}^d$, and $\sigma_0 > 0$ is a constant. We leave their verifications to Section EC.2.8 in the electronic companion.

Example 6 (Stochastic Linear Optimization). Let Ω be a polytope and $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{\theta}$. Suppose $\sup_{\boldsymbol{\theta} \in \Omega} \|\boldsymbol{z}^\top \boldsymbol{\theta}\|_{\psi_2} \le \sigma_0$ and $\mathbb{E}(\boldsymbol{z}_n \boldsymbol{z}_n^\top) \preceq \sigma_0^2 \boldsymbol{I}_d$. Then, Assumption 4 holds with $\sigma = 4\sigma_0$ and $\lambda = \sigma_0$.

Example 7 (Quantile Regression). Each sample $\boldsymbol{z}_n \sim \mathcal{P}_n$ takes the form $\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, where $\boldsymbol{x}_n$ is the covariate vector, and y_n is the response. Let $\nu \in [0,1]$ and define the check loss $\rho_\nu(z) = (1-\nu)(-z)_+ + \nu z_+$. In quantile regression for the ν-th conditional quantile of y given $\boldsymbol{x}$, we use the loss $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \rho_\nu(y - \boldsymbol{x}^\top \boldsymbol{\theta})$. Suppose $\|\boldsymbol{x}_n\|_{\psi_2} \le \sigma_0$. Then, Assumption 4 holds with $\sigma \asymp M\sigma_0$ and $\lambda \asymp \sigma_0$.

Example 8 (Newsvendor Problem). Let $d = 1$. The sample $z_n \sim \mathcal{P}_n$ represents the demand, and the decision θ represents the stocking quantity. The loss function is $\ell(\theta, z) = h(\theta - z)_+ + b(z - \theta)_+$, where h is the holding/overage cost, and b is the backorder/underage cost. Suppose $\|z_n\|_{\psi_2} \le \sigma_0$. Then, Assumption 4 holds with $\sigma \asymp (h+b)M\sigma_0$ and $\lambda \asymp (h+b)\sigma_0$. We note that the newsvendor problem can be cast as a special case of quantile regression in Example 7 with $\nu = b/(h+b)$.

Example 9 (Support Vector Machine). Let $\Omega = B(\boldsymbol{0}, M/2)$. Each sample $\boldsymbol{z}_n \sim \mathcal{P}_n$ takes the form $\boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$, where $\boldsymbol{x}_n$ is the feature vector, and y_n is the label. The loss function for the soft-margin support vector machine is given by $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = (1 - y\boldsymbol{x}^\top \boldsymbol{\theta})_+$. Suppose $\|\boldsymbol{x}_n\|_{\psi_2} \le \sigma_0$. Then, Assumption 4 holds with $\sigma \asymp M\sigma_0$ and $\lambda = \sigma_0$.

As in the strongly convex case in Section 4.1, we will decompose the underlying sequence $\{F_n\}_{n=1}^N$ into quasistationary segments. In general, the Lipschitz population loss F_n does not have a unique minimizer, so the quantity $\|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_k^*\|_2$ in Definition 1 is not well defined. Moreover, even if each F_n has a unique minimizer, in the absence of strong convexity, a large distance $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n^*\|_2$ does not necessarily imply a large suboptimality gap $F_n(\boldsymbol{\theta}) - F_n(\boldsymbol{\theta}_n^*)$. Therefore, instead of the distance between minimizers, it is more suitable to measure the difference in function values. We will use $\|F_i - F_k\|_\infty$ to quantify the distribution shift between two periods, i and k.

To motivate the segmentation criterion, consider a one-dimensional example ($d = 1$). It is easily seen that the stochastic error $|f_{n,k}(\boldsymbol{\theta}) - F_{n,k}(\boldsymbol{\theta})|$ is of order $1/\sqrt{Bk}$ for every fixed $\boldsymbol{\theta} \in \Omega$. If $\{F_i\}_{i=n-k}^{n-1}$ differ by at most $\mathcal{O}(1/\sqrt{Bk})$, then the bias $F_{n-1}(\widehat{\boldsymbol{\theta}}_{n,k}) - \min_{\boldsymbol{\theta} \in \Omega} F_{n-1}(\boldsymbol{\theta}) \lesssim \|F_{n-1} - F_{n,k}\|_\infty$ is at most comparable to the stochastic error. In this case, we can ignore the distribution shift over the past k periods. In the general case, we think of a length-k segment of $\{F_n\}_{n=1}^N$ as stationary if its variation does not exceed $\mathcal{O}(\sqrt{\frac{d}{Bk}})$. This leads to Definition 2.

Definition 2 (Segmentation). The function sequence $\{F_n\}_{n=1}^N$ is said to consist of J *quasistationary segments* if there exist $0 = N_0 < N_1 < \cdots < N_J = N - 1$ such that for each $j \in [J]$,

$$\max_{N_{j-1} < i, k \le N_j} \|F_i - F_k\|_\infty \le \frac{\sigma}{2} \sqrt{\frac{d}{B(N_j - N_{j-1})}}.$$

As in Section 4.1, for each sequence $\{F_n\}_{n=1}^N$, we will take a segmentation that leads to the minimal J. In Lemma 2 below, we bound J in terms of the path variation $\sum_{n=1}^{N-1} \|F_{n+1} - F_n\|_\infty$. Its proof is given in Section EC.2.9 in the electronic companion.

Lemma 2 (From Path Variation to Segmentation). *Let $\{F_n\}_{n=1}^N$ consist of J quasistationary segments, and define $V = \sum_{n=1}^{N-1} \|F_{n+1} - F_n\|_\infty$. Then, $J \le 1 + C(BN/d)^{1/3} V^{2/3}$, where $C > 0$ is a constant depending on σ.*

In Theorem 2, we give a regret bound for Algorithm 4 in the Lipschitz case. Its proof can be found in Section EC.2.10 in the electronic companion and contains a more refined bound.

Theorem 2 (Regret Bound). *Let Assumptions 1 and 4 hold. Let J_N be the number of quasistationary segments in $\{F_n\}_{n=1}^N$. Choose any $\alpha \in (0,1]$. There exists a constant $\bar{C}_\tau > 0$ such that if we choose $C_\tau \geq \bar{C}_\tau$ and run Algorithm 4 with*

$$\tau(n,k) = C_\tau \sqrt{\frac{d}{Bk} \log(\alpha^{-1} + B + n)},$$

then with probability at least $1-\alpha$, the output $\{\boldsymbol{\theta}_n\}_{n=1}^N$ of Algorithm 4 satisfies

$$\sum_{n=1}^{N} [F_n(\boldsymbol{\theta}_n) - \min_{\boldsymbol{\theta}'_n \in \Omega} F_n(\boldsymbol{\theta}'_n)] \lesssim \min\left\{ J_N + \sqrt{J_N N \frac{d}{B}},\ N \right\}, \quad \forall N \in \mathbb{Z}_+. \tag{5}$$

Here, $\lesssim$ only hides a polylogarithmic factor of B, N, and α^{-1}.

Theorem 2 shows that the dynamic regret of Algorithm 4 in the Lipschitz case is $\widetilde{\mathcal{O}}(\sqrt{J_N N})$. As in the strongly convex case, the algorithm attains the Bound (5) without any prior knowledge of the nonstationarity. In Section 6.2, we provide a minimax lower bound that matches (5) up to logarithmic factors, which shows that our algorithm automatically adapts to the unknown nonstationarity. For a stationary environment where $F_1 = \cdots = F_N$, we have $J_N = 1$, which yields a regret bound of $\widetilde{\mathcal{O}}(\sqrt{N})$. We remark that Theorem 2 continues to hold when $\widehat{\boldsymbol{\theta}}_{n,k}$ is only an approximate minimizer of $f_{n,k}$ satisfying $f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) - \min_{\boldsymbol{\theta}\in\Omega} f_{n,k}(\boldsymbol{\theta}) = \mathcal{O}(\sqrt{\frac{d}{Bk}})$.

As a corollary of Theorem 2, we derive the following PV-based regret bound. Its proof is deferred to Section EC.2.13 in the electronic companion.

Corollary 2 (PV-Based Regret Bound). *Consider the setting of Theorem 2 and define $V_N = \sum_{n=1}^{N-1} \|F_{n+1} - F_n\|_\infty$. With probability at least $1-\alpha$, the output $\{\boldsymbol{\theta}_n\}_{n=1}^N$ of Algorithm 4 satisfies*

$$\sum_{n=1}^{N} [F_n(\boldsymbol{\theta}_n) - \min_{\boldsymbol{\theta}'_n \in \Omega} F_n(\boldsymbol{\theta}'_n)] \lesssim 1 + \sqrt{\frac{Nd}{B}} + N^{2/3}\left(\frac{V_N d}{B}\right)^{1/3} + V_N, \quad \forall N \in \mathbb{Z}_+. \tag{6}$$

Here, $\lesssim$ only hides a polylogarithmic factor of B, N, and α^{-1}.

We note that the PV-based regret bound in Corollary 2 exhibits an $\widetilde{\mathcal{O}}(V_N^{1/3} N^{2/3})$ dependence on V_N and N, which also appears in Besbes et al. (2015) for the setting of convex losses with noisy first-order feedback. In Section 6, we provide a minimax lower bound that matches the PV-based regret bound up to logarithmic factors.

5. A General Theory of Learning Under Nonstationarity

In this section, we will develop a general framework for analyzing Algorithms 3 and 4. It contains, as special cases, the regret bounds in Section 4. Our theory comprises two major components: a novel measure of similarity between functions and a general segmentation technique for dividing a nonstationary sequence into quasistationary pieces.

5.1. Overview

We begin with an overview of the main ideas to motivate our new notions. Recall that at time n, we seek to minimize F_n based on noisy observations $\{f_i\}_{i=1}^{n-1}$ of its predecessors $\{F_i\}_{i=1}^{n-1}$. Each look-back window $k \in [n-1]$ induces an estimated loss function $f_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} f_i$ and a candidate solution $\widehat{\boldsymbol{\theta}}_{n,k} \in \arg\min_{\boldsymbol{\theta}\in\Omega} f_{n,k}(\boldsymbol{\theta})$. Because $f_{n,k}$ is an empirical approximation of a surrogate $F_{n,k} = \frac{1}{k}\sum_{i=n-k}^{n-1} F_i$ for F_n, we can apply statistical learning theory to bound their discrepancies, ensuring that any approximate minimizer of $f_{n,k}$ is also near optimal for $F_{n,k}$, and vice versa.

