OPERATIONS RESEARCH
Vol. 63, No. 5, September–October 2015, pp. 1227–1244
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2015.1408

Non-Stationary Stochastic Optimization

Omar Besbes
Columbia University, New York, New York 10027, ob2105@gsb.columbia.edu

Yonatan Gur
Stanford University, Stanford, California 94305, ygur@stanford.edu

Assaf Zeevi
Columbia University, New York, New York 10027, assaf@gsb.columbia.edu

We consider a non-stationary variant of a sequential stochastic optimization problem, in which the underlying cost functions may change along the horizon. We propose a measure, termed *variation budget*, that controls the extent of said change, and study how restrictions on this budget impact achievable performance. We identify sharp conditions under which it is possible to achieve long-run average optimality and more refined performance measures such as rate optimality that fully characterize the complexity of such problems. In doing so, we also establish a strong connection between two rather disparate strands of literature: (1) adversarial online convex optimization and (2) the more traditional stochastic approximation paradigm (couched in a non-stationary setting). This connection is the key to deriving well-performing policies in the latter, by leveraging structure of optimal policies in the former. Finally, tight bounds on the minimax regret allow us to quantify the "price of non-stationarity," which mathematically captures the added complexity embedded in a temporally changing environment versus a stationary one.

Keywords: stochastic approximation; non-stationary; minimax regret, online convex optimization.
Subject classifications: decision analysis: sequential; statistics: nonparametric; probability: stochastic model applications; computers/computer science: artificial intelligence.
Area of review: Stochastic Models.
History: Published online in *Articles in Advance* Setember 14, 2015.

1. Introduction and Overview

Background and motivation. In the prototypical setting of sequential stochastic optimization, a decision maker selects at each epoch $t \in \{1, \ldots, T\}$, a point X_t that belongs (typically) to some convex compact action set $\mathscr{X} \subset \mathbb{R}^d$, and incurs a *cost* $f(X_t)$, where $f(\cdot)$ is an a priori unknown convex *cost function*. Subsequent to that, a *feedback* $\phi_t(X_t, f)$ is given to the decision maker; representative feedback structures include a noisy realization of the cost and/or gradient of the cost. When the cost function is assumed to be strongly convex, a typical objective is to minimize the mean squared error (MSE), $\mathbb{E}\|X_T - x^*\|^2$, where x^* denotes the minimizer of $f(\cdot)$ in $\mathscr{X}$. When $f(\cdot)$ is only assumed to be *weakly convex*, a more reasonable objective is to minimize $\mathbb{E}[f(X_T) - f(x^*)]$, the expected difference between the cost incurred at the terminal epoch T and the minimal achievable cost. (This objective reduces to the MSE criterion, up to a multiplicative constant, in the strongly convex case.) The study of such problems originates with the pioneering work of Robbins and Monro (1951), which focuses on stochastic estimation of a level crossing, and its counterpart studied by Kiefer and Wolfowitz (1952), which focuses on stochastic estimation of the point of maximum; these methods are collectively known as stochastic approximation (SA), and with some abuse of terminology, we will use this term to refer to the methods as well as the problem area. Since the publication of these seminal papers, SA has been widely studied and applied to diverse problems in a variety of fields, including Economics, Statistics, Operations Research, Engineering and Computer Science; cf. books by Benveniste et al. (1990) and Kushner and Yin (2003), and a survey by Lai (2003).

A fundamental assumption in SA, which has been adopted by almost all of the relevant literature (exceptions to be noted in what follows), is that the cost function does not change throughout the horizon over which we seek to (sequentially) optimize it. Departure from this stationarity assumption brings forward many fundamental questions. Primarily: (1) How do we model temporal changes in a manner that is "rich" enough to capture a broad set of scenarios while still being mathematically tractable? and (2) What is the performance that can be achieved in such settings in comparison to the stationary SA environment? Our paper is concerned with these questions.

The non-stationary SA problem. Consider the stationary SA formulation outlined above with the following modifications: rather than a single unknown cost function, there is now a *sequence* of convex functions $\{f_t\colon t = 1, \ldots, T\}$; like the stationary setting, in every epoch $t = 1, \ldots, T$, the decision maker selects a point $X_t \in \mathscr{X}$ (this will be referred to as "action" or "decision" in what follows), and then

observes a feedback, only now this signal, $\phi_t(X_t, f_t)$, will depend on the particular function within the sequence. In this paper, we consider two canonical feedback structures alluded to earlier, namely, (1) noisy access to the function value $f(X_t)$ and (2) noisy access to the gradient $\nabla f(X_t)$. Let $\{x_t^*\colon t = 1, \ldots, T\}$ denote the sequence of minimizers corresponding to the sequence of cost functions.

In this "moving target" formulation, a natural objective is to minimize the *cumulative* counterpart of the performance measure used in the stationary setting, for example, $\sum_{t=1}^{T} \mathbb{E}[f_t(X_t) - f_t(x_t^*)]$ in the general convex case. This is often referred to in the literature as the *regret*. It measures the quality of a policy, and the sequence of actions $\{X_1, \ldots, X_T\}$ it generates, by comparing its performance to a clairvoyant who knows the sequence of functions in advance, and hence selects the minimizer x_t^* at each step t; we refer to this benchmark as a *dynamic oracle* for reasons that will become clear soon.[1]

To constrain temporal changes in the sequence of functions, this paper introduces the concept of a *temporal uncertainty set* $\mathscr{V}$, which is driven by a *variation budget* V_T:

$$\mathscr{V} := \{\{f_1, \ldots, f_T\}\colon \mathrm{Var}(f_1, \ldots, f_T) \leqslant V_T\}.$$

The precise definition of the variation functional $\mathrm{Var}(\cdot)$ will be given in §2; roughly speaking, it measures the extent to which functions can change from one time step to the next, and adds this up over the horizon T. As will be seen in §2, the notion of variation we propose allows for a broad range of temporal changes in the sequence of functions and minimizers. Note that the variation budget is allowed to depend on the length of the horizon, and therefore measures the scales of variation relative to the latter.

For the purpose of outlining the flavor of our main analytical findings and key insights, let us further formalize the notion of *regret* of a policy π relative to the above-mentioned dynamic oracle:

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

In this set up, a policy π is chosen and then nature (playing the role of the adversary) selects the sequence of functions $f := \{f_t\}_{t=1,\ldots,T} \in \mathscr{V}$ that maximizes the regret; here we have made explicit the dependence of the regret and the expectation operator on the policy π, as well as its dependence on the feedback mechanism ϕ, which governs the observations. The first-order characteristic of a "good" policy is that it achieves *sublinear* regret, namely,

$$\frac{\mathscr{R}_\phi^\pi(\mathscr{V}, T)}{T} \to 0 \quad \text{as } T \to \infty.$$

A policy π with the above characteristic is called *long-run average optimal*, as the average cost it incurs (per period) asymptotically approaches the one incurred by the clairvoyant benchmark. Differentiating among such policies requires a more refined yardstick. Let $\mathscr{R}_\phi^*(\mathscr{V}, T)$ denote the *minimax regret*: the minimal regret that can be achieved over the space of admissible policies subject to feedback signal ϕ, *uniformly* over nature's choice of cost function sequences within the temporal uncertainty set $\mathscr{V}$. A policy is said to be *rate optimal* if it achieves the minimax regret up to a constant multiplicative factor; this implies that, in terms of growth rate of regret, the policy's performance is essentially best possible.

Overview of the main contributions. Our main results and key qualitative insights can be summarized as follows:

1. *Necessary and sufficient conditions for sublinear regret.* We first show that if the variation budget V_T is *linear* in T, then, as one may expect, sublinear regret *cannot* be achieved by any admissible policy. Conversely, we show that if V_T is *sublinear* in T, long-run average optimal policies exist. So, our notion of temporal uncertainty supports a sharp dichotomy in characterizing first-order optimality in the non-stationary SA problem.

2. *Complexity characterization.* We prove a sequence of results that characterizes the order of the minimax regret for the convex as well as the strongly convex settings. This is done by deriving lower bounds on the regret that hold for *any* admissible policy, and then proving that the order of these lower bounds can be achieved by suitable (rate optimal) policies. The essence of these results can be summarized by the following characterization of the minimax regret:

$$\mathscr{R}_\phi^*(\mathscr{V}, T) \asymp V_T^\alpha T^{1-\alpha},$$

where α is either $1/3$ or $1/2$, depending on the particulars of the problem (namely, whether the cost functions in $\mathscr{V}$ are convex/strongly convex, and whether the feedback ϕ is a noisy observation of the cost/gradient); see below for more specificity, and further details in §4 and §5.

3. *The "price of non-stationarity."* The minimax regret characterization allows, among other things, to contrast the stationary and non-stationary environments, where the "price" of the latter relative to the former is expressed in terms of the "radius" (variation budget) of the temporal uncertainty set. Table 1 summarizes our main findings. Note that even in the most "forgiving" non-stationary environment, where the variation budget V_T is a constant and independent of T, there is a marked degradation in performance between the stationary and non-stationary settings. (The table omits the general convex case with noisy cost observations; this will be explained later in the paper.)

4. *A metaprinciple for constructing optimal policies.* One of the key insights we wish to communicate in this paper pertains to the construction of well-performing policies, either long-run average or rate optimal. The main idea is a result of bridging two relatively disconnected streams of literature that deal with dynamic optimization under uncertainty from very different perspectives: the so-called

Table 1. The price of non-stationarity.

Setting		Order of regret	
Class of functions	Feedback	Stationary	Non-stationary
Convex	Noisy gradient	$\sqrt{T}$	$V_T^{1/3}T^{2/3}$
Strongly convex	Noisy gradient	$\log T$	$\sqrt{V_T T}$
Strongly convex	Noisy function	$\sqrt{T}$	$V_T^{1/3}T^{2/3}$

Note. The rate of growth of the minimax regret in the stationary and non-stationary settings under different assumptions on the cost functions and feedback signal.

adversarial and the *stochastic* frameworks. The former, which in our context is often referred to as online convex optimization (OCO), allows nature to select the worst-possible function at *each* point in time depending on the actions of the decision maker, and with little constraints on nature's choices. This constitutes a more pessimistic environment compared with the traditional stochastic setting, where the function is picked a priori at $t = 0$ and held fixed thereafter, or the setting we propose here, where the *sequence* of functions is chosen by nature subject to a variation constraint. Because of the freedom awarded to nature in OCO settings, a policy's performance is typically measured relative to a rather coarse benchmark, known as the *single best action in hindsight*; the best static action that would have been picked ex post, namely, after having observed all of nature's choices of functions. Though typically a policy that is designed to compete with the single best action benchmark in an adversarial OCO setting does not admit performance guarantees in our non-stationary stochastic problem setting (relative to a dynamic oracle), we establish an important connection between performance in the former and the latter environments, given roughly by the following "metaprinciple":

> If a policy has "good" performance with respect to the single best action in the adversarial framework, it can be adapted in a manner that guarantees "good" performance in the stochastic non-stationary environment subject to the variation budget constraint.

In particular, according to this principle, a policy with sublinear regret in an OCO setting can be adapted to achieve sublinear regret in the non-stationary stochastic setting, and in a similar manner, we can port over the property of rate optimality. It is important to emphasize that although policies that admit these properties have, by and large, been identified in the OCO literature,[2] to the best of our knowledge there are no counterparts to date in a non-stationary stochastic setting, including the one considered in this paper. (It is worthwhile noting that the construction of said policies is mostly done with the intent of providing a relatively simple and unified way to highlight key trade-offs at play.)

Relation to literature. The use of the cumulative performance criterion and regret, even though mostly absent from the traditional SA stream of literature, has been adopted in several occasions (when the cost function does not change over time). Examples include the work of Cope (2009), which is couched in an environment where the feedback structure is noisy observations of the cost and the target function is strongly convex. That paper shows that the estimation scheme of Kiefer and Wolfowitz (1952) is rate optimal and the minimax regret in such a setting is of order $\sqrt{T}$. Considering a convex (and differentiable) cost function, Agarwal et al. (2013) showed that the minimax regret is of the same order, building on estimation methods presented in Nemirovski and Yudin (1983). In the context of gradient-type feedback and strongly convex cost, it is straightforward to verify that the scheme of Robbins and Monro (1951) is rate optimal, and the minimax regret is of order $\log T$.

