Adaptive Algorithms for Online Optimization Problems in Operations

by

Sen Yang

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Stern School of Business

New York University

August, 2025

Professor Jiawei Zhang

Dedication

To Everyone I Love

Acknowledgments

I am deeply grateful to Professors Jiawei Zhang and Divya Singhvi. Their guidance, patience, and encouragement shaped every stage of this work—from the earliest kernel of an idea to the final draft. Jiawei's enthusiasm for new problems and willingness to connect me with the broader research community constantly pushed my thinking forward, while Divya's steady presence and clear advice turned countless stumbling blocks into manageable tasks. Working with them has been the defining privilege of my doctoral years.

I have also benefited immensely from the wisdom of many other faculty members at NYU Stern. Conversations and courses with Professors Zhengyuan Zhou, Joshua Reed, Wenqiang Xiao, Michael Pinedo, Ilan Lobel, Srikanth Jagabathula, Xi Chen, Foster Provost, Elynn Chen, and Mor Armony broadened my perspective and sharpened my judgment. Their questions in seminars, feedback on drafts, and chance hallway discussions found their way, sometimes unexpectedly, into the pages that follow.

Several chapters of this dissertation were built side by side with Jinzhi Bu and Siyi Wang. Our long whiteboard sessions, shared code, and inevitable late-night coffee runs made difficult problems enjoyable and good ideas even better. I could not have asked for better collaborators.

Daily life in the program was sustained by friends who understood its ups and downs. Shixin Wang, Ziran Liu, Haotian Song, Jiashuo Jiang, Xinyi Zhao, Zhuoyi Yang, Ozgecan Gumusbas, Sandeep Chitla, Richard Bryant, Xinyu Zhang, Xiaole Liu, Nipun Thakurele, Xiaoyu Fan, and Zijian Liu turned heavy course loads into shared adventures and setbacks into shared jokes. Their

companionship anchored these years and reminded me that research, at its best, is a communal endeavor.

To everyone named here—and to the many others who offered a reference, a theorem, or simply a listening ear—thank you. Your generosity and example have made me a better researcher and a happier person.

Abstract

This thesis develops a unified framework for online convex optimization in which the pace of environmental change is unknown and potentially time-varying. The central question is how to craft gradient-based algorithms that sense and adapt to drift without any advance estimate of its magnitude. Chapters 1 and 2 introduce two meta-algorithms that address this challenge. The first assumes access to both noisy cost values and noisy gradients; it runs several stochastic-gradient trajectories in parallel, each with a different step size, and restarts with a finer step whenever the cumulative cost of a more conservative trajectory falls behind. The second algorithm relies only on noisy gradients; assuming mild curvature near the optimum, it monitors the distance between parallel iterates and tightens its learning rate as soon as those paths diverge. Both variants automatically match their step size to the unknown level of non-stationarity and achieve regret that is provably near the best possible. The analysis is further extended to a broad family of mixed-norm variation measures, capturing more intricate temporal and spatial patterns of change.

Chapter 3 applies the methodology to inventory control with drifting demand. We design an adaptive stochastic-gradient policy that updates an order-up-to level using the sign of the inventory imbalance and invokes the same restart logic based on cost gaps. Regret is decomposed into a part due to the gradient updates and a part due to inventory carry-over; both terms are shown to remain sublinear, so the policy performs nearly as well as a clairvoyant benchmark even when demand shifts unpredictably.

Chapter 4 turns to universal portfolio selection in markets whose return distributions may

change adversarially over time. We propose an adaptive strategy that runs a small council of entropic mirror-descent experts at different learning rates and promotes a faster expert when cumulative log-loss gaps signal increased market drift. Performance is measured by dynamic regret against the clairvoyant constant-rebalanced portfolio that would, in hindsight, have maximized the one-step log-return each round. When total market drift over the horizon is bounded (but unknown), this dynamic regret grows strictly slower than time; thus average log-return converges to that of the period-wise optimum even through substantial regime shifts. In a perfectly stationary market the rate specializes to the familiar two-thirds power of the horizon, reflecting only a modest cost for robustness.

Overall, the thesis demonstrates that coupling a handful of gradient or mirror-descent learners with a simple, data-driven restart rule yields algorithms that remain stable in tranquil periods yet react swiftly when conditions shift. Across general optimization, inventory management, and portfolio selection, these methods achieve performance guarantees that are essentially as strong as if the variation budget had been known in advance.

Contents

List of Tables

List of Appendices

1 Chapter 1

1.1 Introduction

Real-world decision–making problems often involve making a sequence of decisions under uncertainty and dynamic conditions. In operations research and management, classic examples include inventory control, dynamic pricing, and portfolio selection etc. A retailer must decide how much inventory to stock in each period while facing demand that can evolve due to seasonality or trends. Similarly, in revenue management, dynamic pricing strategies must adapt to shifts in customer behavior or market conditions. Even in finance, the universal portfolio selection problem requires sequentially reallocating assets without knowing future returns, effectively learning and adapting to changing market patterns. In all these scenarios, the underlying data-generating process—whether demand distributions, customer preferences, or asset return profiles—is *non-stationary*, meaning it can change over time. This non-stationarity poses a fundamental challenge: a policy that is optimal under yesterday's conditions may no longer perform well today. The decision-maker needs algorithms that continuously *learn and adapt* to an evolving environment.

One way to characterize the degree of environmental change is through a *variation budget*, which measures how much the underlying cost or reward function can drift over the horizon. Intuitively, the variation budget quantifies the total amount of change in the problem's dynamics; it provides a handle on the "price of non-stationarity," i.e., the performance loss incurred by not

knowing the future compared to a stationary world. For instance, if the demand distribution in an inventory problem shifts drastically (large variation budget), the problem is inherently harder than if it only fluctuates mildly. Prior research in online learning and optimization has leveraged such measures to design algorithms with performance guarantees that degrade gracefully as the environment's variability increases [Besbes et al. 2015]. Notably, Besbes *et al.* (2015) introduced a restarted online gradient descent approach for non-stationary stochastic optimization, showing that if the total variation is bounded by V_T, one can achieve a dynamic regret of order $O(T^{2/3}V_T^{1/3})$. However, a crucial assumption in their framework is that the variation budget V_T (or a reliable upper bound on it) is known to the decision-maker in advance.

In practice, such knowledge is rarely available—the decision-maker typically does *not* know how many changes will occur or how large those changes will be. In the absence of prior information on when and how the environment might change, the learner faces two main difficulties. First, *detecting and reacting to changes* is non-trivial without explicit cues. The algorithm must rely on incoming observations to infer that the underlying distribution or cost function has shifted, all while continuing to make near-optimal decisions. Second, the *magnitude of non-stationarity* (captured by the unknown variation budget) affects what performance is achievable, yet the algorithm cannot directly tune itself to this unknown parameter. Many existing approaches bypass this issue by either assuming a known bound on variation or by tuning hyperparameters in hindsight, leaving a gap for truly adaptive methods.

In this chapter, we address these challenges for *online stochastic convex optimization in non-stationary environments*. We consider a general setting in which, at each time t, the decision-maker incurs a convex cost $f_t(x_t)$, but the cost function $f_t(\cdot)$ may change over time with an unknown variation budget $V_T = B \cdot T^v$ for some $v \in [0, 1]$. Neither the change points nor the exponent v is known to the learner. Our goal is to design policies that attain low *dynamic regret*—that is, total cost close to that of an oracle who, in hindsight, knows the best action x_t^* for each period. Building on insights from both adversarial online learning and stochastic ap-

proximation, we develop a framework of *adaptive* stochastic gradient descent (SGD) algorithms that *do not require any prior knowledge of the variation budget* yet provably achieve near-optimal performance.

1.1.1 Main Results and Contributions

Our contributions can be summarized as follows:

Adaptive SGD with first-order feedback. We propose a restarting SGD algorithm for the case where only noisy gradient information is observed. Under a mild local strong convexity assumption, our method automatically adapts to the unknown variation budget and attains dynamic regret of order

$$\tilde{O}\big(T^{\frac{2}{3}} V_T^{\frac{1}{3}}\big),$$

matching the minimax rate up to logarithmic factors without knowing V_T in advance.

Hybrid feedback algorithm without curvature assumptions. When both function values and gradients are available (*zeroth*– and *first*–order feedback), we design a hybrid adaptive algorithm that dispenses with any strong convexity requirements. By using raw cost differences to detect shifts, it also achieves dynamic regret

$$\tilde{O}\big(T^{\frac{2}{3}} V_T^{\frac{1}{3}}\big)$$

in the general convex setting, again without prior knowledge of the variation.

Fully adaptive, near-optimal performance. Both algorithms attain the optimal trade-off between horizon length and environmental variability—characterized by T and V_T—*without* assuming that V_T (or its exponent) is known. This closes the adaptivity gap in non-stationary online convex optimization.

1.1.2 Literature Review

Online learning in non-stationary environments. This project is also related to a broader area of online learning and optimization in non-stationary environments. Extensive studies have been developed for stochastic optimization with non-stationary rewards (e.g., Besbes et al. 2015, Jiang et al. 2020), non-stationary multi-armed bandits (e.g., Besbes et al. 2019, Cheung et al. 2019a), non-stationary reinforcement learning (e.g., Cheung et al. 2019b, Wei and Luo 2021), and dynamic pricing and demand learning in non-stationary environments (e.g., Besbes and Zeevi 2011, Keskin and Zeevi 2017, Chen et al. 2019c, den Boer and Keskin 2020, Zhu and Zheng 2021, Keskin and Li 2021).

The challenge of non-stationary rewards in multi-armed bandit problems introduces a fundamental trilemma between exploration, exploitation, and adaptation to environmental changes. Besbes et al. 2019 pioneered a non-parametric framework to address this, quantifying the total change in mean rewards over a horizon T with a "temporal variation" budget, V_T. They characterized the complexity of this problem by establishing a minimax dynamic regret of order $\tilde{O}((KV_T)^{1/3}T^{2/3})$ against a powerful dynamic oracle, which knows the best arm at every time step. Their analysis, however, assumes that this variation budget is known beforehand. Building upon this, Cheung et al. 2019a tackles the more practical scenario where the variation budget is unknown. They introduce a novel Bandit-over-Bandit (BOB) framework that adaptively tunes a sliding-window UCB algorithm, effectively learning the rate of change online. This method achieves nearly optimal dynamic regret without prior knowledge of the non-stationarity, successfully "hedging the drift" and extending its applicability to various settings, including linear and combinatorial semi-bandits.

Dynamic pricing and demand learning in non-stationary environments present a significant challenge, as firms must adapt to evolving customer behavior with only partial feedback. Recent literature has explored various models of non-stationarity and proposed tailored learning poli-

cies. For instance, Keskin and Li 2021 study a market where customer preferences for quality-differentiated products shift according to an unknown Markovian process. They design a rate-optimal "bounded learning policy" that carefully balances learning new market states with exploiting existing knowledge, demonstrating a robust approach for structured, recurring changes. In a different vein, Zhu and Zheng 2021 investigate a continuously growing environment where both the mean and variance of demand increase over time. Their work reveals that the optimal regret and policy design are critically dependent on the growth rate of demand variance and, surprisingly, on whether the time horizon is known in advance. Addressing abrupt market shifts, Chen et al. 2019c model demand that evolves linearly but is subject to discrete change-points of different orders (e.g., sudden jumps vs. changes in trend). They propose an algorithm that explicitly detects these change-points and adapts its pricing strategy accordingly, achieving tight regret bounds that scale differently ($T^{1/2}$ vs. $T^{2/3}$) depending on the smoothness of the change, thereby underscoring the need for algorithms that can identify and react to the specific nature of environmental shocks.

Among these works, Besbes et al. 2015 is the most related to this project. They consider a sequential stochastic convex optimization problem, where the underlying unknown cost functions may change over the horizon whose amount of changes is constrained by the variation budget. The authors connect the adversarial online convex optimization (OCO) with the non-stationary stochastic optimization. They establish a restarting framework that adapts any algorithm with a "good" performance with respect to the single best action in the adversarial setting to a policy with a "good" performance with respect to the dynamic benchmark in the stochastic setting. In this paper, we borrow their notion of variation budget to measure the amount of changes in demand distributions and their restarting framework to design the algorithm.

Online learning for inventory models with non-stationary demands. There is a large body of literature in the area of data-driven inventory control that develop online learning algorithms for inventory models with *i.i.d.* demand distribution. Earlier studies analyze the repet-

itive newsvendor problem (see, e.g., Huh and Rusmevichientong 2009, Huh et al. 2011, Besbes and Muharremoglu 2013, Levi et al. 2007, Levi et al. 2015). Later more complicated systems are studied, e.g., lost-sales inventory systems with lead times (Huh et al. 2009, Zhang et al. 2020, Agrawal and Jia 2019), perishable inventory system (Zhang et al. 2018), inventory system with random capacity (Chen et al. 2020b), inventory system with fixed ordering cost (Yuan et al. 2021), and joint pricing and inventory system (Chen et al. 2019a, Chen et al. 2021, Chen et al. 2020a, Chen et al. 2022). Among these works, the stochastic gradient descent method has been adopted by many papers, e.g., Huh and Rusmevichientong 2009. In particular, Shi et al. 2016 considers a multi-product inventory control problem under a warehouse-capacity constraint is the study most related to this project. The authors develop a learning algorithm based on SGD method and prove the regret upper bound $O(\sqrt{T})$. The major difference of this paper compared to Shi et al. 2016 is that they assume demand distributions are *i.i.d.* whereas we consider a non-stationary environment.

There are also a growing body of literature going beyond the *i.i.d.* assumption to study data-driven inventory models with non-stationary demands. Chen 2021 considers an inventory control problem in a shifting demand environment with the number of changes in demand distributions at most $O(\log T)$. The author proves a regret lower bound $\Omega(\sqrt{T})$ and construct a learning algorithm with regret upper bound $\widetilde{O}(\sqrt{T})$. Keskin et al. 2022 studies a joint pricing and inventory problem for a perishable product, and construct algorithms for both the settings with non-parametric and parametric noise distributions, with the regret upper bound $\widetilde{O}(T^{2/3})$ and $\widetilde{O}(T^{1/2})$ respectively. Keskin et al. 2021 studies a repetitive newsvendor problem with a time-varying mean demand level. The authors design a moving window ordering policy and prove the regret upper bound $O(\sqrt{T})$ under constant variation budget and $O(T^{(1+v)/2})$ under $O(T^v)$ variation budget. Cheung et al. 2019b studies the non-stationary reinforcement learning with a single-product lost-sales inventory model as an application. Gong and Simchi-Levi 2021 apply the Q-learning technique to analyze inventory models with unknown cyclic demands and study the single-product lost-

sales model with zero lead time and multi-product backlogging model with positive lead times. Ding et al. 2021 considers a stochastic inventory system where demand distributions are feature-dependent and thus non-stationary. They design two algorithms based on SGD and prove the regret upper bound $O\sqrt{T}$) for both algorithms.

Closest to our work is the recent study of the Nonstationary Newsvendor by An et al. [An et al. 2025], who analyze sequential inventory decisions with unknown, time-varying demand (in a fully discrete model of both demand and order quantities), with and without side predictions. They design policies that are *variation–adaptive* without being given the nonstationarity level: in the no–prediction case they obtain a (near) minimax regret of $\tilde{O}\big(T^{(3+\nu)/4}\big)$, and with generic predictions of accuracy exponent a they achieve $\tilde{O}\big(T^{\min\{(3+\nu)/4,\, a\}}\big)$, matching lower bounds up to logarithmic factors. Methodologically, their policies are not gradient-based; instead, they work directly with the newsvendor loss and employ multi-scale, backward-looking estimators (windowed statistics) whose effective horizon trades off bias and variance as a function of the hypothesized variation level. By contrast, our algorithm is developed in the *general* online convex optimization setting and operates in the gradient-descent paradigm: we run K parallel OGD learners tuned to a grid of variation exponents and promote across learners via an empirical cost–gap test (first-order or hybrid zero/first-order feedback), which requires no distributional assumptions and, in the hybrid case, dispenses with curvature constants. Importantly, we also instantiate our algorithms in two domains—(i) nonstationary inventory control and (ii) universal portfolio selection—and in the inventory specialization we obtain dynamic regret $\tilde{O}\big(T^{(2+\nu)/3}\big)$, improving the T-exponent over [An et al. 2025]'s $\tilde{O}\big(T^{(3+\nu)/4}\big)$ for all $\nu < 1$ (the rates coincide at $\nu = 1$), while still requiring no prior knowledge of ν and retaining applicability to general convex losses with first-/hybrid-order feedback.

1.1.3 Notations

Throughout this paper, we use $\mathbb{R}_+^n$ to denote the set of n-dimensional non-negative vectors. Unless otherwise specified, for any n-dimensional vector x, $||x||$ denotes the Euclidean norm (or 2-norm), i.e, $||x|| := (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$. For two vectors $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ in $\mathbb{R}^m$, $\mathbf{x} \geqslant \mathbf{y}$ is equivalent to $x_i \geqslant y_i$ for each $i = 1, 2, \ldots, m$. For any $n \in \mathbb{N}^+$, we denote $[n]$ as the set $\{1, 2, \ldots, n\}$. For any $x \in \mathbb{R}^n$ and compact set $\mathcal{X} \subseteq \mathbb{R}^n$, let $\text{Proj}_{\mathcal{X}}(x)$ be the projection operator that projects x to $\mathcal{X}$, i.e., $\text{Proj}_{\mathcal{X}}(x) \triangleq \arg\min_{y \in \mathcal{X}} ||y - x||^2$.

1.2 Problem Formulation

Consider a sequential stochastic optimization problem over T time periods. In each period $t \in [T]$, the decision maker is faced with an unknown convex and differentiable function $f_t(x)$ defined on a convex, compact and non-empty action set $\mathcal{X} \subset \mathbb{R}^d$. For convenience, we refer to $f_t(x)$ as the *cost function* for period t and denote the radius of action set $\mathcal{X}$ by $D := \max_{x_1, x_2 \in \mathcal{X}} ||x_1 - x_2||$. The decision maker chooses an action x_t from action set $\mathcal{X}$ and then observes some feedback ϕ_t at the end of period t. For action $x \in \mathcal{X}$, there are forms of feedback:

(i) Noisy access to the cost, denoted by $\phi_t^{(0)}(x, f_t)$, such that $\mathbb{E}[\phi_t^{(0)}(x, f_t)] = f_t(x)$;

(ii) Noisy access to the gradient, denoted by $\phi_t^{(1)}(x, f_t)$, such that $\mathbb{E}[\phi_t^{(1)}(x, f_t)] = \nabla f_t(x)$;

In this work, we consider the following two scenarios of the feedback the decision maker can observe in each period $t \in [T]$:

(i) for all feasible actions in $x \in \mathcal{X}$, only $\phi_t^{(1)}(x, f_t)$ is observed, and in this case, we denote the feedback in period t by $\phi_t^{(1)} := \{\phi_t^{(1)}(x, f_t) : \forall x \in \mathcal{X}\}$;

(ii) for all feasible actions in $x \in \mathcal{X}$, both $\phi_t^{(0)}(x, f_t)$ and $\phi_t^{(1)}(x, f_t)$ are observed, and in this case, we denote the feedback in period t by $\phi_t^{(0,1)} := \{\phi_t^{(i)}(x, f_t) : \forall x \in \mathcal{X}, \forall i = 0, 1\}$.

Let $\mathscr{F}$ be a class of sequences $f := (f_1, f_2, \ldots, f_T)$ of convex cost functions satisfying the following conditions: there exists a finite and positive constant G such that

$$|f_t(x)| \leqslant G, \quad ||\nabla f_t(x)|| \leqslant G, \quad \forall t \in [T], \forall x \in \mathscr{X}.$$

We also assume that for each $t \in [T]$, $x_t^* := \arg\min_{x \in \mathscr{X}} f_t(x)$ is an interior point of $\mathscr{X}$.

Assumption 1. *We assume for any t,*

(1) ***Sub-Gaussian noise:***

$$\left\|\phi_t^{(0)}(x, f_t) - f_t(x)\right\|$$

is a σ_0-sub-Gaussian random variable (Definition 1.1)

(2) ***Sub-Gaussian noise:***

$$\left\|\phi_t^{(1)}(x, f_t) - \nabla f_t(x)\right\|$$

is a σ_1-sub-Gaussian random variable.

There are several equivalent definitions of sub-Gaussian random variables up to an absolute constant scaling. For convenience, we use the following property as the definition.

Definition 1.1 (Sub-Gaussian random variable)**.** A random variable X is called σ-sub-Gaussian if

$$\mathbb{E}\left[\exp(\lambda^2 X^2)\right] \leqslant \exp(\lambda^2 \sigma^2) \quad \text{for all } \lambda \text{ with } |\lambda| \leqslant \frac{1}{\sigma}.$$

The class of admissible policies. Following the definition in [Besbes et al. 2015], an admissible policy $\pi = \{\pi_t : t \in [T]\}$ is a sequence of functions, where each π_t maps the history $(x_1, \phi_1, x_2, \phi_2, \ldots, x_{t-1}, \phi_{t-1})$ and possibly some external randomness U_t to a feasible action x_t in set $\mathscr{X}$. Let $\mathscr{P}_\phi$ denote the class of all admissible policies. Note that any $\pi \in \mathscr{P}_\phi$ is non-anticipating, i.e., it depends only on the past history of actions and observations, and allows for randomized strategies via their dependence on U_t in each period t.

Temporal variation and variation budget. For any function sequence $f = (f_1, f_2, \ldots, f_T) \in \mathscr{F}$, following the notion in [Besbes et al. 2015], we consider the following variation based on the sup norm:

$$\mathrm{Var}(f_1, f_2, \ldots, f_T) := \sum_{t=2}^{T} ||f_t - f_{t-1}||,$$

where $||g - h|| := \sup_{x \in \mathscr{X}} |g(x) - h(x)|$ for two bounded functions $g(\cdot)$ and $h(\cdot)$ from $\mathscr{X}$ to $\mathbb{R}$. Let $\{V_t : t \in [T]\}$ be a sequence of non-decreasing numbers with $V_1 = 0$. We refer to V_T as the *variation budget* over T periods and make the assumption that $V_T = B \cdot T^v$ where B is some constant (which we take to be equal to one from here on). Then we define the corresponding *temporal uncertainty set*, as the set of admissible cost function sequences that are subject to the variation budget V_T over T periods:

$$\mathscr{V} := \left\{ (f_1, f_2, \ldots, f_T) \in \mathscr{F} : \sum_{t=2}^{T} ||f_t - f_{t-1}|| \leqslant V_T \right\}. \tag{1.1}$$

It is important to notice that the value of V_T is unknown in practice.

The notion of dynamic regret. The performance of any admissible policy $\pi \in \mathscr{P}_\phi$ is defined against a dynamic oracle:

$$\mathcal{R}_\phi^\pi(T, V_T) := \sup_{f \in \mathscr{V}} \left\{ \mathbb{E}^\pi \Big[\sum_{t=1}^{T} f_t(x_t) \Big] - \sum_{t=1}^{T} f_t(x_t^*) \right\},$$

where the expectation $\mathbb{E}^\pi[\cdot]$ is taken with respect to any randomness in the feedback, as well as in the policy π's actions. In the adversarial online convex optimization context, the performance of a policy is usually evaluated against a single best action benchmark and the objective is usually to minimize the following static regret: $\sup_{f \in \mathscr{V}} \{\mathbb{E}^\pi[\sum_{t=1}^{T} f_t(x_t)] - \min_{x \in \mathscr{X}} \sum_{t=1}^{T} f_t(x)\}$. As discussed in [Besbes et al. 2015], the dynamic oracle used as benchmark in equation (1.1) can be a significantly harder target than the single best action defining the static regret.

1.2.1 Preliminaries in [Besbes et al. 2015] and Challenges from Unknown Variation Budget

In this section, we will review the most relevant result in [Besbes et al. 2015] and discuss the new challenges in our context.

1. Unknown Changing Time and Magnitude of Demand Distribution Shifts. One of the foremost challenges in this setting is the uncertainty surrounding when and how significantly the demand distribution changes over time. The demand distribution may shift due to various external factors such as market trends, seasonal effects, or economic conditions. These shifts can occur at arbitrary times and can vary in magnitude, making it difficult to predict and adapt to changes promptly. Without prior knowledge of the change points or the extent of distributional shifts, the algorithm must be robust enough to detect and respond to these changes in real-time to minimize inventory costs effectively.

2. Unknown Variation Budget B_T. Another significant challenge is the lack of knowledge about the variation budget B_T, which is a crucial quantity measuring the total amount of distributional changes over the planning horizon. We assume that $B_T = GD \cdot T^v$ for some $v \in [0, 1]$, but the exact value of v is not known in advance. On the other side, the magnitude of $B_T = GD \cdot T^v$ will definitely affect the best possible regret bound a learning algorithm is able to achieve. The uncertainty regarding v complicates the design of the learning algorithm. Ideally, the algorithm should achieve a regret upper bound that is tightly dependent on v, even without knowing its exact value. To address this, the algorithm must dynamically adjust its learning rate and decision-making process to effectively respond to varying degrees of non-stationarity in the demand process.

1.3 Adaptive SGD Algorithms and Regret Upper Bounds

As preparations, we introduce a discrete set of hypothetical variation parameters $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ defined as follows:

$$v_k = \frac{k}{\log T}, \quad k = 1, 2, \ldots, K, \tag{1.2}$$

where K is chosen such that $v_{K-1} < 1 \leq v_K$. This discretization ensures that the candidate set $\mathcal{V}$ covers the entire range of possible variation budgets up to $v < 1$. Let k^* be the smallest index in $\{1, \ldots, K\}$ such that $v \leq v_{k^*}$. Given the relation $v_{k^*} = v_{k^*-1} + \frac{1}{\log T}$, it follows that $v \leq v_{k^*} < v + \frac{1}{\log T}$. Consequently, $T^v \leq T^{v_{k^*}} < eT^v$, implying that $T^{v_{k^*}}$ is within a constant factor of T^v. For each $k \in [K]$, we define the following window size under the hypothesis that the variation budget is of order $\Theta(T^{v_k})$:

$$\Delta_T^k = \left\lceil T^{\frac{2(1-v_k)}{3}} \right\rceil.$$

This batch size can be interpreted as the length of the *restarting window* in the restarting OGD framework established by [Besbes et al. 2015] for online convex optimization with non-stationary rewards, when the variation budget is $\Theta(T^{v_k})$. It is important to notice that since the decision maker does not know the exact value of v, neither does he/she know the exact value of k^*. Therefore, one crucial task in designing the algorithm is to detect the true value of k^* and the true variation budget.

1.3.1 With First-Order Feedback

In this subsection, we study the setting that only the first-order feedback is available to the decision maker in each period. In this scenario, we make the following assumption on the local smoothness property of cost functions in each period.

Assumption 2. *There exist constants $\underline{\delta} > 0$ known to the decision maker and $\overline{\delta} > 0$ such that*

$$\underline{\delta}||x - x_t^*||^2 \leqslant f_t(x) - f_t(x_t^*) \leqslant \overline{\delta}||x - x_t^*||^2, \quad \forall x \in \mathcal{X}, \forall t \in [T].$$

Assumption 2 requires that the cost function $f_t(x)$ is bounded above and below by two quadratic functions. When $f_t(x)$ is twice-differentiable, a sufficient condition for this assumption is that the Hessian matrix $\nabla^2 f_t(x)$ of $f_t(x)$ satisfies $2\underline{\delta}\mathbf{I}_d \preceq \nabla^2 f_t(x) \preceq 2\bar{\delta}\mathbf{I}_d$, where $\mathbf{I}_d$ denotes the d-dimensional identity matrix. Note that when the only available feedback is the first-order information, we have to assume that parameter $\underline{\delta}$ is known and the algorithm we design later will use the information of this quantity. If $\underline{\delta}$ is unknown, the learning problem in the unknown changing environment becomes much more difficult. However, if zeroth-order information is available, we can address the problem without imposing Assumption 2 (and thus without requiring the existence or knowledge of $\underline{\delta}$) by modifying the algorithm, and establish a similar regret bound. See Section 1.3.2 for further details.

The description of the adaptive SGD algorithm with first-order feedback is given in Algorithm 1. We next provide an intuitive explanation on the design of this algorithm. At the high level, the algorithm maintains K parallels "hypothetical" learners, indexed by $g = 1, 2, \ldots, K$, corresponding to different guesses $v_1, v_2, \ldots, v_K$ on the unknown variation parameter v, and adaptively chooses the action based on it updated belief on v. Due to the assumption of fully first-order feedback, we are able to update the SGD estimator under each hypothetical variation parameter. Each learner uses a different step size η_T^g (or equivalently, window size Δ_T^g), selected carefully to hypothetically achieve its theoretical optimal regret rate. At each period t, the algorithm monitors the performance of the current active learner (indexed by k) against all high-indexed learners. In Lines 5-6, the LHS term $\sum_{s=t_{\text{if}}}^{t-1} ||\hat{x}_s^k - \hat{x}_s^g||^2$ in the if condition measures the cumulative squared difference between the actions of the current and alternative learners since the last switch. If this term exceeds the threshold in the RHS, it suggests that the current learner is no longer well-tuned

to the environment's variation, and the algorithm switches to the next learner.

Algorithm 1 Adaptive SGD with First-Order Feedback

Input: hypothetical variation parameters $\{v_g : 1 \leqslant g \leqslant K\}$, hypothetical window sizes $\{\Delta_T^g : g = 1, 2, \ldots, K\}$, upper bound G, diameter D, tuning parameter γ
Initialization: Set $k = 1$, $t_{\text{if}} = 1$, arbitrarily set $x_1 \in \mathcal{X}$ and $\hat{x}_1^g \in \mathcal{X}$ for each $g \in \{1, 2, \ldots, K\}$.
for $t = 1$ to T **do**
  if $t \geqslant 2$ **then**
    if for some $g \in \{k+1, k+2, \ldots, K\}$,

$$\sum_{s=t_{\text{if}}}^{t-1} ||\hat{x}_s^k - \hat{x}_s^g||^2 \geq 2T^{\frac{2+v_g}{3}} \left(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + 2GD\right)\frac{1}{\delta}$$

    then $k \leftarrow k+1$ and $t_{\text{if}} \leftarrow t$.
    end if
    for $g = 1$ to K **do**
      Observe the gradient at the g-th hypothetical action $G_t(\hat{x}_{t-1}^g) := \phi_{t-1}^{(1)}(\hat{x}_{t-1}^g, f_{t-1})$;
      Compute the g-th hypothetical step size $\eta_T^g = \frac{\gamma D}{G\sqrt{\Delta_T^g}}$;
      Update the g-th hypothetical action $\hat{x}_t^g = \hat{x}_{t-1}^g - \eta_T^g \cdot G_t(\hat{x}_{t-1}^g)$;
    end for
  end if
  Choose the k-th hypothetical action $x_t = \hat{x}_t^k$ and incur the loss $C_t(\hat{x}_t^k) := \phi_t^{(0)}(\hat{x}_t^k, f_t)$.
end for

The theoretical performance of Algorithm 1 is presented in the following theorem.

Theorem 1.2. *Suppose $\phi_t = \phi_t^{(1)}$ for each $t \in [T]$. Under Assumption 2, Algorithm 1, denoted by π_1, achieves the following regret upper bound:*

$$\mathcal{R}_{\phi^{(1)}}^{\pi_1}(T, V_T) = O\left(T^{\frac{2}{3}} V_T^{\frac{1}{3}} (\log T)^{\frac{3}{2}}\right).$$

For general stochastic convex optimization problems with variation budget V_T and noisy gradient feedback, [Besbes et al. 2014] has proven that the minimax regret lower bound is $\Omega(T^{\frac{2}{3}} V_T^{\frac{1}{3}})$. The upper bound we establish for Algorithm 1 matches this lower bound up to logarithmic factors in T. Meanwhile, we notice that when the Hessian of each $f_t(\cdot)$ is uniformly bounded from below

and from above (see inequality (10) in [Besbes et al. 2014]) and the variation budget V_T is known, [Besbes et al. 2014] shows that the regret can be reduced to $\Omega(T^{\frac{1}{2}}V_T^{\frac{1}{2}})$. Comparing our result with theirs, Assumption 2 is weaker than their condition, as it only requires a local property around the optimum. In addition, we relax the crucial assumption of knowing V_T. As a consequence, our regret upper bound remains at $\tilde{O}(T^{\frac{2}{3}}V_T^{\frac{1}{3}})$. It remains an open question whether it is possible to design an adaptive learning algorithm that does not require knowledge of V_T and can achieve the regret upper bound $O(T^{\frac{1}{2}}V_T^{\frac{1}{2}})$ for strongly convex cost functions.