Let $K \in [n-1]$ be the largest look-back window in which the environment only undergoes negligible changes. Ideally, this is the optimal window to use. However, the window K depends on the unknown nonstationarity, and we wish to use data to find a window $\widehat{k}$ that is comparable to K. To this end, we study basic properties of the window K. By the definition of K, F_n is very close to $\{F_i\}_{i=n-K}^{n-1}$ and thus $\{F_{n,k}\}_{k=1}^K$. This, combined with the fact that $f_{n,k}$ is close to $F_{n,k}$, leads to the following observation.

Fact 1. *For all $k \in [K]$, any point $\boldsymbol{\theta} \in \Omega$ that is near optimal for $f_{n,k}$ is also near optimal for F_n, and vice versa.*

Because $\widehat{\boldsymbol{\theta}}_{n,K} \in \arg\min_{\boldsymbol{\theta}\in\Omega} f_{n,K}(\boldsymbol{\theta})$, Fact 1 implies that $\widehat{\boldsymbol{\theta}}_{n,K}$ is near optimal for F_n. Applying Fact 1 again yields the following:

Fact 2. *For all $k \in [K]$, $\widehat{\boldsymbol{\theta}}_{n,K}$ is near optimal for $f_{n,k}$; that is, $f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,K}) - \min_{\boldsymbol{\theta}\in\Omega} f_{n,k}(\boldsymbol{\theta}) = f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,K}) - f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k})$ is small.*

Algorithm 3 chooses a window $\widehat{k}$ according to a rule that mimics Fact 2. For simplicity, consider its simple version, Algorithm 1, which selects

$$\widehat{k} = \max\{k \in [n-1] : \forall i \in [k], \ f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,k}) - f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,i}) \leq \tau(n,i)\}.$$

When is the performance of $\widehat{k}$ comparable to that of K?

- If $\widehat{k} \geq K$, then the window selection rule implies

$$f_{n,K}(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}}) - \min_{\boldsymbol{\theta}' \in \Omega} f_{n,K}(\boldsymbol{\theta}') = f_{n,K}(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}}) - f_{n,K}(\widehat{\boldsymbol{\theta}}_{n,K}) \leq \tau(n,K).$$

In this case, we can use Fact 1 to translate the bound above into a bound for $F_n(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}}) - \min_{\boldsymbol{\theta}'_n \in \Omega} F_n(\boldsymbol{\theta}'_n)$.

- If $\widehat{k} < K$, then the window selection rule implies the existence of $k \in [K-1]$ such that

$$f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,K}) - \min_{\boldsymbol{\theta}' \in \Omega} f_{n,k}(\boldsymbol{\theta}') = f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,K}) - f_{n,k}(\widehat{\boldsymbol{\theta}}_{n,k}) > \tau(n,k).$$

According to Fact 2, this cannot happen if the thresholds $\{\tau(n,i)\}_{i=1}^{K-1}$ are sufficiently large.

Consequently, it is desirable to have large $\{\tau(n,k)\}_{k=1}^{K-1}$ to keep $\widehat{k}$ from being too small but small $\tau(n,K)$ for bounding the suboptimality of $\widehat{\boldsymbol{\theta}}_{n,\widehat{k}}$. This is similar to controlling type I and type II errors in hypothesis testing. We choose $\tau(n,k)$ using simple bounds on the stochastic error of the empirical loss minimizer given by Bk independent samples. In particular, $\tau(n,k) \asymp \frac{d}{Bk}$ and $\sqrt{\frac{d}{Bk}}$ up to logarithmic factors for strongly convex and Lipschitz population losses, respectively.

To make the above analysis precise, we propose a novel notion of closeness between two functions: f and g with the same domain Ω are regarded as close if the suboptimality gaps $f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}')$ and $g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}')$ can bound each other up to an affine transform (Definition 3). The slope and the intercept of the affine transform provide a quantitative measure. It will help us depict the concentration of the empirical loss $f_{n,k}$ around its population version $F_{n,k}$, as well as the discrepancy between $F_{n,k}$ and F_n caused by the distribution shift over time. Moreover, it has convenient operation rules that enable the following reasoning:

- If $f_{n,k}$ is close to $F_{n,k}$ and if $F_{n,k}$ is close to F_n, then $f_{n,k}$ is close to F_n.
- If $\{F_i\}_{i=n-k}^{n-1}$ are close to F_n, then the average $F_{n,k}$ is also close to F_n.

We have seen that Algorithm 3 selects a window to maximize the utilization of historical data while keeping the cumulative bias under control. In the online setting, Algorithm 4 applies Algorithm 3 in every time period to get a look-back window tailored to the local nonstationarity. If the whole sequence $\{F_n\}_{n=1}^N$ consists of quasistationary segments, then Algorithm 4 is comparable to an oracle online algorithm that restarts at the beginning of each segment and treats data within the same segment as i.i.d. This observation leads to our formal notion of quasistationarity (Definition 4) based on function closeness (Definition 3) and a segmentation technique (Definition 5) for regret analysis.

5.2. A Measure of Closeness Between Two Functions

We now introduce our measure of function closeness.

Definition 3 (Closeness). Suppose that $f, g : \Omega \to \mathbb{R}$ are lower bounded and $\varepsilon, \delta \geq 0$. The functions f and g are said to be (ε,δ)-close if the following inequalities hold for all $\boldsymbol{\theta} \in \Omega$:

$$g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}') \leq e^{\varepsilon}\left(f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}') + \delta\right),$$

$$f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}') \leq e^{\varepsilon}\left(g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}') + \delta\right).$$

In this case, we also say that f is (ε,δ)-close to g.

The closeness measure reflects the conversion between the suboptimality gaps of two functions. We give a more geometric interpretation through a sandwich-type inclusion of sublevel sets.

Fact 3. *For any lower bounded $h : \Omega \to \mathbb{R}$ and $t \in \mathbb{R}$, define the sublevel set*

$$S(h,t) = \left\{\boldsymbol{\theta} \in \Omega : \; h(\boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta}' \in \Omega} h(\boldsymbol{\theta}') + t\right\}.$$

Two lower bounded functions $f, g : \Omega \to \mathbb{R}$ are (ε,δ)-close if and only if

$$S(g, e^{-\varepsilon}t - \delta) \subseteq S(f,t) \subseteq S(g, e^{\varepsilon}(t+\delta)), \quad \forall t \in \mathbb{R}.$$

Intuitively, δ measures the intrinsic discrepancy between two functions, and ε provides some leeway. The latter allows for a large difference between the suboptimality gaps $f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} f(\boldsymbol{\theta}')$ and $g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}' \in \Omega} g(\boldsymbol{\theta}')$ when $\boldsymbol{\theta}$ is highly suboptimal for f or g. After all, we are mainly interested in the behaviors of f and g near their minimizers. Similar ideas are also used in the peeling argument in empirical process theory (van de Geer 2000). Thanks to the scaling factor e^{ε}, our closeness measure gives a more refined characterization than the supremum metric $\|f - g\|_\infty = \sup_{\boldsymbol{\theta} \in \Omega} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta})|$. We illustrate this using the elementary example below.

Example 10. Let $\Omega = [-1,1]$ and $a, b \in \Omega$. If $f(\theta) = |\theta - a|$ and $g(\theta) = 2|\theta - b|$, then f and g are $(\log 2, |a-b|)$-close. In contrast, $\|f - g\|_\infty \geq 1$ always, even when f and g have the same minimizer $a = b$. To see this, because $f(-1) = 1 + a$, $g(-1) = 2 + 2b$, $f(1) = 1 - a$ and $g(1) = 2 - 2b$, then

$$\|f - g\|_\infty \geq \frac{|f(-1) - g(-1)| + |f(1) - g(1)|}{2} = \frac{|1 + 2b - a| + |1 - (2b - a)|}{2} \geq 1.$$

We now provide user-friendly conditions for computing the closeness parameters. The proof is deferred to Section EC.1.2 in the electronic companion.

Lemma 3. *Let $\Omega \subseteq \mathbb{R}^d$ be closed and convex, with* diam $(\Omega) = M < \infty$. *Let $f, g : \Omega \to \mathbb{R}$. Suppose that g is lower bounded.*

1. *If $D_0 = \sup_{\boldsymbol{\theta}\in\Omega} |f(\boldsymbol{\theta}) - g(\boldsymbol{\theta}) - c| < \infty$ for some $c \in \mathbb{R}$, then f and g are $(0, 2D_0)$-close.*
2. *If $D_1 = \sup_{\boldsymbol{\theta}\in\Omega} \|\nabla f(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta})\|_2 < \infty$, then f and g are $(0, 2MD_1)$-close.*
3. *If the assumption in part 2 holds and there exists $\rho > 0$ such that g is ρ-strongly convex over Ω, then f and g are $(\log 2, \frac{2}{\rho}\min\{D_1^2, \rho M D_1\})$-close.*
4. *Suppose there exist $0 < \rho \le L < \infty$ such that f and g are ρ-strongly convex and L-smooth over Ω. In addition, suppose that f and g attain their minima at some interior points $\boldsymbol{\theta}_f^*$ and $\boldsymbol{\theta}_g^*$ of Ω, respectively. Then, f and g are $(\log(\frac{4L}{\rho}), \frac{\rho}{2}\|\boldsymbol{\theta}_f^* - \boldsymbol{\theta}_g^*\|_2^2)$-close.*

For Lipschitz losses in Section 4.2, part 1 of Lemma 3 will be useful for establishing the closeness between the empirical loss $f_{n,k}$ and the population loss $F_{n,k}$ with $D_0 \asymp \sqrt{\frac{d}{Bk}}$, as well as the closeness between two population losses, F_n and F_i. For strongly convex losses in Section 4.1, part 3 of Lemma 3 applies to the pair $f_{n,k}$ and $F_{n,k}$ with $D_1 \asymp \sqrt{\frac{d}{Bk}}$, and part 4 applies to the pair F_n and F_i. We summarize these closeness results in Table 1.

Our notion of closeness shares some similarities with the equivalence relation, including reflexivity, symmetry, and a weak form of transitivity. See Lemma 4 below for its nice properties and Section EC.1.1 in the electronic companion for the proof.