Although temporal changes in the cost function are typically not discussed within the traditional stationary SA literature (see Chapter 3 in Kushner and Yin 2003, and Chapter 4 in Benveniste et al. 1990 for exceptions), the literature on OCO, which has mostly evolved in the machine learning community starting with Zinkevich (2003), allows the cost function to be selected at any point in time by an *adversary*. Discussed above, the performance of a policy in this setting is compared against a relatively weak benchmark, namely, the single best action in hindsight; or, a *static* oracle. These ideas have their origin in game theory with the work of Blackwell (1956) and Hannan (1957), and have since seen significant development in several sequential decision-making settings; cf. Cesa-Bianchi and Lugosi (2006) for an overview. The OCO literature largely focuses on a class of either convex or strongly convex cost functions, and sublinearity and rate optimality of policies have been studied for a variety of feedback structures. The original work of Zinkevich (2003) considered the class of convex functions, and focused on a feedback structure in which the function f_t is *entirely revealed* after the selection of X_t, providing an *online gradient descent* (OGD) algorithm with regret of order $\sqrt{T}$; see also Flaxman et al. (2005). Hazan et al. (2007) achieve regret of order $\log T$ for a class of strongly convex cost functions, when the gradient of f_t, evaluated at X_t is observed. Additional algorithms were shown to be rate optimal under further assumptions on the function class (see, e.g., Kalai and Vempala 2003, Hazan et al. 2007), or other feedback structures such as multipoint access (Agarwal et al. 2010). A closer paper, at least in spirit, is that of Hazan and Kale (2010). It derives upper bounds on the regret with respect to the static single best action, in terms of a measure of dispersion of the cost functions chosen by nature, akin to variance. The cost functions in their setting are restricted to be linear and are revealed to the decision maker after each action.

It is important to draw attention to a significant distinction between the framework we pursue in this paper and the adversarial setting, concerning the quality of the benchmark that is used in each of the two formulations. Recall, in the adversarial setting, the performance of a policy is

compared to the ex post best static feasible solution, while in our setting the benchmark is given by a dynamic oracle (where "dynamic" refers to the sequence of minima $\{f_t(x_t^*)\}$ and minimizers $\{x_t^*\}$ that is changing throughout the time horizon). It is fairly straightforward that the gap between the performance of the static oracle that uses the single best action, and that of the dynamic oracle can be significant, in particular, these quantities may differ by order T; for an illustrative example, see §2, Example 1. Therefore, even if it is possible to show that a policy has a "small" regret relative to the best static action, there is no guarantee on how well such a policy will perform when measured against the best dynamic sequence of decisions. A second potential limitation of the adversarial framework lies in its rather pessimistic assumption of the world in which policies are to operate in, to wit, the environment can change at any point in time in the worst-possible way as a *reaction* to the policy's chosen actions. In most application domains, one can argue, the operating environment is not nearly as harsh.

Key to establishing the connection between the adversarial setting and the non-stationary stochastic framework proposed herein is the notion of a variation budget, and the corresponding temporal uncertainty set, that curtails nature's actions in our formulation. These ideas echo, at least philosophically, concepts that have permeated the robust optimization literature, where uncertainty sets are fundamental predicates; see, e.g., Ben-Tal and Nemirovski (1998) and a survey by Bertsimas et al. (2011).

Another line of research considers sequential stochastic optimization using Kalman filters (Kalman 1960). There, the typical objective is to minimize the MSE when estimating a state, under zero mean Gaussian noise. In the non-stationary variant of this problem, the state may change; in such cases the aforementioned change is typically well structured by some parameterized dynamics. An overview of this research domain is given in Haykin (2001), where Chapters 3, 4, and 6 include a survey of methods and applications for state-dynamic models. The focus, formulation, and analysis in this paper are different from the ones adopted in the literature on Kalman filters in the following key aspects. First, a main interest of the current study is in characterizing the extent of non-stationarity under which one may achieve sublinear regret with respect to the dynamic oracle benchmark. In particular, we show that whenever the variation is a sublinear function of the time horizon T, one may achieve sublinear regret relative to the dynamic oracle, but when variation is at least linear in T sublinear regret is not achievable. While non-stationary instances that are considered in the literature on Kalman filters typically fall under the latter case (linear variation), the focus of the current paper is on characterizing the minimax regret in the former. Second, the formulation in this paper is more general than the one adopted in the literature on Kalman filters; most importantly, we consider very general classes of cost functions, and temporal changes that are constrained only by a budget of variation, and are otherwise arbitrary (and, in particular, nonparametric).

A rich line of work in the literature considers concrete sequential decision problems embedded in an SA setting (namely, noisy observations of the cost or the gradient, where the underlying cost function is unknown). Various studies consider dynamic pricing problems where the demand function is unknown, and noisy cost observations are obtained at each step; see recent works by Broder and Rusmevichientong (2012), den Boer and Zwart (2014), and Keskin and Zeevi (2014), as well as the review by Araman and Caldentey (2011) for parametric and nonparametric approaches. Other studies consider a problem of inventory control with censored demand, where noisy observations of the gradient can be obtained in each step; see, e.g., Huh and Rusmevichientong (2009) and Besbes and Muharremoglu (2013). Other applications arise in queueing networks, online advertisements, wireless communications, and manufacturing systems, among other areas; see Kushner and Yin (2003) for an overview.

Most of the studies mentioned above focus on a setting in which the underlying environment (though unknown) is stationary. Whereas, several papers have considered settings where changes in the environment may occur, these papers typically assume a very specific structure on said changes (for example, considering dynamic pricing in the absence of capacity constrains, Keller and Rady 1999 study a setting where demand is switching between two known demand functions according to a known Markov process; Besbes and Zeevi 2011 consider a similar problem in a setting where the timing of a single (known) change in the demand function is unknown). The current paper suggests a general framework to study stochastic optimization problems while allowing a broad array of changes in the underlying environment. In that sense, special cases of the formulation given in the current paper allow an extension of studies such as the ones mentioned above for a variety of non-stationary settings.

Structure of the paper. Section 2 contains the problem formulation. In §3, we establish a principle that connects achievable regret of policies in the adversarial and non-stationary stochastic settings, in particular, proving that the property of sublinearity of the regret can be carried over from the former to the latter. Sections 4 and 5 include the main rate optimality results for the convex and strongly convex settings, respectively. Section 6 presents concluding remarks. Proofs can be found in Appendix A in the main text, and in Appendices B, C, and D that appear in an electronic companion (available as supplemental material at http://dx.doi.org/10.1287/opre.2015.1408).

2. Problem Formulation

Having already laid out in the previous section the key building blocks and ideas behind our problem formulation, the purpose of the present section is to fill in any gaps and make that exposition more precise where needed; some repetition is expected but is kept to a minimum.

Preliminaries and admissible polices. Let $\mathscr{X}$ be a convex, compact, nonempty *action set*, and $\mathscr{T} = \{1, \ldots, T\}$ be the sequence of decision epochs. Let $\mathscr{F}$ be a class of sequences $f := \{f_t\colon t = 1, \ldots, T\}$ of convex cost functions from $\mathscr{X}$ into $\mathbb{R}$ that submit to the following two conditions:

1. There is a finite number G such that for any action $x \in \mathscr{X}$ and any epoch $t \in \mathscr{T}$,

$$|f_t(x)| \leqslant G, \quad \|\nabla f_t(x)\| \leqslant G. \tag{1}$$

2. There is some $\nu > 0$ such that

$$\{x \in \mathbb{R}^d\colon \|x - x_t^*\| \leqslant \nu\} \subset \mathscr{X} \quad \text{for all } t \in \mathscr{T}, \tag{2}$$

where $x_t^* := x_t^*(f_t) \in \arg\min_{x\in\mathscr{X}} f_t(x)$.

Here, $\nabla f_t(x)$ denotes the gradient of f_t evaluated at point x, and $\|\cdot\|$ the Euclidean norm. In every epoch $t \in \mathscr{T}$, a decision maker selects a point $X_t \in \mathscr{X}$ and then observes a feedback $\phi_t := \phi_t(X_t, f_t)$, which takes one of two forms:

1. noisy access to the cost, denoted by $\phi^{(0)}$, such that $\mathbb{E}[\phi_t^{(0)}(X_t, f_t) \mid X_t = x] = f_t(x)$;

2. noisy access to the gradient, denoted by $\phi^{(1)}$, such that $\mathbb{E}[\phi_t^{(1)}(X_t, f_t) \mid X_t = x] = \nabla f_t(x)$.

For all $x \in \mathscr{X}$ and f_t, $t \in \{1, \ldots, T\}$, we will use $\phi_t(x, f_t)$ to denote the feedback observed at epoch t, conditioned on $X_t = x$, and ϕ will be used in reference to a generic feedback structure. The feedback signal is assumed to possess a second moment uniformly bounded over $\mathscr{F}$ and $\mathscr{X}$.

Example 1 (Independent Noise). A conventional cost feedback structure is $\phi_t^{(0)}(x, f_t) = f_t(x) + \varepsilon_t$, where ε_t are, say, independent Gaussian random variables with zero mean and variance uniformly bounded by σ^2. A gradient counterpart is $\phi_t^{(1)}(x, f_t) = \nabla f_t(x) + \varepsilon_t$, where ε_t are independent Gaussian random vectors with zero mean and covariance matrices with entries uniformly bounded by σ^2. □

We next describe the class of admissible policies. Let U be a random variable defined over a probability space $(\mathbb{U}, \mathscr{U}, \mathbf{P}_u)$. Let $\pi_1\colon \mathbb{U} \to \mathbb{R}^d$ and $\pi_t\colon \mathbb{R}^{(t-1)k} \times \mathbb{U} \to \mathbb{R}^d$ for $t = 2, 3, \ldots$ be measurable functions, such that X_t, the action at time t, is given by

$$X_t = \begin{cases} \pi_1(U) & t = 1, \\ \pi_t(\phi_{t-1}(X_{t-1}, f_{t-1}), \ldots, \phi_1(X_1, f_1), U) & t = 2, 3, \ldots, \end{cases}$$

where $k = 1$ if $\phi = \phi^{(0)}$, namely, the feedback is noisy observations of the cost, and $k = d$ if $\phi = \phi^{(1)}$, namely, the feedback is noisy observations of the gradient. The mappings $\{\pi_t\colon t = 1, \ldots, T\}$ together with the distribution $\mathbf{P}_u$ define the class of admissible policies with respect to feedback ϕ. We denote this class by $\mathscr{P}_\phi$. We further denote by $\{\mathscr{H}_t, t = 1, \ldots, T\}$ the *filtration* associated with a policy $\pi \in \mathscr{P}_\phi$ such that $\mathscr{H}_1 = \sigma(U)$ and $\mathscr{H}_t = \sigma(\{\phi_j(X_j, f_j)\}_{j=1}^{t-1}, U)$ for all $t \in \{2, 3, \ldots\}$. Note that policies in $\mathscr{P}_\phi$ are nonanticipating, i.e., depend only on the past history of actions and observations, and allow for randomized strategies via their dependence on U.

Temporal uncertainty and regret. As indicated already in the previous section, the class of sequences $\mathscr{F}$ is too "rich," insofar as the latitude it affords nature. With that in mind, we further restrict the set of admissible cost function sequences, in particular, the manner in which its elements can change from one period to the other. Define the following notion of *variation* based on the sup norm:

$$\mathrm{Var}(f_1, \ldots, f_T) := \sum_{t=2}^{T} \|f_t - f_{t-1}\|, \tag{3}$$

where for any bounded functions g and h from $\mathscr{X}$ into $\mathbb{R}$ we denote $\|g - h\| := \sup_{x\in\mathscr{X}} |g(x) - h(x)|$. Let $\{V_t\colon t = 1, 2, \ldots\}$ be a nondecreasing sequence of real numbers such that $V_t \leqslant t$ for all t, $V_1 = 0$, and for normalization purposes set $V_2 \geqslant 1$. We refer to V_T as the *variation budget* over $\mathscr{T}$. Using this as a primitive, define the corresponding *temporal uncertainty set*, as the set of admissible cost function sequences that are subject to the variation budget V_T over the set of decision epochs $\{1, \ldots, T\}$:

$$\mathscr{V} = \left\{\{f_1, \ldots, f_T\} \subset \mathscr{F}\colon \sum_{t=2}^{T} \|f_t - f_{t-1}\| \leqslant V_T\right\}. \tag{4}$$

Even though the variation budget places some restrictions on the possible evolution of the cost functions, it still allows for many different temporal patterns: continuous change; discrete shocks; and a nonconstant rate of change. Two possible variations instances are illustrated in Figure 1; other variation patterns are considered in the numerical analysis described in Appendix D.