Sketched Proof of Theorem 1.2. Before delving into the details, we first provide an overview for the key steps to prove the result.

- In the first step, we establish a high-probability bound for some "good event" on the performance of the actions maintained by different hypothetical variation budgets in the two settings. Specifically, the good event refers to that for each $g \in [K]$, the actions maintained by the hypothetical variation budget T^{v_g} in periods $t = 1, 2, \ldots, T$ deviate from the optimal actions $x_1^*, x_2^*, \ldots, x_T^*$ by at most $\tilde{O}(T^{\frac{2+v_g}{3}})$.
- In the second step, we show that in both settings, when the good event happens, the algorithm will never over-estimate the value of f. Let $k(t)$ denote the index for the hypothetical variation parameter used in period t. Then we show that under the good event, $k(t) \leqslant f$ for each $t \in [T]$.
- In the third step, we bound the total regret of adaptive SGD algorithm by bounding the regret under the good event and the bad event, respectively.

Step 1: Construct a high-probability bound for all hypothetical SGD estimators.

Proposition 1.3. *Suppose $\phi_t = \phi_t^{(1)}$ for each $t \in [T]$. Under Assumption 2, the following event*

$\mathcal{G}_{\phi^{(1)}}$ holds with probability at least $1 - K/T$:

$$\sum_{t=1}^{T} ||\hat{x}_t^g - x_t^*||^2 \leq T^{\frac{2+\nu_g}{3}} \left(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + 2GD\right)\frac{1}{\underline{\delta}}$$

Sketched Proof of Proposition 1.3. Recall that for each $g = k^*, k^* + 1, \ldots, K$, the g-th hypothetical restarting window size is defined by $\Delta_T^g \triangleq \lceil T^{\frac{2}{3}(1-\nu_g)} \rceil$. For each $j = 1, 2, \ldots, \lceil T/\Delta_T^g \rceil$, let $\mathcal{T}_{g,j}$ denote the time periods in the j-th batch, i.e.,

$$\mathcal{T}_{g,j} \triangleq \left\{(j-1)\Delta_T^g + 1,\ (j-1)\Delta_T^g + 2,\ \ldots,\ (j\Delta_T^g) \wedge T\right\}.$$

By applying the strongly convex property in Assumption 2, we have the following inequality:

$$\begin{aligned}
\sum_{t=1}^{T} ||\hat{x}_t^g - x_t^*||^2 &\leqslant \frac{1}{\underline{\delta}} \sum_{t=1}^{T} \left(f_t(\hat{x}_t^g) - f_t(x_t^*)\right) \\
&= \frac{1}{\underline{\delta}} \sum_{j=1}^{\lceil T/\Delta_T^g \rceil} \Bigg(\underbrace{\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w)}_{\text{①: regret relative to batch } j\text{'s single best action}} \Bigg) \\
&\quad + \frac{1}{\underline{\delta}} \sum_{j=1}^{\lceil T/\Delta_T^g \rceil} \underbrace{\left(\min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w) - \sum_{t\in\mathcal{T}_{g,j}} f_t(x_t^*)\right)}_{\text{②: regret due to functional changes}}.
\end{aligned} \tag{1.3}$$

We next bound the two terms ① and ② in inequality (1.3). For the second term, it directly follows from the analysis of Proposition 2 in [Besbes et al. 2015] that

$$\underbrace{\sum_{j=1}^{\lceil T/\Delta_T^g \rceil} \left(\min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w) - \sum_{t\in\mathcal{T}_{g,j}} f_t(x_t^*)\right)}_{\text{②: regret due to functional changes}} \leqslant \sum_{j=1}^{\lceil T/\Delta_T^g \rceil} 2\Delta_T^g \cdot \sum_{t\in\mathcal{T}_{g,j}} ||f_t - f_{t-1}|| = 2\Delta_T^g V_T.$$

Bounding ① in inequality (1.3): To bound the first term, we prove the following technical lemma that develops a high-probability bound for SGD algorithm in a general class of online convex optimization (OCO) problems.

Lemma 1.4 (High Probability Bound for Non-Stationary OCO). *Consider a sequential stochastic convex optimization problem* $\min_{y\in\mathcal{Y}} \sum_{t=1}^{T} f_t(x)$. *For each* $t \geqslant 1$, *we assume*

(i) $f_t(x)$ *is convex and differentiable in* $\mathcal{X} \subseteq \mathbb{R}^n$;

(ii) $\max_{x_1,x_2\in\mathcal{X}} ||x_1 - x_2|| \leqslant D$.

Let $\hat{x}_1$ *be an arbitrary point in* $\mathcal{Y}$ *and for each* $t \geqslant 1$, $\hat{x}_{t+1} \triangleq Proj_{\mathcal{Y}}(\hat{x}_t - \eta_t G_t(\hat{x}_t))$, *where*

(i) $\mathbb{E}[G_t(x)] = \partial f_t(x)$ *for any* $x \in \mathcal{Y}$ *and* $\max_{t\in[T],x\in\mathcal{X}} |G_t(x)| \leqslant G$ *with probability one;*

(ii) $\eta_t \triangleq \frac{\gamma D}{G\sqrt{T}}$ *for some tuning parameter* $\gamma > 0$.

Then for any $\delta > 0$, *the following inequality holds with probability at least* $1 - \delta$:

$$\max_{x\in\mathcal{X}} \sum_{t=1}^{T} \Big(f_t(\hat{x}_t) - f_t(x) \Big) \leqslant 2GD\sqrt{2T\log(1/\delta)} + \frac{1}{2}\Big(\frac{1}{\gamma} + \gamma\Big) GD\sqrt{T}.$$

In the classical framework for analyzing the performance of the SGD algorithm, the typical measure of interest is the expected regret, i.e, $\sum_{t=1}^{T}(\mathbb{E}[f_t(\hat{y}_t)] - f_t(y))$. In Lemma 1.4, we establish a stronger high-probability bound on the empirical regret of the SGD algorithm (i.e., $\sum_{t=1}^{T}(f_t(\hat{y}_t) - f_t(y))$), which is especially useful for constructing the triggering event in our algorithm (Line 5 in Algorithm 4). To prove Lemma 1.4, we construct a sequence of martingale differences and invoke the martingale concentration inequality to bound the empirical regret.

We defer the detailed proof to Appendix **??** and next explain how to apply Lemma 1.4 to bound ① in inequality (1.3). For any fixed $g \in [K]$, for the learning horizon with Δ_T^g periods, from Line 10, the step size η_T^g is set to $\gamma D/(G\sqrt{\Delta_T^g})$. Then by applying Lemma 1.4 to ① in inequality (1.3) for

each $j = 1, 2, \ldots, \lceil T/\Delta_T^g \rceil$, we have the following inequality holds with probability at least $1 - \delta$:

$$\begin{aligned}\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w) &\leqslant 2GD\sqrt{2\Delta_T^g \log\frac{1}{\delta}} + \frac{1}{2}\left(\frac{1}{\gamma}+\gamma\right)GD\sqrt{\Delta_T^g} \\ &= GD\sqrt{\Delta_{T,g}}\left(2\sqrt{2\log\frac{1}{\delta}} + \frac{1}{2}\left(\frac{1}{\gamma}+\gamma\right)\right).\end{aligned}$$

Note that $T/\Delta_T^g \leqslant T^{\frac{1+2v_g}{3}} \leqslant T$. Then by letting $\delta = 1/T^2$ and applying the union bound, we have the following inequality holds with probability at least $1 - \lceil T/\Delta_T^g \rceil \delta \geqslant 1 - 1/T$:

$$\sum_{j=1}^{\lceil T/\Delta_T^g \rceil}\left(\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w)\right) \leqslant \left\lceil \frac{T}{\Delta_T^g}\right\rceil GD\sqrt{\Delta_{T,g}}\left(4\sqrt{\log T} + \frac{1}{2}\left(\frac{1}{\gamma}+\gamma\right)\right). \tag{1.4}$$

With inequalities (1.4) and the previous discussion, we have,

$$\begin{aligned}\sum_{t=1}^{T} ||\hat{x}_t^g - x_t^*||^2 &\leqslant \left\lceil \frac{T}{\Delta_T^g}\right\rceil GD\sqrt{\Delta_{T,g}}\left(4\sqrt{\log T} + \frac{1}{2}\left(\frac{1}{\gamma}+\gamma\right)\right)\frac{1}{\underline{\delta}} + \left\lceil \frac{T}{\Delta_T^g}\right\rceil 2\Delta_T^g V_T \frac{1}{\underline{\delta}} &(1.5)\\ &\leq T^{\frac{2+v_g}{3}}\left(4GD\sqrt{\log T} + \frac{1}{2}\left(\frac{1}{\gamma}+\gamma\right)GD + 2GD\right)\frac{1}{\underline{\delta}}, &(1.6)\end{aligned}$$

which completes the proof for proposition 1.3.

Step 2: ASGD never over-estimates the variation budget with high probability. Let $k(t)$ denote the index k applied when updating the post-ordering inventory level y_t in period t.

Proposition 1.5. *Suppose good event $\mathcal{G}_{\phi^{(1)}}$ occurs. Then $k(t) \leqslant k^*$ for each $t \in [T]$.*

Proof of Proposition 1.5. Suppose there exists some period $t_0 \in \{2, 3, \ldots, T\}$ such that, the algorithm adopts k^* in period $t_0 - 1$ and switches to k^*+1 in period t_0. That it, $k(t_0 - 1) = k^*$ and $k(t_0) = k^*+1$.

Then there exists $g \geqslant k^* + 1$ such that

$$\sum_{s=t_{if}}^{t_0} \|\hat{x}_s^{k^*} - \hat{x}_s^g\|^2 > 2T^{\frac{2+\nu_g}{3}} \Big(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + 2GD\Big)\frac{1}{\underline{\delta}}. \tag{1.7}$$

On the other hand, under event $\mathcal{G}_{\phi^{(1)}}$, we have

$$\begin{aligned}
\sum_{s=1}^{t_0} ||\hat{x}_s^{k^*} - \hat{x}_s^g||^2 &\leqslant \sum_{s=1}^{T} ||\hat{x}_s^{k^*} - \hat{x}_s^g||^2 \\
&\leqslant 2\sum_{s=1}^{T} ||\hat{x}_s^{k^*} - x_s^*||^2 + 2\sum_{s=1}^{T} ||x_s^* - \hat{x}_s^g||^2 \\
&\leqslant 2T^{\frac{2+\nu_g}{3}} \Big(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + 2GD\Big)\frac{1}{\underline{\delta}},
\end{aligned}$$

leading to contradiction with (1.7). Therefore, the algorithm never switches to $k^* + 1$, which implies $k(t) \leqslant k^*$ for all $t \in [T]$. □

Step 3: Regret between consecutive switches and on the terminal tail. Let $\Xi_T = \big(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + 2GD\big)$. Condition on the good event $\mathcal{G}_{\phi^{(1)}}$ and Proposition 1.5, so that $k(t) \leqslant k^*$ for all $t \in [T]$. Let the (at most K) switch times be $1 < t_1 < \cdots < t_M \leqslant T$, and set $t_0 := 1$ and $t_{M+1} := T + 1$. For $m = 0, 1, \ldots, M-1$, consider the interval $I_m = [t_m, t_{m+1} - 1]$ where the **if**-condition is *not* triggered. By Assumption 2 (upper quadratic growth with parameter $\bar{\delta}$) and $(a+b)^2 \leqslant 2(a^2 + b^2)$,

$$\begin{aligned}
\sum_{t\in I_m} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) &\leqslant \bar{\delta} \sum_{t\in I_m} \|\hat{x}_t^{\mathrm{ASGD}} - x_t^*\|^2 \\
&\leqslant 2\bar{\delta} \sum_{t\in I_m} \big(\|\hat{x}_t^{\mathrm{ASGD}} - \hat{x}_t^{k^*}\|^2 + \|\hat{x}_t^{k^*} - x_t^*\|^2\big).
\end{aligned} \tag{1.8}$$

Because the **if**-condition is not triggered on I_m and $k(t) \leqslant k^*$, its negation with $g = k^*$ yields

$$\sum_{t \in I_m} \|\hat{x}_t^{\mathrm{ASGD}} - \hat{x}_t^{k^*}\|^2 < 2\, T^{\frac{2+\nu_{k^*}}{3}} \cdot \frac{\Xi_T}{\underline{\delta}} \;\leqslant\; 2e\, T^{\frac{2}{3}} V_T^{\frac{1}{3}} \cdot \frac{\Xi_T}{\underline{\delta}},$$

where the last inequality uses $T^{\nu_{k^*}} \;\leqslant\; e\, V_T$ (as in the discretization argument). Moreover, by Proposition 1.3,

$$\sum_{t=1}^{T} \|\hat{x}_t^{k^*} - x_t^*\|^2 \leqslant T^{\frac{2+\nu_{k^*}}{3}} \cdot \frac{\Xi_T}{\underline{\delta}} \;\leqslant\; e\, T^{\frac{2}{3}} V_T^{\frac{1}{3}} \cdot \frac{\Xi_T}{\underline{\delta}}.$$

Plugging these into (1.8) gives, for every interval I_m between two consecutive switches,

$$\sum_{t \in I_m} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) \;\leqslant\; 6e\, \frac{\bar{\delta}}{\underline{\delta}}\, \Xi_T\, T^{\frac{2}{3}} V_T^{\frac{1}{3}}. \tag{1.9}$$

Terminal tail $[t_{\mathrm{last}}, T]$. Let $t_{\mathrm{last}} := t_M$ be the last switch time (or 1 if no switch occurs). On $[t_{\mathrm{last}}, T]$ the **if**-condition is never triggered, hence the same negation with $g = k^*$ applies on the whole tail. Repeating the argument above yields

$$\sum_{t=t_{\mathrm{last}}}^{T} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) \;\leqslant\; 6e\, \frac{\bar{\delta}}{\underline{\delta}}\, \Xi_T\, T^{\frac{2}{3}} V_T^{\frac{1}{3}}. \tag{1.10}$$

Summing intervals and accounting for the bad event. There are at most $M \leqslant K$ switches, hence at most K between-switch intervals plus the terminal tail. Summing (1.9) over these intervals and adding (1.10) yields, on $\mathcal{G}_{\phi^{(1)}}$,

$$\sum_{t=1}^{T} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) \;\leqslant\; C\, T^{\frac{2}{3}} V_T^{\frac{1}{3}}\, \Xi_T\, (\log T)^{O(1)},$$

where we used $K = (\log T)^{O(1)}$ for the grid size and absorbed constants into C. On $\mathcal{G}^{\complement}_{\phi^{(1)}}$, which occurs with probability at most K/T by Proposition 1.3, the regret is trivially bounded by $T \cdot \max_{t, x \in \mathcal{X}} f_t(x)$, contributing at most $O(K)$ in expectation. Combining the two parts and recalling

$\Xi_T = \Theta\big(GD\sqrt{\log T}\big)$ establishes

$$\mathcal{R}^{\pi_1}_{\phi^{(1)}}(T, V_T) = O\Big(T^{\frac{2}{3}} V_T^{\frac{1}{3}} (\log T)^{\frac{3}{2}}\Big).$$

1.3.2 With Zeroth-Order & First-Order Feedback

In many applications the learner can observe (or actively query) *both* a noisy gradient $\phi_t^{(1)}(x, f_t)$ and a noisy function value $\phi_t^{(0)}(x, f_t)$ for every feasible action. This hybrid feedback supplies strictly more information than the first-order-only setting of Section 1.3. Crucially, realised costs can now be compared *directly*, allowing us to detect distributional shifts without appealing to geometric surrogates that depend on unknown curvature parameters.

Because the switching test in Algorithm 2 is written in terms of empirical cost differences,

$$\sum_{s=t_{\text{if}}}^{t-1} \big| C_s(\hat{x}_s^k) - C_s(\hat{x}_s^g) \big|,$$

the procedure no longer needs knowledge of the lower-curvature constant $\underline{\delta}$ (cf. Assumption 2). Hence we can work with *arbitrary* convex losses that satisfy the boundedness condition in (1.1), while retaining the same regret guarantees.

As before, we maintain K parallel SGD learners, one for each hypothetical variation exponent v_g. Learner g uses window size Δ_T^g and constant step size $\eta_T^g \propto 1/\sqrt{\Delta_T^g}$ so that, if $v = v_g$ were the true exponent, its standalone regret would achieve the minimax rate $O\big(T^{2/3} V_T^{1/3}\big)$. The algorithm starts with the most aggressive learner ($k = 1$) and monitors its *realised* cost relative to each more conservative learner $g > k$. If the cumulative gap exceeds the data-dependent RHS threshold (Lines 5–6), we interpret this as evidence that the environment is drifting faster than T^{v_k} and immediately promote the next learner, resetting the comparison window at $t_{\text{if}} \leftarrow t$.

Algorithm 2 Adaptive SGD with Hybrid Feedback

Input: hypothetical variation parameters $\{v_g : 1 \leqslant g \leqslant K\}$, hypothetical window sizes $\{\Delta_T^g : g = 1, 2, \ldots, K\}$, upper bound G, diameter D, tuning parameter γ
Initialization: Set $k = 1$, $t_{\mathbf{if}} = 1$, arbitrarily set $x_1 \in \mathcal{X}$ and $\hat{x}_1^g \in \mathcal{X}$ for each $g \in \{1, 2, \ldots, K\}$.
for $t = 1$ to T **do**
  if $t \geqslant 2$ **then**
    if for some $g \in \{k+1, k+2, \ldots, K\}$,

$$\sum_{s=t_{\mathbf{if}}}^{t-1} |\phi^{(0)}(\hat{x}_s^k, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s)| \geq 2T^{\frac{2+v_g}{3}} \left(4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma + \gamma^{-1})GD + 2GD\right) + 8\sigma_0\sqrt{T \log T},$$

    then $k \leftarrow k + 1$ and $t_{\mathbf{if}} \leftarrow t$.
    end if
    for $g = 1$ to K **do**
      Observe the gradient at the g-th hypothetical action $G_t(\hat{x}_{t-1}^g) := \phi_{t-1}^{(1)}(\hat{x}_{t-1}^g, f_{t-1})$;
      Compute the g-th hypothetical step size $\eta_T^g = \frac{\gamma D}{G\sqrt{\Delta_T^g}}$;
      Update the g-th hypothetical action $\hat{x}_t^g = \hat{x}_{t-1}^g - \eta_T^g \cdot G_t(\hat{x}_{t-1}^g)$;
      Query the cost at the g-th hypothetical action $C_t(\hat{x}_t^g) := \phi_t^{(0)}(\hat{x}_t^g, f_t)$.
    end for
  end if
  Choose the k-th hypothetical action $x_t = \hat{x}_t^k$ and incur the loss $C_t(\hat{x}_t^k) := \phi_t^{(0)}(\hat{x}_t^k, f_t)$.
end for

Theorem 1.6. *Suppose $\phi_t = \phi_t^{(0,1)}$ for all $t \in [T]$. Then Algorithm 2, denoted π_2, achieves*

$$\mathcal{R}^{\pi_2}_{\phi^{(0,1)}}(T, V_T) \;=\; O\Big(T^{\frac{2}{3}}\, V_T^{\frac{1}{3}}\, (\log T)^{\frac{3}{2}}\Big).$$

The hybrid procedure matches the dynamic regret rate of the first-order algorithm while dispensing with curvature information. Whether the additional zeroth-order feedback can be leveraged to achieve the faster $O\big(T^{1/2}V_T^{1/2}\big)$ bound *without* knowing V_T remains an intriguing open question.

Sketched Proof of Theorem 1.6. The argument mirrors the first–order case, but leverages the extra zeroth–order feedback and dispenses with local strong convexity.

- *Step 1: Good–event construction.* We show that, with high probability, every hypothetical learner indexed by $g \in [K]$ incurs a total noisy–cost deviation from the dynamic oracle that grows no faster than the optimal rate predicted for variation budget T^{v_g}. The bound follows from a batchwise application of a martingale concentration inequality combined with the regret guarantee of stochastic gradient descent.

- *Step 2: No over–estimation of the variation exponent.* Conditioning on the good event, the **if**–test in the algorithm can never promote an index larger than the true one k^*. The proof inserts the oracle trajectory as an intermediate reference and applies the triangle inequality together with the good–event bounds to reach a contradiction whenever an over–estimate is hypothetically assumed.

- *Step 3: Regret decomposition over trigger intervals.* Between any two successive trigger epochs the noisy–cost gap between the active learner and the oracle is dominated by the good–event threshold. Summing over at most $K = \lceil \log T \rceil$ such intervals, and adding the negligible contribution from the low–probability complement of the good event, yields the stated dynamic-regret upper bound, matching the minimax rate up to logarithmic factors.

Step 1: A high–probability bound for the cumulative noisy costs of all hypothetical learners.

Proposition 1.7. *Assume $\phi_t = \phi_t^{(0,1)}$ for every $t \in [T]$ and fix $g \in \{k^*, k^*+1, \dots, K\}$. Let $\Delta_T^g = \lceil T^{\frac{2}{3}(1-\nu_g)} \rceil$ and define the event*

$$\mathcal{G}_{\phi^{(0,1)}} := \left\{ \left| \sum_{t=1}^{T} \left[\phi^{(0)}(\hat{x}_t^g, f_t) - \phi^{(0)}(x_t^*, f_t) \right] \right| \leqslant T^{\frac{2+\nu_g}{3}} \left(4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma + \gamma^{-1})GD + 2GD \right) + 4\sigma_0 \sqrt{T \log T} \right\}.$$

Then $\Pr\left(\mathcal{G}_{\phi^{(0,1)}}\right) \geqslant 1 - \dfrac{3K}{T}$.

Proof sketch. For the g-th learner write

$$\sum_{t=1}^{T} \left[\phi^{(0)}(\hat{x}_t^g, f_t) - \phi^{(0)}(x_t^*, f_t) \right] = \underbrace{\sum_{t=1}^{T} \left[\phi^{(0)}(\hat{x}_t^g, f_t) - f_t(\hat{x}_t^g) \right]}_{\text{(noise at } \hat{x}_t^g)} + \underbrace{\sum_{t=1}^{T} \left[f_t(\hat{x}_t^g) - f_t(x_t^*) \right]}_{\text{(true loss gap)}} + \underbrace{\sum_{t=1}^{T} \left[f_t(x_t^*) - \phi^{(0)}(x_t^*, f_t) \right]}_{\text{(noise at } x_t^*)}.$$

To bound the first and the third terms in the above, we establish the following high-probability bound: with probability at least $1 - \frac{1}{T^2}$.

$$\sup_{x \in \mathcal{X}} \left| \sum_{t=1}^{T} \left(\phi^{(0)}(x, f_t) - f_t(x) \right) \right| \leqslant \sigma_0 \sqrt{2T \log(T^2)}, \tag{1.11}$$

by sub-Gaussian tail bound, which can be found from Appendix. The failure probability for both terms over $K - k^* + 1 \leqslant K$ learners should be at most $2K/T$

The middle sum, $\sum_{t=1}^{T} [f_t(\hat{x}_t^g) - f_t(x_t^*)]$, is controlled by the estimate

$$\sum_{t=1}^{T} \left[f_t(\hat{x}_t^g) - f_t(x_t^*) \right] \leqslant T^{\frac{2+\nu_g}{3}} \left(4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma + \gamma^{-1})GD + 2GD \right),$$

proved in proposition 1.3. A union bound over the $K - k^* + 1 \leqslant K$ relevant learners completes the argument, giving failure probability at most $3K/T$ and establishing the proposition. □

Step 2: The hybrid ASGD never over-estimates the variation budget.

For every index $g \in \{k^*, k^*+1, \dots, K\}$ define

$$B_g \;:=\; T^{\frac{2+v_g}{3}}\left(4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma+\gamma^{-1})GD + 2GD\right) + 4\sigma_0\sqrt{T\log T}.$$

The **if**-statement in Algorithm 2 uses the threshold

$$\mathrm{RHS}(g) \;:=\; B_{k^*} + B_g, \quad \text{with } B_{k^*} \leqslant B_g \text{ (since } v_{k^*} \leqslant v_g).$$

Good event. Proposition 1.7 implies that, with probability at least $1 - \frac{3K}{T}$, the following bounds hold simultaneously for every $g \geqslant k^*$:

$$\sum_{t=1}^{T}\left|\phi^{(0)}(\hat{x}_t^g, f_t) - \phi^{(0)}(x_t^*, f_t)\right| \leqslant B_g \quad \text{and} \quad \sum_{t=1}^{T}\left|\phi^{(0)}(\hat{x}_t^{k^*}, f_t) - \phi^{(0)}(x_t^*, f_t)\right| \leqslant B_{k^*}. \tag{1.12}$$

Denote this joint event by $\mathcal{G}_{\phi^{(0,1)}}$.

Proposition 1.8. *On $\mathcal{G}_{\phi^{(0,1)}}$ the index used by the algorithm never exceeds k^*, i.e. $k(t) \leqslant k^*$ for all $t \in [T]$.*

Proof. Assume toward a contradiction that the algorithm first switches from k^* to k^*+1 at time t_0 ($\geqslant 2$). Let t_{if} be the start of the current monitoring window. Then, by the switch rule, there exists $g \geqslant k^*+1$ such that

$$\sum_{s=t_{\mathrm{if}}}^{t_0-1}\left|\phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^{g}, f_s)\right| > 2B_g. \tag{1.13}$$

Under the good event we may insert x_s^* between the two trajectories and apply the triangle inequality:

$$\left|\phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^{g}, f_s)\right| \leqslant \left|\phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(x_s^*, f_s)\right| + \left|\phi^{(0)}(\hat{x}_s^{g}, f_s) - \phi^{(0)}(x_s^*, f_s)\right|.$$

Summing over $s \in [t_{\text{if}}, t_0 - 1]$ and using (1.12) yields

$$\sum_{s=t_{\text{if}}}^{t_0-1} \left| \phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^{g}, f_s) \right| \leqslant B_{k^*} + B_g \leq 2B_g,$$

contradicting (1.13). Hence the algorithm can never select an index strictly larger than k^*, completing the proof. □

Step 3: Regret between consecutive switches and on the terminal tail. Condition on the good event $\mathcal{G}_{\phi^{(0,1)}}$ and Proposition 1.8, so $k(t) \leqslant k^*$ for all t. Let the (at most K) switch times be $1 < t_1 < \cdots < t_M \leqslant T$, and set $t_0 := 1$ and $t_{M+1} := T+1$. For $m = 0, 1, \ldots, M-1$, define the interval with no switches

$$I_m \;:=\; [\, t_m,\; t_{m+1} - 1 \,].$$

Since the **if**-test never triggers on I_m and $k(t) \leqslant k^*$, taking $g = k^*$ in Lines 5–6 of Algorithm 2 yields the negation

$$\sum_{t \in I_m} \left| \phi^{(0)}(\hat{x}_t^{k}, f_t) - \phi^{(0)}(\hat{x}_t^{k^*}, f_t) \right| \;<\; 2B_{k^*}, \tag{1.14}$$

where B_{k^*} is as in Step 2.

Between-switch intervals. By the triangle inequality,

$$\phi^{(0)}(\hat{x}_t^{k}, f_t) - \phi^{(0)}(x_t^*, f_t) = \left[\phi^{(0)}(\hat{x}_t^{k}, f_t) - \phi^{(0)}(\hat{x}_t^{k^*}, f_t)\right] + \left[\phi^{(0)}(\hat{x}_t^{k^*}, f_t) - \phi^{(0)}(x_t^*, f_t)\right].$$

Summing over $t \in I_m$ and using (A14) together with the good–event bound for k^* (Step 1) gives

$$\sum_{t \in I_m} \left[\phi^{(0)}(\hat{x}_t^{k}, f_t) - \phi^{(0)}(x_t^*, f_t)\right] \;\leqslant\; 2B_{k^*} + B_{k^*} \;=\; 3B_{k^*}. \tag{1.15}$$

Taking expectations and using $\mathbb{E}[\phi^{(0)}(x, f_t)] = f_t(x)$ yields

$$\mathbb{E}\left[\sum_{t\in I_m}\left(f_t(\hat{x}_t^k) - f_t(x_t^*)\right)\,\middle|\,\mathcal{G}_{\phi^{(0,1)}}\right] \;\leqslant\; 3B_{k^*}.$$

Terminal tail $[t_{\text{last}}, T]$. Let $t_{\text{last}} := t_M$ (or 1 if $M = 0$). On $[t_{\text{last}}, T]$ the **if**-test never triggers; repeating the argument above gives

$$\mathbb{E}\left[\sum_{t=t_{\text{last}}}^{T}\left(f_t(\hat{x}_t^k) - f_t(x_t^*)\right)\,\middle|\,\mathcal{G}_{\phi^{(0,1)}}\right] \;\leqslant\; 3B_{k^*}.$$

Summation and conclusion. There are at most $M \leqslant K$ between-switch intervals plus the terminal tail, so on $\mathcal{G}_{\phi^{(0,1)}}$,

$$\sum_{t=1}^{T}\left(f_t(\hat{x}_t^{\text{ASGD}}) - f_t(x_t^*)\right) \;\leqslant\; 3\,(K+1)\,B_{k^*}.$$

Using $K = \Theta(\log T)$ and $B_{k^*} = \Theta\left(T^{\frac{2}{3}} V_T^{\frac{1}{3}} \sqrt{\log T}\right)$ (Step 2) gives

$$\mathcal{R}^{\pi_2}_{\phi^{(0,1)}}(T, V_T) \;=\; O\left(T^{\frac{2}{3}} V_T^{\frac{1}{3}} (\log T)^{\frac{3}{2}}\right),$$

and the (measure-$\leqslant 3K/T$) bad-event contribution is negligible compared to this bound.

1.4 Numerical Study

We illustrate the upper bounds on the regret by numerical experiments measuring the average regret that is incurred in the presence of various patterns of changing costs, and under different feedback structures and noise. We compare the performance of the Adaptive SGD with First-Order Feedback (ASGD 1) and Adaptive SGD with Hybrid Feedback (ASGD 2) against the performance achieved by applying the ordinary SGD, Restarting SGD algorithm with correctly specified environment variation v and with wrongly specified v. We note that direct implemen-

tation of the original ASGD 1 and ASGD 2 can barely trigger the conditions about ν_k updating, so instead we introduce a tunable multiplicative factor c_{thr} to rescale the trigger condition thresholds. More details are presented in the following.

Variation, feedback, and performance

All experiments are carried out on the quadratic family

$$f_t(x) \;=\; \frac{x^2}{2} - b_t x + 1, \qquad x \in \mathcal{X} := [-2, 2],$$

so that the instantaneous minimizer is $x_t^* = b_t$, $\nabla f_t(x) = x - b_t$, and $\underline{\delta} = \frac{1}{2}$. The drift sequence $\{b_t\}_{t=1}^T$ is generated according to two non-stationary patterns that realise a prescribed variation budget $V_T = GD{*}T^\nu$ for a chosen exponent $\nu \in (0, 1]$, where G is the upper bound for gradient and D is the diameter of decision space. In this section, we would fix $T = 10,000$ and $\nu_1 = 0.11$, $\nu_2 = 0.22$, ..., $\nu_9 = 0.99$.

1. Smooth power–law drift. For given diameter $D > 0$ and exponent ν we set

$$|b_t - b_{t-1}| \;=\; \frac{D\,\nu}{t^{\,1-\nu}} \quad (t \geqslant 2),$$

starting at $b_1 = 2$ and flipping direction whenever the path would cross the bounds $[-2, 2]$. The cumulative variation satisfies $\sum_{t=2}^T |b_t - b_{t-1}| = D{*}T^\nu$ and $\sum_{t=2}^T \max_{x\in\mathcal{X}} |f_t(x) - f_{t-1}(x)| = DG{*}T^\nu$.

2. Piecewise–constant jumps. Fix an inter-jump spacing S (integer). Let $J := D\,T^\nu / \lceil T/S \rceil$ so that exactly $\lceil T/S \rceil$ jumps of magnitude J again give total variation $D\,T^\nu$. Starting at $b_1 = 2$, every S^{th} time step we add $\pm J$ (sign alternates) and clip to $[-2, 2]$; between jumps b_t stays constant. We set $S = 20$ for this study.