Lemma 4. *Let $f, g, h : \Omega \to \mathbb{R}$ be lower bounded. Then,*

1. *f and f are $(0,0)$-close.*
2. *If f and g are (ε, δ)-close, then f and g are (ε', δ')-close for any $\varepsilon' \ge \varepsilon$ and $\delta' \ge \delta$.*
3. *If f and g are (ε, δ)-close and $a, b \in \mathbb{R}$, $f + a$ and $g + b$ are (ε, δ)-close.*
4. *If f and g are (ε, δ)-close, then g and f are (ε, δ)-close.*
5. *If f and g are $(\varepsilon_1, \delta_1)$-close and g and h are $(\varepsilon_2, \delta_2)$-close, then f and h are $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-close.*
6. *If $\sup_{\boldsymbol{\theta}\in\Omega} f(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in\Omega} f(\boldsymbol{\theta}) < F < \infty$ and $\sup_{\boldsymbol{\theta}\in\Omega} g(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in\Omega} g(\boldsymbol{\theta}) < G < \infty$, then f and g are $(0, \max\{F, G\})$-close.*
7. *Suppose that $\{f_i\}_{i=1}^m : \Omega \to \mathbb{R}$ are lower bounded and (ε, δ)-close to g. If $\{\lambda_i\}_{i=1}^m \subseteq [0,1]$ and $\sum_{i=1}^m \lambda_i = 1$, then $\sum_{i=1}^m \lambda_i f_i$ and g are $(\varepsilon, (e^\varepsilon + 1)\delta)$-close.*

Table 1. Results of (ε, δ)-Closeness for Section 4

Function pair	Strongly convex case	Lipschitz case
$f_{n,k}$ and $F_{n,k}$	$\varepsilon \asymp 1, \quad \delta \asymp \frac{d}{Bk}$	$\varepsilon \asymp 1, \quad \delta \asymp \sqrt{\frac{d}{Bk}}$
F_n and F_i	$\varepsilon \asymp 1, \quad \delta \asymp \|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_i^*\|_2^2$	$\varepsilon \asymp 1, \quad \delta \asymp \|F_n - F_i\|_\infty$

Note. Here, $\asymp$ may hide constants such as smoothness parameters.

5.3. Regret Analysis via Segmentation

To study the regret of Algorithm 4, we first investigate its subroutine Algorithm 3 at any given time n. We make the following assumption.

Assumption 5 (Stochastic Error). *There exist $\varepsilon \ge 0$ and $\{\psi(n,k)\}_{n\in\mathbb{Z}_+,\, k\in[n-1]}$ such that for all $n \in \mathbb{Z}_+$, $\psi(n,1) \ge \cdots \ge \psi(n, n-1) \ge 0$, and for all $k \in [n-1]$, $f_{n,k}$ and $F_{n,k}$ are $(\varepsilon, \psi(n,k))$-close.*

Assumption 5 states that at time n, the stochastic error of pooling data from the most recent k periods is characterized by $\psi(n,k)$. That $\psi(n,k)$ is decreasing in k is consistent with the intuition that pooling more data reduces the stochastic error. In Table 1, we have seen the closeness between $f_{n,k}$ and $F_{n,k}$ in the settings of Section 4: $\psi(n,k) \asymp \frac{d}{Bk}$ in the strongly convex case and $\psi(n,k) \asymp \sqrt{\frac{d}{Bk}}$ in the Lipschitz case, up to logarithmic factors.

We also impose the following conditions on the thresholds $\tau(n,k)$ used in Algorithm 4.

Condition 1. *For all $n \in [N]$, $\tau(n,1) \ge \cdots \ge \tau(n, n-1) \ge 0$, and for all $k \in [n-1]$, $\tau(n,k) \ge 6e^{5\varepsilon}\psi(n,k)$. There exists $C \ge 1$ such that for any $n \in [N]$ and $k \in [n-1]$, $\tau(n,k) \le C\tau(n, (2k) \wedge n)$. Finally, for every $k \in [N-1]$, it holds that $\tau(k+1, k) \le \cdots \le \tau(N, k)$.*

Based on Condition 1, we can choose the threshold $\tau(n,k)$ as a constant multiple of the stochastic error $\psi(n,k)$. Therefore, for the strongly convex losses and the Lipschitz losses in Section 4, we take $\tau(n,k) \asymp \frac{d}{Bk}$ and $\tau(n,k) \asymp \sqrt{\frac{d}{Bk}}$ up to some logarithmic factors, respectively. Both choices satisfy Condition 1 with $C = 2$.

We now present an excess risk bound for Algorithm 3. We provide a sketch of proof in Appendix A and a full proof in Section EC.1.3 in the electronic companion.

Theorem 3 (Excess Risk Bound). *Fix $n \in [N]$. Consider Algorithm 3 as a subroutine of Algorithm 4, with $k_{s+1} \le 2k_s$ for each $s \in [m-1]$. Let Assumption 5 and* Condition 1 *hold. Define*

$$\bar{k} = \max\{k \in [n-1] : F_{n-k}, F_{n-k+1} \ldots, F_{n-1} \text{ are } (\varepsilon, \psi(n,k)) - \text{close to } F_{n-1}\}.$$

Then, the output $\boldsymbol{\theta}_n$ of Algorithm 3 satisfies

$$F_{n-1}(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}\in\Omega} F_{n-1}(\boldsymbol{\theta}) \le 2e^{2\varepsilon} C\tau(n, \bar{k} \wedge k_m).$$

The window $\bar{k}$ is the precise mathematical formulation of the ideal window size K in Section 5.1. It is the

largest k for which the bias between F_{n-1} and each of $F_{n-k}, F_{n-k+1}, \ldots, F_{n-1}$ is no more than the stochastic error $\psi(n,k)$. It balances the bias and stochastic error, both of which are of the order $\psi(n,\bar{k})$. Consequently, the associated decision $\widehat{\boldsymbol{\theta}}_{n,\bar{k}}$ has excess risk of the order $\psi(n,\bar{k})$. Theorem 3 shows that the window $k_{\hat{s}}$ chosen by Algorithm 3 is a good approximation of $\bar{k}$ in the sense that the excess risk for $\boldsymbol{\theta}_n = \widehat{\boldsymbol{\theta}}_{n,k_{\hat{s}}}$ has order $\tau(n,\bar{k}\wedge k_m)$, which is approximately $\psi(n,\bar{k})$.

We proceed to analyze Algorithm 4 by approximating the sequence $\{F_n\}_{n=1}^N$ with approximately stationary pieces. We first define a concept of quasistationarity through our notion of function closeness and then introduce a general definition of segmentation.

Definition 4 (Quasistationarity). Let $n \in \mathbb{Z}_+$, $\varepsilon \geq 0$ and $\delta \geq 0$. A sequence of functions $\{g_i\}_{i=1}^n$ is said to be (ε,δ)-*quasistationary* if for all $i,j \in [n]$, g_i and g_j are (ε,δ)-close.

Definition 5 (Segmentation). The function sequence $\{F_n\}_{n=1}^N$ is said to consist of J *quasistationary segments* if there exist $\varepsilon \geq 0$, integers $0 = N_0 < N_1 < \cdots < N_J = N-1$, and nonnegative numbers $\{\delta_j\}_{j=1}^J$ such that for every $j \in [J]$,

- The sequence $\{F_i\}_{i=N_{j-1}+1}^{N_j}$ is $(\varepsilon, \min_{N_{j-1}<n\leq N_j} \psi(n, n-N_{j-1}))$-quasistationary.
- F_{N_j} and F_{N_j+1} are (ε,δ_j)-close.

We call $\{N_j\}_{j=1}^J$ the *knots* and $\{\delta_j\}_{j=1}^J$ the *jumps*.

In Definition 5, we characterize the nonstationarity of the environment by the number of quasisegments J as well as the scales of the jumps $\{\delta_j\}_{j=1}^J$ between consecutive segments. It is a generalization of Definition 1 and Definition 2 in Section 4.

We are now ready to present the regret bound for Algorithm 4. We provide a sketch of proof for Theorem 4 in Appendix B and a full proof in Section EC.1.4 in the electronic companion.

Theorem 4 (Regret Bound). *Let Assumption 5 and Condition 1 hold. Suppose $\{F_n\}_{n=1}^N$ consist of J quasistationary segments with knots $\{N_j\}_{j=1}^J$ and jumps $\{\delta_j\}_{j=1}^J$. Define $U = \max_{n\in[N]}[\sup_{\boldsymbol{\theta}\in\Omega} F_n(\boldsymbol{\theta}) - \inf_{\boldsymbol{\theta}\in\Omega} F_n(\boldsymbol{\theta})]$ and $T(n) = \sum_{i=1}^n \min\{\tau(N,i), U\}$. Then, the output $\{\boldsymbol{\theta}_n\}_{n=1}^N$ of Algorithm 4 satisfies*

$$\sum_{n=1}^N \left[F_n(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}'_n\in\Omega} F_n(\boldsymbol{\theta}'_n)\right] \leq \left[F_1(\boldsymbol{\theta}_1) - \inf_{\boldsymbol{\theta}\in\Omega} F_1(\boldsymbol{\theta})\right] + 3e^{3\varepsilon}C^2\sum_{j=1}^J T(N_j - N_{j-1}) + e^{\varepsilon}\sum_{j=1}^J \delta_j.$$

Theorem 4 contains Theorem 1 and Theorem 2 as special cases. Its regret bound consists of three terms. The first term results from our initial guess $\boldsymbol{\theta}_1$. In the second term, each summand is the regret incurred in the interior of a quasistationary segment. The third term is the cost of approximating F_{N_j+1} by F_{N_j} at the boundary between quasistationary segments.

6. Minimax Lower Bounds and Adaptivity

In this section, we present minimax lower bounds that match the regret bounds in Section 4 up to logarithmic factors. Because SAWS (Algorithm 4) is agnostic to the amount of distribution shift, our results show its adaptivity to the unknown nonstationarity.

6.1. Strongly Convex Population Losses

To prove the sharpness of Theorem 1 and Corollary 1, we consider simple classes of online Gaussian mean estimation problems described in Example 1. Fix a time horizon $N \in \mathbb{Z}_+$.