As described in §1, the performance metric we adopt pits a policy π against a dynamic oracle:

$$\mathscr{R}_\phi^\pi(\mathscr{V}, T) = \sup_{f\in\mathscr{V}} \left\{\mathbb{E}^\pi\left[\sum_{t=1}^{T} f_t(X_t)\right] - \sum_{t=1}^{T} f_t(x_t^*)\right\}, \tag{5}$$

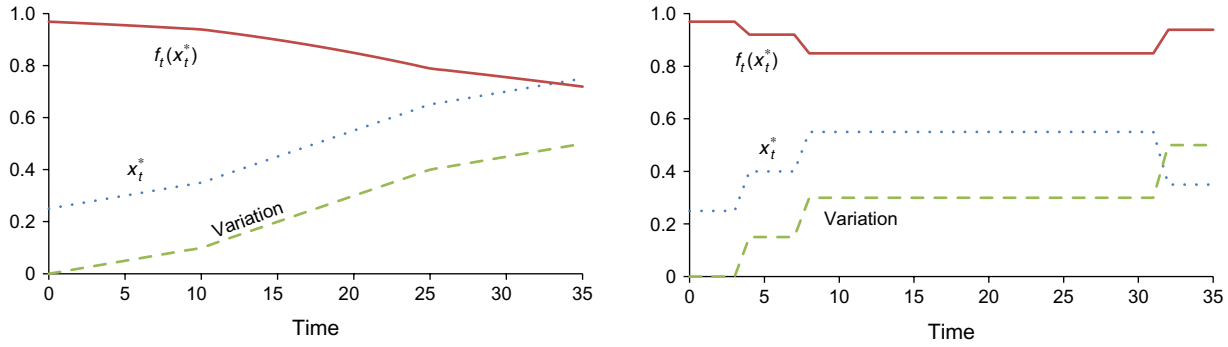
where the expectation $\mathbb{E}^\pi[\cdot]$ is taken with respect to any randomness in the feedback, as well as in the policy's actions. Assuming a setup in which first a policy π is chosen and then nature selects $f \in \mathscr{V}$ to maximize the regret, our formulation allows nature to select the worst-possible sequence of cost functions for that policy, subject to the variation budget.[3] Recall that a policy π is said to have *sublinear* regret if $\mathscr{R}_\phi^\pi(\mathscr{V}, T) = o(T)$, where for sequences $\{a_t\}$ and $\{b_t\}$ we write $a_t = o(b_t)$ if $a_t/b_t \to 0$ as $t \to \infty$. Recall also that the *minimax regret*, being the minimal worst-case regret that can be guaranteed by an admissible policy $\pi \in \mathscr{P}_\phi$, is given by

$$\mathscr{R}_\phi^*(\mathscr{V}, T) = \inf_{\pi\in\mathscr{P}_\phi} \mathscr{R}_\phi^\pi(\mathscr{V}, T).$$

We refer to a policy π as *rate optimal* if it achieves the lowest possible growth rate of regret: there exists a constant $\bar{C} \geqslant 1$, independent of V_T and T such that for any $T \geqslant 1$,

$$\mathscr{R}_\phi^\pi(\mathscr{V}, T) \leqslant \bar{C} \cdot \mathscr{R}_\phi^*(\mathscr{V}, T).$$

Figure 1. (Color online) Variation instances within a temporal uncertainty set.



Notes. Assume $\mathscr{X} = [0, 1]$ and consider a sequence of quadratic cost functions of the form $f_t(x) = \frac{1}{2}x^2 - b_t x + 1$. The change in the minimizer $x_t^* = b_t$, the optimal performance $f_t(x_t^*) = 1 - \frac{1}{2}b_t^2$, and the variation measured by (3), is illustrated for cases characterized by continuous changes (left), and "jump" changes (right) in b_t. In both instances, the variation budget is $V_T = 1/2$.

Contrasting with the adversarial OCO paradigm. An OCO problem consists of a convex set $\mathscr{X} \subset \mathbb{R}^d$ and an a priori unknown sequence $f = \{f_1, \ldots, f_T\} \in \mathscr{F}$ of convex cost functions. At any epoch t, the decision maker selects a point $X_t \in \mathscr{X}$, and observes some feedback ϕ_t. The efficacy of a policy over a given time horizon T is typically measured relative to a benchmark, which is defined by the *single best action in hindsight*: the best *static* action fixed throughout the horizon, and chosen with benefit of having observed the sequence of cost functions. We use the notions of admissible, long-run average optimal, and rate optimal policies in the adversarial OCO context as defined in the stochastic non-stationary context laid out before. Under the single best action benchmark, the objective is to minimize the regret incurred by an admissible online optimization algorithm $\mathscr{A}$:

$$\mathscr{G}_\phi^{\mathscr{A}}(\mathscr{F}, T) = \sup_{f\in\mathscr{F}}\left\{\mathbb{E}^\pi\left[\sum_{t=1}^T f_t(X_t)\right] - \min_{x\in\mathscr{X}}\left\{\sum_{t=1}^T f_t(x)\right\}\right\}, \quad (6)$$

where the expectation is taken with respect to possible randomness in the feedback and in the actions of the policy. (We use the term "algorithm" to distinguish this from what we have defined as a "policy," and this distinction will be important in what follows.)[4] Interchanging the sum and $\min\{\cdot\}$ operators in the right-hand side of (6), we obtain the definition of regret in the non-stationary stochastic setting, as in (5). As the next example shows, the dynamic oracle used as benchmark in the latter can be a significantly harder target than the single best action defining the static oracle in (6).

Example 2 (Contrasting the Static and Dynamic Oracles). Assume an action set $\mathscr{X} = [-1, 2]$, and variation budget $V_T = 1$. Set

$$f_t(x) = \begin{cases} x^2 & \text{if } t \leqslant T/2 \\ x^2 - 2x & \text{otherwise} \end{cases}$$

for any $x \in \mathscr{X}$. Then, the single best action is suboptimal at each decision epoch, and

$$\min_{x\in\mathscr{X}}\left\{\sum_{t=1}^T f_t(x)\right\} - \sum_{t=1}^T \min_{x\in\mathscr{X}}\{f_t(x)\} = \frac{T}{4}. \quad \square$$

Hence, algorithms that achieve performance that is "close" to the static oracle in the adversarial OCO setting may perform quite poorly in the non-stationary stochastic setting (in particular, they may, as the example above suggests, incur linear regret in that setting). Nonetheless, as the next section unravels, we will see that algorithms designed in the adversarial OCO context can, in fact, be adapted to perform well in the non-stationary stochastic setting laid out in this paper.

3. A General Principle for Designing Efficient Policies

In this section, we will develop policies that operate well in non-stationary environments with given budget of variation V_T. Before exploring the question of what performance one may aspire to in the non-stationary variation constrained world, we first formalize what cannot be achieved.

Proposition 1 (Linear Variation Budget Implies Linear Regret). *Assume a feedback structure $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$. If there exists a positive constant C_1 such that $V_T \geqslant C_1 T$ for any $T \geqslant 1$, then there exists a positive constant C_2 such that for any admissible policy $\pi \in \mathscr{P}_\phi$, and for any $T \geqslant 1$,*

$$\mathscr{R}_\phi^\pi(\mathscr{V}, T) \geqslant C_2 T.$$

The proposition states that whenever the variation budget is at least of order T, *any* policy that is admissible (with respect to the feedback) must incur a regret of order T, so under such circumstances, it is not possible to have long-run average optimality relative to the dynamic oracle benchmark. With that in mind, hereon, we will focus on the case in which the variation budget is sublinear in T. We will show that, in this case, sublinear regret is achievable, and study the behavior of the minimax regret as a function of V_T and T, when T is large. We note that when V_T is sublinear in T, the set $\mathscr{V}$ defined in (4) is still very rich and includes many general patterns, such as sequences of functions $\{f_t\}$ that might change significantly from one period

to the next but only do so rarely (a special case of which is a single change point), or sequences in which functions change often (even infinitely many times) but do so only locally. For example, consider the setting described in Figure 1 with a sequence of coefficients $\{b_t\}$ that does not converge, yet satisfies $|b_{t-1} - b_t| = t^{-1/2}$. Then, the variation budget V_T is of order $\sqrt{T}$.

A class of candidate policies. We introduce a class of policies that leverages existing algorithms designed for fully adversarial environments. We denote by $\mathscr{A}$ an online optimization algorithm that, given a feedback structure ϕ achieves a regret $\mathscr{G}^{\mathscr{A}}_{\phi}(\mathscr{F}, T)$ (see (6)) with respect to the *static* benchmark of the single best action. Consider the following generic "restarting" procedure, which takes as input $\mathscr{A}$ and a batch size Δ_T, with $1 \leqslant \Delta_T \leqslant T$, and consists of restarting $\mathscr{A}$ every Δ_T periods. To formalize this idea, we first refine our definition of history-adapted policies and the actions they generate. Given a feedback ϕ and epochs $t' \geqslant 1$, and $t > t'$, we define the history from t' to t by $\mathscr{H}_{t',t} = \sigma(\{\phi_j(X_j, f_j)\}_{j=t'}^{t-1}, U)$. Then, for each restarting epoch $\tau \geqslant 1$, we have $X_t = \mathscr{A}_{t-\tau}(\mathscr{H}_{\tau+1,t})$ for each $\tau + 1 < t \leqslant \min\{\tau + \Delta_T, T\}$, and $X_{\tau+1} = \mathscr{A}_1(\mathscr{H}_{\tau-\Delta_T+1,\tau})$. Indeed, X_t is $\mathscr{H}_{\tau+1,t}$-measurable for each $\tau + 1 < t \leqslant \min\{\tau + \Delta_T + 1, T\}$, and $X_{\tau+1}$ is $\mathscr{H}_{\tau-\Delta_T+1,\tau}$-measurable. The sequence of measurable mappings $\mathscr{A}_t$, $t = 1, 2, \ldots$ is prescribed by the algorithm $\mathscr{A}$, where we allow the initial action $\mathscr{A}_1$ to be based on information from the previous batch (if such exists). The following procedure restarts $\mathscr{A}$ every Δ_T epochs. In what follows, let $\lceil \cdot \rceil$ denote the ceiling function (rounding its argument to the nearest larger integer).

Restarting procedure. Inputs: an algorithm $\mathscr{A}$, and a batch size Δ_T.

1. Set $j = 1$.
2. Repeat while $j \leqslant \lceil T/\Delta_T \rceil$:
   (a) Set $\tau = (j-1)\Delta_T$.
   (b) If $\tau = 0$, set $X_{\tau+1} = \mathscr{A}_1(U)$, otherwise set $X_{\tau+1} = \mathscr{A}_1(\mathscr{H}_{\tau-\Delta_T+1,\tau})$.
   For any $t = \tau + 2, \ldots, \min\{T, \tau + \Delta_T\}$, select $X_t = \mathscr{A}_{t-\tau}(\mathscr{H}_{\tau+1,t})$.
   (c) Set $j = j + 1$.

Clearly, $\pi \in \mathscr{P}_{\phi}$. Next, we analyze the performance of policies defined via the restarting procedure with suitable subroutine $\mathscr{A}$.

First-order performance. The next result establishes a close connection between $\mathscr{G}^{\mathscr{A}}_{\phi}(\mathscr{F}, T)$, the performance that is achievable in the adversarial environment by $\mathscr{A}$, and $\mathscr{R}^{\pi}_{\phi}(\mathscr{V}, T)$, the performance in the non-stationary stochastic environment under temporal uncertainty set $\mathscr{V}$ of the restarting procedure that uses $\mathscr{A}$ as input.

THEOREM 1 (LONG-RUN AVERAGE OPTIMALITY). *Set a feedback structure $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$. Let $\mathscr{A}$ be an OCO algorithm with $\mathscr{G}^{\mathscr{A}}_{\phi}(\mathscr{F}, T) = o(T)$. Let π be the policy defined by the restarting procedure that uses $\mathscr{A}$ as a subroutine with batch size Δ_T. If $V_T = o(T)$, then for any Δ_T such that $\Delta_T = o(T/V_T)$ and $\Delta_T \to \infty$ as $T \to \infty$,*

$$\mathscr{R}^{\pi}_{\phi}(\mathscr{V}, T) = o(T).$$

In other words, the theorem establishes the following metaprinciple: whenever the variation budget is a sublinear function of the horizon length T, it is possible to construct a long-run average optimal policy in the stochastic non-stationary SA environment by a suitable adaptation of an algorithm that achieves sublinear regret in the adversarial OCO environment. For a given structure of a function class and feedback signal, Theorem 1 is meaningless unless there exists an algorithm with sublinear regret with respect to the single best action in the adversarial setting, under such structure. To that end, for the structures $(\mathscr{F}, \phi^{(0)})$ and $(\mathscr{F}, \phi^{(1)})$, an OGD policy was shown to achieve sublinear regret in Flaxman et al. (2005). We will see in the next sections that, surprisingly, the simple restarting mechanism introduced above allows to carry over not only first-order optimality, but also rate optimality from the OCO paradigm to the non-stationary SA setting.