Feedback and noise. At each round an action $x_t \in \mathcal{X}$ is played, incurring cost $f_t(x_t)$. Noise is i.i.d. Gaussian with standard deviations σ_0 and σ_1:

$$\phi_t^{(0)}(x_t, f_t) = f_t(x_t) + \varepsilon_t^{(0)}, \quad \phi_t^{(1)}(x_t, f_t) = \nabla f_t(x_t) + \varepsilon_t^{(1)}, \quad \varepsilon_t^{(i)} \sim \mathcal{N}(0, \sigma_i^2).$$

We consider three noise levels $\sigma_0 = \sigma_1 \in \{0.1,\ 0.3,\ 1\}$.

Performance. For the ASGD 1 and ASGD 2, we set $\gamma = 1$. Given a policy π, feedback type ϕ, horizon T and drift path $b_{1:T}$, the dynamic regret is

$$R_\phi^\pi(f, T) \ = \ \sum_{t=1}^{T} \Big(f_t(x_t) \ - \ f_t(x_t^*) \Big).$$

We denote the loss relative to the oracle:

$$L_\phi^\pi(f, T) \ = \ \frac{R_\phi^\pi(f, T)}{\sum_{t=1}^{T} f_t(x_t^*)}.$$

1.4.1 Limitation of ASGD: too conservative thresholds

The original trigger rules in Algorithms 1 and 2 require the left–hand statistics

$$\text{(First–order)} \ \sum_{s=t_{\text{if}}}^{t-1} \big\| \hat{x}_s^k - \hat{x}_s^g \big\|^2, \qquad \text{(Hybrid)} \ \sum_{s=t_{\text{if}}}^{t-1} \big| \phi^{(0)}(\hat{x}_s^k, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s) \big|$$

to exceed an upper-confidence bound of the generic form $\kappa_g = GD\,T^{\frac{2+\nu_g}{3}}$ (For simplicity, the true thresholds can only be larger than this κ_g). These bounds are derived from worst–case OCO regret bound, as well time–uniform martingale inequalities; hence they are intentionally *pessimistic*. In moderate horizons and with realistic noise levels the accumulated empirical discrepancies stay far below κ_g, so the algorithm never promotes its variation index k.

To make this gap visible we ran ASGD 1 (first–order trigger) and ASGD 2 (hybrid trigger) with the initial index fixed at $v_1 = 0.11$ and recorded, for every competing index $v_g \in \{0.22, 0.33, \ldots, 0.99\}$,

$$\text{ratio1}_g = \frac{\sum_{s=1}^{T} \|\hat{x}_s^1 - \hat{x}_s^g\|^2}{\kappa_g}, \qquad \text{ratio2}_g = \frac{\sum_{s=1}^{T} \left|\phi^{(0)}(\hat{x}_s^1, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s)\right|}{\kappa_g},$$

and we report their maxima over g. Table 1.1 displays the results for the two drift patterns and three noise levels. Each table entry is written as $(v_g^{\max}, \text{ratio})$: the index v_g that attains the maximum and the value of the maximum (in %).[1]

Table 1.1: Maximum trigger-condition ratios for different variation patterns and noise levels.

	Smooth power–law drift			Piecewise–constant jumps		
σ	0.1	0.3	1.0	0.1	0.3	1.0
$T^{0.22}$						
$\max_g(\text{ratio1}_g)$	(0.55, 0.093)	(0.88, 0.286)	(0.88, 2.429)	(0.55, 0.081)	(0.88, 0.277)	(0.88, 2.418)
$\max_g(\text{ratio2}_g)$	(0.33, 0.094)	(0.55, 0.175)	(0.77, 1.454)	(0.33, 0.070)	(0.77, 0.157)	(0.77, 1.451)
$T^{0.33}$						
$\max_g(\text{ratio1}_g)$	(0.55, 0.332)	(0.55, 0.459)	(0.88, 2.540)	(0.55, 0.269)	(0.66, 0.394)	(0.88, 2.510)
$\max_g(\text{ratio2}_g)$	(0.33, 0.382)	(0.33, 0.395)	(0.77, 1.474)	(0.33, 0.306)	(0.33, 0.290)	(0.77, 1.454)

Across every scenario we tested—even the most adverse setting with $\sigma_0 = \sigma_1 = 1.0$ and a genuine variation exponent of $v = 0.33$—the empirical sums never came close to their theoretical ceilings: the first–order statistic topped out at roughly 2.5% of κ_g and its hybrid counterpart at about 1.5%. As expected, larger noise levels and faster environmental drift push the ratios upward, yet they remain comfortably an order of magnitude beneath the bound. In practice the gap widens further because the constant G must itself over-estimate the largest observed gradient; any conservatism in that estimate inflates κ_g and delays switching even more.

To restore the algorithm's ability to react, we scale the confidence bound by a user-chosen factor $c_{thr} \in (0, 1]$ and replace every occurrence of trigger condition thresholds with $c_{thr}\kappa_g$. Extensive experiments (Section 1.4.2) indicate that setting c_{thr} between 0.05 and 0.2 brings the trigger

[1] All experiments use $T = 10\,000$, $\gamma = 1$, and the noise levels $\sigma_0 = \sigma_1 \in \{0.1, 0.3, 1.0\}$.

frequency back to a sensible range without sacrificing the regret guarantees derived in the noisier regimes.

1.4.2 Practical Implementation of ASGD: Tuning the trigger scale c_{thr}

To quantify how the multiplicative scale on the switching threshold affects performance, we focus on a single representative scenario:

$$T = 10{,}000, \quad \text{piecewise–constant jumps, } \nu = 0.33, \ \sigma_0 = \sigma_1 = 0.3.$$

Three non–adaptive baselines anchor the comparison:

(i) ordinary OGD without restarts; (ii) Restarted OGD with a *misspecified* variation exponent ($\nu = 0.66$); (iii) Restarted OGD with the *true* exponent ($\nu = 0.33$), which represents the *oracle* strategy that ASGD should ideally match.

Table 1.2 reports the relative loss $L_\phi^\pi(f, T)$ and the (rounded) switch times produced by **ASGD 1** as the scale c_{thr} is varied (in %).

Table 1.2: Effect of threshold scale on ASGD 1 for $\nu = 0.33$, piecewise–constant drift, $\sigma = 0.3$.

Method	$L_\phi^\pi(f, T)$	Mean switch points
OGD (no restarts)	0.633	—
Restarted OGD, wrong $\nu = 0.66$	0.031	—
Restarted OGD, correct $\nu = 0.33$	**0.021**	—
ASGD 1, $c_{thr} = 0.50$	0.0311	[ ]
ASGD 1, $c_{thr} = 0.20$	0.0272	[4694]
ASGD 1, $c_{thr} = 0.105$	**0.0238**	[2035, 5741]
ASGD 1, $c_{thr} = 0.040$	0.0245	[278, 1729, 3467, 5644, 8600]

When the scale is left at κ_g's value ($c_{thr} = 1$; not shown) or even halved ($c_{thr} = 0.5$), ASGD never fires and simply mimics the misspecified restarted OGD, incurring a loss of 3.11% relative to the oracle. Reducing the threshold by a factor of five ($c_{thr} = 0.2$) allows a *single* switch around

$t \approx 4.7\times10^3$; this already trims the loss by roughly 12.5%. The best performance in this grid arises at $c_{thr} \approx 0.10$: two well–timed switches bring the regret within 13% of the oracle benchmark. Pushing c_{thr} lower still (0.04) triggers five resets; the extra variance from excessive restarts starts to outweigh bias reduction and the loss drifts upward again.

Overall, the experiment confirms the qualitative bias–variance trade-off: large thresholds under-react to drift, small thresholds over-react to noise. In this setting, scaling the theoretical bound by $c_{thr} \in [0.05, 0.15]$ is sufficient to recover nearly oracle performance without hand-tuning to the true v.

Sensitivity to variation intensity and noise level Table 1.3 extends the tuning experiment to (i) two variation exponents, $v \in \{0.22, 0.33\}$, and (ii) three noise levels $\sigma_0 = \sigma_1 \in \{0.1, 0.3, 1.0\}$. For each configuration we report

(a) ordinary OGD, (b) restarted OGD with *misspecified* exponent (over estimated v), (c) restarted OGD with the *true* exponent, (d) ASGD 1 with the theoretical (non-firing) threshold, (e) ASGD 1 and ASGD 2 with a *data-driven* scale c_{thr}. The table lists $(c_{thr}, L_\phi^\pi(f, T))$; the scale that minimizes the loss over a dense grid $c_{thr} \in [0.02, 2]$ (in %). Note that theoretical ASGD 2 is omitted—it coincides with ASGD 1 when the trigger never fires.

When the drift is mild $\big(v = 0.22\big)$ the theoretical ASGD never leaves its most aggressive track $v_1 = 0.11$. Because the resulting bias is already negligible, adjusting the threshold pays off only marginally: the best-tuned scale reduces the loss by at most one part in a thousand. In this low-variation regime the preferred c_{thr} clusters around 0.1 and is virtually insensitive to the level of observation noise.

The situation is markedly different at $v = 0.33$. Here a single well-timed switch can halve the bias, so lowering c_{thr} cuts the loss of ASGD 1 by 22%. The hybrid version ASGD 2, which relies on cost differences rather than decision distances, tracks the oracle a little more closely, especially for the practical noise levels $\sigma = 0.1$ and 0.3.

Table 1.3: Relative loss $L_{\phi}^{\pi}(f, T)$ for piecewise–constant jumps. Each ASGD entry shows $(c_{thr}^{\star}, \text{loss})$.

	Piecewise–constant jumps		
σ	0.1	0.3	1.0
$\mathbf{V_T = T^{0.22}}$			
OGD	0.138	0.137	0.143
Restarted OGD, wrong	0.006	0.021	0.195
Restarted OGD, correct	0.006	0.012	0.073
ASGD 1 (theo)	0.008	0.012	0.055
ASGD 1 (c_{thr})	(0.04, 0.007)	(0.14, 0.012)	(1.2, 0.064)
ASGD 2 (c_{thr})	(0.04, 0.007)	(0.10, 0.012)	(1.0, 0.061)
$\mathbf{V_T = T^{0.33}}$			
OGD	0.634	0.633	0.636
Restarted OGD, wrong	0.032	0.031	0.204
Restarted OGD, correct	**0.013**	**0.021**	0.109
ASGD 1 (theo)	0.028	0.031	0.076
ASGD 1 (c_{thr})	(0.04, 0.014)	(0.105, 0.024)	(0.8, 0.087)
ASGD 2 (c_{thr})	(0.03, 0.014)	(0.065, 0.023)	(0.7, 0.081)

Optimal thresholds depend systematically on noise. The scale that minimizes regret grows from about 0.04 at $\sigma = 0.1$ through $0.1 - 0.14$ at $\sigma = 0.3$ and reaches 1 when $\sigma = 1$. Yet, for any fixed σ the same c_{thr} works well across both variation levels studied, showing that one pilot calibration is sufficient for a broad range of environments.

Overall, reducing the theoretical bound with $c_{thr} \in [0.05, 0.15]$ when $\sigma \leqslant 0.3$—is enough for ASGD 1 and ASGD 2 to approach the oracle performance. Under heavy noise larger thresholds, and thus fewer switches, are preferable, echoing the bias–variance discussion in Section 1.4.1.

2 | Chapter 2: Extension to $L_{p,q}$-Variation Measure

2.1 Introduction

We consider a non-stationary sequential optimization problem where the loss (cost) functions $f_t(x)$ change over time. In this setting, it is natural to measure the magnitude of change of the environment by a variation budget. In particular, we follow the framework of [Chen et al. 2019b] defining an $L_{p,q}$-variation of the function sequence to quantify both spatial and temporal changes. Our focus is on first-order (gradient) feedback in a smooth, strongly convex setting. We design an adaptive stochastic gradient descent (SGD) algorithm that does *not* know the true variation rate in advance. Our main result (Theorem 2.1 below) shows that this algorithm achieves a regret of order

$$O\big(T^{\frac{4p+d}{6p+d}}\, V_T^{\frac{2p}{6p+d}}\, (\log T)^{3/2}\big),$$

where V_T is the total $L_{p,1}$-variation budget. In the limit $p \to \infty$, this recovers the $O(T^{2/3} V_T^{1/3} \log^{3/2} T)$ rate , consistent with the classic bounded-variation ($p = \infty$) case. Our bound is to be compared with the result of Chen et al. (2019) for noisy-gradient feedback, who showed an $O(T V_T^{2p/(4p+d)} \log T)$ regret under strong convexity . Thus we obtain a comparable bound, up to logarithmic factors, via our adaptive SGD scheme under the $L_{p,1}$-variation model.

2.2 Problem Setting

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact convex domain of finite volume. At each time $t = 1, 2, \ldots, T$, the learner selects an action $x_t \in \mathcal{X}$ and incurs loss $f_t(x_t)$, where $f_t : \mathcal{X} \to \mathbb{R}$ is a convex, smooth, and strongly convex cost function (unknown to the learner in advance). We assume an adversarial setting where the sequence $f_1, \ldots, f_T$ may change over time, but the total amount of change is controlled. Following the notation in Chen et al. (2019), we adopt the volume-normalized L_p norm for any measurable $f : \mathcal{X} \to \mathbb{R}$:

$$\|f\|_p = \begin{cases} \left(\frac{1}{\mathrm{vol}(\mathcal{X})} \int_{\mathcal{X}} |f(x)|^p \, dx \right)^{1/p}, & 1 \leqslant p < \infty, \\ \sup_{x \in \mathcal{X}} |f(x)|, & p = \infty. \end{cases}$$

Here $\mathrm{vol}(\mathcal{X})$ is finite since $\mathcal{X}$ is compact. For a sequence of functions $f = (f_1, \ldots, f_T)$, Chen et al. define the $L_{p,q}$-variation as:

$$\mathrm{Var}_{p,q}(f) = \begin{cases} \left(\frac{1}{T} \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right)^{1/q}, & 1 \leqslant q < \infty, \\ \sup_{1 \leqslant t \leqslant T-1} \|f_{t+1} - f_t\|_p, & q = \infty. \end{cases}$$

Intuitively, $\mathrm{Var}_{p,q}(f)$ measures how much the functions change in both the domain (through the L_p norm) and over time (through the ℓ_q norm of differences). In this chapter we specialize to $q = 1$, so that

$$\mathrm{Var}_{p,1}(f) = \frac{1}{T} \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p.$$

We assume a total variation budget of the form

$$V_T = \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p = F_{\max} \cdot T^{\nu},$$

for some exponent $0 \leqslant \nu \leqslant 1$, with $F_{\max} := \sup_{t,x} |f_t(x)|$. Equivalently we set $U_T = V_T/T = F_{\max} T^{\nu-1}$, so that $\mathrm{Var}_{p,1}(f) \leqslant U_T$. Define the function class

$$\mathcal{F}_{p,1}(U_T) := \{ f = (f_1, \ldots, f_T) : \mathrm{Var}_{p,1}(f) \leqslant U_T \}.$$

This captures all loss sequences whose average per-step L_p change is at most U_T. Under this constraint (with $\nu < 1$), sub-linear regret is achievable. In summary, we consider online convex optimization with strongly convex costs, noisy gradient feedback $\phi_t^{(1)}$, and the assumption that $f \in \mathcal{F}_{p,1}(U_T)$.

2.3 Algorithm Description

Our algorithm is an adaptive variant of SGD that does not know the variation exponent ν a priori. We create a grid of hypothetical exponents $\{\nu_k\}_{k=1}^K$ with $\nu_k = k/\log T$, for $k = 1, \ldots, K$, where K is chosen so that $\nu_{K-1} < 1 \leqslant \nu_K$. Each hypothesis k corresponds to an assumed variation level $U_T^k = F_{\max} T^{\nu_k - 1}$ (equivalently $V_T^k = F_{\max} \cdot T^{\nu_k}$) and a window size Δ_T^k. The algorithm maintains K parallel SGD instances ("hypothetical actions") $\{\hat{x}_t^g\}_{g=1}^K$. Each instance g uses a step size $\eta_T^g = \frac{\gamma D}{G\sqrt{\Delta_T^g}}$, where D bounds the diameter of $\mathcal{X}$ and G bounds the gradient norm. At each time t, we update each hypothetical iterate by gradient descent:

$$\hat{x}_t^g = \hat{x}_{t-1}^g - \eta_T^g \, \phi_{t-1}^{(1)}(\hat{x}_{t-1}^g, f_{t-1}), \quad g = 1, \ldots, K,$$

where $\phi_{t-1}^{(1)}(\cdot, f_{t-1})$ denotes the (noisy) gradient of f_{t-1}. We maintain an index k which indicates the current guessed variation level. After each update, we perform a "testing" step: we check if for some $g > k$, the cumulative squared difference between the k-th and g-th trajectories exceeds

a threshold. Concretely, if

$$\sum_{s=t_{\text{if}}}^{t-1} \|\hat{x}_s^k - \hat{x}_s^g\|^2 \;\geqslant\; 2T^{\frac{2pvg+4p+d}{6p+d}} \Big(4GD\sqrt{\log T} + \frac{1}{2}(\tfrac{1}{\gamma}+\gamma)GD + F_{\max}^{\frac{2p}{2p+d}}\Big)\frac{1}{\underline{\delta}},$$

for some $g > k$ (with parameters chosen as in Alg. 3), then we increase $k \leftarrow k+1$ and reset $t_{\text{if}} \leftarrow t$. Intuitively, this test detects when the true variation seems larger than the current hypothesis: if the k-th SGD iterate drifts too far from a higher-variation hypothesis g, we switch to assume a larger v. Finally, after the update and possible switch, we play the action $x_t = \hat{x}_t^k$ and observe its loss $f_t(x_t)$. Algorithm 3 summarizes the procedure.

This multi-hypothesis scheme is inspired by the idea of meta-algorithms for unknown variation: by running parallel learners for different v and monitoring their divergence, we automatically adapt to the true variation without knowing it in advance. The careful choice of threshold ensures that once k exceeds the true underlying variation exponent, further switches will be unlikely (see analysis). The algorithm incurs only polylogarithmic overhead from managing multiple sequences.

With the algorithm defined, we now state the main regret guarantee, which parallels Theorem 3.1 of Chen et al. (2019) for the $L_{p,1}$-variation model. Let $\mathcal{R}^{\pi}_{\phi^{(1)}}(T, U_T)$ denote the worst-case regret of policy π under first-order feedback and variation U_T. We show:

Theorem 2.1. *Suppose the cost functions are smooth and strongly convex (as per Assumptions (A1)–(A5) of Chen et al. (2019)), and the algorithm receives noisy gradient feedback $\phi_t = \phi_t^{(1)}$. Then Algorithm 3 (denoted π_3) guarantees*

$$\mathcal{R}^{\pi_3}_{\phi^{(1)}}(T, U_T) \;=\; O\Big(T \cdot U_T^{\frac{2p}{6p+d}} (\log T)^{3/2}\Big) \;=\; O\Big(T^{\frac{4p+d}{6p+d}} \cdot V_T^{\frac{2p}{6p+d}} (\log T)^{3/2}\Big),$$

for any variation class $\mathcal{F}_{p,1}(U_T)$. In other words, the dynamic regret scales as $O(T^{(4p+d)/(6p+d)} V_T^{2p/(6p+d)} \log^{3/2} T)$.

Note that when $p \to \infty$, we recover the exponent $2p/(6p+d) \to 1/3$, yielding a bound

Algorithm 3 Adaptive SGD with First-Order Feedback

Input: hypothetical variation parameters $\{v_g : 1 \leqslant g \leqslant K\}$, hypothetical window sizes $\{\Delta_T^g : g = 1, 2, \ldots, K\}$, upper bound G, diameter D, tuning parameter γ
Initialization: Set $k = 1$, $t_{\mathbf{if}} = 1$, arbitrarily set $x_1 \in \mathscr{X}$ and $\hat{x}_1^g \in \mathscr{X}$ for each $g \in \{1, 2, \ldots, K\}$.
for $t = 1$ to T **do**
  if $t \geqslant 2$ **then**
    if for some $g \in \{k+1, k+2, \ldots, K\}$,

$$\sum_{s=t_{\mathbf{if}}}^{t-1} ||\hat{x}_s^k - \hat{x}_s^g||^2 \geq 2T^{\frac{2pv_g+4p+d}{6p+d}} \left(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma} + \gamma)GD + F_{\max}^{\frac{2p}{2p+d}}\right) \frac{1}{\underline{\delta}}$$

    then $k \leftarrow k+1$ and $t_{\mathbf{if}} \leftarrow t$.
    end if
    for $g = 1$ to K **do**
      Observe the gradient at the g-th hypothetical action $G_t(\hat{x}_{t-1}^g) := \phi_{t-1}^{(1)}(\hat{x}_{t-1}^g, f_{t-1})$;
      Compute the g-th hypothetical step size $\eta_T^g = \frac{\gamma D}{G\sqrt{\Delta_T^g}}$;
      Update the g-th hypothetical action $\hat{x}_t^g = \hat{x}_{t-1}^g - \eta_T^g \cdot G_t(\hat{x}_{t-1}^g)$;
    end for
  end if
  Choose the k-th hypothetical action $x_t = \hat{x}_t^k$ and incur the loss $C_t(\hat{x}_t^k) := \phi_t^{(0)}(\hat{x}_t^k, f_t)$.
end for

$O(T^{2/3}V_T^{1/3}(\log T)^{3/2})$, which matches the known $L_{\infty,1}$-variation result. This is consistent with the special case result of Besbes et al. (2015) as recovered by Chen et al.. (For finite p, our exponent $(4p+d)/(6p+d)$ is larger than $1/2$, indicating a dependence on the domain dimension; this "curse of dimensionality" is noted in [Chen et al.] for $p < \infty$.)

2.4 Regret Analysis

As preparations, we state the following result established in Lemma 4.3, [Chen et al. 2019b].

Lemma 2.2. *[Lemma 4.3, [Chen et al. 2019b]] With $r = \frac{2p}{2p+d}$, suppose*

$$\max_{1\leqslant \ell\leqslant J}|B_\ell| \leqslant \Delta_T + 1, \quad 1 \leqslant q \leqslant \infty, \quad \textit{and} \quad \mathrm{Var}_{p,q}(f) \leqslant U_T.$$

Then

$$\sum_{\ell=1}^{J}\Big(\sum_{t=\underline{b}_\ell}^{\bar{b}_{\ell-1}}\|f_{t+1}-f_t\|_p\Big)^r \leqslant \Delta_T^{r-\frac{r}{q}}\, J^{1-\frac{r}{q}}\, T^{\frac{r}{q}}\, U_T^r.$$

Proof of Theorem 2.1. Compared with the proof of Theorem 1.2, the major changes lie in the analysis of Step 1, which we elaborate as follows.

We Similar to equation (1.3), for each for each $g = k^*, k^*+1, \ldots, K$ and $j = 1, 2, \ldots, \lceil T/\Delta_T^g \rceil$,

we start from the following decomposition:

$$\begin{aligned}\sum_{t=1}^{T}||\hat{x}_t^g - x_t^*||^2 &\leqslant \frac{1}{\underline{\delta}}\sum_{t=1}^{T}\Big(f_t(\hat{x}_t^g) - f_t(x_t^*)\Big)\\ &= \frac{1}{\underline{\delta}}\sum_{j=1}^{\lceil T/\Delta_T^g\rceil}\Bigg(\underbrace{\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}}\sum_{t\in\mathcal{T}_{g,j}} f_t(w)}_{\text{①: regret relative to batch } j\text{'s single best action}}\Bigg)\\ &\quad + \frac{1}{\underline{\delta}}\sum_{j=1}^{\lceil T/\Delta_T^g\rceil}\underbrace{\Big(\min_{w\in\mathcal{X}}\sum_{t\in\mathcal{T}_{g,j}} f_t(w) - \sum_{t\in\mathcal{T}_{g,j}} f_t(x_t^*)\Big)}_{\text{②: regret due to functional changes}}. \end{aligned} \tag{2.1}$$

For the first term, we still have the following high-probability event as (1.4): with probability at least $1 - \lceil T/\Delta_T^g\rceil\delta \geqslant 1 - 1/T$,

$$\sum_{j=1}^{\lceil T/\Delta_T^g\rceil}\Big(\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}}\sum_{t\in\mathcal{T}_{g,j}} f_t(w)\Big) \leqslant \left\lceil\frac{T}{\Delta_T^g}\right\rceil GD\sqrt{\Delta_T^g}\left(4\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma}+\gamma)\right). \tag{2.2}$$

By applying ineuqality (16) and Lemma 2.2 to the second term (2.1), we have the following inequality:

$$\sum_{j=1}^{\lceil T/\Delta_T^g\rceil}\left(\min_{w\in\mathcal{X}}\sum_{t\in\mathcal{T}_{g,j}} f_t(w) - \sum_{t\in\mathcal{T}_{g,j}} f_t(x_t^*)\right) \leqslant \left(\left(\frac{T}{\Delta_T^g}\right)^{1-\frac{r}{q}}(\Delta_T^g)^{r-\frac{r}{q}}T^{\frac{r}{q}}(U_T^g)^r\right).$$

By setting $\Delta_T^g := (T^{v_g-1})^{-r/(r+\frac{1}{2})} = (T^{v_g-1})^{-\frac{4p}{6p+d}} = T^{-\frac{4p(v_g-1)}{6p+d}}$ to balance the two terms in the above inequalities, we obtain the following high-probability event:

$$\begin{aligned}\sum_{t=1}^{T}||\hat{x}_t^g - x_t^*||^2 &\leq T\cdot(U_T^g)^{\frac{2p}{6p+d}}\big(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma}+\gamma)GD + F_{\max}^{\frac{2p}{2p+d}}\big)\frac{1}{\underline{\delta}},\\ &= T^{\frac{2pv_g+4p+d}{6p+d}}\big(4GD\sqrt{\log T} + \frac{1}{2}(\frac{1}{\gamma}+\gamma)GD + F_{\max}^{\frac{2p}{2p+d}}\big)\frac{1}{\underline{\delta}}.\end{aligned}$$

With the above inequality, the subsequent analysis in Steps 2 and 3 is similar to those in the proof of Theorem 1.2. We omit the details here. □

3 | Chapter 3: Application to Inventory Problem

3.1 Introduction

Modern supply chains operate in environments where demand can fluctuate unpredictably due to seasonality, marketing campaigns, macro-economic shifts, or rapid changes in consumer preferences. For a firm managing inventory, failing to account for such non-stationarity can lead to chronic overstocking or frequent stockouts—both costly outcomes. Consequently, there is a pressing need for *data-driven* inventory policies that (i) learn demand patterns as they evolve, (ii) respond swiftly to distributional shifts, and (iii) remain robust even when the *extent* and *timing* of those shifts are unknown.

Motivation. Traditional models often assume demand is independent and identically distributed over time. While analytically convenient, this assumption rarely holds in practice and can severely degrade performance when demand drifts. A principled way to capture non-stationarity is to bound the *total variation* of demand distributions over a planning horizon. In this work we measure variation via the Wasserstein distance, which provides an intuitive, geometry-aware metric for how much successive demand distributions differ. Crucially, we do *not* assume the decision maker knows the magnitude of this variation budget in advance. The challenge, therefore, is to design an online algorithm that adapts automatically—achieving strong performance in tranquil

periods yet remaining agile when demand becomes volatile. In the remainder of this section, we first summarize our main results and contributions, and present a brief literature review in Section 3.1.2.

3.1.1 Main Results and Contributions.

We develop an *Adaptive Stochastic Gradient Descent (ASGD)* policy for a single non-perishable product over a finite horizon of T periods, and later extend it to a multi-product system with a shared capacity constraint. Our key findings are:

(i) *Algorithmic innovation.* ASGD blends classical stochastic gradient steps with a data-driven restarting rule that automatically calibrates its learning rate to the (unknown) pace of demand change. No prior tuning for the variation budget is required.

(ii) *Regret guarantee.* Without any knowledge of the true budget $B_T = O(T^\nu)$, ASGD attains a worst-case regret of $\widetilde{O}\big(T^{\frac{2+\nu}{3}}\big)$ against a clairvoyant policy that foresees every demand realization. The bound is *sublinear* in T for every $\nu < 1$, proving that per-period regret vanishes asymptotically even under polynomially growing non-stationarity.

3.1.2 Literature Review

There is a large body of literature in the area of data-driven inventory control that develop online learning algorithms for inventory models with *i.i.d.* demand distribution. Earlier studies analyze the repetitive newsvendor problem (see, e.g., Huh and Rusmevichientong 2009, Huh et al. 2011, Besbes and Muharremoglu 2013, Levi et al. 2007, Levi et al. 2015). Later more complicated systems are studied, e.g., lost-sales inventory systems with lead times (Huh et al. 2009, Zhang et al. 2020, Agrawal and Jia 2019), perishable inventory system (Zhang et al. 2018), inventory system with random capacity (Chen et al. 2020b), inventory system with fixed ordering cost (Yuan et al. 2021), and joint pricing and inventory system (Chen et al. 2019a, Chen et al. 2021, Chen et al.

Table 3.1: Comparison with Literature on Online Learning for Inventory Models with Non-stationary Demands
(For papers studying both settings without and with inventory carryover, we only list the results for the latter.)

Literature	Single Product	Multiple Products	Inventory Carryover	Censored Demand	Changes in Distribution	Assumption of Non-stationarity	Regret Upper Bound
[Cheung et al. 2019b]	✓		✓	✗	✓	Unknown variation budget T^{ν}	$\tilde{O}(T^{\frac{2+\nu}{3}})$
[Chen 2021]	✓		✓	✓	✓	Unknown change-points[1]	$\tilde{O}(T^{\frac{1}{2}})$
[Keskin et al. 2021]	✓		✗	✗	✗	Known variation budget T^{ν}	$\tilde{O}(T^{\frac{1+\nu}{2}})$
[Gong and Simchi-Levi 2021]	✓	✓	✓ ✓	✗ ✗	✓	Cyclic demand	$\tilde{O}(T^{\frac{1}{2}})$ $\tilde{O}(T^{\frac{5}{6}})$
[Keskin et al. 2022]	✓		✓	✗	✗	Unknown change-points[2]	$\tilde{O}(T^{\frac{2}{3}})^{3}$ $\tilde{O}(T^{\frac{1}{2}})^{4}$
[An et al. 2025]	✓		✗	✗	✗	Unknown variation budget T^{ν}	$\tilde{O}(T^{\frac{3+\nu}{4}})$
This paper	✓		✓	✗	✓	Unknown variation budget T^{ν}	$\tilde{O}(T^{\frac{2+\nu}{3}})$
This paper		✓	✓	✗	✓	Unknown variation budget T^{ν}	$\tilde{O}(T^{\frac{2+\nu}{3}})$

[1] The paper also assumes that there are at most $O(\log T)$ changes of demand distributions;
[2] The paper also assumes that there is a minimum shift each time in the changes;
[3] This bound is proven under the non-parametric setting;
[4] This bound is proven under the parametric setting.

2020a, Chen et al. 2022). Among these works, the stochastic gradient descent method has been adopted by many papers, e.g., [Huh and Rusmevichientong 2009]. In particular, [Shi et al. 2016] considering a multi-product inventory control problem under a warehouse-capacity constraint is the study most related to this project. The authors develop a learning algorithm based on SGD method and prove the regret upper bound $O(\sqrt{T})$. The major difference of this paper compared to [Shi et al. 2016] is that they assume demand distributions are *i.i.d.* whereas we consider a non-stationary environment.

There are also a growing body of literature going beyond the *i.i.d.* assumption to study data-driven inventory models with non-stationary demands. [Chen 2021] considers an inventory control problem in a shifting demand environment with the number of changes in demand distribu-

tions at most $O(\log T)$. The author proves a regret lower bound $\Omega(\sqrt{T})$ and construct a learning algorithm with regret upper bound $\widetilde{O}(\sqrt{T})$. [Keskin et al. 2022] study a joint pricing and inventory problem for a perishable product, and construct algorithms for both the settings with non-parametric and parametric noise distributions, with the regret upper bound $\widetilde{O}(T^{2/3})$ and $\widetilde{O}(T^{1/2})$ respectively. [Keskin et al. 2021] study a repetitive newsvendor problem with a time-varying mean demand level. The authors design a moving window ordering policy and prove the regret upper bound $O(\sqrt{T})$ under constant variation budget and $O(T^{(1+v)/2})$ under $O(T^v)$ variation budget. [Cheung et al. 2019b] study the non-stationary reinforcement learning with a single-product lost-sales inventory model as an application. [Gong and Simchi-Levi 2021] apply the Q-learning technique to analyze inventory models with unknown cyclic demands and study the single-product lost-sales model with zero lead time and multi-product backlogging model with positive lead times. [Ding et al. 2021] consider a stochastic inventory system where demand distributions are feature-dependent and thus non-stationary. They design two algorithms based on SGD and prove the regret upper bound $O\sqrt{T})$ for both algorithms. [An et al. 2025] study XXXX

Among these studies, [Chen 2021], [Cheung et al. 2019b], [Gong and Simchi-Levi 2021] (in their first model), and [Ding et al. 2021] assume censored demand, whereas [Keskin et al. 2022], [Keskin et al. 2021] and [Gong and Simchi-Levi 2021] (in their second model) assume full demand observations. Except [Gong and Simchi-Levi 2021] (in their second model), all papers focus on single product problems. The paper is the first one that considers multi-product inventory control in a non-stationary environment.