Definition 6 (Problem Classes). Let $\Omega = B(\mathbf{0},1)$. Define $\mathcal{Z}$, ℓ, and c as in Example 1. For $J \in [N-1]$, define the problem class

$$\mathscr{P}(J) = \Bigg\{(\mathcal{P}_1,\ldots,\mathcal{P}_N):\ \mathcal{P}_n = N(\boldsymbol{\theta}_n^*, \boldsymbol{I}_d),\ \boldsymbol{\theta}_n^* \in B(\mathbf{0},1/2),\ \forall n\in[N],\ \text{there exist } 0 = N_0 < \cdots < N_J = N-1 \text{ such that } \max_{N_{j-1}<i,k\leq N_j} \|\boldsymbol{\theta}_i^* - \boldsymbol{\theta}_k^*\|_2 \leq \sqrt{\frac{8c^2 d}{B(N_j - N_{j-1})}},\ \forall j\in[J]\Bigg\}.$$

In addition, for any $V \geq 0$, define

$$\mathcal{Q}(V) = \Bigg\{(\mathcal{P}_1,\ldots,\mathcal{P}_N):\ \mathcal{P}_n = N(\boldsymbol{\theta}_n^*, \boldsymbol{I}_d),\ \boldsymbol{\theta}_n^* \in B(\mathbf{0},1/2),\ \sum_{n=1}^{N-1}\|\boldsymbol{\theta}_{n+1}^* - \boldsymbol{\theta}_n^*\|_2 \leq V\Bigg\}.$$

For every problem instance in $\mathscr{P}(J)$ or $\mathcal{Q}(V)$, Assumptions 1, 2, and 3 hold with $M = 2$, $\sigma_0 = \rho = L = \lambda = 1$, and $\sigma = c$. The set $\mathscr{P}(J)$ consists of minimizer sequences $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ with at most J quasistationary segments, and $\mathcal{Q}(V)$ consists of minimizer sequences $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ with path variation at most V.

Theorem 5 below shows that for any algorithm, there exists a problem instance in the class such that the expected regret is at least comparable to the upper bound in Theorem 1. See Section EC.3.1 in the electronic companion for a stronger version and its proof.

Theorem 5 (Lower Bound). *Assume $N \geq 2$ and that $J \in [N-1]$ divides $N-1$. There exists a universal constant*

$C > 0$ such that

$$\inf_{\mathcal{A}} \sup_{(\mathcal{P}_1,\dots,\mathcal{P}_N)\in\mathscr{P}(J)} \mathbb{E}\left[\sum_{n=1}^{N}(F_n(\boldsymbol{\theta}_n) - F_n(\boldsymbol{\theta}_n^*))\right] \geq C\min\left\{J\left(\frac{d}{B}+1\right), N\right\}.$$

The infimum is taken over all online algorithms $\mathcal{A}$ for Problem 1, and $\{\boldsymbol{\theta}_n\}_{n=1}^N$ is the output of $\mathcal{A}$.

Comparing the upper bound in Theorem 1 and the matching lower bound in Theorem 5, we see that Algorithm 4 achieves the minimax optimal regret up to polylogarithmic factors for every J, adapting to the unknown nonstationarity.

From the stronger version of Theorem 5 in Section EC.3.1 in the electronic companion, we can easily derive a lower bound expressed using the path variation. The proof is deferred to Section EC.3.2.

Corollary 3 (PV-Based Lower Bound). *Assume $N \geq \max\{2, d/B\}$ and $V \leq N\min\{B/d, \sqrt{d/B}\}$. There is a universal constant $C > 0$ such that*

$$\inf_{\mathcal{A}} \sup_{(\mathcal{P}_1,\dots,\mathcal{P}_N)\in\mathscr{Q}(V)} \mathbb{E}\left[\sum_{n=1}^{N}(F_n(\boldsymbol{\theta}_n) - F_n(\boldsymbol{\theta}_n^*))\right] \geq C\left[1+\frac{d}{B}+N^{1/3}\left(\frac{Vd}{B}\right)^{2/3}\right].$$

The infimum is taken over all online algorithms $\mathcal{A}$ for Problem 1, and $\{\boldsymbol{\theta}_n\}_{n=1}^N$ is the output of $\mathcal{A}$.

When $V \leq N(d/B)^2$, we have $V \leq N^{1/3}(Vd/B)^{2/3}$, and the regret bound in Corollary 1 simplifies to $1+\min\{d/B, N\} + N^{1/3}(Vd/B)^{2/3}$. Therefore, Corollary 3 shows that Algorithm 4 adapts to the unknown path variation when $0 \leq V \leq N\min\{B/d, (d/B)^2\}$.

6.2. Lipschitz Population Losses

Finally, we present minimax lower bounds that match the regret bounds in Theorem 2 and Corollary 2 up to logarithmic factors. We consider a class of stochastic linear optimization problems in Example 6.

Definition 7 (Stochastic Linear Optimization). Define $B_\infty(\boldsymbol{x}, r) = \{\boldsymbol{y}\in\mathbb{R}^d : \|\boldsymbol{y}-\boldsymbol{x}\|_\infty \leq r\}$ for any $\boldsymbol{x}\in\mathbb{R}^d$ and $r \geq 0$. For any $\boldsymbol{\mu}\in B_\infty(\mathbf{0}, 1/2)$, denote by $\mathcal{P}(\boldsymbol{\mu})$ the distribution of $\boldsymbol{z} = \sqrt{d}\boldsymbol{x}\circ\boldsymbol{y}$, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, the entries $\{x_j\}_{j=1}^d$ of $\boldsymbol{x}$ are independent, $\mathbb{P}(x_j = \pm 1) = \frac{1}{2}\pm\mu_j$, $\boldsymbol{y}$ is uniformly distributed over $\{\boldsymbol{e}_j\}_{j=1}^d$, and $\circ$ denotes the entry-wise product. Let $\mathcal{Z} = \mathbb{R}^d$, $\Omega = B_\infty(\mathbf{0}, 1/\sqrt{d})$, $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \boldsymbol{z}^\top\boldsymbol{\theta}$, and $F_{\boldsymbol{\mu}}(\cdot) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}(\boldsymbol{\mu})}\ell(\cdot, \boldsymbol{z})$.

When $\boldsymbol{y} = \boldsymbol{e}_j$, $\ell(\boldsymbol{\theta}, \boldsymbol{z}) = \sqrt{d}\theta_j x_j$. We have $|\ell(\boldsymbol{\theta}, \boldsymbol{z})| \leq 1$ for all $\boldsymbol{\theta}\in\Omega$, and $\mathbb{E}(\boldsymbol{z}\boldsymbol{z}^\top) = \boldsymbol{I}_d$. Hence, $\mathcal{P}(\boldsymbol{\mu})$ satisfies the conditions in Example 6 with $\sigma_0 = 1$. Note that $\mathbb{E}\boldsymbol{z} = \boldsymbol{\mu}/\sqrt{d}$, $F_{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \boldsymbol{\mu}^\top\boldsymbol{\theta}/\sqrt{d}$ and $\|F_{\boldsymbol{\mu}} - F_{\boldsymbol{\nu}}\|_\infty = \|\boldsymbol{\mu}-\boldsymbol{\nu}\|_1/d$.

We now construct two classes of learning problems similar to those in Definition 6.

Definition 8 (Problem Classes). For $J\in[N-1]$, define the problem class

$$\mathscr{P}(J) = \Bigg\{(\mathcal{P}_1,\dots,\mathcal{P}_N) :\ \mathcal{P}_n = \mathcal{P}(\boldsymbol{\mu}_n^*),\ \boldsymbol{\mu}_n^*\in B_\infty(\mathbf{0}, 1/2),\ \forall n\in[N],\ \text{there exist } 0 = N_0 < \cdots < N_J = N-1 \text{ such that } \frac{1}{d}\sum_{n=N_{j-1}+1}^{N_j-1}\|\boldsymbol{\mu}_{n+1}^* - \boldsymbol{\mu}_n^*\|_1 \leq \sqrt{\frac{d}{B(N_j - N_{j-1})}},\ \forall j\in[J]\Bigg\}.$$

In addition, for any $V \geq 0$, define

$$\mathscr{Q}(V) = \left\{(\mathcal{P}_1,\dots,\mathcal{P}_N) :\ \mathcal{P}_n = \mathcal{P}(\boldsymbol{\mu}_n^*),\ \boldsymbol{\mu}_n^*\in B_\infty(\mathbf{0}, 1/2),\ \frac{1}{d}\sum_{n=1}^{N-1}\|\boldsymbol{\mu}_{n+1}^* - \boldsymbol{\mu}_n^*\|_1 \leq V\right\}.$$

For every problem instance in $\mathscr{P}(J)$ or $\mathcal{Q}(V)$, Assumption 4 holds with $\sigma = 4$ and $\lambda = 2$. The set $\mathscr{P}(J)$ corresponds to function sequences $\{F_n\}_{n=1}^N$ with at most J quasistationary segments, and $\mathcal{Q}(V)$ corresponds to function sequences $\{F_n\}_{n=1}^N$ with path variation at most V. We are now ready to present our lower bounds. See Section EC.3.3 in the electronic companion for the proof.

Theorem 6 (Lower Bound). *Assume $N \geq 2$ and that $J\in[N-1]$ divides $N-1$. There exists a universal constant $C > 0$ such that*

$$\inf_{\mathcal{A}} \sup_{(\mathcal{P}_1,\dots,\mathcal{P}_N)\in\mathscr{P}(J)} \mathbb{E}\left[\sum_{n=1}^{N}(F_n(\boldsymbol{\theta}_n) - \min_{\boldsymbol{\theta}_n'\in\Omega}F_n(\boldsymbol{\theta}_n'))\right] \geq C\min\left\{J + \sqrt{\frac{JNd}{B}}, N\right\}.$$

The infimum is taken over all online algorithms $\mathcal{A}$ for Problem 1, and $\{\boldsymbol{\theta}_n\}_{n=1}^N$ is the output of $\mathcal{A}$.

As the upper bound in Theorem 2 matches the lower bound in Theorem 6, we see that Algorithm 4 achieves the minimax optimal regret up to polylogarithmic factors for every J and thus adapts to the unknown nonstationarity.

Finally, we present a lower bound based on the path variation. The proof is given in Section EC.3.4 in the electronic companion.

Corollary 4 (PV-Based Lower Bound). *When* $N \geq \max\{2, d/B\}$ *and* $0 \leq V \leq N\min\{B/d, \sqrt{d/B}\}/6$, *it holds that*

$$\inf_{\mathcal{A}} \sup_{(\mathcal{P}_1,\ldots,\mathcal{P}_N)\in Q(V)} \mathbb{E}\left[\sum_{n=1}^{N}\left(F_n(\boldsymbol{\theta}_n) - \min_{\boldsymbol{\theta}'_n\in\Omega} F_n(\boldsymbol{\theta}'_n)\right)\right] \geq C\left[1 + \sqrt{\frac{Nd}{B}} + N^{2/3}\left(\frac{Vd}{B}\right)^{1/3}\right].$$

The infimum is taken over all online algorithms $\mathcal{A}$ for Problem 1, and $\{\boldsymbol{\theta}_n\}_{n=1}^N$ is the output of $\mathcal{A}$.

When $V \leq N\sqrt{d/B}$, we have $V \leq N^{2/3}(Vd/B)^{1/3}$, and the regret bound in Corollary 2 simplifies to $1 + \sqrt{Nd/B} + N^{2/3}(Vd/B)^{1/3}$. Therefore, Algorithm 4 adapts to the unknown path variation when $0 \leq V \leq N\min\{\sqrt{d/B}, B/d\}/6$.