Key ideas behind the proof. Theorem 1 is driven directly by the next proposition that connects the performance of the restarting procedure with respect to the dynamic benchmark in the stochastic non-stationary environment, and the performance of the input subroutine algorithm $\mathscr{A}$ with respect to the single best action in the adversarial setting.

PROPOSITION 2 (CONNECTING PERFORMANCE IN OCO AND NON-STATIONARY SA). *Set $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$. Let π be the policy defined by the restarting procedure that uses $\mathscr{A}$ as a subroutine with batch size Δ_T. Then, for any $T \geqslant 1$,*

$$\mathscr{R}^{\pi}_{\phi}(\mathscr{V}, T) \leqslant \left\lceil \frac{T}{\Delta_T} \right\rceil \cdot \mathscr{G}^{\mathscr{A}}_{\phi}(\mathscr{F}, \Delta_T) + 2\Delta_T V_T. \tag{7}$$

We next describe the high-level arguments. The main idea of the proof lies in analyzing the difference between the dynamic oracle and the static oracle benchmarks used, respectively, in the OCO and the non-stationary SA contexts. We define a partition of the decision horizon into batches $\mathscr{T}_1, \ldots, \mathscr{T}_m$ of size Δ_T each (except, possibly the last batch):

$$\mathscr{T}_j = \{t\colon (j-1)\Delta_T + 1 \leqslant t \leqslant \min\{j\Delta_T, T\}\} \quad \text{for all } j = 1, \ldots, m, \tag{8}$$

where $m = \lceil T/\Delta_T \rceil$ is the number of batches. Then, one may write

$$\mathscr{R}^{\pi}_{\phi}(\mathscr{V}, T) = \sup_{f \in \mathscr{V}} \Bigg\{ \underbrace{\sum_{j=1}^{m} \left(\mathbb{E}^{\pi}\left[\sum_{t \in \mathscr{T}_j} f_t(X_t) \right] - \min_{x \in \mathscr{X}} \left\{ \sum_{t \in \mathscr{T}_j} f_t(x) \right\} \right)}_{J_{1,j}} + \underbrace{\sum_{j=1}^{m} \left(\min_{x \in \mathscr{X}} \left\{ \sum_{t \in \mathscr{T}_j} f_t(x) \right\} - \sum_{t \in \mathscr{T}_j} f_t(x_t^*) \right)}_{J_{2,j}} \Bigg\}.$$

The regret with respect to the dynamic benchmark is represented as two sums. The first, $\sum_{j=1}^{m} J_{1,j}$, sums the regret terms with respect to the single best action within each batch $\mathcal{T}_j$, which are each bounded by $\mathcal{G}_{\phi}^{\mathcal{A}}(\mathcal{F}, \Delta_T)$. Noting that there are $\lceil T/\Delta_T \rceil$ batches, this gives rise to the first term on the right-hand side of (7). The second sum, $\sum_{j=1}^{m} J_{2,j}$, is the sum of differences between the performances of the single best action benchmark and the dynamic benchmark within each batch. The latter is driven by the rate of functional change in the batch. Although locally this gap can be large, we show that, given the variation budget, the second sum is at most of order $\Delta_T V_T$. This leads to the result of the proposition. Intuitively, Proposition 2 highlights the following trade-off. When Δ_T is large, the performance of a "good" subroutine policy approaches the one of the static oracle; when Δ_T is small, the sequence of static oracles approaches the dynamic oracle. Theorem 1 follows by balancing this trade-off. □

Remark (Alternative Forms of Feedback). The principle laid out in Theorem 1 can also be derived for other forms of feedback using Proposition 2. For example, the proof of Theorem 1 holds for settings with richer feedback structures, such as noiseless access to the full cost function (Zinkevich 2003) or a multipoint access (Agarwal et al. 2010).

4. Rate Optimality: The General Convex Case

A natural question arising from the analysis of §3 is what type of performance one may achieve in non-stationary environments and how does such performance depend on the variation one may face. We first focus on the feedback structure $\phi^{(1)}$, for which rate optimal polices are known in the OCO setting (as these will serve as inputs for the restarting procedure). To answer such a question, we first develop a lower bound for a subclass of problems and then establish that such a lower bound is achievable.

A lower bound on achievable performance. We will establish a fundamental bound on the performance of any admissible policy under the following technical assumption on the structure of the gradient feedback signal (a cost feedback counterpart will be provided in the next section).

Assumption 1 (Gradient Feedback Structure). 1. *$\phi_t^{(1)}(x, f_t) = \nabla f_t(x) + \varepsilon_t$ for any $f \in \mathcal{F}$, $x \in \mathcal{X}$, and $t \in \mathcal{T}$, where ε_t, $t \geqslant 1$ are iid random vectors with zero mean and covariance matrix with bounded entries.*

2. *Let $G(\cdot)$ be the cumulative distribution function of ε_t. There exists a constant $\tilde{C}$ such that for any $a \in \mathbb{R}^d$, $\int \log(dG(y)/dG(y+a))\, dG(y) \leqslant \tilde{C}\|a\|^2$.*

For the sake of concreteness, we impose an additive noise feedback structure, given in the first part of the assumption. This simplifies notation and streamlines proofs, but otherwise is not essential. The key properties that are needed are: $\mathbb{P}(\phi_t^{(1)}(x, f_t) \in A) > 0$ for any $f \in \mathcal{F}$, $t \in \mathcal{T}$, $x \in \mathcal{X}$, and $A \subset \mathbb{R}^d$; and that the feedback observed at any epoch t, conditioned on the action X_t, is independent of the history that is available at that epoch. Given the structure imposed in the first part of the assumption, the second part implies that if gradients of two cost functions are "close" to each other, the probability measures of the observed feedbacks are also "close." The structure imposed by Assumption 1 is satisfied in many settings. For instance, it applies to Example 1 (with $\mathcal{X} \subset \mathbb{R}$) with $\tilde{C} = 1/(2\sigma^2)$.

Theorem 2 (Lower Bound on Achievable Performance). *Let Assumption 1 hold. Then, there exists a constant $C > 0$, independent of T and V_T, such that for any policy $\pi \in \mathcal{P}_{\phi^{(1)}}$ and for all $T \geqslant 1$,*

$$\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}, T) \geqslant C \cdot V_T^{1/3} T^{2/3}.$$

Key ideas in the proof of Theorem 2. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a probability space $\mathcal{Y}$, let

$$\mathcal{K}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}\left[\log\left(\frac{d\,\mathbb{P}\{Y\}}{d\,\mathbb{Q}\{Y\}}\right)\right], \tag{9}$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to $\mathbb{P}$, and Y is a random variable defined over $\mathcal{Y}$. This quantity is known as the Kullback-Leibler (KL) divergence. To establish the result, we consider sequences from a subset of $\mathcal{V}$ defined in the following way: in the beginning of each batch of size $\tilde{\Delta}_T$ (nature's decision variable), one of two "almost flat" functions is independently drawn according to a uniform distribution, and set as the cost function throughout the next $\tilde{\Delta}_T$ epochs. Then, the distance between these functions, and the batch size $\tilde{\Delta}_T$ are tuned such that (1) any drawn sequence must maintain the variation constraint; and (2) the functions are chosen to be "close" enough whereas the batches are sufficiently short, such that distinguishing between the two functions over the batch is subject to a significant error probability, yet the two functions are sufficiently "separated" to maximize the incurred regret. (Formally, the KL divergence is bounded throughout each batch, and hence any admissible policy trying to identify the current cost function can only do so with a strictly positive error probability.)

Upper bound on performance. To establish that the lower bound is achievable, we show that the restarting procedure introduced in §3 enables to carry over the property of rate optimality from the adversarial setting to the non-stationary stochastic setting. As a subroutine algorithm, we will use an adaptation of the OGD algorithm introduced by Zinkevich (2003).

OGD algorithm. Input: a decreasing sequence of nonnegative real numbers $\{\eta_t\}_{t=2}^{T}$.

1. Select some $X_1 \in \mathcal{X}$.
2. For any $t = 1, \ldots, T-1$, set $X_{t+1} = P_{\mathcal{X}}(X_t - \eta_{t+1}\phi_t^{(1)}(X_t, f_t))$, where $P_{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{X}} \|x - y\|$ is the Euclidean projection operator on $\mathcal{X}$.

For any value of τ that is dictated by the restarting procedure, the OGD algorithm can be defined via the sequence of mappings $\{\mathcal{A}_{t-\tau}\}$, $t \geqslant \tau+1$ as follows:

$$\mathcal{A}_{t-\tau}(\mathcal{H}_{\tau,t}) = \begin{cases} \text{some } X_1 \in \mathcal{X} & \text{if } t = \tau+1 \\ P_{\mathcal{X}}(X_{t-1} - \eta_{t-\tau}\phi^{(1)}_{t-1}) & \text{if } t > \tau+1 \end{cases}$$

for any epoch $t \geqslant \tau+1$. For the structure $(\mathcal{F}, \phi^{(1)})$ of convex cost functions and noisy gradient access, Flaxman et al. (2005) consider the OGD algorithm with $X_1 = 0$ and the selection $\eta_t = r/(G\sqrt{T})$, $t = 2, \ldots, T$. Here, r denotes the radius of the action set: $r = \inf\{y > 0\colon \mathcal{X} \subseteq \mathbf{B}_y(x) \text{ for some } x \in \mathbb{R}^d\}$, where $\mathbf{B}_y(x)$ is a ball with radius y, centered at point x, and show that this algorithm achieves a regret of order $\sqrt{T}$ in the adversarial setting. For completeness, we prove in Lemma 4 (given in Appendix C) that under Assumption 1, this performance cannot be improved upon in the adversarial OCO setting.

We next characterize the regret of the restarting procedure that uses the OGD policy as an input.

Theorem 3 (Performance of Restarted OGD Under Noisy Gradient Access). *Consider the feedback setting $\phi = \phi^{(1)}$, and let π be the policy defined by the restarting procedure with a batch size $\Delta_T = \lceil (T/V_T)^{2/3} \rceil$, and the OGD algorithm parameterized by $\eta_t = r/(G\sqrt{\Delta_T})$, $t = 2, \ldots, \Delta_T$ as a subroutine. Then, there is some finite constant $\bar{C}$, independent of T and V_T, such that for all $T \geqslant 2$,*

$$\mathcal{R}^{\pi}_{\phi}(\mathcal{V}, T) \leqslant \bar{C} \cdot V_T^{1/3} T^{2/3}.$$

Recalling the connection between the regret in the adversarial setting and the one in the non-stationary SA setting (Proposition 2), the result of the theorem is essentially a direct consequence of bounds in the OCO literature. In particular, Flaxman et al. (2005, Lemma 3.1) provide a bound on $\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}, \Delta_T)$ of order $\sqrt{\Delta_T}$, and the result follows by balancing the terms in (7) by a proper selection of Δ_T.

When selecting a large batch size, the ability to track the single best action within each batch improves, but the single best action within a certain batch may have substantially worse performance than that of the dynamic oracle. In contrast, when selecting a small batch size, the performance of tracking the single best action within each batch gets worse, but over the whole horizon, the series of single best actions (one for each batch) achieves a performance that approaches the dynamic oracle.

We note that Theorem 3 holds for any (deterministic or random) initial action of the subroutine OGD algorithm; a practical special case is one in which the initial action of any batch $j > 1$ is determined by taking one further gradient step from the last action of batch $j-1$.

Recalling the lower bound in Theorem 2, Theorem 3 implies that the performance of restarted OGD is rate optimal, and the minimax regret under structure $(\mathcal{V}, \phi^{(1)})$ is

$$\mathcal{R}^{*}_{\phi^{(1)}}(\mathcal{V}, T) \asymp V_T^{1/3} T^{2/3}.$$

Roughly speaking, this characterization provides a mapping between the variation budget V_T and the minimax regret under noisy gradient observations. For example, when $V_T = T^{\alpha}$ for some $0 \leqslant \alpha \leqslant 1$, the minimax regret is of order $T^{(2+\alpha)/3}$, hence we obtain the minimax regret in a full spectrum of variation scales, from order $T^{2/3}$ when the variation is a constant (independent of the horizon length), up to order T that corresponds to the case where V_T scales linearly with T (consistent with Proposition 1).