3.2 Model Formulation

3.2.1 Notation

Throughout this paper, $\mathbb{R}_+^n$ denotes the set of n-dimensional non-negative vectors. For $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m$, we write $\mathbf{x} \geqslant \mathbf{y}$ if and only if $x_i \geqslant y_i$ for every $i \in [m]$, where $[n] := \{1, \ldots, n\}$. For any compact set $\mathcal{X} \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, let $\mathrm{Proj}_{\mathcal{X}}(x) := \arg\min_{y \in \mathcal{X}} \|y - x\|_2^2$ be the Euclidean projection of x onto $\mathcal{X}$.

3.2.2 Single-Product Inventory System

We investigate a finite-horizon stochastic inventory problem for a *single, non-perishable* product over T periods. Demand in period $t \in [T]$ is an independent random variable D_t with (possibly distinct) cumulative distribution function (c.d.f.) $F_t(\cdot)$. The sequence $\{F_t\}_{t=1}^T$ may evolve arbitrarily over time; precise notions of non-stationarity are introduced in Section 3.2.3.

Timing within each period. For every $t = 1, \ldots, T$ the following events occur:

(i) Review. The firm observes the on-hand inventory $x_t \in \mathbb{R}_+$ at the start of period t. We set $x_1 = 0$ without loss of generality.

(ii) Replenishment. An order quantity $q_t \in \mathbb{R}_+$ is placed and received instantaneously, yielding post-order inventory $y_t := x_t + q_t$.

(iii) Demand realization. Demand D_t is revealed and met up to the available stock y_t. Unsatisfied demand is lost.[1] We assume full demand observation, consistent with Keskin et al. [2021], Keskin et al. [2022], and others.

[1] Our results also hold under backlogging with zero lead time.

(iv) Cost and carry-over. Surplus inventory incurs holding cost h per unit, while lost sales incur shortage cost b per unit. Remaining inventory is carried into the next period, i.e.,

$$x_{t+1} \;=\; \big(y_t - D_t\big)^+.$$

Single-period cost. Let $C(y, D) := h\,(y-D)^+ + b\,(D-y)^+$ be the cost when inventory is y and realized demand is D. Define the *expected* cost in period t under inventory level y by

$$C_t(y) \;:=\; \mathbb{E}_{D_t \sim F_t}\big[C(y, D_t)\big].$$

Information structure and policies. At the start of period t, the decision maker observes the history

$$H_t := \big(x_1, q_1, D_1, \ldots, x_{t-1}, q_{t-1}, D_{t-1}, x_t\big).$$

An *admissible policy* π is a sequence of measurable maps $\{\pi_t\}_{t=1}^T$ with $\pi_t : H_t \mapsto q_t^\pi \in \mathbb{R}_+$. Let Π be the set of all such policies. Given $\pi \in \Pi$, the total expected cost over the horizon equals

$$\mathbb{E}\Big[\sum_{t=1}^{T} C\big(y_t^\pi, D_t\big)\Big] \;=\; \mathbb{E}\Big[\sum_{t=1}^{T} h\big(y_t^\pi - D_t\big)^+ + b\big(D_t - y_t^\pi\big)^+\Big],$$

where expectations are taken with respect to all randomness up to T. The firm's objective is to choose $\pi \in \Pi$ minimizing this quantity.

3.2.3 Non-Stationary Demand and the Variation Budget

Demand distributions may drift over time. We measure this evolution by the *temporal variation*

$$TV(F_{1:T}) \;:=\; \sum_{t=2}^{T} \mathcal{W}\big(F_t, F_{t-1}\big),$$

where $\mathcal{W}$ is the 1-Wasserstein distance: for any c.d.f.'s μ and ν,

$$\mathcal{W}(\mu, \nu) \;:=\; \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\big[|X - Y|\big],$$

and $\Gamma(\mu, \nu)$ is the set of couplings of (μ, ν). The Wasserstein metric is widely used in distributionally-robust optimization and captures the "transport cost" of moving mass from one distribution to another.

Variation budget. Fix $B_T > 0$ (possibly growing with T) and define the *uncertainty set*

$$\mathcal{V}(T, B_T) \;:=\; \big\{F_{1:T} : TV(F_{1:T}) \leqslant B_T\big\}.$$

We assume $B_T = M \cdot T^v$ for some $v \in [0, 1]$ but stress that its exact value is *unknown* to the decision maker, where M is the upper bound for order-up-to level (introduced later in Assumption 3).

Why Wasserstein? Compared with ℓ^p distances, the Wasserstein metric (i) aligns with the economic intuition of transporting demand "mass" and (ii) is sensitive to the geometry of the distributions, making it well suited for inventory settings. All results extend to other metrics; we focus on Wasserstein for concreteness.

Lemma 3.1. *For any* $1 \leqslant t_1 < t_2 \leqslant T$,

$$\sup_{y \geqslant 0} \big| C_{t_2}(y) - C_{t_1}(y) \big| \;\leqslant\; \max(h, b) \sum_{t=t_1+1}^{t_2} \mathcal{W}\big(F_t, F_{t-1}\big).$$

Lemma 3.1 shows that larger cumulative distributional change translates directly into greater divergence between single-period cost functions. Its proof (Appendix B) relies on the closed-form expression of the one-dimensional Wasserstein distance together with the piecewise-linear structure of the holding–shortage cost.

3.2.4 Worst–Case Regret

Let π^* denote the *clairvoyant* policy that knows the entire sequence of demand distributions $\{F_t\}_{t=1}^T$ in advance and solves the corresponding dynamic program. For any admissible policy $\pi \in \Pi$, its (instance-dependent) regret is

$$R^\pi(T, \mathbf{F}_{1:T}) \;:=\; \mathbb{E}\Big[\, \textstyle\sum_{t=1}^T \big(C(y_t^\pi, D_t) - C(y_t^{\pi^*}, D_t)\big)\Big],$$

where the expectation is over all randomness up to period T. The *worst-case regret* under an unknown variation budget is

$$R^\pi(T, B_T) \;:=\; \sup_{\mathbf{F}_{1:T} \in \mathcal{V}(T, B_T)} R^\pi(T, \mathbf{F}_{1:T}),$$

with $\mathcal{V}(T, B_T)$ defined in (**??**). Our goal is to design a policy π that keeps $R^\pi(T, B_T)$ as small as possible—even though the exact magnitude of B_T is *unknown* to the firm.

Standing Assumptions

Let $q_t^* := \arg\min_{y \geqslant 0} C_t(y)$ be the myopic newsvendor solution in period t.

Assumption 3. *(i)* ***Bounded order-up-to level.*** *There exists a known constant $M > 0$ such that $q_t^* \in [0, M]$ for all $t \in [T]$.*

(ii) ***Bounded demand moments.*** *There are constants $\underline{D}, \overline{D} > 0$ with $\min_{t \in [T]} \mathbb{E}[D_t] \geqslant \underline{D}$ and $\max_{t \in [T]} \mathbb{E}[D_t^4] \leqslant \overline{D}$.*

Assumption 3 (i) is standard in data-driven inventory control and can be enforced by selecting a sufficiently large M. Condition (ii) imposes mild moment bounds satisfied by most light-tailed demand models and is needed only for concentration arguments in the regret analysis.

3.3 Algorithm Description

We now address the single–product inventory problem under the non-stationary demand model of Section 3.2. Before detailing the learning algorithm we outline the key design challenges.

3.3.1 Design Challenges

(i) Unknown change points and shift magnitudes. Demand distributions may alter at arbitrary times and by arbitrary amounts. Because the timing and size of each shift are unobservable ex-ante, any policy must *detect* and *react* to changes purely from cost feedback.

(ii) Unknown variation budget B_T. Although the total variation satisfies $B_T = O(T^{v})$ for some $v \in [0, 1]$, the firm does not know v a priori. Thus the policy must perform well simultaneously across a continuum of possible non-stationarity levels.

3.3.2 Adaptive SGD Policy

To hedge against the unknown v, we consider a geometric grid

$$\mathcal{V} = \left\{ v_k : \; v_k = \tfrac{k}{\log T}, \; k = 1, \ldots, K \right\},$$

where K is the smallest index with $v_{K-1} < 1 \leqslant v_K$. At each time t the algorithm assumes some candidate $v_k \in \mathcal{V}$ is correct and sets the SGD step size

$$\eta_t^k \;=\; \frac{\gamma\, M}{\max\{h, b\}}\, T^{-\frac{1-v_k}{3}},$$

where M is a known upper bound on inventory and $\gamma > 0$ is a tuning constant. If the cumulative squared deviation between the current iterate and that of a more reactive grid point $v_g \,(> v_k)$

exceeds a carefully chosen threshold, the algorithm "switches" to v_g. Full pseudocode appears in Algorithm 4.

Algorithm 4 Adaptive Stochastic Gradient Descent (ASGD) Algorithm

Input: candidate variation parameters $\{v_g : 1 \leqslant g \leqslant K\}$, candidate step sizes $\{\eta_T^g : g = 1, 2, \ldots, K\}$, inventory upper bound M, cost parameters h and b, tuning parameter $\gamma > 0$.
Initialization: Set $k = 1$, $t_{\text{if}} = 1$, $\hat{y}_1 = 0$ and $\hat{y}_1^g = 0$ for each $g \in \{1, 2, \ldots, K\}$.
for $t = 1$ **to** T **do**
  if $t \geq 2$ **then**
    if for some $g \in \{k+1, k+2, \ldots, K\}$,

$$\sum_{s=t_{\text{if}}}^{t-1} C(\hat{y}_t^k, D_t) - C(\hat{y}_t^g, D_t) \geq T^{\frac{2+v_g}{3}} \left[4\max\{h, b\}M + 2\left(C_0\sqrt{\log T} + C_1\right)\right] \; + \; 4\,C_3\,\sqrt{2T\log(T^2)},$$

    then $k \leftarrow k+1$ and $t_{\text{if}} \leftarrow t$.
    for $g = k$ to K **do**
      Compute the g-th candidate gradient $G_{t-1}(\hat{y}_{t-1}^g)$:

$$G_{t-1}(\hat{y}_{t-1}^g) = \begin{cases} h, & \text{if } \hat{y}_{t-1}^g > D_{t-1}, \\ -b, & \text{otherwise.} \end{cases}$$

      Compute the g-th candidate target order-up-to level $\hat{y}_t^g$:

$$\hat{y}_t^g = \hat{y}_{t-1}^g - \eta_T^g G_{t-1}(\hat{y}_{t-1}^g).$$

    end for
  end if
  Implement the k-th target order-up-to level and raise the inventory to $y_t = \max\{\hat{y}_t^k, x_t\}$.
  Observe demand D_t and update the inventory level $x_{t+1} = (y_t - D_t)^+$.
end for

Theorem 3.2. *Fix any horizon $T \geqslant 1$ and variation budget $B_T = O(T^v)$ with unknown $v \in [0, 1]$. Under Assumptions 3, Algorithm 4 achieves*

$$R^\pi(T, B_T) \;\leqslant\; C\,T^{\frac{2+v}{3}}\log^{\frac{3}{2}} T,$$

for some constant C depending only on $h, b, \underline{\delta}, \gamma, M$.

Rate optimality. The leading term $T^{(2+\nu)/3}$ matches the lower bounds known for non-stationary stochastic optimization with unknown variation (cf. Besbes et al. 2015). Thus ASGD is near-minimax up to logarithmic factors.

Adaptivity. By monitoring the squared gap between candidate trajectories, the algorithm automatically tightens its step size when demand proves volatile and relaxes when the environment is stable—without ever estimating ν explicitly.

Efficiency. Each iteration uses only the sign of the inventory error (via the gradient), requires $O(K)$ book-keeping, and performs a single projection, making the method scalable in practice.

Generality. The grid-based restart scheme extends seamlessly to other convex cost structures and, as shown later, to a multi-product system with a capacity constraint while preserving the same regret order.

3.4 Analysis of Regret Bound

We first notice the following upper bound on the regret of ASGD algorithm:

$$\begin{aligned} R^{ASGD}(T, B_T, \mathbf{F}_{1:T}) &= \mathbb{E}\Big[\sum_{t=1}^{T}\Big(C\big(y_t^{ASGD}, D_t\big) - C\big(y_t^{\pi^*}, D_t\big)\Big)\Big] \\ &\leqslant \mathbb{E}\left[\sum_{t=1}^{T}\Big(C\big(y_t^{ASGD}, D_t\big) - C\big(q_t^*, D_t\big)\Big)\right] \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\Big(C_t(\hat{y}_t^{ASGD}) - C_t(q_t^*)\Big)\right]}_{\text{Regret due to non-stationary SGD}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\Big(C_t(y_t^{ASGD}) - C_t(\hat{y}_t^{ASGD})\Big)\right]}_{\text{Regret due to inventory carry-over}}, \end{aligned} \tag{3.1}$$

where the inequality holds because for each period $t \in [T]$, the minimum possible expected cost incurred by any policy, including the true optimal policy π^*, is at least the newsvendor cost $\min_{y\geqslant 0}\mathbb{E}[C(y, D_t)]$, which, from the definition of q_t^*, equals $\mathbb{E}[C(q_t^*, D_t^*)]$. Therefore, it suffices to bound the two terms in the RHS of (4.1). We establish the upper bounds on these two terms

for the ASGD algorithm in the following two propositions. Combining these two propositions with (4.1), we complete the proof of Theorem 3.2.

Proposition 3.3 (Regret due to Non-stationary SGD). *Under Assumption 3 (i),*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(C_t(\hat{y}_t^{ASGD}) - C_t(q_t^*)\right)\right] = O\left(T^{\frac{2+v}{3}}\log^{\frac{3}{2}} T\right).$$

Proposition 3.4 (Regret due to Inventory Carryover). *Under Assumption 3 (i) and (ii),*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(C_t(y_t^{ASGD}) - C_t(\hat{y}_t^{ASGD})\right)\right] = O\left(T^{\frac{2+v}{3}}\log T\right).$$

3.4.1 Sketched Proof of Proposition 3.3.

We begin by defining a discrete set of candidate variation parameters $\mathcal{V} = \{v_1, \ldots, v_K\}$ as before, and at each period t, the policy assumes a variation parameter $v_k \in \mathcal{V}$ and sets the corresponding Online Gradient Descent (OGD) step size:

$$\eta_t^k = \frac{\gamma M}{\max\{h, b\} \cdot T^{\frac{1-v_k}{3}}} \quad \text{for some } \gamma > 0.$$

Consider a realization path of demands $\{D_1, D_2, \ldots, D_t, \ldots\}$. The constrained target inventory levels for each candidate v_k are denoted by $\{\hat{y}_t^k\}_{t=1}^T$.

We observe that $K = O(\log T)$. Define k^* as the smallest index in $\{1, \ldots, K\}$ such that $v \le v_{k^*}$. Given the relation $v_{k^*} = v_{k^*-1} + \frac{1}{\log T}$, it follows that $v \le v_{k^*} < v + \frac{1}{\log T}$. Consequently, $T^v \le T^{v_{k^*}} < eT^v$, implying that $T^{v_{k^*}}$ is within a constant factor of T^v. Therefore, it is sufficient to bound the total regret by

$$O\left(T^{\frac{2+v_{k^*}}{3}}\log^{\frac{3}{2}} T\right).$$

To establish this, we first demonstrate that (with high probability) throughout the algorithm,

the running index k never exceeds k^*. To illustrate this, consider the following steps:

Step 1. For every $k \geq k^*$, with high probability for T time periods (refer to the high probability proof for OGD in a later section), we have

$$\sum_{t=1}^{T}\left[C(\hat{y}_t^k, D_t) - C(q_t^*, D_t)\right] \leqslant T^{\frac{2+\nu_k}{3}}\left[2\max\{h,b\}M + \left(C_0\sqrt{\log T} + C_1\right)\right] + 2\,C_3\,\sqrt{2T\log(2T)}.$$

We may assume that this event holds from this point onward.

Step 2. Contradiction via empirical-cost differences. Suppose, for the sake of contradiction, that at some period t we have $k \geqslant k^*$ and the **if**-condition fires with some $g > k$. Then by the triangle inequality

$$\sum_{s=t_{\text{if}}}^{t}\left|C(\hat{y}_s^k, D_s) - C(\hat{y}_s^g, D_s)\right| \leqslant \sum_{s=t_{\text{if}}}^{t}\left|C(\hat{y}_s^k, D_s) - C(q_s^*, D_s)\right| + \sum_{s=t_{\text{if}}}^{t}\left|C(\hat{y}_s^g, D_s) - C(q_s^*, D_s)\right|.$$

But from our high-probability bound on each "arm" $j \in \{k, g\}$,

$$\sum_{s=1}^{T}\left|C(\hat{y}_s^j, D_s) - C(q_s^*, D_s)\right| \leqslant T^{\frac{2+\nu_g}{3}}\left[2\max\{h,b\}M + \left(C_0\sqrt{\log T} + C_1\right)\right] + 2\,C_3\,\sqrt{2T\log(2T)},$$

it follows that even summing only from $s = t_{\text{if}}$ to t cannot exceed

$$T^{\frac{2+\nu_g}{3}}\left[4\max\{h,b\}M + 2\left(C_0\sqrt{\log T} + C_1\right)\right] + 4\,C_3\,\sqrt{2T\log(2T)},$$

This is exactly the threshold the **if**-condition would require, so no such trigger can occur when $k \geqslant k^*$. Hence k never surpasses k^*, completing the contradiction.

Step 3. Bounding regret between successive triggers. Let t' and t'' be two consecutive times when the **if**-condition fires. On the intervening interval $t' + 1 \leqslant t \leqslant t'' - 1$, the **if**-test fails even

for $g = k^*$, so by hypothesis

$$\sum_{t=t'+1}^{t''-1} \left[C(\hat{y}_t^k, D_t) - C(\hat{y}_t^{k^*}, D_t)\right] \;<\; T^{\frac{2+v_{k^*}}{3}} \left[4\max\{h,b\}M + 2\big(C_0\sqrt{\log T} + C_1\big)\right] \;+\; 4\,C_3\,\sqrt{2T\log(2T)}.$$

Now split the extra regret against the oracle q_t^* by the triangle inequality:

$$\sum_{t=t'+1}^{t''-1} \left[C(\hat{y}_t^k, D_t) - C(q_t^*, D_t)\right] \;\leqslant\; \sum_{t=t'+1}^{t''-1} \left|C(\hat{y}_t^k, D_t) - C(\hat{y}_t^{k^*}, D_t)\right| \;+\; \sum_{t=t'+1}^{t''-1} \left|C(\hat{y}_t^{k^*}, D_t) - C(q_t^*, D_t)\right|.$$

The first sum is bounded by the failed-trigger threshold above. The second sum, over at most T periods, is bounded by the high-probability empirical-regret of trajectory k^*:

$$\sum_{t=1}^{T} \left|C(\hat{y}_t^{k^*}, D_t) - C(q_t^*, D_t)\right| \;\leqslant\; T^{\frac{2+v_{k^*}}{3}} \left[2\max\{h,b\}M + \big(C_0\sqrt{\log T} + C_1\big)\right] \;+\; 2\,C_3\,\sqrt{2T\log(2T)}.$$

Putting these together shows that on each interval between triggers,

$$\sum_{t=t'+1}^{t''-1} \left[C(\hat{y}_t^k, D_t) - C(q_t^*, D_t)\right] \leq T^{\frac{2+v_{k^*}}{3}} \left[6\max\{h,b\} + 3\big(C_0\sqrt{\log T} + C_1\big)\right] \;+\; 6\,C_3\,\sqrt{2T\log(2T)}.$$

By taking expectation of both sides, we have

$$\sum_{t=t'+1}^{t''-1} \left[C_t(\hat{y}_t^k) - C_t(q_t^*)\right] = O\big(T^{\frac{2+v_{k^*}}{3}}\sqrt{\log T}\big).$$

Since the **if**-condition can fire at most $K = \log T$ times, the total interval-wise regret (ignoring carry-over) is

$$O\big(T^{\frac{2+v_{k^*}}{3}}\log^{\frac{3}{2}} T\big).$$

Step 4. Suppose that the last **if** condition happens at time t_{last}, then the regret between time 1 and t_{last} is well controlled by above steps. In this step, we can upper bound the regret from time t_{last} to T similarly to step 3, and hence finish the proof for Proposition 3.3.

3.4.2 Sketched Proof of Proposition 3.4.

Due to inventory carryover, the target order-up-to level by SGD estimators may not always be achieved in all periods. In other words, the true inventory level y_t^{ASGD} may be strictly higher than the target order-up-to level $\hat{y}_t^{ASGD}$ and we refer to their difference as the overshooting. From the Lipschitz continuity of the newsvendor cost, we have the following upper bound on the regret due to inventory carryover:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(C_t(y_t^{ASGD}) - C_t(\hat{y}_t^{ASGD})\right)\right] \leqslant (h \vee b)\mathbb{E}\left[\sum_{t=1}^{T}\left(y_t^{ASGD} - \hat{y}_t^{ASGD}\right)\right]. \tag{3.2}$$

Therefore, it suffices to bound the expected overshooting $\mathbb{E}[\sum_{t=1}^{T}(y_t^{ASGD} - \hat{y}_t^{ASGD})]$.

Similar to the analysis of Theorem 6 in [Huh and Rusmevichientong 2009], we have the following inequality for the overshooting:

$$y_{t+1}^{ASGD} - \hat{y}_{t+1}^{ASGD} \leq \left(y_t^{ASGD} - \hat{y}_t^{ASGD} + h\eta_t^{k(t)} - D_t\right)^+. \tag{3.3}$$

Here, we denote $\eta_t^{k(t)} = \frac{\gamma M}{h \vee b} T^{\frac{\nu_{k(t)}-1}{3}}$. Following the analysis of [Huh and Rusmevichientong 2009], we next construct an auxiliary overshooting process $\{W_t : t \geqslant 1\}$ to bound the original overshooting. Let $W_1 = 0$ and for each $t \geqslant 2$, define $W_t = (W_{t-1} + \gamma - D_t)^+$. Note that W_t can be interpreted as the waiting time of the t-th customer in a GI/D/1 queue with the inter-arrival time between the t-th and the $(t-1)$-th customer being D_t and the service time being constant γ. Also, let $\tau_0 \triangleq 1$ and $\tau_i \triangleq \inf_s\{s > \tau_{i-1} : W_s = 0\}$ for each $i \geq 1$. Further, for $i \geq 1$, we define $J_i \triangleq \{t : \tau_{i-1} < t \leq \tau_i\}$ as the i-th busy cycle, and define $|J_i|$ as the length of busy cycle J_i. For any $s \geqslant 1$, let $i(s)$ to be the index of the busy cycle containing s, i.e. $s \in J_{i(s)}$.

The following lemma establishes an upper bound on the total overshooting by leveraging the auxiliary overshooting process. The proof can be found in Appendix **??**.

Lemma 3.5. *The following inequality holds under any demand sample path:*

$$\sum_{t=1}^{T}\left(y_t^{ASGD}-\hat{y}_t^{ASGD}\right)\leqslant\gamma\sum_{s=1}^{T}T^{\frac{v_{k(s)}-1}{3}}|J_{i(s)}|.$$

To proceed, we also establish an upper bound on the length of each busy cycle. The proof of Proposition 3.6 can be found from Appendix **??**.

Proposition 3.6. *Under Assumption 3(i) and (ii), the following inequality holds for any $s\in[T]$:*

$$\mathbb{E}[|J_{i(s)}|]\leq\frac{63\bar{D}}{(\bar{D}-\underline{D})^4}\log s.$$

We now bound the regret due to inventory carryover by applying Lemma 3.5 and Proposition 3.6. Note the following inequalities:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\left(y_t^{ASGD}-\hat{y}_t^{ASGD}\right)\right]&\leqslant\mathbb{E}\left[\sum_{t=1}^{T}\left(y_t^{ASGD}-\hat{y}_t^{ASGD}\right)\,\middle|\,\mathcal{G}\right]\mathbb{P}(\mathcal{G})+TM\mathbb{P}(\mathcal{G}^c)\\
&\leqslant\gamma\mathbb{E}\left[\sum_{s=1}^{T}T^{\frac{v_{k(s)}-1}{3}}|J_{i(s)}|\,\middle|\,\mathcal{G}\right]\mathbb{P}(\mathcal{G})+TM\mathbb{P}(\mathcal{G}^c)\\
&\leqslant\gamma T^{\frac{v_f-1}{3}}\mathbb{E}\left[\sum_{s=1}^{T}|J_{i(s)}|\,\middle|\,\mathcal{G}\right]\mathbb{P}(\mathcal{G})+TM\mathbb{P}(\mathcal{G}^c)\\
&\leqslant\gamma T^{\frac{v_f-1}{3}}\mathbb{E}\left[\sum_{s=1}^{T}|J_{i(s)}|\right]+TM\mathbb{P}(\mathcal{G}^c)\\
&\leqslant\frac{63\bar{D}\gamma}{(\bar{D}-\underline{D})^4}T^{\frac{v_f-1}{3}}\sum_{s=1}^{T}\log s+TM\mathbb{P}(\mathcal{G}^c)\\
&\leqslant\frac{63\bar{D}\gamma e^{1/3}}{(\bar{D}-\underline{D})^4}\,T^{\frac{2+v}{3}}\,\log T+TM\mathbb{P}(\mathcal{G}^c),
\end{aligned}$$

where the first inequality follows from $y_t^{ASGD}\leqslant M$ and $\hat{y}_t^{ASGD}\geqslant 0$, the second inequality follows from Lemma 3.5, the third inequality follows from Lemma 1.5, the fifth inequality follows from Proposition 3.6, and the last inequality holds because the definition of f implies $T^{v_f}\leqslant e^v T^v$ and

$\nu \leqslant 1$. Combining the above inequality with (3.2), we complete the proof of Proposition 3.4. □

3.5 Numerical Study

We next test **ASGD 4** in the single-period newsvendor setting

$$C(y, D) \;=\; h\,(y - D)^{+} + b\,(D - y)^{+},$$

where $y \in [0,\, y_{\max}]$ is the order quantity and D is the random demand. Because the realised cost is observed at the end of each period and the sub-gradient $\partial_y C(y, D) \in \{-b,\, h\}$ is trivial to compute, the hybrid oracle is natural.

Experimental design

Horizon and cost parameters. We fix $T = 10{,}000$, under- and overage costs $(b, h) = (1, 4)$, and a capacity $y_{\max} = 9$. The benchmark mean demand starts at $\mu_0 = 5$.

Non-stationary demand generator. Demand evolves as $D_t = \mu_t + \varepsilon_t$, where the mean path $\{\mu_t\}$ is driven by a *piecewise, sign-alternating random walk* that expends a prescribed variation budget $y_{\max}\, T^{\nu}$ ($\nu \in \{0.22, 0.33\}$), so that the total variation B_T in the cost function level is $B_T = \max(b, h) * y_{\max} * T^{\nu}$: (i) draw $T - 1$ i.i.d. uniform gaps, normalize and scale so the non-negative increments sum to B_T; (ii) add the increments to μ_{t-1} while alternately cycling upward and downward between $0.5\mu_0$ and $1.6\mu_0$; (iii) discard any overshoot when a boundary is hit. Independent Gaussian shocks $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.5, 1.0, 2.5\}$ model observation noise.

Policies compared.

- *OGD*: one fixed learning rate, no restarts.

- *Restarted OGD (oracle)*: window size tuned to the true variation exponent ν.
- *ASGD 4 (theoretical)*: original confidence thresholds, which (as before) rarely fire.
- *ASGD 4 (c_{thr})*: scaled threshold with c_{thr} chosen on a dense grid $[0.02, 2]$ to minimize the empirical loss (one scale per σ, ν pair).

Performance metric. Dynamic regret and the loss relative to the moving oracle are defined exactly as in Section **??**:

$$R_\phi^\pi(f,T) = \sum_{t=1}^{T} C\big(y_t, D_t\big) - C\big(y_t^*, D_t\big), \qquad L_\phi^\pi(f,T) = \frac{R_\phi^\pi(f,T)}{\sum_{t=1}^{T} C(y_t^*, D_t)}.$$

Results

Table 3.2 summarises the relative losses for all methods across the six (ν, σ) configurations. Each ASGD entry lists the best threshold scale and the resulting loss.

A clear noise–dependence is visible in the optimal trigger scale. For a fixed variation exponent, the best c_{thr} grows steadily with the observation noise: when $\nu = 0.22$ the scale rises from 1.65 at $\sigma = 0.5$ to 3.35 at $\sigma = 2.5$; for $\nu = 0.33$ it moves from 0.60 to 1.90 over the same noise range. Intuitively, a larger threshold is required to prevent noise-induced false switches in the hybrid trigger when the gradients and costs are highly corrupted.

Crucially, once the scale is tuned, **ASGD 4** matches the oracle *restarted OGD* remarkably well. Across all six (ν, σ) combinations the tuned scheme reduces the theoretical ASGD loss by factors between 0.5 and 3.5. The optimal c_{thr} depends far more on the noise level than on the variation exponent, suggesting that a single calibration based on σ can be reused across a range of non-stationarity profiles without noticeable performance loss.

Table 3.2: Relative loss $L_{\phi}^{\pi}(f, T)$ in the newsvendor experiment. Each ASGD line shows $(c_{\text{thr}}^{\star}, \text{loss})$.

	Random walk jumps		
σ	0.5	1	2.5
$\boldsymbol{V_T = T^{0.22}}$			
OGD	2.193	0.915	0.236
Restarted OGD, correct	0.156	0.077	0.048
ASGD 4 (theo)	0.242	0.122	0.068
ASGD 4 (c_{thr})	(1.65, 0.221)	(2.15, 0.099)	(3.35, 0.048)
$\boldsymbol{V_T = T^{0.33}}$			
OGD	2.542	0.998	0.241
Restarted OGD, correct	0.258	0.162	0.099
ASGD 4 (theo)	0.798	0.416	0.166
ASGD 4 (c_{thr})	(0.6, 0.235)	(1.2, 0.150)	(1.9, 0.079)

4 | Chapter 4: Application to Universal Portfolio Selection

4.1 Introduction to Universal Portfolio Selection

Modern financial markets can exhibit highly non-stationary behavior. Asset returns may shift unpredictably due to regime changes, macroeconomic events, or evolving industry trends. An investment strategy that assumes stationary or i.i.d. returns can perform suboptimally when faced with such drifting market conditions. For instance, a fixed "buy-and-hold" portfolio calibrated to past data might suffer prolonged losses if a new sector suddenly outperforms the previously dominant assets. This motivates the need for *universal portfolio selection* algorithms that (i) remain competitive with the best fixed portfolio in hindsight (the classical goal of Cover's universal portfolios), while (ii) adapting swiftly to changing return distributions over time. In particular, we seek a data-driven investment policy that automatically detects and responds to distributional shifts in asset returns, without prior knowledge of the *extent* or *timing* of those shifts.

Motivation. Traditional portfolio selection theory often relies on either stochastic i.i.d. market models or worst-case static assumptions. In a stochastic setting, if asset returns are drawn i.i.d. from a fixed distribution, one can invoke the Kelly criterion [Kelly 1956] to find a constant rebalanced portfolio (CRP) that maximizes long-run growth; indeed, with unknown distribution, the optimal CRP can be learned over time with vanishing per-round regret (e.g., via empirical

estimation). However, real markets rarely remain stationary indefinitely. In contrast, in the adversarial formulation in the worst case, the seminal work of Cover [1991] introduced a universal portfolio strategy that guarantees sublinear regret of order $O(n \log T)$ (where n is the number of assets and T the number of trading periods) against the best fixed portfolio in hindsight.

Follow-up research produced more efficient algorithms and improved bounds. For example, Helmbold et al. [1996] proposed the *Exponentiated Gradient (EG)* method for online portfolio selection, achieving a regret of $O(\sqrt{T \log n})$ under assumptions of bounded returns.