7. Numerical Experiments

In this section, we test the practical performance of our algorithm SAWS (Algorithm 4) on synthetic and real data. To illustrate the adaptivity of our algorithm, we will compare it against fixed-window algorithms MA(k) that only use a fixed look-back window k in every period $n \in [N]$. The detailed description of MA(k) is given in Algorithm 5.

Algorithm 5 (Fixed-Window Moving Average MA(k))

Input: Window size k.

Choose any $\boldsymbol{\theta}_1 \in \Omega$.

For $n = 2, \ldots, N$**:**

Let $r = k \wedge (n-1)$, and compute a minimizer $\boldsymbol{\theta}_n$ of $f_{n,r} = \frac{1}{r}\sum_{i=n-r}^{n-1} f_i$.

Output: $\{\boldsymbol{\theta}_n\}_{n=1}^N$.

7.1. Synthetic Data

In the synthetic data experiment, we take one problem instance from the strongly convex case (Section 4.1) and one from the Lipschitz case (Section 4.2). For both instances, we consider time horizons $N \in \mathcal{N} = \{250, 500, 1{,}000, 2{,}000, 4{,}000, 8{,}000\}$. We will compare against benchmarks MA(k) with $k \in \{\lceil N^p\rceil : p = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1\}$.

7.1.1. Strongly Convex Instance.

We consider online linear regression (Example 2) under nonstationarity, with $d = 10$, $M = 12$, $\sigma_0 = 1$, $B = 1$, and $N \in \mathcal{N}$. In each period $n \in [N]$, a sample $(\boldsymbol{x}_n, y_n)$ is generated from $y_n = \boldsymbol{x}_n^\top \boldsymbol{\theta}_n^* + \varepsilon_n$, with $\boldsymbol{x}_n \sim N(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\varepsilon_n \sim N(0,1)$ independent. The minimizer sequence $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$ is piecewise constant and has the following pattern. Let $n_1 = 5\lceil N^{1/3}\rceil$, $n_2 = 5\lceil N^{1/6}\rceil$, and $n_3 = 5\lceil N^{1/2}\rceil$. The horizon is divided into segments of equal length $n_1 + n_2 + 2n_3$. Within each segment, in the n_1-th, $(n_1 + n_3)$-th, $(n_1 + n_3 + n_2)$-th, and $(n_1 + n_3 + n_2 + n_3)$-th periods, $\boldsymbol{\theta}_n^*$ switches to a point sampled uniformly at random from $B(\boldsymbol{0}, M/4) \subseteq \mathbb{R}^d$.

7.1.2. Lipschitz Instance.

We consider online stochastic linear optimization (Example 6) under nonstationarity, with $\Omega = \{\boldsymbol{\theta} \in \mathbb{R}_+^d : \|\boldsymbol{\theta}\|_1 \leq 1\}$, $d = 10$, $B = 1$, and $N \in \mathcal{N}$. For each $n \in [N]$, $\mathcal{P}_n = N(\boldsymbol{\mu}_n, \boldsymbol{I}_d)$. The sequence $\{\boldsymbol{\mu}_n\}_{n=1}^N$ is piecewise constant and has the following pattern. Let $n_1 = \lceil N^{1/2}\rceil$, $n_2 = \lceil N^{1/6}\rceil$, and $n_3 = \lceil N^{1/3}\rceil$. The horizon is divided into segments of equal length $n_1 + n_2 + 2n_3$. Within each segment, in the n_1-th, $(n_1 + n_3)$-th, $(n_1 + n_3 + n_2)$-th, and $(n_1 + n_3 + n_2 + n_3)$-th periods, $\boldsymbol{\mu}_n$ switches to a point generated by randomly picking half of the entries to be uniform over $\{-1, 1\}^{d/2}$ and the other half uniform over $[-1, 1]^{d/2}$.

We choose the thresholds $\{\tau(n,k)\}_{n\in\mathbb{Z}_+, k\in[n-1]}$ for SAWS according to Theorem 1 and Theorem 2. For the strongly convex instance, we take $\alpha = 0.1$ and $C_\tau = 0.3$. For the Lipschitz instance, we take $\alpha = 0.1$ and $C_\tau = 0.5$.

In Figure 3, we present the log-log plots for the dynamic regrets of SAWS and the benchmarks MA(k). The curves and error bands show the means and 1.96 times the standard errors over 50 random seeds, respectively. The latter gives 95% confidence intervals for the mean dynamic regrets of the methods, which have small widths.

In both instances, SAWS consistently outperforms the fixed-window benchmarks. The slopes of its curves are generally smaller than those of the benchmarks, indicating smaller orders of dynamic regrets. This demonstrates the adaptivity of SAWS to unknown nonstationarity.

7.2. Real Data: Electricity Demand Prediction

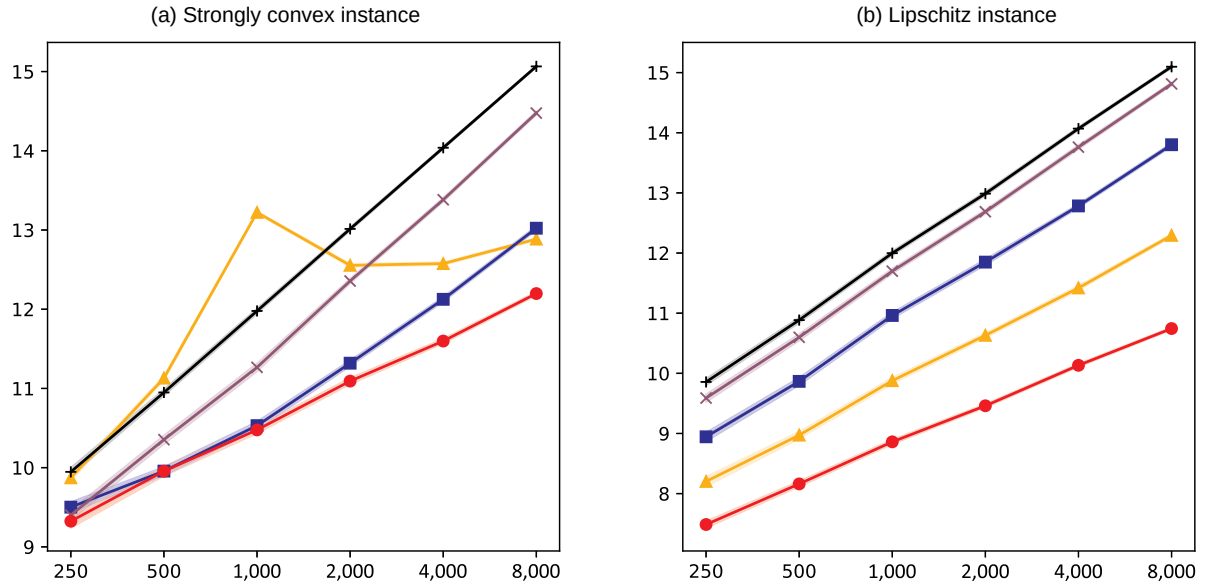
Our first real data experiment uses a electricity demand data set maintained by the Australian Bureau of Meteorology and collected by Kozlov (2020). We study the daily electricity demand in Victoria, Australia, from January 1, 2016 to October 6, 2020. In Figure EC.1 of Section EC.5 in the electronic companion, we plot the pattern of the electricity demand over time.

Our task is to use linear regression (Example 2) to predict the daily electricity demand y_n given features $\boldsymbol{x}_n$ on the same day, including minimum and maximum temperatures, rainfall, and solar exposure. Along with an additional intercept term, this yields a feature vector of length $d = 5$. Each day is treated as a time period, so there are $N = 1{,}760$ periods in total. We consider the setting where $M = \mathrm{diam}(\Omega)$ is large, and for simplicity, we will set $\Omega = \mathbb{R}^d$. We set $C_\tau = 10$ in SAWS and will compare it with MA(k), $k \in \{1, 7, 14, 30, 180, 365, 1{,}826\}$.

In Figure 4(a), we plot the per-period losses of SAWS and MA(k), given by $\frac{1}{N}\sum_{n=1}^N \left[\frac{1}{2}(y_n - \boldsymbol{x}_n^\top \boldsymbol{\theta}_n)^2\right]$. Among the fixed-window benchmarks MA(k) considered, the optimal fixed window is $k^* = 30$ days. We see that the performance of SAWS is comparable to that of MA(30).

In Figure 5(a), we visualize the rolling window picked by SAWS. We observe that SAWS adaptively

Figure 3. (Color online) Log-Log Plots of Dynamic Regrets of SAWS and Fixed-Window Benchmarks on Synthetic Data



Notes. Horizontal axis: time horizon $N \in \mathcal{N}$. Vertical axis: logarithm of dynamic regret $\log_2\{\sum_{n=1}^{N}[F_n(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}' \in \Omega} F_n(\theta')]\}$. Red circles: SAWS (Algorithm 4). Orange triangles: MA($\lceil N^{1/3} \rceil$). Blue squares: MA($\lceil N^{1/2} \rceil$). Purple ×'s: MA($\lceil N^{2/3} \rceil$). Black +'s: MA(N).