Alternative algorithms. Although the restarting procedure (together with suitable balancing of the batch size) can be used as a template for deriving "good" policies in non-stationary stochastic settings, it serves mainly as a tool to articulate a general and unified principle for designing rate optimal policies. Indeed, rate optimal performance may also be achieved by taking alternative paths that may be considered as more appealing from practical points of view. One of these may rely on attempting to directly retune the parameters of the subroutine OCO algorithm. Though not surprisingly, OGD-type policies with classical step size selections (such as $1/t$ or $1/\sqrt{t}$) may perform poorly in non-stationary environments (see Example 1 in Appendix B), we establish next that one may fine tune such a policy to achieve rate optimality, matching the lower bound given in Theorem 2.

Proposition 3 (Optimal Tuning of OGD). *Assume $\phi = \phi^{(1)}$, and let π be the OGD algorithm with $\eta_t = (r/G) \cdot (V_T/T)^{1/3}$, $t = 2, \ldots, T$. Then, there exists a finite constant $\bar{C}$, independent of T and V_T, such that for all $T \geqslant 2$,*

$$\mathcal{R}^{\pi}_{\phi}(\mathcal{V}, T) \leqslant \bar{C} \cdot V_T^{1/3} T^{2/3}.$$

The key in tuning the OGD algorithm to achieves rate optimal performance in the non-stationary SA setting is a suitable adjustment of the stepsize sequence as a function of the variation budget V_T. A sequence of "larger" steps that converge "slower" to zero allows the policy to respond efficiently to potential changes in the environment; the larger the variation budget is (relative to the horizon length T), the larger the stepsizes that are required to "keep up" with the potential changes.

Noisy access to the function value. Considering the feedback structure $\phi^{(0)}$ and the class $\mathcal{F}$, Flaxman et al. (2005) show that, in the adversarial OCO setting, a modification of the OGD algorithm can be tuned to achieve regret of order $T^{3/4}$; see also Kleinberg (2004). There is no indication that this regret rate is the best possible, and to the best of our knowledge, under cost observations and general convex cost functions, the question of rate optimality is an open problem in the adversarial OCO setting. By Proposition 2, the regret of order $T^{3/4}$ that is achievable in the OCO setting implies that a regret of order $V_T^{1/5} T^{4/5}$ is achievable in the non-stationary SA setting, by applying the restarting procedure. At present, we are not aware of any algorithm that guarantees a lower regret rate for arbitrary action spaces of dimension d, we conjecture that a rate optimal algorithm in

the OCO setting can be lifted to a rate optimal procedure in the non-stationary stochastic setting by applying the restarting procedure.[5] The next section supports this conjecture examining the case of strongly convex cost functions.

5. Rate Optimality: The Strongly Convex Case

Preliminaries. We now focus on the class of strongly convex functions $\mathcal{F}_s \subseteq \mathcal{F}$, defined such that in addition to the conditions that are stipulated by membership in $\mathcal{F}$, for a finite number $H > 0$, the sequence $\{f_t\}$ satisfies

$$H\mathbf{I}_d \preceq \nabla^2 f_t(x) \preceq G\mathbf{I}_d \quad \text{for all } x \in \mathcal{X} \text{ and all } t \in \mathcal{T}, \tag{10}$$

where $\mathbf{I}_d$ denotes the d-dimensional identity matrix. Here, for two square matrices of the same dimension A and B, we write $A \preceq B$ to denote that $B - A$ is positive semidefinite, and $\nabla^2 f(x)$ denotes the Hessian of $f(\cdot)$, evaluated at point $x \in \mathcal{X}$; for the sake of simplicity, we assume that G is a unified bound that also appears in (1). In the presence of strongly convex cost functions, it is well known that local properties of the functions around their minimum play a key role in the performance of sequential optimization procedures. To localize the analysis, we adapt the functional variation definition so that it is measured by the uniform norm over the *convex hull of the minimizers*, denoted by

$$\mathcal{X}^* = \left\{ x \in \mathbb{R}^d \colon x = \sum_{t=1}^{T} \lambda_t x_t^*, \sum_{t=1}^{T} \lambda_t = 1, \lambda_t \geqslant 0 \text{ for all } t \in \mathcal{T} \right\}.$$

Using the above, we measure variation by

$$\mathrm{Var}_s(f_1, \ldots, f_T) := \sum_{t=2}^{T} \sup_{x \in \mathcal{X}^*} |f_t(x) - f_{t-1}(x)|. \tag{11}$$

Given the class $\mathcal{F}_s$ and a variation budget V_T, we define the temporal uncertainty set as follows:

$$\mathcal{V}_s = \{ f = \{f_1, \ldots, f_T\} \subset \mathcal{F}_s \colon \mathrm{Var}_s(f_1, \ldots, f_T) \leqslant V_T \}.$$

We note that the proof of Proposition 2 effectively holds without change under the above structure. Hence first-order optimality is carried over from the OCO setting, as long as V_T is sublinear. We next examine rate optimality results.

5.1. Noisy Access to the Gradient

For the class $\mathcal{F}_s$ and gradient feedback $\phi_t(x, f_t) = \nabla f_t(x)$, Hazan et al. (2007) consider the OGD algorithm with a tuned selection of $\eta_t = 1/Ht$ for $t = 2, \ldots T$, and provide in the OCO framework a regret guarantee of order $\log T$ (relative to the single best action benchmark). For completeness, we provide in Appendix C (Lemma 2) a simple adaptation of this result to the case of *noisy* gradient access and an arbitrary random X_1. Hazan and Kale (2011) show that this algorithm is rate optimal in the OCO setting under strongly convex functions and a class of unbiased gradient feedback.[6]

Theorem 4 (Rate Optimality for Strongly Convex Functions and Noisy Gradient Access). 1. *Consider the feedback structure $\phi = \phi^{(1)}$, and let π be the policy defined by the restarting procedure with a batch size $\Delta_T = \lceil \sqrt{T \log T / V_T} \rceil$, and the OGD algorithm parameterized by $\eta_t = (Ht)^{-1}$, $t = 2, \ldots, \Delta_T$ as a subroutine. Then, there exists a finite positive constant $\bar{C}$, independent of T and V_T, such that for all $T \geqslant 2$,*

$$\mathcal{R}_\phi^\pi(\mathcal{V}_s, T) \leqslant \bar{C} \cdot \log\left(\frac{T}{V_T} + 1\right) \sqrt{V_T T}.$$

2. *Let Assumption 1 hold. Then, there exists a constant $C > 0$, independent of T and V_T, such that for any policy $\pi \in \mathcal{P}_{\phi^{(1)}}$ and for all $T \geqslant 1$,*

$$\mathcal{R}_\phi^\pi(\mathcal{V}_s, T) \geqslant C \cdot \sqrt{V_T T}.$$

Up to a logarithmic term, Theorem 4 establishes rate optimality in the non-stationary SA setting of the policy defined by the restarting procedure with the tuned *OGD* algorithm as a subroutine. We further note that by directly tuning the OGD algorithm (in a manner similar to the one described in Proposition 3) one may achieve a performance of $O(\sqrt{V_T T})$. Hence the *minimax regret* under structure $(\mathcal{V}_s, \phi^{(0)})$ is

$$\mathcal{R}^*_{\phi^{(1)}}(\mathcal{V}_s, T) \asymp \sqrt{V_T T}.$$

Theorem 4 further validates the "metaprinciple" in the case of strongly convex functions and noisy gradient feedback: rate optimality in the adversarial setting (relative to the single best action benchmark) can be adapted by the restarting procedure to guarantee an essentially optimal regret rate in the non-stationary stochastic setting (relative to the dynamic benchmark).

The first part of Theorem 4 is derived directly from Proposition 2, by plugging in a bound on $\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}_s, \Delta_T)$ of order $\log T$ (given by Lemma 2 in the case of noisy gradient access), and a tuned selection of Δ_T. The proof of the second part follows by arguments similar to the ones used in the proof of Theorem 2, adjusting for strongly convex cost functions.

5.2. Noisy Access to the Cost

We now consider the structure $(\mathcal{V}_s, \phi^{(0)})$, in which the cost functions are strongly convex and the decision maker has noisy access to the cost. To show that rate optimality is carried over from the adversarial setting to the non-stationary stochastic setting, we first need to introduce an algorithm that is rate optimal in the adversarial setting under the structure $(\mathcal{F}_s, \phi^{(0)})$.

Estimated gradient step (EGS). For a small δ, we denote by $\mathcal{X}_\delta$ the δ-interior of the action set $\mathcal{X}$:

$$\mathcal{X}_\delta = \{ x \in \mathcal{X} \colon \mathbf{B}_\delta(x) \subseteq \mathcal{X} \}.$$

We assume access to the projection operator $P_{\mathscr{X}_\delta}(y) = \arg\min_{x\in\mathscr{X}_\delta}\|x-y\|$ on the set $\mathscr{X}_\delta$.

For $k=1,\ldots,d$, let $e^{(k)}$ denote the unit vector with 1 at the kth coordinate. The EGS algorithm is defined through three sequences of real numbers $\{h_t\}$, $\{a_t\}$, and $\{\delta_t\}$, where[7] $\nu \geqslant \delta_t \geqslant h_t$ for all $t\in\mathscr{T}$.

EGS algorithm. Inputs: decreasing sequences of real numbers $\{a_t\}_{t=1}^{T-1}$, $\{h_t\}_{t=1}^{T-1}$, $\{\delta_t\}_{t=1}^{T-1}$.

1. Select some initial point $X_1 = Z_1$ in $\mathscr{X}$.
2. For each $t=1,\ldots,T-1$:
   (a) Draw ψ_t uniformly over the set $\{\pm e^{(1)},\ldots,\pm e^{(d)}\}$.
   (b) Compute stochastic gradient estimate $\hat{\nabla}_{h_t} f_t(Z_t) = h_t^{-1}\phi_t^{(0)}(Z_t + h_t\psi_t)\psi_t$.
   (c) Update $Z_{t+1} = P_{\mathscr{X}_{\delta_t}}(Z_t - a_t\hat{\nabla}_{h_t} f_t(Z_t))$.
   (d) Select the action $X_{t+1} = Z_{t+1} + h_{t+1}\psi_t$.

For any value of τ dictated by the restarting procedure, the EGS policy can be formally defined by

$$\mathscr{A}_{t-\tau} = \begin{cases} \text{some } Z_1 & \text{if } t=\tau+1 \\ Z_{t-\tau} + h_{t-\tau}\psi_{t-\tau-1} & \text{if } t>\tau+1. \end{cases}$$

Note that $\mathbb{E}[\hat{\nabla}_h f_t(Z_t) \mid X_t] = \nabla f_t(Z_t)$ (cf. Nemirovski and Yudin 1983, Chapter 7), and that the EGS algorithm essentially consists of estimating a stochastic direction of improvement and following this direction. In Lemma 1 (Appendix C), we show that when tuned by $a_t = 2d/Ht$ and $\delta_t = h_t = a_t^{1/4}$ for all $t\in\{1,\ldots,T-1\}$, the EGS algorithm achieves a regret of order $\sqrt{T}$ compared to a single best action in the adversarial setting under structure $(\mathscr{F}_s, \phi^{(0)})$. In Lemma 3 (Appendix C), we establish that under Assumption 2 (given below), this performance is rate optimal in the adversarial setting.

Before analyzing the minimax regret in the non-stationary SA setting, let us introduce a counterpart to Assumption 1 for the case of cost feedback, that will be used in deriving a lower bound on the regret.