Subsequent advances leveraged the convexity and exp-concavity of the log-loss to attain even lower regret: Hazan et al. [2007] developed an online Newton method yielding $O(\log T)$ regret for arbitrary sequences of strictly convex portfolio losses, matching Cover's order but with greatly improved computational efficiency. More recently, Hazan and Kale [2015] achieved the first guarantee beyond $O(\log T)$ by bounding regret in terms of market variability (specifically, $O(\log Q)$ where Q is the quadratic variation of price relatives) Hazan and Kale [2015], and demonstrated improved performance under stochastic geometric Brownian motion models as well.

While the literature above focuses on static benchmark portfolios, far less has addressed *non-stationary* benchmarks, wherein the optimal CRP itself may shift over time. In general online learning, competing with a time-varying comparator incurs the notion of *dynamic regret*, which grows with the "variation" of the comparator sequence (e.g., the total movement of the best portfolio between periods). Existing universal portfolio algorithms do not explicitly account for unknown changes in the best asset allocation; indeed, their regret guarantees can deteriorate to linear order if the market undergoes frequent or abrupt regime shifts. Some recent works have begun exploring this direction. For example, Das et al. [2014] introduces a lazy-update online portfolio method and obtains sublinear regret not only against the best fixed portfolio but also against the best shifting portfolio (allowing a limited number of changes) in hindsight. However, to our knowledge, no prior result provides a theoretical regret guarantee that scales optimally with an arbitrary budget of non-stationarity in an adversarial market. This gap is what we aim

to fill. In this work, we quantify market non-stationarity via a variation budget and develop an adaptive algorithm that remains provably competitive even under polynomially growing changes in the return distribution.

4.1.1 Main Results and Contributions.

We propose an *Adaptive Universal Portfolio (AUP)* strategy for an investor allocating wealth across n assets over T rounds. The policy is fully online and combines mirror-descent updates with a dynamic change-point detection mechanism. Our key contributions are as follows:

(i) *Algorithmic innovation.* We design a multi-expert mirror descent scheme that simultaneously runs K candidate portfolio strategies with different learning rates. A data-driven triggering rule (akin to a restart mechanism) selects and switches to the appropriate expert when evidence of a distributional change in asset returns emerges. This approach automatically tunes the effective step-size to the (unknown) pace of market change, without requiring any prior knowledge of the variation budget.

(ii) *Regret guarantee.* We measure performance through *dynamic regret*, the cumulative gap between the log-return our strategy earns on each round and the log-return of the constant-rebalanced portfolio that, with hindsight, would have been optimal *for that same round.* When the total amount of market drift up to time T is bounded by an (unknown) budget B_T, we show that the dynamic regret grows strictly slower than the horizon. More precisely, it is bounded by a polynomial in T whose exponent depends smoothly on the unknown drift parameter ν in $B_T = \Theta(T^\nu)$. Consequently, for every $\nu < 1$ the average per-period regret vanishes, so the algorithm's long-run growth rate converges to that of the period-wise best constant portfolios. In the stationary limit the bound specialises to the classical two-thirds power of T — slower than the logarithmic rate that exp-concave methods can attain in a perfectly stable market, but achieved here by a single, drift-robust strategy that retains sublinear regret even when the

market becomes highly volatile.

(iii) *Comparison with literature.* To our knowledge, this is the first result on online portfolio selection with *unknown* non-stationary returns that achieves sublinear regret with theoretical guarantees. Table 4.1 summarizes how our setting and performance compare to prior works. Unlike classical universal portfolio algorithms [e.g., Cover 1991; Helmbold et al. 1996; Hazan et al. 2007] which assume a fixed comparator, our approach handles adversarially changing markets without foreknowledge of change times or magnitudes. Compared to existing dynamic or specialized models (such as Das et al. 2014 who require structural assumptions like sector group sparsity), our AUP is a general-purpose method and yields regret scaling near-optimally in T and B_T.

4.1.2 Literature Review

Online portfolio selection, first formalized by Cover [1991], has since developed along two main threads: worst-case (universal) strategies and stochastic or structural models. Below we review representative results from each category, emphasizing their regret bounds. Table 4.1 provides a high-level comparison.

Table 4.1: Selected literature on online portfolio selection (universal portfolios). Regret bounds are against the best constant rebalanced portfolio (CRP) or a shifting sequence as noted. Here $d = n-1$ is the simplex dimension and L denotes an upper bound on the number of comparator shifts.

Literature	Market model	Comparator	Regret upper bound	Notes
Cover [1991]	Adversarial	Fixed CRP	$O(d \log T)$	Info-theoretic mix
Helmbold et al. [1996]	Adversarial	Fixed CRP	$O(\sqrt{T \ln n})$	Multiplicative updates
Blum and Kalai [1999]	Adversarial	Fixed CRP	$O(d \log T)$	Handles transaction costs
Kalai and Vempala [2003]	Adversarial	Fixed CRP	$O(d \log T)$	Poly-time approximation
Hazan et al. [2007]	Adversarial	Fixed CRP	$O(\log T)$	Online Newton
Hazan and Kale [2015]	Adversarial	Fixed CRP	$O(\log Q)$	Variance-dependent
Das et al. [2014]	Adversarial	Shifting ($\leqslant L$ switches)	$O(\sqrt{T})$	Dynamic regret $O(\sqrt{TL})$
This work	Stochastic	Dynamic	$\widetilde{O}(T^{\frac{2+\nu}{3}})$	Adapts to unknown variation B_T

Stationary or i.i.d. market models. In a classical stochastic setting, asset prices are often modeled by i.i.d. returns or stationary processes (e.g., geometric Brownian motion). The celebrated Kelly criterion [Kelly 1956] prescribes the optimal constant rebalanced portfolio for maximizing expected log-growth when the return distribution is known; Breiman [1961] showed that this strategy also almost surely outperforms any other fixed portfolio in the long run. When the distribution is unknown, an investor can learn the Kelly portfolio over time. For example, Cover [1984] and Györfi et al. [2006] study nonparametric and data-driven approaches to converge towards the optimal i.i.d. portfolio. In general, under a stationary distribution, the regret (difference in total log-wealth) of an adaptive strategy can often be bounded by a constant or grow sub-logarithmically, reflecting the law of large numbers. These probabilistic results, however, rely heavily on the i.i.d. assumption and do not provide worst-case guarantees.

Adversarial static environment (universal portfolios). Cover's universal portfolio algorithm inaugurated the adversarial paradigm, ensuring $O(n \log T)$ worst-case regret with respect to the best fixed portfolio. The regret here is typically measured as the excess log-wealth of the comparator over the algorithm, and $O(n \log T)$ was shown to be achievable via a Bayesian mixture over all possible CRPs [Cover 1991]. This guarantee is sublinear in T (for fixed n) and thus the average per-round regret vanishes. Subsequent research sought to improve the efficiency and tighten the dependence on problem parameters. **?** (an expanded journal version of Helmbold et al. 1996) introduced the EG algorithm, which uses multiplicative weight updates and attained a regret of $O\left(\frac{1}{r}\sqrt{T \ln n}\right)$ assuming returns are bounded in $[r, 1]$ for some $r > 0$. Although the $\sqrt{T}$ dependence is larger for long horizons, EG has the advantage of simple updates in $O(n)$ time per round, versus Cover's method which is computationally intensive (exponential in n in the worst case). Further improvements came from exploiting the curvature of the loss function $f_t(x) = -\ln(x^\top r_t)$. This loss is 1-exp-concave (and also strongly convex in the simplex domain), enabling algorithms with logarithmic regret. Notably, Hazan et al. [2007] present an efficient

online Newton algorithm with $O(\log T)$ regret for strictly convex costs, which when applied to portfolio selection yields $O(\log T)$ regret independent of n (essentially matching Cover's bound up to constant factors). Around the same time, Blum and Kalai [1999] and Kalai and Vempala [2003] investigated alternative "universal" algorithms and information-theoretic limits; Kalai and Vempala 2003 in particular provided a polynomial-time randomized algorithm that achieves the optimal $O(d \log T)$ regret (with $d = n - 1$ degrees of freedom on the simplex) while improving computational complexity over Cover's exhaustive integration. Hazan and Kale [2015] further refined the analysis by introducing data-dependent regret bounds: they prove that the worst-case regret can be tightened to $O(\log Q)$, where $Q = \sum_{t=1}^{T} \sum_{i=1}^{n} \left(\frac{r_t(i)}{\mathbf{r}_t^\top x^*} - 1 \right)^2$ is the quadratic variation of the sequence of price relatives under the hindsight-optimal portfolio x^*. Since $Q \leqslant O(T)$, this result subsumes the $O(\log T)$ bound and can be significantly smaller when market fluctuations are mild. We note that all these methods assume a fixed comparator x^* that maximizes total wealth in hindsight.

Non-stationary or time-varying environment. In practice, the identity of the best investment portfolio may change over time—e.g., as different sectors go through boom and bust cycles. Traditional static-regret algorithms are not designed to track such shifts. The problem of *dynamic* or *adaptive* regret minimization has been studied in the broader online learning literature: for instance, Herbster and Warmuth [1998] developed an expert-tracking algorithm with regret $O(\sqrt{TL \ln N})$ when competing against a sequence of experts with at most L switches, and Zinkevich [2003] analyzed gradient descent for convex problems, showing regret $O(\sqrt{T(1 + P_T)})$ where P_T is the path-length of the comparator sequence. However, applying these general results to portfolio selection either requires discretizing the continuum of possible portfolios or yields loose bounds. Some specialized progress has been made: Das et al. [2014] incorporate a group-sparsity regularizer and prove regret bounds $O(\log T)$ for static and $O(\sqrt{T})$ for shifting portfolios in a sector-rotation scenario. Nevertheless, these results assume either known struc-

ture or do not achieve the optimal dependence on the number of changes. By contrast, our work addresses a fully adversarial, unstructured market where the total variation of the optimal portfolio is bounded by $B_T = O(T^\nu)$, and we obtain a regret of order $\widetilde{O}(T^{(2+\nu)/3})$ without knowing B_T in advance. This matches the best-known rates for dynamic OCO in the absence of exp-concavity assumptions, and is the first such guarantee in the context of online portfolio selection.

4.2 Model Formulation

We consider an investor who allocates her wealth among n assets repeatedly over a horizon of T rounds. In each round $t = 1, 2, \ldots, T$, the following events occur:

1. **Portfolio selection:** The investor chooses a portfolio vector

$$\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,n})^\top \in \Delta_n, \qquad \Delta_n := \Big\{\mathbf{x} \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\Big\},$$

   where $x_{t,i}$ is the fraction of current wealth invested in asset i at round t.

2. **Market outcome:** After the portfolio is chosen, the market returns for all assets are realized and revealed. We denote by P_t the (unknown) distribution of the price-relative vector in round t, and let

$$\mathbf{r}_t = \big(r_t(1), \ldots, r_t(n)\big)^\top \in \mathbb{R}^n_+$$

   be the actual price-relative vector drawn from P_t. Here

$$r_t(i) := \frac{\text{price of asset } i \text{ at time } t+1}{\text{price of asset } i \text{ at time } t}$$

   is the gross return factor of asset i during round t. The investor observes $\mathbf{r}_t$ at the end of the round (full-information feedback).

3. **Wealth update:** Ignoring transaction costs or taxes, the investor's wealth is updated multiplicatively based on the portfolio return $\mathbf{r}_t^\top \mathbf{x}_t$. If W_t denotes the total wealth at the start of round t, then

$$W_{t+1} \;=\; W_t \left(\mathbf{r}_t^\top \mathbf{x}_t\right).$$

Cumulative wealth. Over T rounds, the cumulative wealth relative to the initial wealth W_1 is given by the product of one-period returns:

$$\frac{W_{T+1}}{W_1} \;=\; \prod_{t=1}^{T} \left(\mathbf{r}_t^\top \mathbf{x}_t\right) \quad\Longrightarrow\quad \log \frac{W_{T+1}}{W_1} \;=\; \sum_{t=1}^{T} \log\left(\mathbf{r}_t^\top \mathbf{x}_t\right).$$

Thus, maximizing long-term wealth growth is equivalent to maximizing the sum of log-returns $\sum_{t=1}^{T} \log(\mathbf{r}_t^\top \mathbf{x}_t)$.

OCO formulation and regret. We cast this problem in an online convex optimization framework with full information. The per-round realized loss (negative log-return) in round t is

$$f_t(\mathbf{x}) \;=\; -\ln\left(\mathbf{r}_t^\top \mathbf{x}\right), \qquad \mathbf{x} \in \Delta_n.$$

Minimizing $f_t(\mathbf{x}_t)$ is equivalent to maximizing the portfolio's log-return in round t. The decision set is the probability simplex Δ_n. Since $\mathbf{r}_t$ is drawn from P_t, we may also define the expected loss for a decision $\mathbf{x}$ in round t as

$$L_t(\mathbf{x}) \;=\; \mathbb{E}_{\mathbf{r}_t \sim P_t}\left[-\ln\left(\mathbf{r}_t^\top \mathbf{x}\right)\right],$$

which is a convex function of $\mathbf{x}$. However, the investor only observes the single-sample loss $f_t(\mathbf{x}_t)$ each round.

The performance of a portfolio strategy is now evaluated against the *best period-wise action*

instead of a single fixed portfolio. For each round $t = 1, 2, \ldots, T$, define

$$\mathbf{x}_t^* := \arg\max_{\mathbf{x} \in \Delta_n} \ln\left(\mathbf{r}_t^\top \mathbf{x}\right),$$

the clairvoyant constant-rebalanced portfolio that would have maximized the one-step log-return on round t. The *dynamic regret* up to time T is

$$R_T^{\mathrm{dyn}} := \sum_{t=1}^{T} \left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*) \right) = \sum_{t=1}^{T} \left[-\ln\left(\mathbf{r}_t^\top \mathbf{x}_t\right) + \ln\left(\mathbf{r}_t^\top \mathbf{x}_t^*\right) \right].$$

A portfolio-selection algorithm is called *dynamically universal* if it achieves sublinear dynamic regret $R_T^{\mathrm{dyn}} = o(T)$ (in expectation or with high probability), ensuring its average per-round log-return approaches that of the period-wise optimal portfolios.

4.2.1 Non-Stationary Returns and the Variation Budget

In general, the return distribution may drift over time. We quantify this drift by a notion of temporal variation in the sequence of return distributions. Let P_t denote the (unknown) distribution of $\mathbf{r}_t$ at round t, as defined above. We measure the cumulative distributional change over T rounds by

$$TV(P_{1:T}) := \sum_{t=2}^{T} \mathcal{W}\left(P_t, P_{t-1}\right),$$

where $\mathcal{W}$ is the 1-Wasserstein distance on $\mathbb{R}_+^n$. Namely, for any two probability distributions μ and ν on $\mathbb{R}_+^n$, the Wasserstein metric is defined as

$$\mathcal{W}(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\left[\|X - Y\|_1 \right],$$

and $\Gamma(\mu, \nu)$ is the set of all couplings of μ and ν. Intuitively, $\mathcal{W}(\mu, \nu)$ represents the minimum "transport cost" of moving probability mass from distribution μ to ν, using the L^1 norm as the

ground metric. In our setting, $\mathcal{W}(P_t, P_{t-1})$ captures the magnitude of distributional shift in the market's returns from round $t-1$ to round t.

Variation budget. Fix a budget $B_T \geqslant 0$ (possibly growing with T), and consider the set of plausible environment sequences

$$\mathcal{V}(T, B_T) := \left\{ P_{1:T} : \ TV(P_{1:T}) \leqslant B_T \right\}.$$

We assume $B_T = O(T^\nu)$ for some $\nu \in [0, 1]$, but note that the investor does not know the exact value of B_T. Intuitively, $\mathcal{V}(T, B_T)$ is the class of all non-stationary return processes whose total distributional shift over T rounds does not exceed B_T. Smaller ν (or smaller B_T) corresponds to a more slowly changing, near-stationary environment, whereas $\nu = 1$ allows the possibility of abrupt changes in distribution each round (linear variation in T).

Why Wasserstein? Compared to simpler metrics (e.g. coordinate-wise ℓ^p distances), the Wasserstein distance (i) aligns with the intuition of transporting "mass" between return distributions with minimal cost, and (ii) is sensitive to the geometry of distributions in $\mathbb{R}^n_+$, making it well suited for capturing changes in the joint distribution of asset returns. We focus on the Wasserstein metric for concreteness; analogous results hold under other reasonable choices for measuring distributional variation.

Lemma 4.1. *For any $1 \leqslant t_1 < t_2 \leqslant T$,*

$$\sup_{\mathbf{x} \in \Delta_n} \left| L_{t_2}(\mathbf{x}) - L_{t_1}(\mathbf{x}) \right| \ \leqslant \ \frac{1}{m_{\min}} \sum_{t=t_1+1}^{t_2} \mathcal{W}\big(P_t, \, P_{t-1}\big),$$

where $L_t(\mathbf{x}) := \mathbb{E}_{\mathbf{r}_t \sim P_t}\left[-\ln(\mathbf{r}_t^\top \mathbf{x}) \right]$ is the expected loss in round t.

4.2.2 Worst-Case Dynamic Regret

Let π be an admissible online strategy. Given a return-distribution sequence $P_{1:T}$, its (expected) dynamic regret is

$$R^{\pi}_{\mathrm{dyn}}(T, P_{1:T}) := \mathbb{E}\left[\sum_{t=1}^{T}\left(f_t(\mathbf{x}^{\pi}_t) - f_t(\mathbf{x}^{*}_t)\right)\right],$$

where the expectation is over $\mathbf{r}_t \sim P_t$ and any internal randomness of π. Equivalently,

$$R^{\pi}_{\mathrm{dyn}}(T, P_{1:T}) = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\ln(\mathbf{r}_t^{\top}\mathbf{x}^{\pi}_t) + \ln(\mathbf{r}_t^{\top}\mathbf{x}^{*}_t)\right)\right].$$

Under the Wasserstein-variation class $\mathcal{V}(T, B_T)$ from (**??**), the *worst-case dynamic regret* is

$$R^{\pi}_{\mathrm{dyn}}(T, B_T) := \sup_{P_{1:T}\in\mathcal{V}(T,B_T)} R^{\pi}_{\mathrm{dyn}}(T, P_{1:T}).$$

Our goal is to design π so that $R^{\pi}_{\mathrm{dyn}}(T, B_T)$ grows sublinearly in T (ideally $o(T)$ or $O(\sqrt{T})$) without knowing B_T in advance.

4.2.3 Standing Assumptions

We impose the following assumptions on the return sequences and their distributions, which are standard in the literature and ensure the losses and gradients remain well-behaved:

1. **Bounded asset returns.** There exist known constants $m_{\min} > 0$ and $m_{\max} > 0$ such that for all rounds t and all assets $i \in [n]$,

$$m_{\min} \leqslant r_t(i) \leqslant m_{\max}.$$

In other words, each period's return vector $\mathbf{r}_t$ lies in a bounded subset of $\mathbb{R}^n_+$. In particular,

no asset's price can drop to zero in one period, and there is some uniform upper bound on one-period returns. This ensures that $\mathbf{r}_t^\top \mathbf{x}$ is always bounded between $m_{\min}$ and $m_{\max}$ for any portfolio $\mathbf{x} \in \Delta_n$. Consequently, the loss $-\ln(\mathbf{r}_t^\top \mathbf{x})$ is well-defined (no division by zero) and uniformly bounded. These bounds also imply that the gradient of the loss (which is $-\mathbf{r}_t/(\mathbf{r}_t^\top \mathbf{x})$ for portfolio $\mathbf{x}$) has bounded magnitude, a key requirement for our regret analysis.

4.3 Algorithm Description

We now describe an alternative, fully–online adaptive algorithm for universal portfolio selection. Like Algorithm 5, we maintain K candidate portfolios ("experts") each run with a different mirror–descent step–size, and we dynamically track which expert would have performed best so far.

Notation.

- Let $\mathbf{r}_t \in \mathbb{R}_+^n$ be the price–relative vector at time t.
- Define the instantaneous loss

$$f_t(\mathbf{x}) \;=\; -\log(\mathbf{r}_t^\top \mathbf{x}), \quad \mathbf{x} \in \Delta_n \,.$$

- The gradient of the loss at time t is

$$\nabla f_t(\mathbf{x}_t) \;=\; -\frac{\mathbf{r}_t}{\mathbf{r}_t^\top \mathbf{x}_t} \quad \Longrightarrow \quad \big[\nabla f_t(\mathbf{x}_t)\big]_i = -\frac{r_t(i)}{\mathbf{r}_t^\top \mathbf{x}_t} \,.$$

- We let

$$G \;:=\; \sup_{t,\,\mathbf{x}\in\Delta_n} \big\|\nabla f_t(\mathbf{x})\big\|_\infty = \sup_{t,\,\mathbf{x}\in\Delta_n} \max_i \frac{r_t(i)}{\mathbf{r}_t^\top \mathbf{x}} \,.$$

Under the bounded-returns assumption $m_{\min} \leqslant r_t(i) \leqslant m_{\max}$ and $\mathbf{r}_t^\top \mathbf{x} \geqslant m_{\min}$, it follows that we can set

$$G \;=\; \frac{m_{\max}}{m_{\min}}\,.$$

- As before, we have a discrete set of hypothetical variation parameters $\mathcal{V} = \{\nu_1, \nu_2, \ldots, \nu_K\}$ defined as follows:

$$\nu_k = \frac{k}{\log T}, \quad k = 1, 2, \ldots, K\,,$$

where K is chosen such that $\nu_{K-1} < 1 \leqslant \nu_K$.

- Fix K candidate step–sizes $\{\eta_T^g : g = 1, \ldots, K\}$. We will run K mirror–descent updates in parallel, with $\Delta_T^g = T^{\frac{2}{3}(1-\nu_g)}$. Then the step–size for expert g is

$$\eta_T^g \;=\; \sqrt{\frac{\log n}{2\,\Delta_T^g\,G^2}}\,.$$

Mirror–Descent Update. Each expert g maintains a portfolio $\mathbf{x}_t^g \in \Delta_n$, updated by the entropic mirror step:

$$x_{t+1}^g(i) \;=\; \frac{x_t^g(i)\;\exp\Bigl(-\,\eta_T^g\,\nabla_t(i)\Bigr)}{\displaystyle\sum_{j=1}^{n} x_t^g(j)\;\exp\Bigl(-\,\eta_T^g\,\nabla_t(j)\Bigr)}\,.$$

Meta–Selection with Change–Point Triggering. We maintain a single active index k that only increases when a change–point condition is met. Recall t_{if} is the last time we switched. For each $g > k$, define the inter–expert loss gap over $[t_{\text{if}},\, t-1]$:

$$\Delta_{t-1}^{k,g} \;=\; \sum_{s=t_{\text{if}}}^{t-1} \Bigl| f_s(\hat{\mathbf{x}}_s^k) \;-\; f_s(\hat{\mathbf{x}}_s^g) \Bigr|\,.$$

$$B_g = T^{\frac{2+\nu_g}{3}} \left[\, \frac{4}{m_{\min}} \;+\; 2\Bigl(C_0\,\sqrt{\log T} \;+\; C_1\,\sqrt{\log n}\,\Bigr) \right] \;+\; 4\,C_3\,\sqrt{2T\,\log(2T)}\,,$$

If for any $g \in \{k+1, \ldots, K\}$ we have $\Delta_{t-1}^{k,g} \geqslant B_g$, we increment $k \leftarrow k+1$ and reset $t_{\text{if}} \leftarrow t$. We then play

$$\mathbf{x}_t \;=\; \hat{\mathbf{x}}_t^k \quad \text{and incur} \quad f_t(\mathbf{x}_t)\,.$$

Algorithm 5 Adaptive Universal Portfolio with Change-Point Triggering

Input: $\{\eta_T^g\}_{g=1}^K$ (step–sizes), $\{B_g\}_{g=1}^K$ (trigger thresholds), horizon T.
Initialize: $k \leftarrow 0$, $t_{\text{if}} \leftarrow 1$. For each $g = 0, 1, \ldots, K$, pick $\hat{\mathbf{x}}_1^g \in \Delta_n$ (e.g. uniform), and set $L_0^g = 0$.
for $t = 1$ **to** T **do**
  if $t \geqslant 2$ **then**
    for $g = k+1$ **to** K **do**
      if $\sum_{s=t_{\text{if}}}^{t-1} \left| f_s(\hat{\mathbf{x}}_s^k) - f_s(\hat{\mathbf{x}}_s^g) \right| \geqslant B_g$ **then**
        $k \leftarrow k+1, \quad t_{\text{if}} \leftarrow t$
        break
      end if
    end for
  end if
  for $g = k$ **to** K **do**
    Compute gradient $\nabla_t^g = -\,\mathbf{r}_t \,/\, \big(\mathbf{r}_t^\top \hat{\mathbf{x}}_t^g\big)$.
    Update expert g:

$$\hat{x}_{t+1}^g(i) \;=\; \frac{\hat{x}_t^g(i)\, \exp\Big(-\eta_T^g\, \nabla_t^g(i)\Big)}{\sum_{j=1}^n \hat{x}_t^g(j)\, \exp\Big(-\eta_T^g\, \nabla_t^g(j)\Big)}\,.$$

    Accumulate loss $L_t^g = L_{t-1}^g + f_t(\hat{\mathbf{x}}_t^g)$.
  end for
  Play $\mathbf{x}_t = \hat{\mathbf{x}}_t^k$, incur $f_t(\mathbf{x}_t)$.
end for

Discussion. This change–point rule ensures that we only switch to a finer step–size (expert) once there is sufficient evidence—measured by the accumulated loss discrepancy—that the current mirror–descent trajectory is underperforming. By controlling thresholds $\{B_g\}$, the algorithm adapts its effective step–size on the fly, balancing stability against responsiveness to market shifts.

4.4 Analysis of Regret Bound

Note that we have the following regret definition:

$$R^{\pi}_{\text{dyn}}(T, P_{1:T}) = \mathbb{E}\left[\sum_{t=1}^{T}\left(-\ln(\mathbf{r}_t^{\top}\mathbf{x}_t^{\pi}) + \ln(\mathbf{r}_t^{\top}\mathbf{x}_t^{*})\right)\right]. \tag{4.1}$$

And we need to show the following proposition.

Proposition 4.2 (Regret due to Non-stationary Mirror Descent)**.** *Under Assumption (i), we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_t(\hat{x}_t^{AUP}) - f_t(x_t^{*})\right)\right] \;=\; O\Big(T^{\frac{2+v_{k^*}}{3}}(\log^{3/2} T + \log T \log^{1/2} n)\Big)\,.$$

4.4.1 Sketched Proof of Proposition 4.2.

As before, we define a discrete set of hypothetical variation parameters $\mathcal{V} = \{v_1, v_2, \ldots, v_K\}$ with

$$v_k \;=\; \frac{k}{\log T}\,, \qquad k = 1, 2, \ldots, K\,,$$

where K is chosen such that $v_{K-1} < 1 \leqslant v_K$. We run K mirror–descent updates in parallel, treating each $v_k \in \mathcal{V}$ as a candidate variation exponent. In particular, let $\Delta_T^g := T^{\frac{2}{3}(1-v_g)}$ for each expert g, and set the step–size

$$\eta_T^g \;=\; \sqrt{\frac{\log n}{2\,\Delta_T^g\,G^2}}\,, \qquad g = 1, 2, \ldots, K\,,$$

where G is the bound on the gradient (or subgradient) norm in the mirror descent updates.

We observe that $K = O(\log T)$. Define k^* as the smallest index in $\{1, \ldots, K\}$ such that $v \leqslant v_{k^*}$. Given the relation $v_{k^*} = v_{k^*-1} + \frac{1}{\log T}$, it follows that $v \leqslant v_{k^*} < v + \frac{1}{\log T}$. Consequently, $T^{v} \leqslant T^{v_{k^*}} < e\,T^{v}$, implying that $T^{v_{k^*}}$ is within a constant factor of T^{v}. Therefore, it is sufficient to bound the

total regret by

$$O\Big(T^{\frac{2+v_{k^*}}{3}}\log^{3/2}T\Big).$$

To establish this bound, we proceed in four steps:

Step 1. For each expert $g \in \{1,\dots,K\}$, with high probability over T time periods we have the following empirical dynamic regret bound:

$$\sum_{t=1}^{T}\Big[f_t(\hat{x}_t^g) - f_t(x_t^*)\Big] \;\leqslant\; T^{\frac{2+v_g}{3}}\left[\frac{2}{m_{\min}} \;+\; \Big(C_0\sqrt{\log T} \;+\; C_1\sqrt{\log n}\Big)\right] \;+\; 2\,C_3\sqrt{2T\,\log(2T)}\,.$$

Here $m_{\min}$ is the positive lower bound (and $m_{\max}$ the upper bound) for each decision coordinate under Assumption (i), and C_0, C_1, C_3 are the corresponding constants (for completeness: $C_3 := \ln(m_{\max}/m_{\min})$, $C_0 := 8\,\frac{m_{\max}}{m_{\min}}$, $C_1 := 2\sqrt{2}\,\frac{m_{\max}}{m_{\min}}$). We may assume that this high-probability event holds from this point onward.

Step 2. Contradiction via trigger condition. Suppose, for the sake of contradiction, that at some period t we have $k \geqslant k^*$ and the **if**-condition in our algorithm fires with some $g > k$. Then by the triangle inequality, on the interval from the time t_{if} of that trigger to t we have

$$\sum_{s=t_{\text{if}}}^{t}\Big|f_s(\hat{x}_s^k) - f_s(\hat{x}_s^g)\Big| \;\leqslant\; \sum_{s=t_{\text{if}}}^{t}\Big[f_s(\hat{x}_s^k) - f_s(x_s^*)\Big] \;+\; \sum_{s=t_{\text{if}}}^{t}\Big[f_s(\hat{x}_s^g) - f_s(x_s^*)\Big].$$

But from our high-probability bound on each "arm" $j \in \{k, g\}$ (from Step 1),

$$\sum_{s=1}^{T}\Big[f_s(\hat{x}_s^j) - f_s(x_s^*)\Big] \;\leqslant\; T^{\frac{2+v_g}{3}}\left[\frac{2}{m_{\min}} \;+\; \Big(C_0\sqrt{\log T} \;+\; C_1\sqrt{\log n}\Big)\right] \;+\; 2\,C_3\sqrt{2T\,\log(2T)},$$

so it follows that even summing only from $s = t_{\text{if}}$ to t cannot exceed

$$T^{\frac{2+v_g}{3}}\left[\frac{4}{m_{\min}} \;+\; 2\Big(C_0\sqrt{\log T} \;+\; C_1\sqrt{\log n}\Big)\right] \;+\; 4\,C_3\sqrt{2T\,\log(2T)},$$

which is exactly the threshold required for the **if**-condition to trigger. This is a contradiction,

implying that no such trigger can occur when $k \geqslant k^*$. Hence the running index k never surpasses k^*.