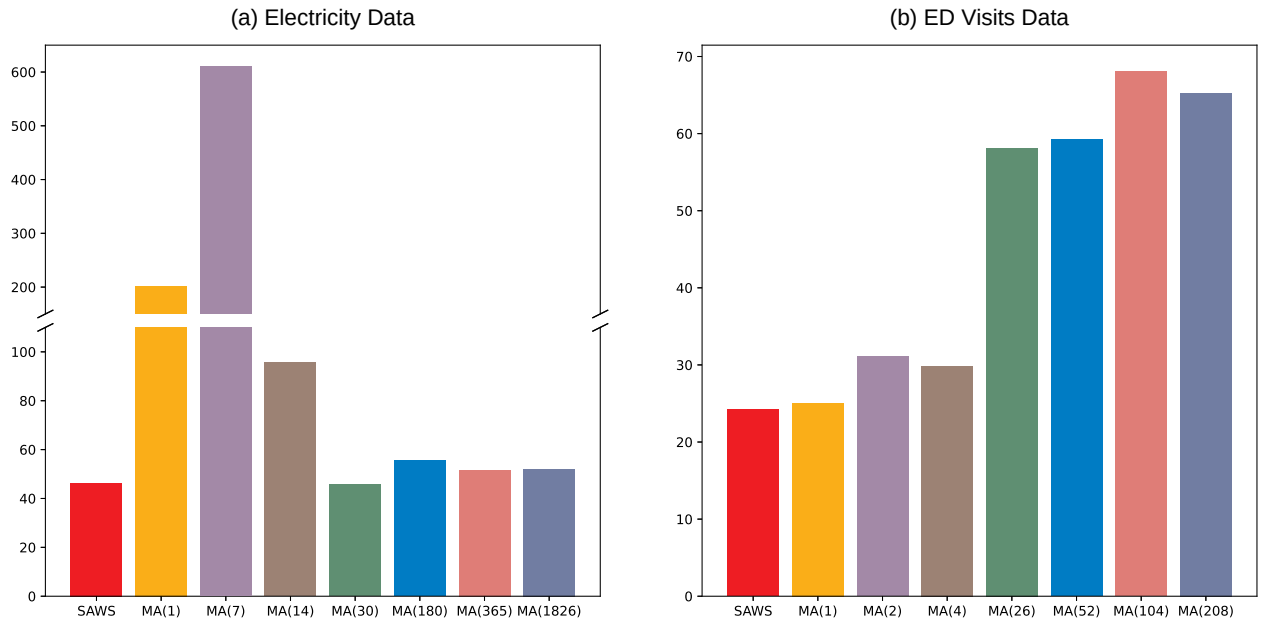
selects rolling windows, which roughly align with the nonstationarity pattern in Figure EC.1 in the electronic companion.

7.3. Real Data: Hospital Nurse Staffing

Finally, we test our method on an emergency department (ED) visits data set maintained by the New York City (NYC) government (NYC Health 2024). The data set contains daily and weekly ED visit counts over time in NYC for various syndromes. In Figure EC.2 of Section EC.5 in the electronic companion, we plot the weekly ED visit counts for vomiting from January 7, 2019 to December 31, 2023.

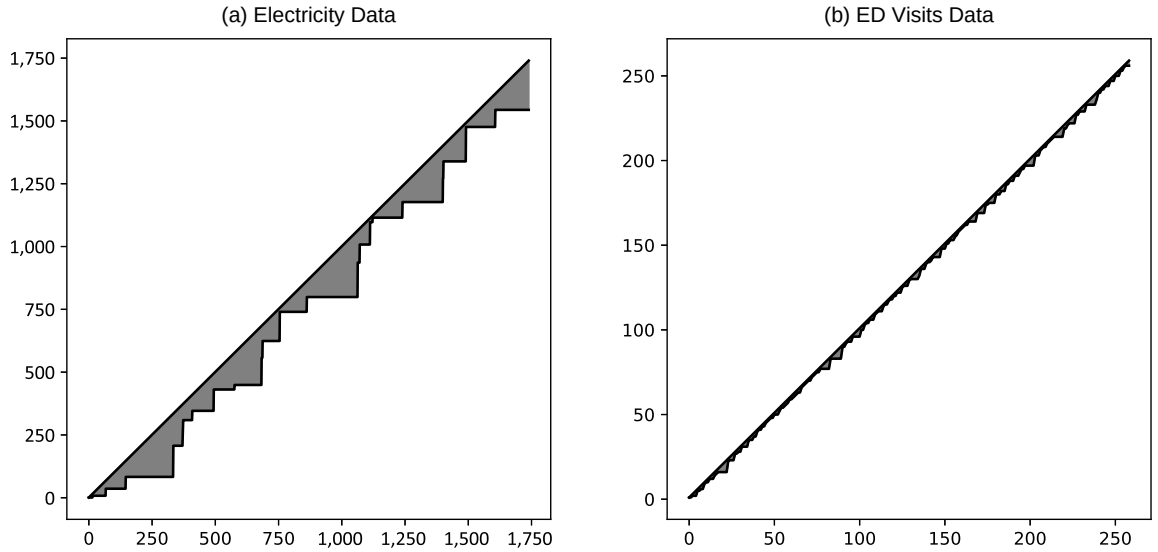
We study the problem of nurse staffing for this date range, where the goal is to decide the appropriate number of nurses to schedule each week. Following Keskin et al. (2023), we formulate it as a newsvendor problem (Example 8), take the weekly demand for

Figure 4. (Color online) Per-Period Losses of SAWS and Fixed-Window Benchmarks on Real Data



Notes. Horizontal axis: algorithms. Vertical axis: per-period loss. For the electricity data, the predicted and true demand (unit: megawatt-hour) is scaled by 5×10^{-4}.

Figure 5. Rolling Windows Selected by SAWS on Real Data



Notes. Horizontal axis: time period n. Vertical axis: endpoints of look-back windows. Lower black curve: left endpoints. Upper black curve: right endpoints $(n-1)$.

nurse staffing to be the weekly patient visits divided by three, and set the critical ratio as $b/(b+h) = 0.7$. For simplicity, we take $b = 0.7$ and $h = 0.3$, and set $\Omega = \mathbb{R}$. We take $C_\tau = 5$ in SAWS and compare it with MA(k), $k \in \{1, 2, 4, 26, 52, 104, 208\}$.

In Figure 4(b), we plot the per-period losses $\frac{1}{N}\sum_{n=1}^{N}[h(\theta_n - z_n)_+ + b(z_n - \theta_n)_+]$ of SAWS and MA(k). In Figure 5(b), we also visualize the rolling windows selected by SAWS. We observe that by adaptively varying the window size, SAWS achieves a lower loss than all fixed-window benchmarks considered.

7.4. Summary of Experiments

In our synthetic and real data experiments, the problem instances exhibit different patterns of nonstationarity, which lead to different optimal windows. In practice, as the nonstationarity pattern is generally unknown beforehand, it is not clear a priori what the best window should be or even what candidate windows to choose from. Our experiments show that without any prior knowledge of the nonstationarity, SAWS *adaptively* selects look-back windows for learning and achieves performance comparable to or even better than the best fixed-window benchmark *in hindsight*.

8. Discussions

Based on a stability principle, we developed an adaptive approach to learning under unknown nonstationarity. Our algorithm attains optimal dynamic regrets in common problems. As by-products of our analysis, we develop a novel measure of function similarity and a segmentation technique.

A number of future directions are worth pursuing. First, we do not assume any structure of the underlying nonstationarity. In practice, some prior knowledge or forecast of the dynamics is available. Incorporating them into our method may further boost its performance. Second, the threshold sequence in our algorithm relies on knowledge of the function class, smoothness parameters, and noise levels. It would be interesting to develop adaptive thresholds for handling these parameters. Third, it is also worth investigating whether our approach enjoys good theoretical guarantees with respect to other performance measures, such as the strongly adaptive regret (Daniely et al. 2015). Finally, an important future direction is to extend our framework to sequential decision-making problems with partial feedback, including bandit problems and reinforcement learning, where the learner only receives feedback on the chosen decisions. This requires understanding the interplay between the nonstationarity and the exploration-exploitation trade-off.

9. Code and Data Disclosure

The code and data to support the numerical experiments in this paper can be found at https://github.com/ch3702/SAWS.

Acknowledgments

The authors are grateful to the anonymous referees, the associate editor, and the area editor for their insightful comments.

Appendix. Proof Sketches for Main Theorems

In the appendices, we provide proof sketches for the main results, namely, Theorem 3 (excess risk bound in a specific time period), Theorem 4 (general regret bound), and Theorem 1 (regret bound in the strongly convex case). The proof sketch for Theorem 2 (regret bound in the Lipschitz

case) parallels that of Theorem 1 and is thus omitted. For ease of exposition, we will analyze the simpler Algorithms 1 and 2 instead of their more complicated counterparts Algorithms 3 and 4.

The key property we will use is: if f and g are (ε,δ)-close, then for all, $\boldsymbol{\theta}\in\Omega$ and $R\geq 0$,

$$\begin{aligned} & f(\boldsymbol{\theta})-\inf_{\boldsymbol{\theta}'\in\Omega} f(\boldsymbol{\theta}')\lesssim R \\ \Rightarrow\quad & g(\boldsymbol{\theta})-\inf_{\boldsymbol{\theta}'\in\Omega} g(\boldsymbol{\theta}')\lesssim R+\delta. \end{aligned}$$

Appendix A. A Proof Sketch for Theorem 3

The full proof is given in Section EC.1.3 in the electronic companion and uses some ideas from Mathé (2006). Recall that

$$\begin{aligned} \bar{k}=\max\{k\in[n-1]: & F_{n-k},F_{n-k+1}\ldots,F_{n-1} \text{ are} \\ & (\varepsilon,\psi(n,k))\text{-close to } F_{n-1}\} \end{aligned}$$

and

$$\begin{aligned} \widehat{k}=\max\Big\{ & k\in[n-1]:\ \forall i\in[k], \\ & f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,k})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,i}(\boldsymbol{\theta})\leq\tau(n,i)\Big\}. \end{aligned}$$

We will first prove that $\widehat{k}\geq\bar{k}$. To this end, it suffices to show that for all $i\in[\bar{k}]$, $f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,i}(\boldsymbol{\theta})\leq\tau(n,i)$. For all $i\in[\bar{k}]$, because $F_{n-i},\ldots,F_{n-1}$ are $(\varepsilon,\psi(n,\bar{k}))$-close to F_{n-1} and $\psi(n,\bar{k})\leq\psi(n,i)$, then by part 7 of Lemma 4, $F_{n,i}$ is $(\varepsilon,c\psi(n,i))$-close to F_{n-1}, with $c=e^{\varepsilon}+1$. Then, for all $i\in[\bar{k}]$,

$$\begin{aligned} & f_{n,\bar{k}}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,\bar{k}}(\boldsymbol{\theta})=0 \\ \Rightarrow\quad & F_{n,\bar{k}}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n,\bar{k}}(\boldsymbol{\theta})\lesssim\psi(n,\bar{k}) \\ & \quad (f_{n,\bar{k}} \text{ and } F_{n,\bar{k}} \text{ are } (\varepsilon,\psi(n,\bar{k}))\text{-close}) \\ \Rightarrow\quad & F_{n-1}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n-1}(\boldsymbol{\theta})\lesssim\psi(n,\bar{k}) \\ & \quad (F_{n,\bar{k}} \text{ and } F_{n-1} \text{ are } (\varepsilon,c\psi(n,\bar{k}))\text{-close}) \\ \Rightarrow\quad & F_{n,i}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n,i}(\boldsymbol{\theta})\lesssim\psi(n,\bar{k})+\psi(n,i) \\ & \qquad\qquad\lesssim\psi(n,i) \\ & \quad (F_{n,i} \text{ and } F_{n-1} \text{ are } (\varepsilon,c\psi(n,i))\text{-close}) \\ \Rightarrow\quad & f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,i}(\boldsymbol{\theta})\lesssim\psi(n,i). \\ & \quad (f_{n,i} \text{ and } F_{n,i} \text{ are } (\varepsilon,\psi(n,i))\text{-close}). \end{aligned}$$

The condition $\tau(n,k)\geq 6e^{5\varepsilon}\psi(n,k)$ in Condition 1 is used to ensure that the last inequality above implies $f_{n,i}(\widehat{\boldsymbol{\theta}}_{n,\bar{k}})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,i}(\boldsymbol{\theta})\leq\tau(n,i)$. This shows that $\widehat{k}\geq\bar{k}$.