Assumption 2 (Cost Feedback Structure).
1. *$\phi_t^{(0)}(x, f_t) = f_t(x) + \varepsilon_t$ for any $f\in\mathscr{F}$, $x\in\mathscr{X}$, and $t\in\mathscr{T}$, where ε_t, $t\geqslant 1$ are iid random variables with zero mean and bounded variance.*
2. *Let $G(\cdot)$ be the cumulative distribution function of ε_t. Then, there exists a constant $\tilde{C}$ such that for any $a\in\mathbb{R}$, $\int \log(dG(y)/dG(y+a))\,dG(y) \leqslant \tilde{C}\cdot a^2$.*

Theorem 5 (Rate Optimality for Strongly Convex Functions and Noisy Cost Access). 1. *Consider the feedback structure $\phi = \phi^{(0)}$, and let π be the policy defined by the restarting procedure with EGS parameterized by $a_t = 2d/(Ht)$, $h_t = \delta_t = (2d/(Ht))^{1/4}$, $t=1,\ldots,T-1$, as subroutine, and a batch size $\Delta_T = \lceil (T/V_T)^{2/3}\rceil$. Then, there exists a finite constant $\bar{C} > 0$, independent of T and V_T, such that for all $T\geqslant 2$,*

$$\mathscr{R}_\phi^\pi(\mathscr{V}_s, T) \leqslant \bar{C}\cdot V_T^{1/3}T^{2/3}.$$

2. *Let Assumption 2 hold. Then, there exists a constant $C>0$, independent of T and V_T, such that for any policy $\pi\in\mathscr{P}_{\phi^{(0)}}$ and for all $T\geqslant 1$,*

$$\mathscr{R}_\phi^\pi(\mathscr{V}_s, T) \geqslant C\cdot V_T^{1/3}T^{2/3}.$$

Theorem 5 again establishes the ability to "port over" rate optimality from the adversarial OCO setting to the non-stationary stochastic setting, this time under structure $(\mathscr{F}_s, \phi^{(0)})$. The theorem establishes a characterization of the minimax regret under structure $(\mathscr{V}_s, \phi^{(0)})$:

$$\mathscr{R}^*_{\phi^{(0)}}(\mathscr{V}_s, T) \asymp V_T^{1/3}T^{2/3}.$$

Illustrative Numerical Results. In Appendix D, 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 of fixed variation, different feedback structures, and noise. Under noisy gradient access ($\phi^{(1)}$), our results support a regret of order $\sqrt{T}$ achieved by the restarted OGD, where the multiplicative factor ranges in the interval $[0.05, 0.94]$. Under noisy cost access ($\phi^{(0)}$), our results support a regret of order $T^{2/3}$ achieved by the restarted EGS, where the multiplicative factor ranges in the interval $[2.09, 2.88]$. In both cases when observations are more noisy, the multiplicative constant increases. While the above policies were introduced mainly as a tool to study the minimax regret rates in the non-stationary stochastic optimization problem, and, in particular, were not designed to optimize performance in practical settings, we note that, in most of the instances that we considered, these restarting policies perform at least "on par" with policies that use fixed stepsizes; whereas such fixed-step policies are considered as possible heuristics in many practical instances (see, e.g., Chapter 4 of Benveniste et al. 1990), they have no performance guarantees relative to the dynamic oracle considered here.

6. Concluding Remarks

On the transition from stationary to non-stationary settings. Throughout the paper we address "significant" variation in the cost function, and for the sake of concreteness, assume $V_T \geqslant 1$. Nevertheless, one may show (following the proofs of Theorems 2–5) that under each of the different cost and feedback structures, the established bounds hold for "smaller" variation scales, and if the variation scale is sufficiently "small," the minimax regret rates coincide with the ones in the classical stationary SA settings. We refer to the variation scales at which the stationary and the non-stationary complexities coincide as "critical variation scales." Not surprisingly, these transition points between the stationary and the non-stationary regimes differ across cost and feedback structures. Table 2 summarizes the minimax regret rates for a variation budget of the form $V_T = T^\alpha$, and documents the critical variation scales in different

Table 2. Critical variation scales.

Setting		Order of regret		
Class of functions	Feedback	Stationary	Non-stationary	Critical variation scale
Convex	Noisy gradient	$T^{1/2}$	$\max\{T^{1/2}, T^{(2+\alpha)/3}\}$	$T^{-1/2}$
Strongly convex	Noisy gradient	$\log T$	$\max\{\log T, T^{(1+\alpha)/2}\}$	$(\log T)^2 T^{-1}$
Strongly convex	Noisy function	$T^{1/2}$	$\max\{T^{1/2}, T^{(2+\alpha)/3}\}$	$T^{-1/2}$

Note. The growth rates of the minimax regret in different settings for $V_T = T^\alpha$ (where $\alpha \leqslant 1$) and the variation scales that separate the stationary and the non-stationary regimes.

settings. In all cases highlighted in the table, the transition point occurs for variation scales that diminish with T; this critical quantity therefore measures how "small" should the temporal variation be, relative to the horizon length, to make non-stationarity effects insignificant relative to other problem primitives insofar as the regret measure goes.

Inaccurate or no information on the variation budget. The policies introduced in this paper rely on prior knowledge of the variation budget V_T, but predictions of V_T may underestimate or overestimate it. Denoting the "real" variation budget by V_T and the estimate that is used by the agent when tuning the restarting procedure by $\hat{V}_T$, one may observe that Proposition 2 holds with V_T (and the respective class $\mathscr{V}$), but Δ_T is tuned (e.g., in Theorems 2, 4, and 5) using the estimate $\hat{V}_T$. This implies that, in all the settings that have been considered here, when the "real" budget is close enough to the estimate $\hat{V}_T$, the restarting procedure still guarantees long-run average optimality (naturally, the respective performance is dominated by the one achieved with an accurate knowledge of V_T).

Since there are essentially no restrictions on the rate at which the variation budget can be consumed (in particular, nature is not constrained to sequences with epoch-homogenous variation), an interesting and potentially challenging open problem is to delineate to what extent it is possible to design adaptive policies that do not have a priori knowledge of the variation budget, yet have performance "close" to the order of the minimax regret characterized in this paper. Moreover, for known or unknown variation budgets, characterizing the minimax regret more finely, including the multiplicative constants, remains an important open research avenue of clear practical importance.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2015.1408.

Acknowledgments

The authors are grateful to the associate editor and two reviewers for their feedback that helped to improve the paper. This work was supported by the National Science Foundation [NSF Grant 0964170] and U.S.-Israel Binational Science Foundation [BSF Grant 2010466].

Appendix A. Proofs of Main Results

Proof of Proposition 1. See Appendix B.

Proof of Theorem 1. Fix $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$, and assume that $V_T = o(T)$. Let $\mathscr{A}$ be a policy such that $\mathscr{G}^{\mathscr{A}}_\phi(\mathscr{F}, T) = o(T)$, and let $\Delta_T \in \{1, \ldots, T\}$. Let π be the policy defined by the restarting procedure that uses $\mathscr{A}$ as a subroutine with batch size Δ_T. Then, by Proposition 2,

$$\frac{\mathscr{R}^\pi_\phi(\mathscr{V}, T)}{T} \leqslant \frac{\mathscr{G}^{\mathscr{A}}_\phi(\mathscr{F}, \Delta_T)}{\Delta_T} + \frac{\mathscr{G}^{\mathscr{A}}_\phi(\mathscr{F}, \Delta_T)}{T} + 2\Delta_T \cdot \frac{V_T}{T}$$

for any $1 \leqslant \Delta_T \leqslant T$. Since $V_T = o(T)$, for any selection of Δ_T such that $\Delta_T = o(T/V_T)$ and $\Delta_T \to \infty$ as $T \to \infty$, the right-hand side of the above converges to zero as $T \to \infty$, concluding the proof. □

Proof of Proposition 2. Fix $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$, $T \geqslant 1$, and $1 \leqslant V_T \leqslant T$. For $\Delta_T \in \{1, \ldots, T\}$, we break the horizon $\mathscr{T}$ into a sequence of batches $\mathscr{T}_1, \ldots, \mathscr{T}_m$ of size Δ_T each (except possibly the last batch) according to (8). Fix $\mathscr{A} \in \mathscr{P}_\phi$, and let π be the policy defined by the restarting procedure that uses $\mathscr{A}$ as a subroutine with batch size Δ_T. Let $f \in \mathscr{V}$. We decompose the regret in the following way: $R^\pi(f, T) = \sum_{j=1}^m R^\pi_j$, where

$$\begin{aligned} R^\pi_j &:= \mathbb{E}^\pi\left[\sum_{t\in\mathscr{T}_j}(f_t(X_t) - f_t(x^*_t))\right] \\ &= \underbrace{\mathbb{E}^\pi\left[\sum_{t\in\mathscr{T}_j} f_t(X_t)\right] - \min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\}}_{J_{1,j}} \\ &\quad + \underbrace{\min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\} - \sum_{t\in\mathscr{T}_j} f_t(x^*_t)}_{J_{2,j}}. \end{aligned} \tag{12}$$

The first component, $J_{1,j}$, is the regret with respect to the single best action of batch j, and the second component, $J_{2,j}$, is the difference in performance along batch j between the single best action of the batch and the dynamic benchmark. We next analyze $J_{1,j}$, $J_{2,j}$, and the regret throughout the horizon.

Step 1 (*Analysis of* $J_{1,j}$). By taking the sup over all sequences in $\mathscr{F}$ (recall that $\mathscr{V} \subseteq \mathscr{F}$) and using the regret with respect to the single best action in the adversarial setting, one has

$$\begin{aligned} J_{1,j} &\leqslant \sup_{f\in\mathscr{F}}\left\{\mathbb{E}^\pi\left[\sum_{t\in\mathscr{T}_j} f_t(X_t)\right] - \min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\}\right\} \\ &\leqslant \mathscr{G}^{\mathscr{A}}_\phi(\mathscr{F}, \Delta_T), \end{aligned} \tag{13}$$

where the last inequality holds using (6), and since in each batch decisions are dictated by $\mathscr{A}$, and since in each batch there are at most Δ_T epochs (recall that $\mathscr{G}_\phi^{\mathscr{A}}$ is nondecreasing in the number of epochs).

Step 2 (*Analysis of* $J_{2,j}$). Defining $f_0(x) = f_1(x)$, we denote by $V_j = \sum_{t\in\mathscr{T}_j} \|f_t - f_{t-1}\|$ the variation along batch $\mathscr{T}_j$. By the variation constraint (3), one has

$$\sum_{j=1}^{m} V_j = \sum_{j=1}^{m} \sum_{t\in\mathscr{T}_j} \sup_{x\in\mathscr{X}} |f_t(x) - f_{t-1}(x)| \leqslant V_T. \tag{14}$$

Let $\tilde{t}$ be the first epoch of batch $\mathscr{T}_j$. Then,

$$\begin{aligned}\min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\} - \sum_{t\in\mathscr{T}_j} f_t(x_t^*) &\leqslant \sum_{t\in\mathscr{T}_j} (f_t(x_{\tilde{t}}^*) - f_t(x_t^*)) \\ &\leqslant \Delta_T \cdot \max_{t\in\mathscr{T}_j}\{f_t(x_{\tilde{t}}^*) - f_t(x_t^*)\}. \end{aligned}\tag{15}$$

We next show that $\max_{t\in\mathscr{T}_j}\{f_t(x_{\tilde{t}}^*) - f_t(x_t^*)\} \leqslant 2V_j$. Suppose otherwise. Then, there is some epoch $t_0 \in \mathscr{T}_j$ at which $f_{t_0}(x_{\tilde{t}}^*) - f_{t_0}(x_{t_0}^*) > 2V_j$, implying

$$f_t(x_{t_0}^*) \overset{(a)}{\leqslant} f_{t_0}(x_{t_0}^*) + V_j < f_{t_0}(x_{\tilde{t}}^*) - V_j \overset{(b)}{\leqslant} f_t(x_{\tilde{t}}^*) \quad \text{for all } t \in \mathscr{T}_j,$$

where (a) and (b) follow from the fact that V_j is the maximal variation along batch $\mathscr{T}_j$. In particular, the above holds for $t = \tilde{t}$, contradicting the optimality of $x_{\tilde{t}}^*$ at epoch $\tilde{t}$. Therefore, one has from (15)

$$\min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\} - \sum_{t\in\mathscr{T}_j} f_t(x_t^*) \leqslant 2\Delta_T V_j. \tag{16}$$

Step 3 (*Analysis of the regret over* T *periods*). Summing (16) over batches and using (14), one has

$$\sum_{j=1}^{m}\left(\min_{x\in\mathscr{X}}\left\{\sum_{t\in\mathscr{T}_j} f_t(x)\right\} - \sum_{t\in\mathscr{T}_j} f_t(x_t^*)\right) \leqslant \sum_{j=1}^{m} 2\Delta_T V_j \leqslant 2\Delta_T V_T. \tag{17}$$

Therefore, by the regret decomposition in (12), and following (13) and (17), one has

$$R^\pi(f,T) \leqslant \sum_{j=1}^{m} \mathscr{G}_\phi^{\mathscr{A}}(\mathscr{F},\Delta_T) + 2\Delta_T V_T.$$

Since the above holds for any $f \in \mathscr{V}$, and recalling that $m = \lceil T/\Delta_T \rceil$, we have

$$\mathscr{R}_\phi^\pi(\mathscr{V},T) = \sup_{f\in\mathscr{V}} R^\pi(f,T) \leqslant \left\lceil \frac{T}{\Delta_T} \right\rceil \cdot \mathscr{G}_\phi^{\mathscr{A}}(\mathscr{F},\Delta_T) + 2\Delta_T V_T.$$

This concludes the proof. □

Proof of Theorem 2. Fix $T \geqslant 1$ and $1 \leqslant V_T \leqslant T$. We will restrict nature to a specific class of function sequences $\mathscr{V}' \subset \mathscr{V}$. In any element of $\mathscr{V}'$, the cost function is limited to be one of two known quadratic functions, selected by nature in the beginning of every batch of $\tilde{\Delta}_T$ epochs, and applied for the following $\tilde{\Delta}_T$ epochs. Then, we will show that any policy in $\mathscr{P}_{\phi^{(1)}}$ must incur regret of order $V_T^{1/3}T^{2/3}$.