Step 3. Bounding regret between successive triggers. Let t' and t'' be two consecutive times when the **if**-condition fires. On the intervening interval $t' + 1 \leqslant t \leqslant t'' - 1$, the **if**-test fails even for $g = k^*$ by assumption. Therefore, we have

$$\sum_{t=t'+1}^{t''-1} \left[f_t(\hat{x}_t^k) - f_t(\hat{x}_t^{k^*}) \right] \;<\; T^{\frac{2+v_{k^*}}{3}} \left[\frac{4}{m_{\min}} \;+\; 2\Big(C_0 \sqrt{\log T} \;+\; C_1 \sqrt{\log n} \Big) \right] \;+\; 4\, C_3 \sqrt{2T \,\log(2T)}\,.$$

Now split the excess regret against the comparator x_t^* by the triangle inequality:

$$\sum_{t=t'+1}^{t''-1} \left[f_t(\hat{x}_t^k) - f_t(x_t^*) \right] \;\leqslant\; \sum_{t=t'+1}^{t''-1} \left| f_t(\hat{x}_t^k) - f_t(\hat{x}_t^{k^*}) \right| \;+\; \sum_{t=t'+1}^{t''-1} \left| f_t(\hat{x}_t^{k^*}) - f_t(x_t^*) \right|.$$

The first summation is bounded by the failed-trigger threshold above. The second summation, over at most T periods, is bounded by the high-probability regret of trajectory k^* (from Step 1):

$$\sum_{t=1}^{T} \left[f_t(\hat{x}_t^{k^*}) - f_t(x_t^*) \right] \;\leqslant\; T^{\frac{2+v_{k^*}}{3}} \left[\frac{2}{m_{\min}} \;+\; \Big(C_0 \sqrt{\log T} \;+\; C_1 \sqrt{\log n} \Big) \right] \;+\; 2\, C_3 \sqrt{2T \,\log(2T)}\,.$$

Putting these together, we conclude that on each interval between triggers,

$$\sum_{t=t'+1}^{t''-1} \left[f_t(\hat{x}_t^k) - f_t(x_t^*) \right] \;\leqslant\; T^{\frac{2+v_{k^*}}{3}} \left[\frac{6}{m_{\min}} \;+\; 3\Big(C_0 \sqrt{\log T} \;+\; C_1 \sqrt{\log n} \Big) \right] \;+\; 6\, C_3 \sqrt{2T \,\log(2T)}\,.$$

By taking expectations on both sides (over the random draw of losses), we obtain

$$\mathbb{E}\left[\sum_{t=t'+1}^{t''-1} \Big(f_t(\hat{x}_t^k) - f_t(x_t^*) \Big) \right] \;=\; O\Big(T^{\frac{2+v_{k^*}}{3}} \big(\sqrt{\log T} + \sqrt{\log n} \big) \Big)\,.$$

Since the **if**-condition can fire at most $K = \log T$ times, the total interval-wise regret (ignoring

the small residual at the end) is

$$O\Big(T^{\frac{2+v_{k^*}}{3}}(\log^{3/2} T + \log T \log^{1/2} n)\Big).$$

Step 4. Suppose that the last if-condition triggers at time t_{last}. The regret from time t_{last} to T can be bounded in a similar fashion (using the high-probability guarantee for the final active expert $k \leqslant k^*$). Adding this final segment to the interval-wise regret derived in Step 3, and recalling that $T^{v_{k^*}}$ differs from T^v only by a constant factor, we conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T}\Big(f_t(\hat{x}_t^k) - f_t(x_t^*)\Big)\right] = O\Big(T^{\frac{2+v_{k^*}}{3}}(\log^{3/2} T + \log T \log^{1/2} n)\Big),$$

which is of the same order as stated in Proposition 4.2. This completes the proof.

A | Appendix A: Supplementary Material For Chapter 1

Appendix A: Supplementary Material For Chapter 1

With First-Order Feedback 1.3.1

Proof of Lemma 1.4. Let $x \in \mathcal{X}$ be arbitrary and denote $g_t(\hat{x}_t) := \partial f_t(\hat{x}_t)$. By convexity of f_t,

$$f_t(\hat{x}_t) - f_t(x) \;\leqslant\; \langle g_t(\hat{x}_t), \hat{x}_t - x \rangle.$$

Insert the stochastic gradient $G_t(\hat{x}_t)$ and define the noise term $\xi_t := g_t(\hat{x}_t) - G_t(\hat{x}_t)$ to obtain

$$f_t(\hat{x}_t) - f_t(x) \;\leqslant\; \langle \xi_t, \hat{x}_t - x \rangle + \langle G_t(\hat{x}_t), \hat{x}_t - x \rangle.$$

The second inner product can be controlled by the standard projection argument: because $\hat{x}_{t+1} = \mathrm{Proj}_{\mathcal{Y}}\big(\hat{x}_t - \eta_t G_t(\hat{x}_t)\big)$,

$$\langle G_t(\hat{x}_t), \hat{x}_t - x \rangle \leqslant \frac{\|\hat{x}_t - x\|^2 - \|\hat{x}_{t+1} - x\|^2}{2\eta_t} + \frac{\eta_t}{2}\|G_t(\hat{x}_t)\|^2.$$

Summing this inequality from $t = 1$ to T, using $\|G_t(\hat{x}_t)\| \leqslant G$ and the diameter bound $\|\hat{x}_t - x\| \leqslant$

D, gives

$$\sum_{t=1}^{T} \langle G_t(\hat{x}_t), \hat{x}_t - x \rangle \leqslant \frac{D^2}{2\eta_T} + \frac{\eta_T T}{2} G^2.$$

With the prescribed constant stepsize $\eta_T = \gamma D / \left(G\sqrt{T}\right)$ this becomes $\frac{1}{2}\left(\frac{1}{\gamma} + \gamma\right) GD\sqrt{T}$.

Turn now to the martingale term $Z_t := \langle \xi_t, \hat{x}_t - x \rangle$. Conditioned on the past, $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ because $\mathbb{E}[G_t(\hat{x}_t) \mid \mathcal{F}_{t-1}] = g_t(\hat{x}_t)$. Moreover $|Z_t| \leqslant \|\xi_t\| \, \|\hat{x}_t - x\| \leqslant 2GD$. Freedman's inequality therefore yields, for any $\delta \in (0, 1)$,

$$\Pr\left\{ \sum_{t=1}^{T} Z_t > 2GD\sqrt{2T \log(1/\delta)} \right\} \leqslant \delta.$$

Combining the deterministic and stochastic bounds and using a union bound gives, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \left(f_t(\hat{x}_t) - f_t(x) \right) \leqslant 2GD\sqrt{2T \log(1/\delta)} + \frac{1}{2}\left(\frac{1}{\gamma} + \gamma\right) GD\sqrt{T}.$$

Because the right-hand side is independent of x, taking the maximum over $x \in \mathcal{X}$ proves the lemma.

□

Proof of Proposition 1.3. Fix $g \in \{1, \dots, K\}$ and set the (hypothetical) restart window $\Delta_T^g \triangleq \lceil T^{\frac{2}{3}(1-\nu_g)} \rceil$. Partition the horizon into $J_g \triangleq \lceil T / \Delta_T^g \rceil$ batches $\{\mathcal{T}_{g,j}\}_{j=1}^{J_g}$ with

$$\mathcal{T}_{g,j} \;=\; \left\{ (j-1)\Delta_T^g + 1, \dots, (j\Delta_T^g) \wedge T \right\}.$$

By Assumption 2 (strong convexity with parameter $\underline{\delta}$) we have, for all t, $f_t(\hat{x}_t^g) - f_t(x_t^*) \geqslant \underline{\delta} \, \|\hat{x}_t^g -$

$x_t^*\|^2$. Summing over t and decomposing batchwise yields

$$\sum_{t=1}^{T} \|\hat{x}_t^g - x_t^*\|^2 \leqslant \frac{1}{\underline{\delta}} \sum_{t=1}^{T} \left(f_t(\hat{x}_t^g) - f_t(x_t^*)\right)$$
$$= \frac{1}{\underline{\delta}} \sum_{j=1}^{J_g} \Bigg(\underbrace{\sum_{t\in\mathcal{T}_{g,j}} f_t(\hat{x}_t^g) - \min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w)}_{①} + \underbrace{\min_{w\in\mathcal{X}} \sum_{t\in\mathcal{T}_{g,j}} f_t(w) - \sum_{t\in\mathcal{T}_{g,j}} f_t(x_t^*)}_{②} \Bigg). \quad \text{(A1)}$$

Bounding ② (functional drift). By the standard drifting-comparator bound (e.g., Proposition 2 in [Besbes et al. 2015]),

$$\sum_{j=1}^{J_g} ② \;\leqslant\; 2\,\Delta_T^g \sum_{t=2}^{T} \|f_t - f_{t-1}\|_\infty \;=\; 2\,\Delta_T^g\, V_T.$$

Using $\Delta_T^g = T^{\frac{2}{3}(1-\nu_g)}$ (up to ceilings) and the variation budget assumption $V_T = O(T^\nu)$, together with the grid choice (which guarantees T^{ν_g} is within a universal constant factor of T^ν), we obtain

$$\sum_{j=1}^{J_g} ② \;\leqslant\; c\,T^{\frac{2+\nu_g}{3}},$$

for a universal constant $c > 0$. Moreover, since $\|f_t - f_{t-1}\|_\infty \leqslant GD$ by $\sup_{x\in\mathcal{X}} \|\nabla f_t(x)\| \leqslant G$ and $\operatorname{diam}(\mathcal{X}) \leqslant D$, we may take $c = 2GD$ so that

$$\sum_{j=1}^{J_g} ② \;\leqslant\; 2\,GD\,T^{\frac{2+\nu_g}{3}}. \quad \text{(A2)}$$

Bounding ① (within-batch static regret) with high probability. On each batch $\mathcal{T}_{g,j}$ we run SGD with stepsize $\eta_T^g = \gamma D/(G\sqrt{\Delta_T^g})$ (Algorithm 1, Line 10). Applying Lemma 1.4 to the sub-horizon $\mathcal{T}_{g,j}$ and choosing $\delta = T^{-2}$ gives, with probability at least $1-\delta$,

$$① \;\leqslant\; 2GD\sqrt{2\,\Delta_T^g \log(T^2)} + \tfrac{1}{2}\Big(\tfrac{1}{\gamma} + \gamma\Big) GD\sqrt{\Delta_T^g} \;\leqslant\; GD\sqrt{\Delta_T^g}\left(4\sqrt{\log T} + \tfrac{1}{2}\big(\tfrac{1}{\gamma} + \gamma\big)\right).$$

A union bound over the $J_g \leqslant T$ batches implies that, with probability at least $1 - 1/T$, the above holds simultaneously for all $j = 1, \dots, J_g$; summing over j and using $J_g \sqrt{\Delta_T^g} \leqslant T/\sqrt{\Delta_T^g}$ yields

$$\sum_{j=1}^{J_g} ① \;\leqslant\; \frac{T}{\sqrt{\Delta_T^g}}\, GD\left(4\sqrt{\log T} + \tfrac{1}{2}\left(\tfrac{1}{\gamma} + \gamma\right)\right) \;=\; GD\, T^{\frac{2+\nu_g}{3}}\left(4\sqrt{\log T} + \tfrac{1}{2}\left(\tfrac{1}{\gamma} + \gamma\right)\right). \tag{A3}$$

Combine. Plugging (A2) and (A3) into (A1) gives, with probability at least $1 - 1/T$,

$$\sum_{t=1}^{T} \|\hat{x}_t^g - x_t^*\|^2 \;\leqslant\; \frac{T^{\frac{2+\nu_g}{3}}}{\underline{\delta}}\left(4GD\sqrt{\log T} + \tfrac{1}{2}\left(\tfrac{1}{\gamma} + \gamma\right)GD + 2GD\right).$$

Finally, applying a union bound over $g = 1, \dots, K$ yields that the event $\mathcal{G}_{\phi^{(1)}}$ (the intersection of the per-g events) holds with probability at least $1 - K/T$, establishing the claim.

□

Proof of Proposition 1.5: ASGD never over-estimates the variation budget. Let

$$\Xi_T \;:=\; 4GD\sqrt{\log T} \;+\; \tfrac{1}{2}\left(\gamma + \tfrac{1}{\gamma}\right) GD \;+\; 2GD,$$

so that Proposition 1.3 (the high-probability estimate on $\mathcal{G}_{\phi^{(1)}}$) can be written compactly as

$$\sum_{t=1}^{T} \|\hat{x}_t^g - x_t^*\|^2 \;\leqslant\; \frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+\nu_g}{3}} \qquad \text{for every } g \in \{1, \dots, K\}. \tag{A4}$$

We argue by contradiction. Assume there exists a (first) time $t_0 \in \{2, \dots, T\}$ such that the algorithm switches to an index strictly larger than k^*, i.e.,

$$k(t_0 - 1) \leqslant k^* \quad \text{and} \quad k(t_0) = k^* + 1.$$

By Line 5 of Algorithm 1 (the **if**-condition evaluated at time t_0), there must exist some $g \in \{k^* +$

$1, \dots, K\}$ such that

$$\sum_{s=t_{\text{if}}}^{t_0-1} \|\hat{x}_s^{k^*} - \hat{x}_s^g\|^2 \;\geqslant\; 2\,\frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+v_g}{3}}, \tag{A5}$$

where t_{if} is the last time at which a switch occurred (initialized to 1).

On the other hand, on the good event $\mathcal{G}_{\phi^{(1)}}$ we may bound the *full* pairwise deviation via the elementary inequality $\|a-b\|^2 \leqslant 2\|a-c\|^2 + 2\|c-b\|^2$ (with $c = x_t^*$), followed by (A4):

$$\begin{aligned}\sum_{s=1}^{T} \|\hat{x}_s^{k^*} - \hat{x}_s^g\|^2 &\leqslant 2\sum_{s=1}^{T} \|\hat{x}_s^{k^*} - x_s^*\|^2 + 2\sum_{s=1}^{T} \|x_s^* - \hat{x}_s^g\|^2 \\ &\leqslant 2\,\frac{\Xi_T}{\underline{\delta}}\left(T^{\frac{2+v_{k^*}}{3}} + T^{\frac{2+v_g}{3}}\right) \;\leqslant\; 2\,\frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+v_g}{3}},\end{aligned}$$

where the last inequality uses $v_{k^*} \leqslant v_g$. Since $\sum_{s=t_{\text{if}}}^{t_0-1} \cdot \;\leqslant\; \sum_{s=1}^{T} \cdot$, we obtain

$$\sum_{s=t_{\text{if}}}^{t_0-1} \|\hat{x}_s^{k^*} - \hat{x}_s^g\|^2 \;\leqslant\; 2\,\frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+v_g}{3}},$$

which contradicts the trigger condition (A5) that is necessary for the switch at time t_0.

Therefore no such t_0 can exist, and the algorithm never selects an index larger than k^*. Hence $k(t) \leqslant k^*$ for all $t \in [T]$.

□

Proof of Theorem 1.2. Recall the grid $\{v_g\}_{g=1}^K$ and let

$$k^* \;:=\; \min\{g \in \{1, \dots, K\} :\; v \leqslant v_g\}, \qquad \Xi_T \;:=\; 4GD\sqrt{\log T} + \tfrac{1}{2}\left(\gamma + \tfrac{1}{\gamma}\right) GD + 2GD.$$

By Proposition 1.3, on the good event $\mathcal{G}_{\phi^{(1)}}$ we have, for every g,

$$\sum_{t=1}^{T} \|\hat{x}_t^g - x_t^*\|^2 \;\leqslant\; \frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+v_g}{3}}. \tag{A6}$$

By Proposition 1.5, $k(t) \leqslant k^*$ for all t. Let the (at most K) switch times of Algorithm 1 be

$$1 < t_1 < \cdots < t_M \leqslant T, \qquad (M \leqslant K),$$

and set $t_0 := 1$, $t_{M+1} := T + 1$. For $m = 0, 1, \ldots, M$ denote the closed interval between two consecutive switches by

$$I_m := [\, t_m,\ t_{m+1} - 1 \,],$$

on which the **if**-condition is *not* triggered. Write $\hat{x}_t^{\mathrm{ASGD}} := \hat{x}_t^{k(t)}$.

Regret on a single no-switch interval. By Assumption 2 (upper quadratic growth with curvature parameter $\bar{\delta}$) and the inequality $\|a - b\|^2 \leqslant 2\|a - c\|^2 + 2\|c - b\|^2$ with $c = x_t^*$,

$$\begin{aligned}\sum_{t \in I_m} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) &\leqslant \bar{\delta} \sum_{t \in I_m} \|\hat{x}_t^{\mathrm{ASGD}} - x_t^*\|^2 \\ &\leqslant 2\bar{\delta} \sum_{t \in I_m} \|\hat{x}_t^{\mathrm{ASGD}} - \hat{x}_t^{k^*}\|^2 + 2\bar{\delta} \sum_{t \in I_m} \|\hat{x}_t^{k^*} - x_t^*\|^2. \end{aligned} \tag{A7}$$

Because the **if**-condition is *not* triggered on I_m and $k(t) \leqslant k^*$, its negation with $g = k^*$ yields

$$\sum_{t \in I_m} \|\hat{x}_t^{\mathrm{ASGD}} - \hat{x}_t^{k^*}\|^2 \ \leqslant\ 2\,\frac{\Xi_T}{\underline{\delta}}\, T^{\frac{2+v_{k^*}}{3}}. \tag{A8}$$

Using (A6) with $g = k^*$ and combining with (A7)–(A8), we get

$$\sum_{t \in I_m} \big(f_t(\hat{x}_t^{\mathrm{ASGD}}) - f_t(x_t^*)\big) \ \leqslant\ 6\,\frac{\bar{\delta}}{\underline{\delta}}\,\Xi_T\, T^{\frac{2+v_{k^*}}{3}}. \tag{A9}$$

From v_{k^*} to V_T. By the standard discretization used to define the grid, $T^{v_{k^*}} \leqslant e\, V_T$ (i.e., the surrogate budget indexed by k^* approximates the true V_T within a constant factor), hence

$$T^{\frac{2+v_{k^*}}{3}} \ \leqslant\ e^{1/3}\, T^{\frac{2}{3}}\, V_T^{\frac{1}{3}}.$$

Substituting into (A9),

$$\sum_{t\in I_m}\big(f_t(\hat{x}_t^{\mathrm{ASGD}})-f_t(x_t^*)\big)\ \leqslant\ C_1\,\frac{\bar{\delta}}{\underline{\delta}}\,\Xi_T\,T^{\frac{2}{3}}\,V_T^{\frac{1}{3}},$$

for an absolute constant $C_1>0$.

Summing intervals and the terminal tail. There are $M+1\leqslant K+1$ no-switch intervals $\{I_m\}_{m=0}^{M}$. Summing the above bound and using that the grid size satisfies $K\leqslant C_0\log T$ (geometric spacing in the window/variation grid), we obtain, on $\mathcal{G}_{\phi^{(1)}}$,

$$\sum_{t=1}^{T}\big(f_t(\hat{x}_t^{\mathrm{ASGD}})-f_t(x_t^*)\big)\ \leqslant\ C_2\,\frac{\bar{\delta}}{\underline{\delta}}\,\Xi_T\,T^{\frac{2}{3}}\,V_T^{\frac{1}{3}}\ \log T.$$

Recalling $\Xi_T=\Theta(GD\sqrt{\log T})$, the right-hand side is

$$O\Big(T^{\frac{2}{3}}V_T^{\frac{1}{3}}(\log T)^{\frac{3}{2}}\Big).$$

Accounting for the bad event. Finally, $\mathbb{P}(\mathcal{G}_{\phi^{(1)}}^{\complement})\leqslant K/T\leqslant C_0(\log T)/T$ by Proposition 1.3. On $\mathcal{G}_{\phi^{(1)}}^{\complement}$ the regret is trivially bounded by $T\cdot\max_{t,x\in\mathcal{X}}f_t(x)$, so its expected contribution is $O(\log T)$ and is dominated by the main term above. This completes the proof.

□

With Zeroth-Order & First-Order Feedback 1.3.2

Proof of Proposition 1.7. Write the canonical decomposition

$$\sum_{t=1}^{T}\big[\phi^{(0)}(\hat{x}_t^g,f_t)-\phi^{(0)}(x_t^*,f_t)\big]=\underbrace{\sum_{t=1}^{T}\big[\phi^{(0)}(\hat{x}_t^g,f_t)-f_t(\hat{x}_t^g)\big]}_{\text{(I)}}+\underbrace{\sum_{t=1}^{T}\big[f_t(\hat{x}_t^g)-f_t(x_t^*)\big]}_{\text{(II)}}+\underbrace{\sum_{t=1}^{T}\big[f_t(x_t^*)-\phi^{(0)}(x_t^*,f_t)\big]}_{\text{(III)}}.\tag{A10}$$

Noise terms. Let $\varepsilon_t(x) := \phi^{(0)}(x, f_t) - f_t(x)$. By the zeroth–order feedback assumption, for each fixed x the sequence $\{\varepsilon_t(x)\}_{t=1}^T$ is mean–zero and σ_0–sub–Gaussian (conditionally on the past). Moreover, for any predictable action sequence $\{a_t\}$, the process $\{\varepsilon_t(a_t)\}$ is a martingale difference sequence with the same sub–Gaussian proxy. Hence, by Azuma–Hoeffding (or the standard sub–Gaussian tail bound),

$$\Pr\left(\left|\sum_{t=1}^{T} \varepsilon_t(a_t)\right| > 2\sigma_0\sqrt{T\log T}\right) \leqslant \frac{2}{T^2}.$$

Apply this once with $a_t = \hat{x}_t^g$ to control (I), and once with $a_t = x_t^*$ to control (III). A union bound over the at most K relevant learners (for (I)) together with the single comparator sequence (for (III)) yields that, with probability at least $1 - \frac{2K}{T}$,

$$|(\mathrm{I})| \leqslant 2\sigma_0\sqrt{T\log T} \qquad \text{and} \qquad |(\mathrm{III})| \leqslant 2\sigma_0\sqrt{T\log T}, \tag{A11}$$

simultaneously for all $g \in \{k^*, \dots, K\}$.

Define

$$\Xi_T := 4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma + \gamma^{-1})\, GD + 2GD.$$

By Proposition 1.3 (proved earlier for the first–order estimator run with window size Δ_T^g and step size $\eta_T^g = \gamma D/(G\sqrt{\Delta_T^g})$), we have for each fixed g

$$\sum_{t=1}^{T}\left[f_t(\hat{x}_t^g) - f_t(x_t^*)\right] \leqslant T^{\frac{2+\nu_g}{3}}\,\Xi_T$$

with failure probability at most $1/T$. A union bound over $g \in \{k^*, \dots, K\}$ gives that, with probability at least $1 - \frac{K}{T}$, the above bound holds *simultaneously* for all such g:

$$(\mathrm{II}) \leqslant T^{\frac{2+\nu_g}{3}}\,\Xi_T \qquad \text{for all } g \in \{k^*, \dots, K\}. \tag{A12}$$

Intersecting the events in (A11) and (A12) and using (A10), we obtain, for every $g \in \{k^*, \dots, K\}$,

$$\left| \sum_{t=1}^{T} \left[\phi^{(0)}(\hat{x}_t^g, f_t) - \phi^{(0)}(x_t^*, f_t) \right] \right| \leqslant T^{\frac{2+v_g}{3}} \Xi_T + 4\sigma_0 \sqrt{T \log T},$$

with probability at least $1 - \frac{3K}{T}$. This is exactly the event $\mathcal{G}_{\phi^{(0,1)}}$, completing the proof.

□

Proof of Proposition 1.8: No over-estimation under hybrid feedback. Suppose, for contradiction, that the algorithm first switches from k^* to k^*+1 at some time $t_0 \geqslant 2$. Let t_{if} denote the start of the current monitoring window. By the switch rule (Lines 5–6), there exists $g \in \{k^*+1, \dots, K\}$ such that

$$\sum_{s=t_{\text{if}}}^{t_0-1} \left| \phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s) \right| \geqslant \text{RHS}(g) = B_{k^*} + B_g. \tag{A13}$$

On $\mathcal{G}_{\phi^{(0,1)}}$ we may insert the oracle action x_s^* and apply the triangle inequality:

$$\left| \phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s) \right| \leqslant \left| \phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(x_s^*, f_s) \right| + \left| \phi^{(0)}(\hat{x}_s^g, f_s) - \phi^{(0)}(x_s^*, f_s) \right|.$$

Summing over $s \in [t_{\text{if}}, t_0 - 1]$ and invoking the bounds in (1.12) yields

$$\sum_{s=t_{\text{if}}}^{t_0-1} \left| \phi^{(0)}(\hat{x}_s^{k^*}, f_s) - \phi^{(0)}(\hat{x}_s^g, f_s) \right| \leqslant B_{k^*} + B_g = \text{RHS}(g),$$

which contradicts (A13). Hence no switch to an index $> k^*$ can occur, and $k(t) \leqslant k^*$ for all $t \in [T]$.

□

Proof of Theorem 1.6. Fix the variation grid $\{v_g\}_{g=1}^K$ and let

$$B_g := T^{\frac{2+v_g}{3}} \left(4GD\sqrt{\log T} + \tfrac{1}{2}(\gamma + \gamma^{-1})GD + 2GD \right) + 4\sigma_0 \sqrt{T \log T}\,.$$

By Proposition 1.7, with probability at least $1 - \frac{3K}{T}$ the joint "good" event

$$\mathcal{G}_{\phi^{(0,1)}} = \left\{ \sum_{t=1}^{T} \left| \phi^{(0)}(\hat{x}_t^g, f_t) - \phi^{(0)}(x_t^*, f_t) \right| \leqslant B_g \ \text{ for all } g \in \{k^*, \dots, K\} \right\}$$

holds. Condition on $\mathcal{G}_{\phi^{(0,1)}}$ for the remainder of the argument; we remove this conditioning at the end.

No over-estimation and interval decomposition. By Proposition 1.8, $k(t) \leqslant k^*$ for all $t \in [T]$. Let the (at most K) switch times be $1 < t_1 < \cdots < t_M \leqslant T$, and set $t_0 := 1$ and $t_{M+1} := T+1$. Define the no-switch intervals $I_m := [t_m,\ t_{m+1} - 1]$ for $m = 0, 1, \dots, M$ (the case $m = M$ is the terminal tail). Since the **if**-test never triggers on I_m and $k(t) \leqslant k^*$, taking $g = k^*$ in Lines 5–6 of Algorithm 2 yields the negation

$$\sum_{t \in I_m} \left| \phi^{(0)}(\hat{x}_t^k, f_t) - \phi^{(0)}(\hat{x}_t^{k^*}, f_t) \right| \ < \ 2B_{k^*}. \tag{A14}$$

Regret on a no-switch interval. For each $t \in I_m$,

$$\phi^{(0)}(\hat{x}_t^k, f_t) - \phi^{(0)}(x_t^*, f_t) = \underbrace{\left[\phi^{(0)}(\hat{x}_t^k, f_t) - \phi^{(0)}(\hat{x}_t^{k^*}, f_t) \right]}_{\text{"model gap"}} + \underbrace{\left[\phi^{(0)}(\hat{x}_t^{k^*}, f_t) - \phi^{(0)}(x_t^*, f_t) \right]}_{\text{"oracle gap"}}.$$

Summing over $t \in I_m$ and applying (A14) together with the good–event bound for $g = k^*$ (from Proposition 1.7) gives

$$\sum_{t \in I_m} \left[\phi^{(0)}(\hat{x}_t^k, f_t) - \phi^{(0)}(x_t^*, f_t) \right] \ \leqslant \ 2B_{k^*} + B_{k^*} \ = \ 3B_{k^*}. \tag{A15}$$

Taking expectations and using $\mathbb{E}[\phi^{(0)}(x, f_t)] = f_t(x)$ yields

$$\mathbb{E}\left[\sum_{t \in I_m} \left(f_t(\hat{x}_t^k) - f_t(x_t^*) \right) \middle| \, \mathcal{G}_{\phi^{(0,1)}} \right] \ \leqslant \ 3B_{k^*}. \tag{A16}$$

Summation over intervals. There are $M+1 \leqslant K+1$ intervals $\{I_m\}_{m=0}^{M}$, hence on $\mathcal{G}_{\phi^{(0,1)}}$

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_t(\hat{x}_t^{\pi_2}) - f_t(x_t^*)\right) \,\middle|\, \mathcal{G}_{\phi^{(0,1)}}\right] \;\leqslant\; 3\,(K+1)\,B_{k^*}.$$

Removing the conditioning and simplifying. Let $C_{\max} := \sup_{t,x\in\mathcal{X}} |f_t(x)| < \infty$ (boundedness as in the dynamic-regret definition). On $\mathcal{G}^{\complement}_{\phi^{(0,1)}}$, the regret is at most $2TC_{\max}$, so

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_t(\hat{x}_t^{\pi_2}) - f_t(x_t^*)\right)\right] \;\leqslant\; 3\,(K+1)\,B_{k^*} \;+\; \Pr(\mathcal{G}^{\complement}_{\phi^{(0,1)}})\,2TC_{\max} \;\leqslant\; 3\,(K+1)\,B_{k^*} \;+\; \tfrac{6K}{T}\,TC_{\max}.$$

By construction of the exponent grid and the definition of k^*, $T^{v_{k^*}} \leqslant e\,V_T$, hence

$$B_{k^*} \;\leqslant\; c_1\,T^{\frac{2}{3}}V_T^{\frac{1}{3}}\sqrt{\log T} \;+\; c_2\,\sqrt{T\log T}$$

for absolute constants c_1, c_2 depending only on (G, D, γ, σ_0). Since $K = \Theta(\log T)$, the leading term is

$$3\,(K+1)\,B_{k^*} \;=\; O\Big(T^{\frac{2}{3}}V_T^{\frac{1}{3}}(\log T)^{\frac{3}{2}}\Big),$$

and the $\sqrt{T\log T}$ and $O(K)$ contributions are lower order. This proves

$$\mathcal{R}^{\pi_2}_{\phi^{(0,1)}}(T, V_T) \;=\; O\Big(T^{\frac{2}{3}}\,V_T^{\frac{1}{3}}\,(\log T)^{\frac{3}{2}}\Big).$$

□

B | Appendix B: Supplementary Material For Chapter 3

Appendix B: Supplementary Material For Chapter 3

Proof of Lemma 3.1

Proof. For one–dimensional distributions $F(\cdot)$ and $G(\cdot)$, the Wasserstein distance admits the closed form

$$\mathcal{W}(F,G) \;=\; \int_0^{\infty} \big|\, F(x) - G(x) \,\big|\, dx.$$

For each $t \geqslant 1$ and $y \geqslant 0$, the newsvendor cost can be written as

$$C_t(y) \;=\; h\,\mathbb{E}\big[(y - D_t)^+\big] + b\,\mathbb{E}\big[(D_t - y)^+\big] \;=\; h\int_0^{y} \mathbb{P}(D_t < x)\,dx + b\int_y^{\infty} \mathbb{P}(D_t > x)\,dx.$$

Let $\bar{F}_t(x) := 1 - F_t(x)$ be the complementary CDF of D_t. For any $1 \leqslant t_1 < t_2 \leqslant T$, we have

$$
\begin{aligned}
&\left| C_{t_2}(y) - C_{t_1}(y) \right| \\
&= \left| \sum_{t=t_1+1}^{t_2} \left[C_t(y) - C_{t-1}(y) \right] \right| \\
&\leqslant h \sum_{t=t_1+1}^{t_2} \int_0^y \left| F_t(x) - F_{t-1}(x) \right| dx \;+\; b \sum_{t=t_1+1}^{t_2} \int_y^\infty \left| \bar{F}_t(x) - \bar{F}_{t-1}(x) \right| dx \\
&= h \sum_{t=t_1+1}^{t_2} \int_0^y \left| F_t(x) - F_{t-1}(x) \right| dx \;+\; b \sum_{t=t_1+1}^{t_2} \int_y^\infty \left| F_t(x) - F_{t-1}(x) \right| dx \\
&\leqslant \max\{h, b\} \sum_{t=t_1+1}^{t_2} \int_0^\infty \left| F_t(x) - F_{t-1}(x) \right| dx \\
&= \max\{h, b\} \sum_{t=t_1+1}^{t_2} \mathcal{W}\big(F_t, F_{t-1}\big).
\end{aligned}
$$

Taking the supremum over $y \geqslant 0$ preserves the bound, which proves the claim. □

Inventory Application: High Probability Bound of OGD

In this part, we build the high probability bound for any trajectory of $\{\hat{y}_t^g\}_{t=1}^T$ induced by the stepsize η_t^g with $g \geq f$.

Lemma B.1. *(High Probability Bound) Suppose the true variation budget of the environment is v, and we have $v_g \geq v_{k^*} \geq v$. Then, with probability at least $1 - \frac{1}{T}$, the gradient descent decisions $\{\hat{y}_t^g\}$ using step-size implied by v_g satisfy*

$$
\sum_{t=1}^{T} \left[C_t(\hat{y}_t^g) - C_t(q_t^*) \right] \leq T^{\frac{2+v_g}{3}} \cdot \left[2\max\{h, b\}M + \left(C_0\sqrt{\log T} + C_1 \right) \right].
$$

with $C_0 = \frac{4\sqrt{6}}{3} \max\{h, b\}M$, $C_1 = \frac{1}{2}(\gamma + \frac{1}{\gamma}) \max\{h, b\}M$.