Because $\widehat{k}\geq\bar{k}$, then by the definition of $\widehat{k}$,

$$\begin{aligned} & f_{n,\bar{k}}(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}})-\inf_{\boldsymbol{\theta}\in\Omega} f_{n,\bar{k}}(\boldsymbol{\theta})\leq\tau(n,\bar{k}) \\ \Rightarrow\quad & F_{n,\bar{k}}(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n,\bar{k}}(\boldsymbol{\theta})\lesssim\tau(n,\bar{k})+\psi(n,\bar{k}) \\ & \qquad\qquad\lesssim\tau(n,\bar{k}) \\ & \quad (f_{n,\bar{k}} \text{ and } F_{n,\bar{k}} \text{ are } (\varepsilon,\psi(n,\bar{k}))\text{-close}) \\ \Rightarrow\quad & F_{n-1}(\widehat{\boldsymbol{\theta}}_{n,\widehat{k}})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n-1}(\boldsymbol{\theta})\lesssim\tau(n,\bar{k})+\psi(n,\bar{k}) \\ & \qquad\qquad\lesssim\tau(n,\bar{k}). \\ & \quad (F_{n,\bar{k}} \text{ and } F_{n-1} \text{ are } (\varepsilon,c\psi(n,\bar{k}))\text{-close}) \end{aligned}$$

As $\boldsymbol{\theta}_n=\widehat{\boldsymbol{\theta}}_{n,\widehat{k}}$, this finishes the proof.

Appendix B. Proof Sketch for Theorem 4

For clarity, we add a time index to the quantity $\bar{k}$ defined in Theorem 3; that is,

$$\begin{aligned} \bar{k}_{n-1}=\max\{k\in[n-1]:\ & F_{n-k},F_{n-k+1}\ldots,F_{n-1} \\ & \text{are } (\varepsilon,\psi(n,k))\text{-close to } F_{n-1}\}. \end{aligned}$$

By Definition 5, if $n\in\{N_{j-1}+1,\ldots,N_j\}$, then $F_{N_{j-1}+1},\ldots,F_n$ are $(\varepsilon,\psi(n,n-N_{j-1}))$-close to F_n, so $\bar{k}_n\geq n-N_{j-1}$. By Theorem 3 and Condition 1,

$$\begin{aligned} & F_n(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_n(\boldsymbol{\theta})\lesssim\tau(n+1,\bar{k}_n) \\ & \lesssim\tau(N,\bar{k}_n)\lesssim\tau(N,n-N_{j-1}). \end{aligned}$$

We now convert this into a bound for $F_{n+1}(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n+1}(\boldsymbol{\theta})$. There are two cases.

- If $n\leq N_j-1$, then by Definition 5, F_n and F_{n+1} are $(\varepsilon,\psi(n+1,n-N_{j-1}+1))$-close, so

$$\begin{aligned} & F_{n+1}(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n+1}(\boldsymbol{\theta}) \\ & \lesssim\tau(N,n-N_{j-1})+\psi(n+1,n-N_{j-1}+1) \\ & \lesssim\tau(N,n-N_{j-1})+\tau(n+1,\bar{k}_n) \\ & \lesssim\tau(N,n-N_{j-1}). \end{aligned}$$

- If $n=N_j$, then by Definition 5, $F_n=F_{N_j}$ and $F_{n+1}=F_{N_j+1}$ are (ε,δ_j)-close, so

$$F_{n+1}(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n+1}(\boldsymbol{\theta})\lesssim\tau(N,n-N_{j-1})+\delta_j.$$

Moreover, $F_{n+1}(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n+1}(\boldsymbol{\theta})\leq U$. Therefore,

$$\begin{aligned} & \sum_{n=2}^{N}\left[F_n(\boldsymbol{\theta}_n)-\inf_{\boldsymbol{\theta}\in\Omega} F_n(\boldsymbol{\theta})\right] \\ & =\sum_{j=1}^{J}\sum_{n=N_{j-1}+1}^{N_j}\left[F_{n+1}(\boldsymbol{\theta}_{n+1})-\inf_{\boldsymbol{\theta}\in\Omega} F_{n+1}(\boldsymbol{\theta})\right] \\ & \lesssim\sum_{j=1}^{J}\left(\sum_{n=N_{j-1}+1}^{N_j}\min\{\tau(N,n-N_{j-1}),U\}+\delta_j\right) \\ & =\sum_{j=1}^{J}T(N_j-N_{j-1})+\sum_{j=1}^{J}\delta_j. \end{aligned}$$

Adding the term $F_1(\boldsymbol{\theta}_1)-\inf_{\boldsymbol{\theta}\in\Omega} F_1(\boldsymbol{\theta})$ to both sides finishes the proof.

Appendix C. Proof Sketch for Theorem 1

For notational convenience, we will drop the subscript of J_N. We will prove the following more refined bound: for every segmentation of $\{\boldsymbol{\theta}_n^*\}_{n=1}^N$, it holds that

$$\begin{aligned} & \sum_{n=1}^{N}\left[F_n(\boldsymbol{\theta}_n)-F_n(\boldsymbol{\theta}_n^*)\right] \\ & \lesssim 1+\sum_{j=1}^{J}\min\left\{\frac{d}{B},\ N_j-N_{j-1}\right\} \\ & \quad +\sum_{j=1}^{J}\|\boldsymbol{\theta}_{N_j+1}^*-\boldsymbol{\theta}_{N_j}^*\|_2^2. \end{aligned} \tag{C.1}$$

Then, (4) follows from

$$\sum_{j=1}^{J} \min\left\{\frac{d}{B},\ N_j - N_{j-1}\right\} \le \min\left\{\frac{Jd}{B}, N\right\},$$
$$\sum_{j=1}^{J} \|\boldsymbol{\theta}^*_{N_j+1} - \boldsymbol{\theta}^*_{N_j}\|_2^2 \le JM^2.$$

To prove (C.1), we will verify that in the strongly convex case, the segmentation in Definition 1 translates to Definition 5, and thus, we can apply Theorem 4. By a concentration bound for subexponential random variables (Lemma EC.4 in the electronic companion), with high probability, up to logarithmic factors,

$$\sup_{\boldsymbol{\theta}\in\Omega} \|\nabla f_{n,k}(\boldsymbol{\theta}) - \nabla F_{n,k}(\boldsymbol{\theta})\|_2 \lesssim \max\left\{\sqrt{\frac{d}{Bk}}, \frac{d}{Bk}\right\}.$$

Because $F_{n,k}$ is strongly convex, then substituting the inequality above into part 3 of Lemma 3 shows that $f_{n,k}$ and $F_{n,k}$ are $(\log 2, \eta)$-close with $\eta \asymp \frac{d}{Bk}$. Thus, we will take

$$\psi(n,k) \asymp \frac{d}{Bk}.$$

Moreover, by part 4 of Lemma 3, F_n and F_i are $(\log(4L/\rho), \frac{\rho}{2}\|\boldsymbol{\theta}^*_n - \boldsymbol{\theta}^*_i\|_2^2)$-close.

Let $\{\boldsymbol{\theta}^*_n\}_{n=1}^N$ be segmented as in Definition 1. Then, for every $j \in [J]$,

$$\max_{N_{j-1}<i,k\le N_j} \|\boldsymbol{\theta}^*_i - \boldsymbol{\theta}^*_k\|_2^2 \lesssim \frac{d}{B(N_j - N_{j-1})} \asymp \min_{N_{j-1}<n\le N_j} \psi(n, n - N_{j-1}),$$

and thus, $F_{N_{j-1}+1}, \dots, F_{N_j}$ are $(\log(4L/\rho), \min_{N_{j-1}<n\le N_j} \psi(n, n - N_{j-1}))$-close. In addition, F_{N_j} and F_{N_j+1} are $(\log(4L/\rho), \delta_j)$-close with $\delta_j = \frac{\rho}{2}\|\boldsymbol{\theta}^*_{N_j+1} - \boldsymbol{\theta}^*_{N_j}\|_2^2$. This shows that the segmentation in Definition 1 is also a segmentation in the sense of Definition 5. Therefore, Theorem 4 is applicable and yields

$$\begin{aligned}
&\sum_{n=1}^{N} [F_n(\boldsymbol{\theta}_n) - \inf_{\boldsymbol{\theta}'_n\in\Omega} F_n(\boldsymbol{\theta}'_n)] \\
&\lesssim \left[F_1(\boldsymbol{\theta}_1) - \inf_{\boldsymbol{\theta}\in\Omega} F_1(\boldsymbol{\theta})\right] + \sum_{j=1}^{J} T(N_j - N_{j-1}) + \sum_{j=1}^{J} \delta_j \\
&\lesssim 1 + \sum_{j=1}^{J} \sum_{n=N_{j-1}+1}^{N_j} \min\{\tau(N, n - N_{j-1}), 1\} \\
&\quad + \sum_{j=1}^{J} \|\boldsymbol{\theta}^*_{N_j+1} - \boldsymbol{\theta}^*_{N_j}\|_2^2 \\
&\lesssim 1 + \sum_{j=1}^{J} \sum_{n=N_{j-1}+1}^{N_j} \min\left\{\frac{d}{B(n - N_{j-1})}, 1\right\} \\
&\quad + \sum_{j=1}^{J} \|\boldsymbol{\theta}^*_{N_j+1} - \boldsymbol{\theta}^*_{N_j}\|_2^2 \\
&\lesssim 1 + \sum_{j=1}^{J} \min\left\{\frac{d}{B}, N_j - N_{j-1}\right\} + \sum_{j=1}^{J} \|\boldsymbol{\theta}^*_{N_j+1} - \boldsymbol{\theta}^*_{N_j}\|_2^2.
\end{aligned}$$

This completes the proof.

References

Auer P, Gajane P, Ortner R (2019) Adaptively tracking the best bandit arm with an unknown number of distribution changes. Beygelzimer A, Hsu D, eds. *Proc. 32nd Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 99 (PMLR, New York), 138–158.