Step 1 (*Preliminaries*). Let $\mathscr{X} = [0,1]$ and consider the following two functions:

$$f^1(x) = \begin{cases} 1/2 + \delta - 2\delta x + \left(x - 1/4\right)^2 & x < 1/4 \\ 1/2 + \delta - 2\delta x & 1/4 \leqslant x \leqslant 3/4\,; \\ 1/2 + \delta - 2\delta x + \left(x - 3/4\right)^2 & x > 3/4 \end{cases}$$
$$f^2(x) = \begin{cases} 1/2 - \delta + 2\delta x + \left(x - 1/4\right)^2 & x < 1/4 \\ 1/2 - \delta + 2\delta x & 1/4 \leqslant x \leqslant 3/4 \\ 1/2 - \delta + 2\delta x + \left(x - 3/4\right)^2 & x > 3/4, \end{cases} \tag{18}$$

for some $\delta > 0$ that will be specified shortly. Denoting $x_k^* = \arg\min_{x\in[0,1]} f^k(x)$, one has $x_1^* = 3/4 + \delta$ and $x_2^* = 1/4 - \delta$. It is immediate that f^1 and f^2 are convex and for any $\delta \in (0, 1/4)$, obtain a global minimum in an interior point in $\mathscr{X}$. For some $\tilde{\Delta}_T \in \{1,\ldots,T\}$ that will be specified below, define a partition of the horizon $\mathscr{T}$ to $m = \lceil T/\tilde{\Delta}_T \rceil$ batches $\mathscr{T}_1,\ldots,\mathscr{T}_m$ of size $\tilde{\Delta}_T$ each (except perhaps $\mathscr{T}_m$), according to (8). Define

$$\begin{aligned}\mathscr{V}' = \big\{f\colon\, & f_t \in \{f^1, f^2\} \text{ and } f_t = f_{t+1} \text{ for } (j-1)\tilde{\Delta}_T + 1 \leqslant t \\ & \leqslant \min\{j\tilde{\Delta}_T, T\} - 1,\, j = 1,\ldots,m\big\}. \end{aligned}\tag{19}$$

In every sequence in $\mathscr{V}'$, the cost function is restricted to the set $\{f^1, f^2\}$, and cannot change throughout a batch. Let $\delta = V_T\tilde{\Delta}_T/2T$. Any sequence in $\mathscr{V}'$ consists of convex functions, with minimizers that are interior points in $\mathscr{X}$. In addition, one has

$$\begin{aligned}\sum_{t=2}^{T} \|f_t - f_{t-1}\| &\leqslant \sum_{j=2}^{m} \sup_{x\in\mathscr{X}} |f^1(x) - f^2(x)| = \left(\left\lceil \frac{T}{\tilde{\Delta}_T} \right\rceil - 1\right)\cdot 2\delta \\ &\leqslant \frac{2T\delta}{\tilde{\Delta}_T} \leqslant V_T, \end{aligned}$$

where the first inequality holds since the function can only change between batches. Therefore $\mathscr{V}' \subset \mathscr{V}$.

Step 2 (*Bounding the relative entropy within a batch*). Fix any policy $\pi \in \mathscr{P}_{\phi^{(1)}}$. At each $t \in \mathscr{T}_j$, the decision maker selects $X_t \in \mathscr{X}$ and observes a noisy feedback $\phi_t^{(1)}(X_t, f_t)$. For any $f \in \mathscr{F}$, denote by $\mathbb{P}_f^\pi$ the probability measure under policy π when f is the sequence of cost functions that is selected by nature, and by $\mathbb{E}_f^\pi$ the associated expectation operator. For any $\tau \geqslant 1$, $A \subset \mathbb{R}^{d\times\tau}$ and $B \subset \mathscr{U}$, denote $\mathbb{P}_f^{\pi,\tau}(A,B) := \mathbb{P}_f^\pi\{\{\phi_t^{(1)}(X_t,f_t)\}_{t=1}^\tau \in A,\, U \in B\}$. In what follows, we make use of the KL divergence defined in (9).

Lemma A.1 (Bound on KL Divergence for Noisy Gradient Observations). *Consider the feedback structure* $\phi = \phi^{(1)}$ *and let Assumption* 1 *holds. Then, for any* $\tau \geqslant 1$ *and* $f,\, g \in \mathscr{F}$,

$$\mathscr{K}(\mathbb{P}_f^{\pi,\tau} \,\|\, \mathbb{P}_g^{\pi,\tau}) \leqslant \tilde{C}\,\mathbb{E}_f^\pi\left[\sum_{t=1}^{\tau} \|\nabla f_t(X_t) - \nabla g_t(X_t)\|^2\right],$$

where $\tilde{C}$ *is the constant that appears in the second part of Assumption* 1.

The proof of Lemma A.1 appears in Appendix B. We also use the following result for the minimal error probability in distinguishing between two distributions.

LEMMA A.2 (THEOREM 2.2 IN TSYBAKOV 2008). *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions on $\mathcal{H}$ such that $\mathcal{K}(\mathbb{P}\,\|\,\mathbb{Q}) \leqslant \beta < \infty$. Then, for any $\mathcal{H}$-measurable real function $\varphi\colon \mathcal{H} \to \{0,1\}$,*

$$\max\{\mathbb{P}(\varphi=1), \mathbb{Q}(\varphi=0)\} \geqslant \frac{1}{4}\exp\{-\beta\}.$$

Set $\tilde{\Delta}_T = \max\{\lfloor (1/(4\tilde{C}))^{1/3}(T/V_T)^{2/3}\rfloor, 1\}$ (where $\tilde{C}$ is the constant that appears in Part 2 of Assumption 1). We next show that for each batch $\mathcal{T}_j$, $\mathcal{K}(\mathbb{P}_{f^1}^{\pi,\tau}\,\|\,\mathbb{P}_{f^2}^{\pi,\tau})$ is bounded for any $1 \leqslant \tau \leqslant |\mathcal{T}_j|$. Fix $j \in \{1,\dots,m\}$. Then,

$$\begin{aligned}
\mathcal{K}(\mathbb{P}_{f^1}^{\pi,|\mathcal{T}_j|}\,\|\,\mathbb{P}_{f^2}^{\pi,|\mathcal{T}_j|}) &\overset{(a)}{\leqslant} \tilde{C}\,\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}(\nabla f_t^1(X_t) - \nabla f_t^2(X_t))^2\right] \\
&= \tilde{C}\,\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j} 16\delta^2 X_t^2\right] \leqslant 16\tilde{C}\tilde{\Delta}_T\delta^2 \\
&\overset{(b)}{=} \frac{4\tilde{C}V_T^2\tilde{\Delta}_T^3}{T^2} \overset{(c)}{\leqslant} \max\left\{1, \frac{2\tilde{C}V_T}{T}\right\} \\
&\overset{(d)}{\leqslant} \max\{1, 2\tilde{C}\},
\end{aligned}$$

where (a) follows from Lemma A.1; (b) and (c) hold given the respective values of δ and $\tilde{\Delta}_T$; and (d) holds by $V_T \leqslant T$. Set $\beta = \max\{1, 2\tilde{C}\}$. Since $\mathcal{K}(\mathbb{P}_{f^1}^{\pi,\tau}\,\|\,\mathbb{P}_{f^2}^{\pi,\tau})$ is nondecreasing in τ throughout a batch, we deduce that $\mathcal{K}(\mathbb{P}_{f^1}^{\pi,\tau}\,\|\,\mathbb{P}_{f^2}^{\pi,\tau})$ is bounded by β throughout each batch. Then, for any $x_0 \in \mathcal{X}$, using Lemma A.2 with $\varphi_t = \{X_t \leqslant x_0\}$, one has

$$\max\{\mathbb{P}_{f^1}\{X_t \leqslant x_0\}, \mathbb{P}_{f^2}\{X_t > x_0\}\} \geqslant \frac{1}{4e^{\beta}} \quad \text{for all } t \in \mathcal{T}. \tag{20}$$

Step 3 (*A lower bound on the incurred regret for $f \in \mathcal{V}'$*). Set $x_0 = \frac{1}{2}(x_1^* + x_2^*) = \frac{1}{2}$. Let $\tilde{f}$ be a random sequence in which in the beginning of each batch $\mathcal{T}_j$, a cost function is independently drawn according to a discrete uniform distribution over $\{f^1, f^2\}$, and applied throughout the whole batch. In particular, note that, for any $1 \leqslant j \leqslant m$, for any epoch $t \in \mathcal{T}_j$, f_t is independent of $\mathcal{H}_{(j-1)\tilde{\Delta}_T+1}$ (the history that is available at the beginning of the batch). Clearly, any realization of $\tilde{f}$ is in $\mathcal{V}'$. In particular, taking expectation over $\tilde{f}$, one has

$$\begin{aligned}
&\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}', T) \\
&\geqslant \mathbb{E}^{\pi,\tilde{f}}\left[\sum_{t=1}^{T}\tilde{f}_t(X_t) - \sum_{t=1}^{T}\tilde{f}_t(x_t^*)\right] \\
&= \mathbb{E}^{\pi,\tilde{f}}\left[\sum_{j=1}^{m}\sum_{t\in\mathcal{T}_j}(\tilde{f}_t(X_t) - \tilde{f}_t(x_t^*))\right] \\
&= \sum_{j=1}^{m}\left(\frac{1}{2}\cdot\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}(f^1(X_t) - f^1(x_1^*))\right]\right. \\
&\qquad \left. + \frac{1}{2}\cdot\mathbb{E}_{f^2}^{\pi}\left[\sum_{t\in\mathcal{T}_j}(f^2(X_t) - f^2(x_2^*))\right]\right) \\
&\overset{(a)}{\geqslant} \sum_{j=1}^{m}\frac{1}{2}\left(\sum_{t\in\mathcal{T}_j}(f^1(x_0) - f^1(x_1^*))\,\mathbb{P}_{f^1}^{\pi}\{X_t > x_0\}\right. \\
&\qquad \left. + \sum_{t\in\mathcal{T}_j}(f^2(x_0) - f^2(x_2^*))\,\mathbb{P}_{f^2}^{\pi}\{X_t \leqslant x_0\}\right) \\
&\geqslant \sum_{j=1}^{m}\frac{\delta}{4}\sum_{t\in\mathcal{T}_j}(\mathbb{P}_{f^1}^{\pi}\{X_t > x_0\} + \mathbb{P}_{f^2}^{\pi}\{X_t \leqslant x_0\}) \\
&\geqslant \sum_{j=1}^{m}\frac{\delta}{4}\sum_{t\in\mathcal{T}_j}\max\{\mathbb{P}_{f^1}^{\pi}\{X_t > x_0\}, \mathbb{P}_{f^2}^{\pi}\{X_t \leqslant x_0\}\} \\
&\overset{(b)}{\geqslant} \sum_{j=1}^{m}\frac{\delta}{4}\sum_{t\in\mathcal{T}_j}\frac{1}{4e^{\beta}} = \sum_{j=1}^{m}\frac{\delta\tilde{\Delta}_T}{16e^{\beta}} \\
&\overset{(c)}{=} \sum_{j=1}^{m}\frac{V_T\tilde{\Delta}_T^2}{32e^{\beta}T} \geqslant \frac{T}{\tilde{\Delta}_T}\cdot\frac{V_T\tilde{\Delta}_T^2}{32e^{\beta}T} = \frac{V_T\tilde{\Delta}_T}{32e^{\beta}},
\end{aligned}$$