Proof.

$$\sum_{t=1}^{T}\left[C_t(\hat{y}_t^g)-C_t(q_t^*)\right]=\sum_{j=1}^{\lceil T/\Delta_{T_g}\rceil}\left(\sum_{t\in\mathcal{T}_j}C_t(\hat{y}_t^g)-\min_{w\in[0,M]}\sum_{t\in\mathcal{T}_j}C_t(w)\right)+\sum_{j=1}^{\lceil T/\Delta_{T_g}\rceil}\left(\min_{w\in[0,M]}\sum_{t\in\mathcal{T}_j}C_t(w)-\sum_{t\in\mathcal{T}_j}C_t(q_t^*)\right)$$

By lemma B.2, we have

$$\leq\left[\lceil T/\Delta_{T_g}\rceil\cdot\left\{2GD\sqrt{2\Delta_{T_g}\log(\Delta_{T_g}^2)}+\frac{1}{2}(\gamma+\frac{1}{\gamma})\max\{h,b\}M\sqrt{\Delta_{T_g}}\right\}\right.$$
$$\left.+2\max\{h,b\}\Delta_{T_g}B_T\right]$$

Take $\Delta_{T_g}=T^{\frac{2(1-\nu_g)}{3}}$, then

$$\leqslant T^{\frac{2+\nu_g}{3}}\left(C_0\sqrt{\log T}+C_1\right)+2\max\{h,b\}M\,T^{\frac{2(1-\nu_g)}{3}}\,T^{\nu}$$

Note that $\nu\leq\nu_f\leq\nu_g$

$$\leq T^{\frac{2+\nu_g}{3}}\cdot\left[2\max\{h,b\}M+\left(C_0\sqrt{\log T}+C_1\right)\right]\tag{B1}$$

□

In the above proof, we would need the following lemma for the second inequality.

Lemma B.2 (High Probability Bound for Online Convex Programming)**.** *For the Online Convex Programming algorithm, where y^* could be any fixed point as well as the optimal fixed point, and for $t\geq 1$, recursively we define:*

$$\hat{y}_{t+1}=P_{[0,M]}(\hat{y}_t-\eta_t\cdot G_t(\hat{y}_t))$$

where $\eta_t=\frac{\gamma M}{\max\{h,b\}\sqrt{T}}$, for some $\gamma>0$ and $G_t(\hat{y}_t)$ is an observed (sub)gradient of $C_t(y)$ at $y=\hat{y}_t$, $D=M$ is the length of decision region and $G=\max\{h,b\}$ is the upper bound for gradient, then with

probability at least $1 - \frac{1}{T^2}$*, we have*

$$\sum_{t=1}^{T} C_t(\hat{y}_t) - \sum_{t=1}^{T} C_t(y^*) \leq 2GD\sqrt{2T\log(T^2)} + \frac{1}{2}(\gamma + \frac{1}{\gamma})GD\sqrt{T}$$

Proof. Proof of Lemma 2 For each time period t, note that $g_t(\hat{y}_t) = \mathbb{E}[G_t(\hat{y}_t)]$, where $G_t(\hat{y}_t)$ is the observed gradient. Then,

$$\begin{aligned} C_t(\hat{y}_t) - C_t(y^*) &\leq \langle g_t(\hat{y}_t), \hat{y}_t - y^* \rangle \\ &= \langle g_t(\hat{y}_t) - G_t(\hat{y}_t), \hat{y}_t - y^* \rangle + \langle G_t(\hat{y}_t), \hat{y}_t - y^* \rangle \\ &\text{Set } \xi_t = g_t(\hat{y}_t) - G_t(\hat{y}_t) \\ &\leq \langle \xi_t, \hat{y}_t - y^* \rangle + \frac{1}{2}\left[\frac{|\hat{y}_t - y^*|^2 - |\hat{y}_{t+1} - y^*|^2}{\eta_t} + \eta_t G^2\right] \end{aligned} \tag{B2}$$

Therefore, summing over all t from 1 to T, we obtain:

$$\begin{aligned} \sum_{t=1}^{T} (C_t(\hat{y}_t) - C_t(y^*)) &\leq \sum_{t=1}^{T} \langle \xi_t, \hat{y}_t - y^* \rangle + \frac{D^2}{2\eta_T} + \frac{1}{2}G^2 \sum_{t=1}^{T} \eta_t \\ &\text{Set } \eta_t = \frac{\gamma}{\sqrt{T}} \cdot \frac{D}{G} \\ &\leq \sum_{t=1}^{T} \langle \xi_t, \hat{y}_t - y^* \rangle + \frac{1}{2}(\gamma + \frac{1}{\gamma})GD\sqrt{T} \end{aligned} \tag{B3}$$

Applying the Freedman Inequality (as shown in Lemma 4), denote $Z_t = \langle \xi_t, \hat{y}_t - y^* \rangle$. Observe that $|Z_t| \leq G \cdot 2D$.

In addition, for each $t \geqslant 1$, let $\mathcal{F}_{t-1}$ be the filtration generated by $\hat{y}_1, \ldots, \hat{y}_{t-1}, G_1(\hat{y}_1), \ldots, G_{t-1}(\hat{y}_{t-1})$. Then we have the following equation:

$$\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \mathbb{E}\big[\langle g_t(\hat{y}_t) - G_t(\hat{y}_t),\ \hat{y}_t - y\rangle \big| \mathcal{F}_{t-1}\big] = \big\langle g_t(\hat{y}_t) - \mathbb{E}[G_t(\hat{y}_t)|\mathcal{F}_{t-1}],\ \hat{y}_t - y\big\rangle = 0,$$

where the second identity holds because given $\mathcal{F}_{t-1}$, $\hat{y}_t$ is fixed, and the last identity follows

from the definition of $G_t(\cdot)$. Therefore, $Z_1, Z_2, \ldots, Z_T$ is a sequence of martingale differences with uniform upper bound $2GD$. Then by applying Lemma B.3, we obtain the following inequality for any $\delta > 0$:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{t=1}^{T} \big(C_t(\hat{y}_t) - C_t(y^*)\big) \geqslant 2GD\sqrt{2T\log(1/\delta)} + \frac{1}{2}\big(\frac{1}{\gamma} + \gamma\big)GD\sqrt{T} \right) \\
&\leqslant \mathbb{P}\Big(\sum_{t=1}^{T} \langle g_t(\hat{y}_t) - G_t(\hat{y}_t),\ y - \hat{y}_t \rangle \geqslant 2GD\sqrt{2T\log(1/\delta)} \Big) \leqslant \delta,
\end{aligned}$$

where the first inequality follows from inequality (B3) and the second inequality follows from Lemma B.3.

□

Lemma B.3 (Azuma-Hoeffding inequality)**.** *Suppose $Z_1, Z_2, \ldots, Z_T$ is a sequence of martingale differences with $|Z_i| \leqslant B$ for each $i = 1, 2, \ldots, T$. Then for any $\delta > 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^{T} Z_i \geqslant B\sqrt{2T\log(1/\delta)} \Big) \leqslant \delta.$$

Lemma B.4 (Deviation of Realized from Expected Cost)**.** *Assume $0 \leqslant D_t \leqslant M$ almost surely and let $C_3 := \max\{h, b\}\, M$. For any predictable sequence $\{y_t\}_{t=1}^{T} \subset [0, M]$ and any $\delta \in (0, 1)$,*

$$\Pr\left(\Big| \textstyle\sum_{t=1}^{T} \big[C(y_t, D_t) - C_t(y_t)\big] \Big| \ > \ C_3\sqrt{2T\log\tfrac{1}{\delta}} \right) \ \leqslant \ \delta. \tag{B4}$$

Consequently, with the probability $1 - \frac{1}{T^2}$,

$$\sum_{t=1}^{T} \big[C(y_t, D_t) - C_t(y_t)\big] \ \leqslant \ C_3\sqrt{2T\log(T^2)}. \tag{B5}$$

Proof. For fixed $y \in [0, M]$ the sequence $\Delta_t(y) = C(y, D_t) - \mathbb{E}[C(y, D_t) \mid \mathcal{F}_{t-1}]$ is a martingale

difference with $|\Delta_t(y_t)| \leqslant C_3$ a.s. Apply Azuma–Hoeffding:

$$\Pr\left(\sum_{t=1}^{T}\Delta_t(y_t) \;>\; C_3\sqrt{2T\log\tfrac{1}{\delta}}\right) \leqslant \delta.$$

. □

Lemma B.5 (High-probability bound on empirical regret)**.** *Suppose*

- *the demands satisfy* $0 \leqslant D_t \leqslant M$ *almost surely;*
- *the one-period cost is*

$$C(y, D) = h\,(y - D)^+ + b\,(D - y)^+, \quad h, b > 0;$$

- *the sequence* $\{\hat{y}_t^g\}$ *is generated by OGD with step-size tuned to a guessed variation budget* $v_g \geqslant v$.

Define

$$C_3 = \max\{h, b\}\,M, \quad C_0 = \frac{4\sqrt{6}}{3}\,C_3, \quad C_1 = \tfrac{1}{2}\left(\gamma + \tfrac{1}{\gamma}\right)C_3.$$

Then with probability at least $1 - \frac{3}{T}$,

$$\sum_{t=1}^{T}\left[C(\hat{y}_t^g, D_t) - C(q_t^*, D_t)\right] \;\leqslant\; T^{\frac{2+v_g}{3}}\left[2\max\{h, b\}M + \left(C_0\sqrt{\log T} + C_1\right)\right] + 2\,C_3\,\sqrt{2T\log(T^2)}.$$

Proof. Write the total regret as

$$\sum_{t=1}^{T}\left[C(\hat{y}_t^g, D_t) - C(q_t^*, D_t)\right] = \underbrace{\sum_{t=1}^{T}\left[C_t(\hat{y}_t^g) - C_t(q_t^*)\right]}_{\text{OGD term}} + \underbrace{\sum_{t=1}^{T}\left[C(\hat{y}_t^g, D_t) - C_t(\hat{y}_t^g)\right] + \sum_{t=1}^{T}\left[C_t(y_t^*) - C(q_t^*, D_t)\right]}_{\text{martingale noise}},$$

where $C_t(y) = \mathbb{E}[C(y, D_t) \mid \mathcal{F}_{t-1}]$.

By Lemma B.1, the OGD term is bounded by

$$T^{\frac{2+\nu_g}{3}}\left[2\max\{h,b\} + \left(C_0\sqrt{\log T} + C_1\right)\right]$$

with probability at least $1 - \frac{1}{T}$. For the two noise sums, note that for any fixed $y \in [0, M]$, $\Delta_t(y) = C(y, D_t) - C_t(y)$ is a mean-zero martingale difference with $|\Delta_t(y)| \leqslant C_3$. By Azuma–Hoeffding, each of $\sum_t \Delta_t(\hat{y}_t^g)$ and $\sum_t \Delta_t(y_t^*)$ exceeds $C_3\sqrt{2T\log(2T)}$ with probability at most $1/T$. Hence both noise sums together are bounded by

$$2\,C_3\,\sqrt{2T\log(T^2)}$$

with probability at least $1 - \frac{2}{T}$. A final union bound across the OGD term and the two noise events shows that all three bounds hold simultaneously with probability at least

$$1 - \left(\tfrac{1}{T} + \tfrac{2}{T}\right) = 1 - \tfrac{3}{T}.$$

Adding the three contributions yields the claimed high-probability regret bound. □

Carry-over Regret Analysis

Proof of Lemma 3.5. For notational convenience set

$$Z_t \;=\; y_t^{\mathrm{ASGD}} - \hat{y}_t^{\mathrm{ASGD}}, \qquad t \geqslant 1.$$

Relation (3.3) together with $\eta_t^{k(t)} = \frac{\gamma}{h\vee b}\,T^{\frac{\nu_{k(t)}-1}{3}}$ implies

$$Z_{t+1} \leqslant \left(Z_t + \gamma\,T^{\frac{\nu_{k(t)}-1}{3}} - D_t\right)^+.$$

Because the exponent $(\nu_{k(t)} - 1)/3$ is non-positive, the factor $T^{(\nu_{k(t)}-1)/3} \leqslant 1$. Hence

$$Z_{t+1} \leqslant \big(Z_t + \gamma - D_t\big)^+. \tag{A.1}$$

Introduce an auxiliary process $\{W_t\}_{t\geqslant 1}$ with $W_1 = 0$ and

$$W_t = (W_{t-1} + \gamma - D_t)^+, \qquad t \geqslant 2. \tag{A.2}$$

The two recursions differ only in that the increment of Z_t never exceeds that of W_t; an induction on t therefore yields

$$Z_t \leqslant W_t \quad \text{for all } t \geqslant 1. \tag{A.3}$$

The process W is the waiting-time sequence of a $GI/D/1$ queue with constant service time γ and inter-arrival times D_t. Let the successive busy cycles of this queue be $J_i = \{t : \tau_{i-1} < t \leqslant \tau_i\}$ for $i \geqslant 1$, where $\tau_0 = 1$ and $\tau_i = \inf\{s > \tau_{i-1} : W_s = 0\}$. For each calendar period t denote by $i(t)$ the index of the busy cycle that contains t.

Pick an arbitrary t and write $i = i(t)$. Because $W_{\tau_{i-1}} = 0$ at the beginning of a busy cycle, repeatedly applying (A.1) from $\tau_{i-1}+1$ up to t and using (A.3) gives

$$Z_t \;\leqslant\; \sum_{s=\tau_{i-1}+1}^{t} \gamma\, T^{\frac{\nu_{k(s)}-1}{3}} \;=\; \sum_{s=1}^{T} \gamma\, T^{\frac{\nu_{k(s)}-1}{3}} \mathbf{1}\{\tau_{i-1} < s \leqslant t\}.$$

Summing this bound over all periods $t = 1, \dots, T$ and interchanging the order of summation

produces

$$\begin{aligned}
\sum_{t=1}^{T} Z_t &\leqslant \sum_{t=1}^{T}\sum_{s=1}^{T} \gamma\, T^{\frac{v_{k(s)}-1}{3}} \mathbf{1}\{\tau_{i(t)-1} < s \leqslant t\} \\
&\leqslant \sum_{t=1}^{T}\sum_{s=1}^{T} \gamma\, T^{\frac{v_{k(s)}-1}{3}} \mathbf{1}\{s \in J_{i(t)}\} \\
&= \gamma \sum_{s=1}^{T} T^{\frac{v_{k(s)}-1}{3}} \underbrace{\sum_{t=1}^{T} \mathbf{1}\{s \in J_{i(t)}\}}_{=|J_{i(s)}|} \\
&= \gamma \sum_{s=1}^{T} T^{\frac{v_{k(s)}-1}{3}} |J_{i(s)}|.
\end{aligned}$$

Recalling the definition of Z_t completes the proof:

$$\sum_{t=1}^{T} \left(y_t^{\mathrm{ASGD}} - \hat{y}_t^{\mathrm{ASGD}}\right) \;\leqslant\; \gamma \sum_{s=1}^{T} T^{\frac{v_{k(s)}-1}{3}} |J_{i(s)}|.$$

□

The proof of Proposition 3.6 makes use of the following three lemmas.

The first lemma is a general result on the 6th moment of the summation of independent random variables:

Lemma B.6. *Let $X_1, X_2, \cdots$ be a sequence of independent random variables such that, for any $i \geq 1$, (i) $\mathbb{E}[X_i] = 0$, (ii) $\mathbb{E}[X_i^6] \leq \gamma$. Then, for any $n \geq 1$,*

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^6\right] \leq 21 n^3 \gamma.$$

Proof of Lemma B.6: Based the fact that $\mathbb{E}[X_i] = 0$ for any $i \geq 1$, we have that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^6\right] =& \frac{6!}{2!2!2!} \sum_{1\leq i_1<i_2<i_3\leq n} \mathbb{E}[X_{i_1}^2]\mathbb{E}[X_{i_2}^2]\mathbb{E}[X_{i_3}^2] + \frac{6!}{3!3!} \sum_{1\leq i_1<i_2\leq n} \mathbb{E}[X_{i_1}^3]\mathbb{E}[X_{i_2}^3] \\ &+ \frac{6!}{4!2!} \sum_{1\leq i_1<i_2\leq n} \mathbb{E}[X_{i_1}^4]\mathbb{E}[X_{i_2}^2] + \frac{6!}{4!2!} \sum_{1\leq i_1<i_2\leq n} \mathbb{E}[X_{i_1}^2]\mathbb{E}[X_{i_2}^4] + \sum_{1\leq i_1\leq n} \mathbb{E}[X_{i_1}^6] \\ \leq& 90\binom{n}{3}\gamma + 20\binom{n}{2}\gamma + 30\binom{n}{2}\gamma + n\gamma \\ \leq& 21n^3\gamma.\end{aligned}$$

Before stepping the second lemma, let us introduce a family of auxiliary nonstationary queues $\{W_t^{(u)}\}_{t\geq u}$ parametrized with $u \in [T]$, where for any $u \in [T]$,

$$W_u^{(u)} = 0 \text{ and } W_{t+1}^{(u)} \coloneqq [W_t^{(u)} + \rho - D_t]^+ \text{ for each } t \geq u. \tag{B6}$$

Then, given $u \in [T]$, we define $\tau_i^{(u)} \coloneqq \inf_s\{s > \tau_{i-1} : W_s^{(u)} = 0\}$ for $i \geq 1$ with $\tau_0^{(u)} = u$. For $k \geq 1$, we denote $J_i^{(u)} \coloneqq \{t : \tau_{i-1}^{(u)} < t \leq \tau_i^{(u)}\}$ to be the i-th busy period, and define $|J_i^{(u)}|$ as the length of busy period $J_i^{(u)}$. Further, we define use the notation $|J_{i^{(u)}(s)}^{(u)}|$ to be the busy period containing time $u + s$ for each $s \in [T]$.

In particular, when considering the special case of $u = 0$, i.e. the queue $\{W_t^{(0)}\}_{t\geq 0}$ we will omit the superscript "(0)" and write W_t, τ_k, J_k and $J_{i(s)}$ for simplicity.

The second lemma gives an uniform upper bound for the tail probability of $|J_1^{(u)}|$.

Lemma B.7. *For all $t \geq u$, if (i) $\mathbb{E}[\tilde{D}_t] \geq \beta > \rho > 0$, and (ii) $\mathbb{E}[\tilde{D}_t^6] \leq \alpha$, then we have that, for any $u \geq 0$ and $l \geq 1$,*

$$\mathbb{P}\left\{|J_1^{(u)}| \geq l]\right\} \leq \frac{21\alpha}{(\beta-\rho)^6}\frac{1}{l^3}.$$

Proof of Lemma B.7: First, we have that

$$\mathbb{P}\left(|J_1^{(u)}| \geq l\right) \leq \mathbb{P}\left(\sum_{t=u}^{u+l}(\rho - \tilde{D}_t) \geq 0\right) = \mathbb{P}\left(\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \tilde{D}_t) \geq \sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \rho)\right).$$

Also, for any $t \geq u$, since $\tilde{D}_t \geq 0$, we have $\mathbb{E}[(\tilde{D}_t - \mathbb{E}[\tilde{D}_t])^6] \leq \max\{\mathbb{E}[\tilde{D}_t^6], \mathbb{E}[(\mathbb{E}[\tilde{D}_t])^6]\} \leq \alpha$. Then, since $\mathbb{E}[\tilde{D}_t] \geq \beta > \rho$, by Markov inequality and Lemma B.6, we have that,

$$\begin{aligned}
\mathbb{P}\left\{\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \tilde{D}_t) \geq \sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \rho)\right\} &\leq \mathbb{P}\left\{\left[\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \tilde{D}_t)\right]^6 \geq \left[\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \rho)\right]^6\right\} \\
&\leq \frac{\mathbb{E}\left[\left(\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \tilde{D}_t)\right)^6\right]}{\left(\sum_{t=u}^{u+l}(\mathbb{E}[\tilde{D}_t] - \rho)\right)^6} \\
&\leq \frac{21l^3\alpha}{l^6(\beta-\rho)^6} = \frac{21\alpha}{l^3(\beta-\rho)^6}.
\end{aligned}$$

That is to say,

$$\mathbb{P}\left(|J_1^{(u)}| \geq l\right) \leq \frac{21\alpha}{(\beta-\rho)^6}\frac{1}{l^3}.$$

□

Based on Lemma B.7, we can derive the following lemma which gives an uniform upper bound for the tail expectation of $|J_1^{(u)}|$.

Lemma B.8. *For all $t \geq 0$, if (i) $\mathbb{E}[\tilde{D}_t] \geq \beta > \rho > 0$, and (ii) $\mathbb{E}[\tilde{D}_t^6] \leq \alpha$, then we have that, for any $u \geq 0$ and $s \geq 1$,*

$$\mathbb{E}\left\{|J_1^{(u)}| \cdot \mathbf{1}[|J_1^{(u)}| \geq s]\right\} \leq \frac{63\alpha}{2(\beta-\rho)^6}\frac{1}{s^2}.$$

Proof of Lemma B.8: First, we have that

$$\begin{aligned}
\mathbb{E}\left\{|J_1^{(u)}| \cdot \mathbf{1}[|J_1^{(u)}| \geq s]\right\} &= \sum_{k=s}^{\infty} k\mathbb{P}\left\{|J_1^{(u)}| = k\right\} \\
&= \sum_{k=s}^{\infty}\left(\sum_{j=1}^{k} 1\right)\mathbb{P}\left\{|J_1^{(u)}| = k\right\} \\
&= \sum_{j=1}^{\infty}\sum_{k=\max\{s,j\}}^{\infty}\mathbb{P}\left\{|J_1^{(u)}| = k\right\} \\
&= \sum_{j=1}^{s}\mathbb{P}\left\{|J_1^{(u)}| \geq s\right\} + \sum_{j=s+1}^{\infty}\mathbb{P}\left\{|J_1^{(u)}| \geq j\right\} \\
&= s\mathbb{P}\left\{|J_1^{(u)}| \geq s\right\} + \sum_{j=s+1}^{\infty}\mathbb{P}\left\{|J_1^{(u)}| \geq j\right\}
\end{aligned}$$

Then, based on Lemma B.7, we further have

$$\mathbb{E}\left\{|J_1^{(u)}| \cdot \mathbf{1}[|J_1^{(u)}| \geq s]\right\} \leq \frac{21\alpha}{(\beta-\rho)^6}\frac{s}{s^3} + \frac{21\alpha}{(\beta-\rho)^6}\sum_{j=s+1}^{\infty}\frac{1}{j^3} \leq \frac{63\alpha}{2(\beta-\rho)^6}\frac{1}{s^2}.$$

□

With Lemma B.8 in hand, we can proceed to the proof of Proposition 3.6.

Proof of Proposition 3.6:

First, conditioning on the time of the first renewal, we have that

$$\begin{aligned}
\mathbb{E}\{|J_{i^{(1)}(l)}^{(1)}|\} &= \sum_{s=1}^{l-1}\mathbb{E}\{|J_{i^{(1)}(l)}^{(1)}| \cdot \mathbf{1}[|J_1^{(1)}| = s]\} + \mathbb{E}\{|J_1^{(1)}| \cdot \mathbf{1}[|J_1^{(1)}| \geq l]\} \\
&= \sum_{s=1}^{l-1}\mathbb{E}\{|J_{i^{(1+s)}(l-s)}^{(1+s)}| \cdot \mathbf{1}[|J_1^{(1)}| = s]\} + \mathbb{E}\{|J_1^{(1)}| \cdot \mathbf{1}[|J_1^{(1)}| \geq l]\} \\
&= \sum_{s=1}^{l-1}\mathbb{E}\{|J_{i^{(1+s)}(l-s)}^{(1+s)}|\} \cdot \mathbb{P}\{\mathbf{1}[|J_1^{(1)}| = s]\} + \mathbb{E}\{|J_1^{(1)}| \cdot \mathbf{1}[|J_1^{(1)}| \geq l]\} \\
&\leq \max_{1\leq s\leq l-1}\{\mathbb{E}\{|J_{i^{(1+s)}(l-s)}^{(1+s)}|\}\}\} + \mathbb{E}\{|J_1^{(1)}| \cdot \mathbf{1}[|J_1^{(1)}| \geq l]\}.
\end{aligned}$$

Let $s_1 = \arg\max_{1\le s\le l-1}\{\mathbb{E}\{|J^{(1+s)}_{i^{(1+s)}(l-s)}|\}\}$.

Then, we have that,

$$\mathbb{E}\{|J^{(1)}_{i^{(1)}(l)}|\} \le \mathbb{E}\{|J^{(1+s_1)}_{i^{(1+s_1)}(l-s_1)}|\}\} + \mathbb{E}\{|J^{(1)}_1| \cdot \mathbf{1}[|J^{(1)}_1| \ge l]\}.$$

Second, we repeat the above argument on $\mathbb{E}\{|J^{(1+s_1)}_{i(l-s_1)}|\}\}$, and we will get

$$\begin{aligned}
\mathbb{E}\{|J^{(1)}_{i^{(1)}(l)}|\} &\le \mathbb{E}\{|J^{(1+s_1)}_{i^{(1+s_1)}(l-s_1)}|\}\} + \mathbb{E}\{|J^{(1)}_1| \cdot \mathbf{1}[|J^{(1)}_1| \ge l]\} \\
&\le \mathbb{E}\{|J^{(1+s_1+s_2)}_{i^{(1+s_1+s_2)}(l-s_1-s_2)}|\}\} + \mathbb{E}\{|J^{(1+s_1)}_1| \cdot \mathbf{1}[|J^{(1+s_1)}_1| \ge l-s_1] + \mathbb{E}\{|J^{(1)}_1| \cdot \mathbf{1}[|J^{(1)}_1| \ge l]\} \\
&\le \mathbb{E}\{|J^{(1+s_1+s_2)}_{i^{(1+s_1+s_2)}(l-s_1-s_2)}|\}\} + \sum_{k=0}^{s_1} \mathbb{E}\{|J^{(1+k)}_1| \cdot \mathbf{1}[|J^{(1+k)}_1| \ge l-k],
\end{aligned}$$

where $s_2 = \arg\max_{1\le s\le l-s_1-1}\{\mathbb{E}\{|J^{(1+s_1+s)}_{i(l-s_1-s)}|\}\}$.

Third, by an iteration of the above argument, we will finally have,

$$\mathbb{E}\{|J^{(1)}_{i(l)}|\} \le \sum_{k=0}^{l-1} \mathbb{E}\{|J^{(1+k)}_1| \cdot \mathbf{1}[|J^{(1+k)}_1| \ge l-k].$$

Forth, based on Lemma B.8, we conclude that

$$\mathbb{E}\{|J^{(1)}_{i(l)}|\} \le \frac{63\alpha}{2(\beta-\rho)^6}\sum_{k=0}^{l-1}\frac{1}{(l-k)^2} \le \frac{63\alpha}{2(\beta-\rho)^6}\frac{\pi^2}{6} \le \frac{7\pi^2\alpha}{4(\beta-\rho)^6}$$

If we have the forth moment condition, the above result will become

$$\mathbb{E}\{|J^{(1)}_{i(l)}|\} \le \frac{63\alpha}{(\beta-\rho)^4}\sum_{k=0}^{l-1}\frac{1}{l-k} \le \frac{63\alpha}{(\beta-\rho)^4}\log(l)$$

□

C | Appendix C: Supplementary Material For Chapter 4

Appendix C: Supplementary Material For Chapter 4

Universal Portfolio Selection Application: High Probability Bound of OGD

Proof of Lemma 4.1. Fix any $\mathbf{x} \in \Delta_n$ and define the real-valued function

$$g_{\mathbf{x}}(\mathbf{r}) \;:=\; -\ln\!\left(\mathbf{r}^\top \mathbf{x}\right), \qquad \mathbf{r} \in \mathbb{R}^n_+.$$

Step 1: Lipschitz constant of $g_{\mathbf{x}}$. Under Assumption 1 we have $\mathbf{r}^\top \mathbf{x} \geqslant m_{\min}$ for every $\mathbf{r}$ in the support of any P_t and every $\mathbf{x} \in \Delta_n$. The gradient of $g_{\mathbf{x}}$ with respect to $\mathbf{r}$ is

$$\nabla_{\mathbf{r}} g_{\mathbf{x}}(\mathbf{r}) \;=\; -\frac{\mathbf{x}}{\mathbf{r}^\top \mathbf{x}},$$

so that $\left\|\nabla_{\mathbf{r}} g_{\mathbf{x}}(\mathbf{r})\right\|_\infty \;\leqslant\; 1/m_{\min}$. Hence for any $\mathbf{r}, \mathbf{r}' \in \mathbb{R}^n_+$

$$\left|g_{\mathbf{x}}(\mathbf{r}) - g_{\mathbf{x}}(\mathbf{r}')\right| \;\leqslant\; \frac{1}{m_{\min}} \left\|\mathbf{r} - \mathbf{r}'\right\|_1,$$

i.e. $g_{\mathbf{x}}$ is $(1/m_{\min})$-Lipschitz w.r.t. the ℓ_1–metric on $\mathbb{R}^n_+$.

Step 2: Bounding the change in expected loss between two consecutive rounds. By the Kantorovich–Rubinstein dual representation of the 1-Wasserstein distance,

$$\left| {}_{P_s}[g_{\mathbf{x}}] - {}_{P_{s-1}}[g_{\mathbf{x}}] \right| \leqslant \frac{1}{m_{\min}} \mathcal{W}\big(P_s, P_{s-1}\big) \qquad \forall s \geqslant 2.$$

Equivalently,

$$\left| L_s(\mathbf{x}) - L_{s-1}(\mathbf{x}) \right| \leqslant \frac{1}{m_{\min}} \mathcal{W}\big(P_s, P_{s-1}\big).$$

Step 3: Telescoping over t_1+1, …, t_2. For any $1 \leqslant t_1 < t_2 \leqslant T$,

$$L_{t_2}(\mathbf{x}) - L_{t_1}(\mathbf{x}) = \sum_{s=t_1+1}^{t_2} \big(L_s(\mathbf{x}) - L_{s-1}(\mathbf{x})\big),$$

so by the triangle inequality and the bound of Step 2,

$$\left| L_{t_2}(\mathbf{x}) - L_{t_1}(\mathbf{x}) \right| \leqslant \frac{1}{m_{\min}} \sum_{s=t_1+1}^{t_2} \mathcal{W}\big(P_s, P_{s-1}\big).$$

Step 4: Taking the supremum over $\mathbf{x} \in \Delta_n$. The right-hand side above is independent of $\mathbf{x}$, hence

$$\sup_{\mathbf{x}\in\Delta_n} \left| L_{t_2}(\mathbf{x}) - L_{t_1}(\mathbf{x}) \right| \leqslant \frac{1}{m_{\min}} \sum_{s=t_1+1}^{t_2} \mathcal{W}\big(P_s, P_{s-1}\big),$$

which is precisely the desired inequality. □

Throughout this section recall the block length $\Delta_{T_g} = T^{\frac{2(1-\nu_g)}{3}}$ used by expert g and the gradient bound $G = m_{\max}/m_{\min}$. We partition the horizon into $J := \lceil T/\Delta_{T_g} \rceil$ consecutive blocks $\mathcal{T}_1, \ldots, \mathcal{T}_J$ with $J \leqslant T/\Delta_{T_g}$ for all sufficiently large T.

Per–block concentration. To control regret *simultaneously* on every block we invoke Lemma C.3 inside each block at confidence level

$$\delta_{\text{blk}} \;=\; \frac{\Delta_{T_g}}{T^2} \quad\Longrightarrow\quad J\,\delta_{\text{blk}} \;\leqslant\; \frac{T}{\Delta_{T_g}}\,\frac{\Delta_{T_g}}{T^2} \;=\; \frac{1}{T} \;\leqslant\; \frac{1}{T}.$$

A union bound therefore preserves an $\left(1-\frac{1}{T}\right)$ overall success probability.

Lemma C.1 (High–Probability EG Regret on a Block)**.** *Fix any block $\mathcal{T}_j$ of length Δ_{T_g}. Run exponentiated gradient with the step size*

$$\eta_T^g \;=\; \sqrt{\frac{\ln n}{2\,\Delta_{T_g}\,G^2}},$$

and call the iterates $\hat{\mathbf{x}}_t^g$ $(t \in \mathcal{T}_j)$. With probability at least $1-\delta_{blk}$,

$$\sum_{t\in\mathcal{T}_j}\Big[L_t(\hat{\mathbf{x}}_t^g) - \min_{x\in\Delta_n} L_t(x)\Big] \;\leqslant\; 4G\sqrt{2\,\Delta_{T_g}\,\log\frac{T^2}{\Delta_{T_g}}} \;+\; 2G\sqrt{2\,\Delta_{T_g}\,\log n}.$$

Proof. Apply Lemma C.3 with horizon $T \leftarrow \Delta_{T_g}$ and confidence $\delta \leftarrow \delta_{\text{blk}}$; note that $\log\!\left(T^2/\Delta_{T_g}\right) \leqslant 2\log T$. □

Lemma C.2 (High–Probability Regret for Expert g)**.** *Let the environment's total Wasserstein variation satisfy $TV(P_{1:T}) \leqslant B_T = T^v$ for some $v \in [0,1]$, and assume $v_g \geqslant v_{k^*} \geqslant v$. Then, with probability at least $1-\frac{1}{T}$, the portfolios $\{\hat{\mathbf{x}}_t^g\}_{t=1}^T$ generated with step size η_T^g satisfy*

$$\sum_{t=1}^{T}\Big[L_t(\hat{\mathbf{x}}_t^g) - L_t(\mathbf{x}_t^\star)\Big] \;\leqslant\; T^{\frac{2+v_g}{3}}\left[\frac{2}{m_{\min}} + C_0\sqrt{\log T} + C_1\sqrt{\log n}\right],$$

where $C_0 = 8G$ and $C_1 = 2\sqrt{2}\,G$.