Baby D, Wang YX (2019) Online forecasting of total-variation-bounded sequences. Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. *NIPS'18: Proc. 32nd Internat. Conf. Neural Inform. Processing Systems*, vol. 32 (Curran Associates, Inc., Red Hook, NY), 11071–11081.

Baby D, Wang YX (2021) Optimal dynamic regret in exp-concave online learning. Belkin M, Kpotufe S, eds. *Proc. 34th Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 134 (PMLR, New York), 359–409.

Baby D, Wang YX (2022) Optimal dynamic regret in proper online learning with strongly convex losses and beyond. Camps-Valls G, Ruiz FJR, Valera I, eds. *Proc. 25th Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 151 (PMLR, New York), 1805–1845.

Bai Y, Zhang YJ, Zhao P, Sugiyama M, Zhou ZH (2022) Adapting to online label shift with provable guarantees. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds. *NIPS'22: Proc. 36th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 29960–29974.

Besbes O, Gur Y, Zeevi A (2015) Non-stationary stochastic optimization. *Oper. Res.* 63(5):1227–1244.

Bilodeau B, Negrea J, Roy DM (2023) Relaxing the i.i.d. assumption: Adaptively minimax optimal regret via root-entropic regularization. *Ann. Statist.* 51(4):1850–1876.

Chen Q, Golrezaei N, Bouneffouf D (2023) Non-stationary bandits with auto-regressive temporal dependency. Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. *NIPS'23: Proc. 37th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 7895–7929.

Chen X, Wang Y, Wang YX (2019a) Technical note—Nonstationary stochastic optimization under $L_{p,q}$-variation measures. *Oper. Res.* 67(6):1752–1765.

Chen N, Wang C, Wang L (2025) Learning and optimization with seasonal patterns. *Oper. Res.* 73(2):894–909.

Chen Y, Lee CW, Luo H, Wei CY (2019b) A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. Beygelzimer A, Hsu D, eds. *Proc. 32nd Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 99 (PMLR, New York), 696–726.

Cheung WC, Simchi-Levi D, Zhu R (2022) Hedging the drift: Learning to optimize under nonstationarity. *Management Sci.* 68(3): 1696–1713.

Clements MP, Hendry DF (2001) *Forecasting Non-Stationary Economic Time Series* (MIT Press, Cambridge, MA).

Cutler J, Drusvyatskiy D, Harchaoui Z (2023) Stochastic optimization under distributional drift. *J. Machine Learn. Res.* 24(147): 1–56.

Dacco R, Satchell S (1999) Why do regime-switching models forecast so badly? *J. Forecasting* 18(1):1–16.

Daniely A, Gonen A, Shalev-Shwartz S (2015) Strongly adaptive online learning. Bach F, Blei D, eds. *Proc. 32nd Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 37 (PMLR, Lille, France), 1405–1411.

Fahrbach M, Javanmard A, Mirrokni V, Worah P (2023) Learning rate schedules in the presence of distribution shift. Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, eds. *Proc. 40th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 202 (PMLR, New York), 9523–9546.

Fan J, Yao Q (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer Series in Statistics, vol. 20 (Springer, New York).

Foussoul A, Goyal V, Gupta V (2023) MNL-bandit in non-stationary environments. Preprint, submitted March 4, https://arxiv.org/abs/2303.02504.

Hall E, Willett R (2013) Dynamical models and tracking regret in online convex programming. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 28 (PMLR, Atlanta), 579–587.

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2):357–384.

Hanneke S, Kanade V, Yang L (2015) Learning with a drifting target concept. Chaudhuri K, Gentile C, Zilles S, eds. *Algorithmic Learning Theory, ALT 2015*, Lecture Notes in Computer Science, vol. 9355 (Springer, Cham, Switzerland), 149–164.

Hazan E (2016) Introduction to online convex optimization. *Foundations Trends Optim.* 2(3–4):157–325.

Hazan E, Seshadhri C (2009) Efficient learning algorithms for changing environments. *ICML '09: Proc. 26th Annual Internat. Conf. Machine Learn.* (Association for Computing Machinery, New York), 393–400.

Huber PJ (1964) Robust estimation of a location parameter. *Ann. Math. Statist.* 35(1):73–101.

Jadbabaie A, Rakhlin A, Shahrampour S, Sridharan K (2015) Online optimization: Competing with dynamic comparators. Lebanon G, Vishwanathan SVN, eds. *Proc. 18th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 38 (PMLR, San Diego), 398–406.

Jia S, Xie Q, Kallus N, Frazier PI (2023) Smooth non-stationary bandits. Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, eds. *Proc. 40th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 202 (PMLR, New York), 14930–14944.

Jiang J, Li X, Zhang J (2025) Online stochastic optimization with Wasserstein-based nonstationarity. *Management Sci.*, ePub ahead of print March 3, https://doi.org/10.1287/mnsc.2020.03850.

Kalman RE (1960) A new approach to linear filtering and prediction problems. *J. Basic Engrg.* 82(1):35–45.

Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Math. Oper. Res.* 42(2):277–307.

Keskin NB, Min X, Song JSJ (2023) The nonstationary newsvendor: Data-driven nonparametric learning. Preprint, submitted June 15, 2021, https://doi.org/10.2139/ssrn.3866171.

Kozlov A (2020) Daily electricity price and demand data. Accessed August 30, 2024, https://www.kaggle.com/dsv/1596730.

Lepskii O (1991) On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* 35(3):454–466.

Liu Y, Van Roy B, Xu K (2023) Nonstationary bandit learning via predictive sampling. Ruiz F, Dy J, van de Meent JW, eds. *Proc. 26th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 206 (PMLR, New York), 6215–6244.

Luo H, Wei CY, Agarwal A, Langford J (2018) Efficient contextual bandits in non-stationary worlds. Bubeck S, Perchet V, Rigollet P, eds. *Proc. 31st Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 75 (PMLR, New York), 1739–1776.

Mania H, Jadbabaie A, Shah D, Sra S (2022) Time varying regression with hidden linear dynamics. Firoozi R, Mehr N, Yel E, Antonova R, Bohg J, Schwager M, Kochenderfer M, eds. *Proc. 4th Annual Learn. Dynamics Control Conf.*, Proceedings of Machine Learning Research, vol. 168 (PMLR, New York), 858–869.

Mathé P (2006) The Lepskii principle revisited. *Inverse Problems* 22(3):L11.

Mazzetto A, Upfal E (2023) An adaptive algorithm for learning with unknown distribution drift. Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, eds. *NIPS'23: Proc. 37th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 10068–10087.

Milly PC, Betancourt J, Falkenmark M, Hirsch RM, Kundzewicz ZW, Lettenmaier DP, Stouffer RJ (2008) Stationarity is dead: Whither water management? *Science* 319(5863):573–574.

Min S, Russo D (2023) An information-theoretic analysis of nonstationary bandit learning. Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, eds. *Proc. 40th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 202 (PMLR, New York), 24831–24849.

Mohri M, Muñoz Medina A (2012) New analysis and algorithm for learning with drifting distributions. Bshouty NH, Stoltz G, Vayatis N, Zeugmann T, eds. *Algorithmic Learning Theory, ALT 2012*, Lecture Notes in Computer Science, vol. 7568 (Springer, Berlin, Heidelberg), 124–138.

Mokhtari A, Shahrampour S, Jadbabaie A, Ribeiro A (2016) Online optimization in dynamic environments: Improved regret rates for strongly convex problems. *2016 IEEE 55th Conf. Decision Control (CDC)* (IEEE Press, Piscataway, NJ), 7195–7201.

Nestor B, McDermott MB, Boag W, Berner G, Naumann T, Hughes MC, Goldenberg A, Ghassemi M (2019) Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *Machine Learn. Healthcare Conf.* (PMLR, New York), 381–405.

Niu YS, Hao N, Zhang H (2016) Multiple change-point detection: A selective overview. *Statist. Sci.* 31(4):611–623.

NYC Health (2024) Syndromic surveillance data. Accessed August 30, 2024, https://a816-health.nyc.gov/hdi/epiquery/visualizations?PageType=ps&PopulationSource=Syndromic.

Slivkins A, Upfal E (2008) Adapting to a changing environment: The Brownian restless bandits. *Proc. 21st Conf. Learn. Theory* (Omnipress, Madison, WI), 343–354.

Spokoiny V (2009) Multiscale local change point detection with applications to value-at-risk. *Ann. Statist.* 37(3):1405–1436.

Suk J, Kpotufe S (2022) Tracking most significant arm switches in bandits. Loh PL, Raginsky M, eds. *Proc. 35th Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 278 (PMLR, New York), 2160–2182.

van de Geer S (2000) *Empirical Processes in M-Estimation*, vol. 6 (Cambridge University Press, Cambridge, UK).

Wang Y (2025) Technical note—On adaptivity in nonstationary stochastic optimization with bandit feedback. *Oper. Res.* 73(2):819–828.

Wei CY, Luo H (2021) Non-stationary reinforcement learning without prior knowledge: an optimal black-box approach. Belkin M, Kpotufe S, eds. *Proc. 32nd Conf. Learn. Theory*, Proceedings of Machine Learning Research, vol. 134 (PMLR, New York), 4300–4354.

Yang T, Zhang L, Jin R, Yi J (2016) Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. Balcan MF, Weinberger KQ, eds. *Proc. 33rd Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 48 (PMLR, New York), 449–457.

Zhang L, Lu S, Zhou ZH (2018) Adaptive online learning in dynamic environments. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *NIPS'18: Proc. 32nd Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Inc., Red Hook, NY), 1330–1340.

Zhao P, Zhang L (2021) Improved analysis for dynamic regret of strongly convex and smooth functions. *Proc. 3rd Conf. Learn. Dynamics Control*, Proceedings of Machine Learning Research, vol. 144 (PMLR, New York), 48–59.

Zhao Z, Jiang F, Yu Y, Chen X (2023) High-dimensional dynamic pricing under non-stationarity: Learning and earning with change-point detection. Preprint, submitted March 14, https://arxiv.org/abs/2303.07570.

Zhao P, Zhang YJ, Zhang L, Zhou ZH (2024) Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *J. Machine Learn. Res.* 25(98):1–52.

Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *ICML'03: Proc. 20th Internat. Conf. Machine Learn.* (AAAI Press, Washington, DC), 928–935.

Chengpiao Huang is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on statistical machine learning and data-driven decision-making.

Kaizheng Wang is an assistant professor in the Department of Industrial Engineering and Operations Research and a member of the Data Science Institute at Columbia University. He works at the intersection of machine learning, statistics, and optimization, with a particular focus on learning from heterogeneous data.