Proof. Partition the horizon into $J = \lceil T/\Delta_{T_g}\rceil$ blocks $\mathcal{T}_1,\dots,\mathcal{T}_J$ of length $\Delta_{T_g} = T^{\frac{2(1-v_g)}{3}}$. Decompose

the regret against the dynamic oracle $\mathbf{x}_t^\star := \arg\min_{x\in\Delta_n} L_t(x)$:

$$\sum_{t=1}^{T}\big[L_t(\hat{\mathbf{x}}_t^g) - L_t(\mathbf{x}_t^\star)\big] = \sum_{j=1}^{J}\underbrace{\Big(\sum_{t\in\mathcal{T}_j} L_t(\hat{\mathbf{x}}_t^g) - \min_{x\in\Delta_n}\sum_{t\in\mathcal{T}_j} L_t(x)\Big)}_{(opt_j)} + \sum_{j=1}^{J}\underbrace{\Big(\min_{x}\sum_{t\in\mathcal{T}_j} L_t(x) - \sum_{t\in\mathcal{T}_j} L_t(\mathbf{x}_t^\star)\Big)}_{(tracking_j)}.$$

Step 1: optimisation term.

By Lemma C.1 and the union bound, with overall probability at least $1-\frac{1}{T}$:

$$\sum_{j=1}^{J}\sum_{t\in\mathcal{T}_j}\Big[L_t(\hat{\mathbf{x}}_t^g) - \min_{x\in\Delta_n} L_t(x)\Big] \;\leqslant\; J\Big[4G\sqrt{2\Delta_{T_g}\,2\log T} + 2G\sqrt{2\Delta_{T_g}\,\log n}\Big].$$

Because $J \leqslant T/\Delta_{T_g}$,

$$(\text{opt}) \;:=\; \frac{T}{\Delta_{T_g}}\,4G\sqrt{2\Delta_{T_g}}\big[\sqrt{2\log T}+\sqrt{\log n}\big] \;\leqslant\; T^{\frac{2+\nu_g}{3}}\Big[8G\sqrt{\log T} + 2\sqrt{2\log n}\,G\Big].$$

Step 2: tracking term. By Lemma 4.1, $|L_t(x) - L_{t-1}(x)| \leqslant \frac{1}{m_{\min}}\mathcal{W}(P_t, P_{t-1})$. Within any block, $\min_x \sum_{t\in\mathcal{T}_j} L_t(x) - \sum_{t\in\mathcal{T}_j} L_t(\mathbf{x}_t^\star) \leqslant \frac{2\Delta_{T_g}}{m_{\min}}\mathcal{W}(P_{\tau_j}, P_{\tau_j-1})$. Summing over blocks:

$$(\text{track}) \;:=\; \sum_{j=1}^{J}(tracking_j) \;\leqslant\; \frac{2\Delta_{T_g}}{m_{\min}}\,TV(P_{1:T}) \;\leqslant\; \frac{2\Delta_{T_g}}{m_{\min}}\,T^{\nu}.$$

Because $\nu \leqslant \nu_g$, $\Delta_{T_g}\,T^{\nu} \leqslant T^{\frac{2(1-\nu_g)}{3}}\,T^{\nu} = T^{\frac{2+\nu_g}{3}}$.

Step 3: combine. With probability at least $1-\frac{1}{T}$,

$$\sum_{t=1}^{T}[L_t(\hat{\mathbf{x}}_t^g) - L_t(\mathbf{x}_t^\star)] \;\leqslant\; (\text{opt}) + (\text{track}) \;\leqslant\; T^{\frac{2+\nu_g}{3}}\Big[\frac{2}{m_{\min}} + 8G\sqrt{\log T} + 2\sqrt{2\log n}\,G\Big],$$

which is the claimed bound. □

Lemma C.3 (High-Probability Bound for Exponentiated Gradient). *Let $\{L_t\}_{t=1}^T$ be a sequence of*

convex, differentiable loss functions on the n-simplex $\Delta_n \subset \mathbb{R}^n$. Assume that for each t, the gradient is uniformly bounded: $\|\nabla L_t(x)\|_\infty \leqslant G$ for all $x \in \Delta_n$. Consider the exponentiated-gradient algorithm with updates

$$x_{t+1}(i) \;=\; \frac{x_t(i)\,\exp\bigl(-\eta\,\nabla_t(i)\bigr)}{\sum_{j=1}^{n} x_t(j)\,\exp\bigl(-\eta\,\nabla_t(j)\bigr)},$$

for $i = 1, \ldots, n$, where $\nabla_t(i)$ denotes the i-th component of $\nabla L_t(x_t)$, and choose

$$\eta \;=\; \sqrt{\frac{\ln n}{2\,T\,G^2}}\,.$$

Then for any $\delta > 0$, with probability at least $1 - \delta$, the sequence of decisions $\{x_t\}$ produced by the above updates satisfies

$$\sum_{t=1}^{T}\Bigl[\,L_t(x_t) \;-\; L_t(x^*)\,\Bigr] \;\leqslant\; 4\,G\sqrt{2\,T\,\ln\frac{1}{\delta}} \;+\; 2\,G\,\sqrt{2\,T\,\ln n},$$

where $x^* \in \Delta_n$ is the offline optimal decision that minimizes $\sum_{t=1}^{T} L_t(x)$).

Proof. By convexity of L_t, for any $x^* \in \Delta_n$ we have

$$L_t(x_t) - L_t(x^*) \;\leqslant\; \langle \nabla L_t(x_t),\, x_t - x^* \rangle\,.$$

Now let $f_t(x)$ be the random (or instantaneous) loss incurred at time t (so that $L_t(x) = \mathbb{E}[f_t(x)]$ is the expected loss), and let $G_t(x_t) := \nabla f_t(x_t)$ be the observed stochastic gradient at x_t. Define the martingale difference $\xi_t := \nabla L_t(x_t) - G_t(x_t)$, which satisfies $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ by definition of

L_t. Using $G_t(x_t)$ to rewrite the above, we obtain

$$\begin{aligned} L_t(x_t) - L_t(x^*) &\leqslant \langle \nabla L_t(x_t),\, x_t - x^* \rangle \\ &= \langle G_t(x_t) + \xi_t,\, x_t - x^* \rangle \\ &= \underbrace{\langle G_t(x_t),\, x_t - x^* \rangle}_{(\dagger)} + \underbrace{\langle \xi_t,\, x_t - x^* \rangle}_{Z_t} . \end{aligned} \tag{B1}$$

We will bound the sum of terms (†) (the "optimization error" due to the algorithm's updates) and the sum of Z_t (the martingale noise term) separately. First, we control the optimization error via a standard analysis of the exponentiated gradient method with the negative entropy regularizer. In particular, one can show that for any comparator $x^* \in \Delta_n$, the following inequality holds for all $T \geqslant 1$:

$$\sum_{t=1}^{T} \langle G_t(x_t),\, x_t - x^* \rangle \;\leqslant\; \frac{1}{\eta} D_{\mathrm{KL}}(x^* \,\|\, x_1) \;+\; \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(x_t)\|_\infty^2 , \tag{B2}$$

where $D_{\mathrm{KL}}(x^* \| x_1) = \sum_{i=1}^{n} x^*(i) \ln \frac{x^*(i)}{x_1(i)}$ is the Kullback–Leibler divergence between x^* and the initial decision x_1. *(This inequality can be derived using the Bregman divergence associated with the entropic regularizer; for completeness, a proof is provided at the end of this lemma.)*

Given that x_1 can be chosen as the uniform distribution on $[n]$, we have $D_{\mathrm{KL}}(x^* \| x_1) \leqslant \ln n$. Also, by the gradient bound assumption, $\|\nabla f_t(x_t)\|_\infty \leqslant G$ (since each realized gradient is bounded by the same G). Thus (B2) implies

$$\sum_{t=1}^{T} \langle G_t(x_t),\, x_t - x^* \rangle \;\leqslant\; \frac{\ln n}{\eta} \;+\; \frac{\eta}{2} G^2 T .$$

Now plugging in $\eta = \sqrt{\dfrac{\ln n}{2\,T\,G^2}}$, we obtain

$$\sum_{t=1}^{T} \langle G_t(x_t),\, x_t - x^* \rangle \;\leqslant\; \frac{\ln n}{\eta} + \frac{\eta}{2} G^2 T \;=\; G\sqrt{2T \ln n} \;+\; \frac{1}{2\sqrt{2}}\, G\sqrt{T \ln n} \;\leqslant\; 2\,G\sqrt{2\,T\,\ln n}\,.$$

In other words, the "deterministic" regret due to the exponentiated-gradient updates is bounded by $2\,G\sqrt{2T \ln n}$.

Next, we bound the stochastic term $Z_t = \langle \xi_t,\, x_t - x^* \rangle$. Since $\{\xi_t\}$ is a martingale difference sequence (with respect to the filtration $\mathcal{F}_{t-1}$ that contains all information up to time $t-1$), we can apply a concentration inequality to $\sum_{t=1}^{T} Z_t$. In particular, note that each Z_t is bounded in absolute value: because x_t and x^* are probability vectors that differ in at most 2 units in ℓ_1-norm, and $\|\xi_t\|_\infty \leqslant \|\nabla L_t(x_t)\|_\infty + \|G_t(x_t)\|_\infty \leqslant 2G$, we have

$$|Z_t| \;=\; |\langle \xi_t,\, x_t - x^* \rangle| \;\leqslant\; \|\xi_t\|_\infty \, \|x_t - x^*\|_1 \;\leqslant\; 2G \cdot 2 \;=\; 4G\,.$$

Moreover, $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \langle \mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}],\, x_t - x^* \rangle = 0$, so $Z_1, Z_2, \ldots, Z_T$ are martingale differences with $|Z_t| \leqslant 4G$. We can therefore invoke the Azuma–Hoeffding inequality (or alternatively, the Freedman inequality for martingales) to conclude that for any $\delta > 0$, with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} Z_t \;=\; \sum_{t=1}^{T} \langle \xi_t,\, x_t - x^* \rangle \;\leqslant\; 4G\sqrt{2\,T\,\ln\frac{1}{\delta}}\,.$$

(This follows since $\mathbb{P}\{\sum_{t=1}^{T} Z_t \geqslant 4G\sqrt{2T\ln(1/\delta)}\} \leqslant \delta$ for bounded martingale differences Z_t by Hoeffding's inequality.)

Finally, combining the two parts and using the decomposition (B1), we obtain that with prob-

ability at least $1-\delta$:

$$\sum_{t=1}^{T}\left[L_t(x_t)-L_t(x^*)\right] \;\leqslant\; \underbrace{\sum_{t=1}^{T}\langle G_t(x_t),\, x_t-x^*\rangle}_{\leqslant 2G\sqrt{2T\ln n}} \;+\; \underbrace{\sum_{t=1}^{T}\langle \xi_t,\, x_t-x^*\rangle}_{\leqslant 4G\sqrt{2T\ln(1/\delta)}}.$$

For the chosen η, the first sum is bounded by $2G\sqrt{2T\ln n}$, and the second sum is bounded by $4G\sqrt{2T\ln(1/\delta)}$. So we have

$$\sum_{t=1}^{T}\left[L_t(x_t)-L_t(x^*)\right] \;\leqslant\; 4\,G\sqrt{2\,T\,\ln\frac{1}{\delta}} \;+\; 2\,G\,\sqrt{2\,T\,\ln n}\,.$$

This is exactly the desired high-probability regret bound. □

Lemma C.4 (Deviation of Realized from Expected Loss)**.** *Assume the asset returns are bounded (Assumption (i)), i.e. there exist known constants $m_{\min} > 0$ and $m_{\max} > 0$ such that for all $t \in [T]$ and $i \in [n]$,*

$$m_{\min} \;\leqslant\; r_t(i) \;\leqslant\; m_{\max}.$$

Let $C_3 := \ln\frac{m_{\max}}{m_{\min}}$. For any predictable portfolio sequence $\{x_t\}_{t=1}^{T} \subset \Delta_n$ (where $\Delta_n = \{x \in [0,1]^n : \sum_{i=1}^{n} x(i) = 1\}$) and any $\delta \in (0,1)$, one has

$$\Pr\left(\left|\textstyle\sum_{t=1}^{T}\left[f_t(x_t)-L_t(x_t)\right]\right| \;>\; C_3\sqrt{2T\log\frac{1}{\delta}}\right) \;\leqslant\; \delta\,.$$

Consequently, with probability $1-\frac{1}{T}$,

$$\sum_{t=1}^{T}\left[f_t(x_t)-L_t(x_t)\right] \;\leqslant\; C_3\sqrt{2T\log(T)}\,. \tag{B3}$$

Proof. For any fixed portfolio $x \in \Delta_n$, define the random variable

$$\Delta_t(x) \;=\; f_t(x) - \mathbb{E}[f_t(x) \mid \mathcal{F}_{t-1}] \;=\; f_t(x) - L_t(x)\,.$$

Since x is $\mathcal{F}_{t-1}$-measurable, $\Delta_t(x)$ is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_t\}$. Moreover, by the boundedness of returns (Assumption (i)), we have

$$m_{\min} \;\leqslant\; x \cdot r_t \;\leqslant\; m_{\max} \qquad \text{for each round } t,$$

so the one-period loss $f_t(x) = -\ln(x \cdot r_t)$ is bounded as

$$-\ln(m_{\max}) \;\leqslant\; f_t(x) \;\leqslant\; -\ln(m_{\min})\,.$$

Therefore $|\Delta_t(x)| = |f_t(x) - L_t(x)| \leqslant \ln\frac{m_{\max}}{m_{\min}} = C_3$ almost surely. Applying the Azuma–Hoeffding inequality, we obtain for any $\delta \in (0,1)$:

$$\Pr\!\left(\sum_{t=1}^{T} \Delta_t(x_t) \;>\; C_3\sqrt{2T\log\tfrac{1}{\delta}}\right) \;\leqslant\; \delta\,,$$

and by symmetry the same bound holds for the negative deviation. This proves the high-probability bound on $|\sum_{t=1}^{T}[f_t(x_t) - L_t(x_t)]|$. Setting $\delta = 1/T$ yields in particular

$$\Pr\!\left(\left|\sum_{t=1}^{T}[f_t(x_t) - L_t(x_t)]\right| \;>\; C_3\sqrt{2T\log(T)}\right) \;\leqslant\; \frac{1}{T}\,,$$

so with probability at least $1 - \frac{1}{T}$ we have the inequality (B3). □

Lemma C.5 (High–probability bound on empirical regret (Portfolio Selection))**.** *Suppose:*

- *The asset returns are bounded:* $m_{\min} \leqslant r_t(i) \leqslant m_{\max}$ *for all* $t \in [T]$, $i \in [n]$ *(Assumption (i)).*

- *The one-period loss (negative log-return) is*

$$f_t(x) \;=\; -\ln\bigl(x \cdot r_t\bigr)\,, \qquad x \in \Delta_n,$$

  where $x \cdot r_t = \sum_{i=1}^{n} x(i)\, r_t(i)$ is the portfolio's gross return at time t.

- *The portfolio sequence $\{\hat{x}_t^g\}_{t=1}^{T}$ is generated by online gradient descent (OGD), with step-size tuned to a guessed variation budget $v_g \geqslant v$.*

Define

$$C_3 := \ln\frac{m_{\max}}{m_{\min}}\,, \qquad C_0 := 8\frac{m_{\max}}{m_{\min}}\,, \qquad C_1 := 2\sqrt{2}\frac{m_{\max}}{m_{\min}}\,.$$

Then with probability at least $1 - \frac{3}{T}$,

$$\sum_{t=1}^{T}\Bigl[f_t(\hat{x}_t^g) - f_t(x_t^*)\Bigr] \;\leqslant\; T^{\frac{2+v_g}{3}}\left[\frac{2}{m_{\min}} \;+\; \Bigl(C_0\sqrt{\log T} \;+\; C_1\sqrt{\log n}\Bigr)\right] \;+\; 2\,C_3\sqrt{2T\log(T)}\,.$$

Proof. Write the total regret (random cumulative loss difference) as

$$\sum_{t=1}^{T}\bigl[f_t(\hat{x}_t^g) - f_t(x_t^*)\bigr] \;=\; \underbrace{\sum_{t=1}^{T}\bigl[L_t(\hat{x}_t^g) - L_t(x_t^*)\bigr]}_{\text{OGD term}} \;+\; \underbrace{\sum_{t=1}^{T}\bigl[f_t(\hat{x}_t^g) - L_t(\hat{x}_t^g)\bigr] \;+\; \sum_{t=1}^{T}\bigl[L_t(x_t^*) - f_t(x_t^*)\bigr]}_{\text{martingale noise}}\,,$$

where $L_t(x) = \mathbb{E}[f_t(x) \mid \mathcal{F}_{t-1}]$ is the conditional expected loss at time t.

By Lemma C.2, the OGD term (the regret with respect to the sequence of instantaneous minimizers $x_t^* := \arg\min_{x\in\Delta_n} L_t(x)$) is bounded by

$$T^{\frac{2+v_g}{3}}\left[\frac{2}{m_{\min}} \;+\; \Bigl(C_0\sqrt{\log T} \;+\; C_1\sqrt{\log n}\Bigr)\right],$$

with probability at least $1 - \frac{1}{T}$.

For the two "noise" sums, note that for any fixed portfolio $x \in \Delta_n$, the sequence $\Delta_t(x) = f_t(x) -$

$L_t(x)$ is a mean-zero martingale difference with $|\Delta_t(x)| \leqslant C_3$ almost surely (by the bounded-return assumption). By Azuma–Hoeffding's inequality (relying only on Assumption (i)), each of the sums $\sum_{t=1}^{T}[f_t(\hat{x}_t^g) - L_t(\hat{x}_t^g)]$ and $\sum_{t=1}^{T}[L_t(x_t^*) - f_t(x_t^*)]$ exceeds $C_3\sqrt{2T\log(2T)}$ with probability at most $1/T$. Hence, both martingale noise terms together are bounded by

$$2\,C_3\sqrt{2T\log(T)}$$

with probability at least $1 - \frac{2}{T}$.

Finally, by a union bound over the three high-probability events, we conclude that all parts (the OGD term and both noise terms) satisfy their bounds simultaneously with probability at least $1 - \frac{3}{T}$. Adding the OGD and noise contributions yields the stated high-probability regret bound. □

Bibliography

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2312–2320.

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2017) Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977* .

Agrawal S, Avadhanula V, Goyal V, Zeevi A (2019) MNL-bandit: A dynamic learning approach to assortment selection. *Operations Research* 67(5):1453–1485.

Agrawal S, Jia R (2019) Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 743–744.

An L, Li AA, Moseley B, Ravi R (2025) The nonstationary newsvendor with (and without) predictions. *Manufacturing & Service Operations Management* .

Auer P, Cesa-Bianchi N, Fischer P (2002) Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

Baardman L, Fata E, Pani A, Perakis G (2019) Learning optimal online advertising portfolios with periodic budgets. *Available at SSRN 3346642* .

Badanidiyuru A, Kleinberg R, Slivkins A (2013) Bandits with knapsacks. *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 207–216 (IEEE).

Ban GY (2019) Confidence intervals for data-driven inventory policies with demand censoring.

Forthcoming in Operations Research .

Ban GY, Keskin NB (2020) Personalized dynamic pricing with machine learning: High dimensional features and heterogeneous elasticity. *Forthcoming, Management Science* .

Bartlett P, Dani V, Hayes T, Kakade S, Rakhlin A, Tewari A (2008) High-probability regret bounds for bandit online linear optimization. *Proceedings of the 21st annual conference on learning theory-COLT 2008*, 335–342 (Omnipress).

Bastani H, Simchi-Levi D, Zhu R (2019) Meta dynamic pricing: Learning across experiments. *Available at SSRN 3334629* .

Besbes O, Gur Y, Zeevi A (2014) Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in neural information processing systems*, 199–207.

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

Besbes O, Gur Y, Zeevi A (2019) Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Stochatic Systems* 9(4):319–337.

Besbes O, Ma W, Mouchtaki O (2022) Beyond iid: data-driven decision-making in heterogeneous environments. *arXiv preprint arXiv:2206.09642* .

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Science* 59(6):1407–1424.

Besbes O, Zeevi A (2011) On the minimax complexity of pricing in a changing environment. *Operations research* 59(1):66–79.

Besbes O, Zeevi A (2015) On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* 61(4):723–739.

Beyer D, Sethi SP, Sridhar R (2002) Average-cost optimality of a base-stock policy for a multi-product inventory model with limited storage. *Decision & Control in Management Science*, 241–260 (Springer).

Blum A, Kalai A (1999) Universal portfolios with and without transaction costs. *Machine Learning* 35(3):193–205.

Bouneffouf D, Parthasarathy S, Samulowitz H, Wistub M (2019) Optimal exploitation of clustering and history information in multi-armed bandit. *arXiv preprint arXiv:1906.03979* .

Breiman L (1961) Optimal gambling systems for favorable games. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 65–78 (Berkeley, CA: University of California Press).

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Operations Research* 60(4):965–980.

Bu J, Simchi-Levi D, Wang L (2022) Offline pricing and demand learning with censored data. *Management Science* .

Cesa-Bianchi N, Lugosi G (2006) *Prediction, learning, and games* (Cambridge university press).

Chen B (2021) Data-driven inventory control with shifting demand. *Production and Operations Management* 30(5):1365–1385.

Chen B, Chao X, Ahn HS (2019a) Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research* 67(4):1035–1052.

Chen B, Chao X, Shi C (2021) Nonparametric learning algorithms for joint pricing and inventory control with lost sales and censored demand. *Mathematics of Operations Research* 46(2):726–756.

Chen B, Chao X, Wang Y (2020a) Data-based dynamic pricing and inventory control with censored demand and limited price changes. *Operations Research* 68(5):1445–1456.

Chen B, Shi C (2019) Tailored base-surge policies in dual-sourcing inventory systems with demand learning. *Available at SSRN 3456834* .

Chen B, Simchi-Levi D, Wang Y, Zhou Y (2022) Dynamic pricing and inventory control with fixed ordering cost and incomplete demand information. *Management Science* 68(8):5684–5703.

Chen W, Shi C, Duenyas I (2020b) Optimal learning algorithms for stochastic inventory systems

with random capacities. *Production and Operations Management* 29(7):1624–1649.

Chen X, Li M (2021) M-natural convexity and its applications in operations. *Operations Research* 69(5):1396–1408.

Chen X, Wang Y, Wang YX (2019b) Nonstationary stochastic optimization under l p, q-variation measures. *Operations Research* 67(6):1752–1765.

Chen Y, Wen Z, Xie Y (2019c) Dynamic pricing in an evolving and unknown marketplace. *Available at SSRN 3382957* .

Cheung WC, Simchi-Levi D (2017) Thompson sampling for online personalized assortment optimization problems with multinomial logit choice models. *Available at SSRN 3075658* .

Cheung WC, Simchi-Levi D (2019) Sampling-based approximation schemes for capacitated stochastic inventory control models. *Mathematics of Operations Research* 44(2):668–692.

Cheung WC, Simchi-Levi D, Zhu R (2019a) Hedging the drift: Learning to optimize under non-stationarity. *arXiv preprint arXiv:1903.01461* .

Cheung WC, Simchi-Levi D, Zhu R (2019b) Non-stationary reinforcement learning: The blessing of (more) optimism. *arXiv preprint arXiv:1906.02922* .

Choi J, Cao JJ, Romeijn HE, Geunes J, Bai SX (2005) A stochastic multi-item inventory model with unequal replenishment intervals and limited warehouse capacity. *IIE Transactions* 37(12):1129–1141.

Correa JR, Dütting P, Fischer F, Schewior K (2018) Prophet inequalities for independent random variables from an unknown distribution. *arXiv preprint arXiv:1811.06114* .

Cover TM (1984) An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory* 30(2):369–373.

Cover TM (1991) Universal portfolios. *Mathematical Finance* 1(1):1–29.

Dani V, Hayes TP, Kakade SM (2008) Stochastic linear optimization under bandit feedback. *Proceedings of the 21st Conference on Learning Theory*.

Das P, Johnson N, Banerjee A (2014) Online portfolio selection with group sparsity. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1185–1191 (AAAI Press).

Degenne R, Perchet V (2016) Anytime optimal algorithms in stochastic multi-armed bandits. *International Conference on Machine Learning*, 1587–1595.

den Boer A, Keskin NB (2017) Dynamic pricing with demand learning and reference effects. *Available at SSRN 3092745* .

den Boer AV (2014) Dynamic pricing with multiple products and partially specified demand distribution. *Mathematics of operations research* 39(3):863–888.

den Boer AV (2015) Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science* 20(1):1–18.

den Boer AV, Keskin NB (2020) Discontinuous demand functions: estimation and pricing. *Management Science* 66(10):4516–4534.

den Boer AV, Zwart B (2013) Simultaneously learning and optimizing using controlled variance pricing. *Management science* 60(3):770–783.

Ding J, Huh WT, Rong Y (2021) Feature-based nonparametric inventory control with censored demand. *Available at SSRN 3803777* .

Domb C (2000) *Phase transitions and critical phenomena*, volume 19 (Elsevier).

Federgruen A, Guetta D, Iyengar G, Liu X (2022) An asymptotically optimal heuristic for multi-item inventory models with joint inventory constraints. *Available at SSRN 4221106* .

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using thompson sampling. *Operations research* 66(6):1586–1602.

Filippi S, Cappe O, Garivier A, Szepesvári C (2010) Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 586–594.

Gill R, Levit B (2001) Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli* 1:59.

Goldenshluger A, Zeevi A, et al. (2009) Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability* 19(4):1603–1633.

Gong XY, Simchi-Levi D (2021) Bandits atop reinforcement learning: Tackling online inventory models with cyclic demands. *Available at SSRN 3637705* .

Gur Y, Momeni A (2018) Adaptive learning with unknown information flows. *Advances in Neural Information Processing Systems*, 7473–7482.

Gur Y, Momeni A (2019) Adaptive sequential experiments with unknown information flows. *arXiv preprint arXiv:1907.00107* .

Györfi L, Lugosi G, Udina F (2006) Nonparametric kernel-based sequential investment strategies. *Mathematical Finance* 16(2):337–357.

Halliday D, Resnick R, Walker J (2013) *Fundamentals of physics* (John Wiley & Sons).

Harrison JM, Keskin NB, Zeevi A (2012) Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* 58(3):570–586.

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* 27(2):83–85.

Hazan E (2022) *Introduction to online convex optimization.*

Hazan E, Agarwal A, Kale S (2007) Logarithmic regret algorithms for online convex optimization. *Machine Learning* 69(2-3):169–192.

Hazan E, Kale S (2015) An online portfolio selection algorithm with regret logarithmic in price variation. *Mathematical Finance* 25(2):288–310.

Hazan E, et al. (2016) Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2(3-4):157–325.

Helmbold DP, Schapire RE, Singer Y, Warmuth MK (1996) On-line portfolio selection using multiplicative updates. *Proceedings of the Thirteenth International Conference on Machine Learning*, 243–251 (Morgan Kaufmann).

Herbster M, Warmuth MK (1998) Tracking the best expert. *Machine Learning* 32(2):151–178.

Hsu CW, Kveton B, Meshi O, Martin M, Szepesvari C (2019) Empirical bayes regret minimization. *arXiv preprint arXiv:1904.02664* .

Huh WT, Janakiraman G, Muckstadt JA, Rusmevichientong P (2009) An adaptive algorithm for finding the optimal base-stock policy in lost sales inventory systems with censored demand. *Mathematics of Operations Research* 34(2):397–416.

Huh WT, Levi R, Rusmevichientong P, Orlin JB (2011) Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research* 59(4):929–941.

Huh WT, Rusmevichientong P (2009) A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research* 34(1):103–123.

Ignall E, Veinott Jr AF (1969) Optimality of myopic inventory policies for several substitute products. *Management Science* 15(5):284–304.

Jiang J, Li X, Zhang J (2020) Online stochastic optimization with wasserstein based non-stationarity. *arXiv preprint arXiv:2012.06961* .

Kalai A, Vempala S (2003) Efficient algorithms for universal portfolios. *Journal of Machine Learning Research* 3:423–440.

Kelly JL (1956) A new interpretation of information rate. *Bell System Technical Journal* 35(4):917–926.

Keskin N, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research* 62(5):1142–1167.

Keskin NB (2014) Optimal dynamic pricing with demand model uncertainty: A squared-coefficient-of-variation rule for learning and earning. *Available at SSRN 2487364* .

Keskin NB, Li M (2021) Selling quality-differentiated products in a markovian market with unknown transition probabilities. *Available at SSRN 3526568* .

Keskin NB, Li Y, Song JS (2022) Data-driven dynamic pricing and ordering with perishable inven-

tory in a changing environment. *Management Science* 68(3):1938–1958.

Keskin NB, Min X, Song JS (2021) The nonstationary newsvendor: Data-driven nonparametric learning. *Available at SSRN 3866171* .

Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research* 42(2):277–307.

Keskin NB, Zeevi A (2018) On incomplete learning and certainty-equivalence control. *Operations Research* 66(4):1136–1167.

Lattimore T, Szepesvári C (2018) Bandit algorithms. *preprint* .

Levi R, Perakis G, Uichanco J (2015) The data-driven newsvendor problem: new bounds and insights. *Operations Research* 63(6):1294–1306.

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research* 32(4):821–839.

Li L, Lu Y, Zhou D (2017) Provably optimal algorithms for generalized linear contextual bandits. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2071–2080 (JMLR. org).

Liang P (2016) Cs229t/stat231: Statistical learning theory (winter 2016).

Miao S, Chao X (2020) Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management* .

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

Nambiar M, Simchi-Levi D, Wang H (2019) Dynamic learning and pricing with model misspecification. *Management Science* .

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Operations research* 56(1):188–203.

Qiang S, Bayati M (2016) Dynamic pricing with demand covariates. *Available at SSRN 2765257* .

Qin H, Simchi-Levi D, Wang L (2019) Data-driven approximation schemes for joint pricing and inventory control models. *Available at SSRN 3354358* .

Rusmevichientong P, Tsitsiklis JN (2010) Linearly parameterized bandits. *Mathematics of Operations Research* 35(2):395–411.

Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Mathematics of Operations Research* 39(4):1221–1243.

Shi C, Chen W, Duenyas I (2016) Nonparametric data-driven algorithms for multiproduct inventory systems with censored demand. *Operations Research* 64(2):362–370.

Shivaswamy P, Joachims T (2012) Multi-armed bandit problems with history. *Artificial Intelligence and Statistics*, 1046–1054.

Simchi-Levi D, Xu Y (2019) Phase transitions in bandits with switching constraints. *arXiv preprint arXiv:1905.10825* .

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Tsybakov A (2009) *Introduction to Nonparametric Estimation* (Springer, New York).

Valiant LG (1984) A theory of the learnable. *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, 436–445 (ACM).

Vallender S (1974) Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18(4):784–786.

Veinott Jr AF (1965) Optimal policy for a multi-product, dynamic, nonstationary inventory problem. *Management science* 12(3):206–222.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Wang Z, Deng S, Ye Y (2014) Close the gaps: A learning-while-doing algorithm for single-product

revenue management problems. *Operations Research* 62(2):318–331.

Wei CY, Luo H (2021) Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *Conference on Learning Theory*, 4300–4354 (PMLR).

Ye L, Lin Y, Xie H, Lui J (2020) Combining offline causal inference and online bandit learning for data driven decisions. *arXiv preprint arXiv:2001.05699* .

Yuan H, Luo Q, Shi C (2021) Marrying stochastic gradient descent with bandits: Learning algorithms for inventory systems with fixed costs. *Management Science* 67(10):6089–6115.

Zhang H, Chao X, Shi C (2018) Perishable inventory systems: Convexity results for base-stock policies and learning algorithms under censored demand. *Operations Research* 66(5):1276–1286.

Zhang H, Chao X, Shi C (2020) Closing the gap: A learning algorithm for lost-sales inventory systems with lead times. *Management Science* 66(5):1962–1980.

Zhu F, Zheng Z (2021) Dynamic pricing in a non-stationary growing environment. *Available at SSRN 3637905* .

Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *Proceedings of the 20th International Conference on Machine Learning*, 928–936 (AAAI Press).