where (a) holds since for any function $g\colon [0,1] \to \mathbb{R}^+$ and $x_0 \in [0,1]$ such that $g(x) \geqslant g(x_0)$ for all $x > x_0$, one has that $\mathbb{E}[g(X_t)] = \mathbb{E}[g(X_t) \mid X_t > x_0]\,\mathbb{P}\{X_t > x_0\} + \mathbb{E}[g(X_t) \mid X_t \leqslant x_0]\cdot \mathbb{P}\{X_t \leqslant x_0\} \geqslant g(x_0)\,\mathbb{P}\{X_t > x_0\}$ for any $t \in \mathcal{T}$, and similarly for any $x_0 \in [0,1]$ such that $g(x) \geqslant g(x_0)$ for all $x \leqslant x_0$, one obtains $\mathbb{E}[g(X_t)] \geqslant g(x_0)\,\mathbb{P}\{X_t \leqslant x_0\}$. In addition, (b) holds by (20) and (c) holds by $\delta = V_T\tilde{\Delta}_T/2T$. Suppose that $T \geqslant 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$. Applying the selected $\tilde{\Delta}_T$, one has

$$\begin{aligned}
\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}', T) &\geqslant \frac{V_T}{32e^{\beta}}\cdot\left\lfloor\left(\frac{1}{4\tilde{C}}\right)^{1/3}\left(\frac{T}{V_T}\right)^{2/3}\right\rfloor \\
&\geqslant \frac{V_T}{32e^{\beta}}\cdot\left(\left(\frac{1}{4\tilde{C}}\right)^{1/3}\left(\frac{T}{V_T}\right)^{2/3} - 1\right) \\
&= \frac{V_T}{32e^{\beta}}\cdot\left(\frac{T^{2/3} - (4\tilde{C})^{1/3}V_T^{2/3}}{(4\tilde{C})^{1/3}V_T^{2/3}}\right) \\
&\geqslant \frac{1}{64e^{\beta}(4\tilde{C})^{1/3}}\cdot V_T^{1/3}T^{2/3},
\end{aligned}$$

where the last inequality follows from $T \geqslant 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$. If $T < 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$ by Proposition 1, there exists a constant C such that $\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}, T) \geqslant C\cdot T \geqslant C\cdot V_T^{1/3}T^{2/3}$. Recalling that $\mathcal{V}' \subseteq \mathcal{V}$, we have

$$\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}, T) \geqslant \mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}', T) \geqslant \frac{1}{64e^{\beta}(4\tilde{C})^{1/3}}\cdot V_T^{1/3}T^{2/3}.$$

This concludes the proof. □

PROOF OF THEOREM 3. Fix $T \geqslant 1$ and $1 \leqslant V_T \leqslant T$. For any $\Delta_T \in \{1,\dots,T\}$, let $\mathcal{A}$ be the OGD algorithm with $\eta_t = \eta = r/(G\sqrt{\Delta_T})$ for any $t = 2,\dots,\Delta_T$ (where r denotes the radius of the action set $\mathcal{X}$), and let π be the policy defined by the restarting procedure with subroutine $\mathcal{A}$ and batch size Δ_T. Flaxman et al. (2005) consider the performance of the OGD algorithm (with a specific deterministic $x_1 = 0$) relative to the single best action in the adversarial setting, and show (Flaxman et al. 2005, Lemma 3.1) that $\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F}, \Delta_T) \leqslant rG\sqrt{\Delta_T}$. Following their analysis, one obtains that for an arbitrary (potentially random) initial action $\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F}, \Delta_T) \leqslant 2rG\sqrt{\Delta_T}$. Therefore, by Proposition 2,

$$\begin{aligned}
\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V}, T) &\leqslant \left(\frac{T}{\Delta_T} + 1\right)\cdot\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F}, \Delta_T) + 2V_T\Delta_T \\
&\leqslant \frac{2rG\cdot T}{\sqrt{\Delta_T}} + 2rG\sqrt{\Delta_T} + 2V_T\Delta_T.
\end{aligned}$$

Selecting $\Delta_T = \lceil (T/V_T)^{2/3} \rceil$, one has

$$\begin{aligned}\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V},T) &\leqslant \frac{2rG\cdot T}{(T/V_T)^{1/3}} + 2rG\left(\left(\frac{T}{V_T}\right)^{1/3}+1\right)\\ &\quad + 2V_T\left(\left(\frac{T}{V_T}\right)^{2/3}+1\right)\\ &\overset{(a)}{\leqslant} (2rG+4)\cdot V_T^{1/3}T^{2/3} + 2rG\cdot\left(\frac{T}{V_T}\right)^{1/3} + 2rG\\ &\overset{(b)}{\leqslant} (6rG+4)\cdot V_T^{1/3}T^{2/3}, \end{aligned} \tag{21}$$

where (a) and (b) follows since $1 \leqslant V_T \leqslant T$. This concludes the proof. □

Proof of Theorem 4. *Part* 1. We begin with the first part of the Theorem. Fix $T \geqslant 1$, and $1 \leqslant V_T \leqslant T$. For any $\Delta_T \in \{1,\ldots,T\}$, let $\mathcal{A}$ be the OGD algorithm with $\eta_t = 1/Ht$ for any $t = 2,\ldots,\Delta_T$, and let π be the policy defined by the restarting procedure with subroutine $\mathcal{A}$ and batch size Δ_T. By Lemma 2 (see Appendix C), one has

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}_s,\Delta_T) \leqslant \frac{G^2+\sigma^2}{2H}(1+\log\Delta_T). \tag{22}$$

Therefore, by Proposition 2,

$$\begin{aligned}\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V}_s,T) &\leqslant \left(\frac{T}{\Delta_T}+1\right)\cdot\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}_s,\Delta_T) + 2V_T\Delta_T\\ &\leqslant \left(\frac{T}{\Delta_T}+1\right)\frac{G^2+\sigma^2}{2H}(1+\log\Delta_T) + 2V_T\Delta_T.\end{aligned}$$

Selecting $\Delta_T = \lceil \sqrt{T/V_T} \rceil$, one has

$$\begin{aligned}&\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V}_s,T)\\ &\leqslant \left(\frac{T}{\sqrt{T/V_T}}+1\right)\frac{G^2+\sigma^2}{2H}\left(1+\log\left(\sqrt{\frac{T}{V_T}}+1\right)\right)\\ &\quad + 2V_T\left(\sqrt{\frac{T}{V_T}}+1\right)\\ &\overset{(a)}{\leqslant} \left(4+\frac{G^2+\sigma^2}{2H}\left(1+\log\left(\sqrt{\frac{T}{V_T}}+1\right)\right)\right)\cdot\sqrt{V_TT}\\ &\quad + \frac{G^2+\sigma^2}{2H}\left(1+\log\left(\sqrt{\frac{T}{V_T}}+1\right)\right)\\ &\overset{(b)}{\leqslant} \left(4+\frac{2G^2+2\sigma^2}{H}\right)\cdot\log\left(\sqrt{\frac{T}{V_T}}+1\right)\sqrt{V_TT},\end{aligned}$$

where (a) and (b) hold since $1 \leqslant V_T \leqslant T$.

Part 2. We next prove the second part of the Theorem. The proof follows steps and notation appearing in the proof of Theorem 2. For strongly convex cost functions, a different choice of δ is used in Step 2 and $\tilde{\Delta}_T$ is modified accordingly in Step 3. The regret analysis in Step 4 is adjusted as well.

Step 1. Let $\mathcal{X} = [0,1]$, and consider the following two quadratic functions:

$$\begin{aligned}f^1(x) &= x^2 - x + \frac{3}{4},\\ f^2(x) &= x^2 - (1+\delta)x + \frac{3}{4} + \frac{\delta}{2}\end{aligned} \tag{23}$$

for some small $\delta > 0$. Note that $x_1^* = \frac{1}{2}$ and $x_2^* = (1+\delta)/2$. We define a partition of $\mathcal{T}$ into batches $\mathcal{T}_1,\ldots,\mathcal{T}_m$ of size $\tilde{\Delta}_T$ each (perhaps except $\mathcal{T}_m$), according to (8), where $\tilde{\Delta}_T$ will be specified below. Define the class $\mathcal{V}'_s$ according to (19), such that in every $f \in \mathcal{V}'_s$, the cost function is restricted to the set $\{f^1, f^2\}$, and cannot change throughout a batch. The sequences in $\mathcal{V}'_s$ consist of strongly convex functions ((10) holds for any $H \leqslant 1$), with minimizers that are interior points in $\mathcal{X}$. Set $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$. Then,

$$\begin{aligned}\sum_{t=2}^{T}\sup_{x\in\mathcal{X}^*}|f_t(x)-f_{t-1}(x)| &\leqslant \sum_{j=2}^{m}\sup_{x\in\mathcal{X}^*}|f^1(x)-f^2(x)|\\ &\leqslant \frac{T}{\tilde{\Delta}_T}\cdot\frac{\delta^2}{2} = V_T,\end{aligned}$$

where the first inequality holds since the function can change only between batches. Therefore $\mathcal{V}'_s \subset \mathcal{V}_s$.

Step 2. Fix $\pi \in \mathcal{P}_{\phi^{(1)}}$, and let $\tilde{\Delta}_T = \max\{\lfloor 1/\sqrt{2\tilde{C}}\cdot\sqrt{T/V_T}\rfloor, 1\}$ ($\tilde{C}$ appears in part 2 of Assumption 1). Fix $j \in \{1,\ldots,m\}$. Then,

$$\begin{aligned}\mathcal{K}(\mathbb{P}^{\pi,|\mathcal{T}_j|}_{f^1}\|\mathbb{P}^{\pi,|\mathcal{T}_j|}_{f^2}) &\overset{(a)}{\leqslant} \tilde{C}\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathcal{T}_j}(\nabla f^1(X_t)-\nabla f^2(X_t))^2\right]\\ &\leqslant \tilde{C}\tilde{\Delta}_T\delta^2 \overset{(b)}{=} \frac{2\tilde{C}V_T\tilde{\Delta}_T^2}{T}\\ &\overset{(c)}{\leqslant} \max\left\{1,\frac{2\tilde{C}V_T}{T}\right\} \overset{(d)}{\leqslant} \max\{1,2\tilde{C}\},\end{aligned} \tag{24}$$

where (a) follows from Lemma A.1, (b) and (c) hold by the selected values of δ and $\tilde{\Delta}_T$; respectively, and (d) holds by $V_T \leqslant T$. Set $\beta = \max\{1,2\tilde{C}\}$. Then, for any $x_0 \in \mathcal{X}$, using Lemma A.2 with $\varphi_t = \{X_t > x_0\}$, one has

$$\max\{\mathbb{P}_{f^1}\{X_t > x_0\}, \mathbb{P}_{f^2}\{X_t \leqslant x_0\}\} \geqslant \frac{1}{4e^{\beta}}, \quad \forall\, t \in \mathcal{T}. \tag{25}$$

Step 3. Set $x_0 = \frac{1}{2}(x_1^* + x_2^*) = 1/2 + \delta/4$. Let $\tilde{f}$ be a random sequence in which in the beginning of each batch $\mathcal{T}_j$, a cost function is independently drawn according to a discrete uniform distribution over $\{f^1, f^2\}$, and applied throughout the batch. Taking expectation over $\tilde{f}$, one has

$$\begin{aligned}&\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V}'_s,T)\\ &\geqslant \sum_{j=1}^{m}\left(\frac{1}{2}\cdot\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathcal{T}_j}(f^1(X_t)-f^1(x_1^*))\right]\right.\\ &\qquad \left. + \frac{1}{2}\cdot\mathbb{E}^{\pi}_{f^2}\left[\sum_{t\in\mathcal{T}_j}(f^2(X_t)-f^2(x_2^*))\right]\right)\\ &\geqslant \sum_{j=1}^{m}\frac{1}{2}\left(\sum_{t\in\mathcal{T}_j}(f^1(x_0)-f^1(x_1^*))\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\}\right.\\ &\qquad \left. + \sum_{t\in\mathcal{T}_j}(f^2(x_0)-f^2(x_2^*))\mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\}\right)\\ &\geqslant \sum_{j=1}^{m}\frac{\delta^2}{16}\sum_{t\in\mathcal{T}_j}(\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\} + \mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\})\\ &\geqslant \sum_{j=1}^{m}\frac{\delta^2}{16}\sum_{t\in\mathcal{T}_j}\max\{\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\}, \mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\}\}\\ &\overset{(a)}{\geqslant} \sum_{j=1}^{m}\frac{\delta^2}{16}\sum_{t\in\mathcal{T}_j}\frac{1}{4e^{\beta}} = \sum_{j=1}^{m}\frac{\delta^2\tilde{\Delta}_T}{64e^{\beta}} \overset{(b)}{=} \sum_{j=1}^{m}\frac{V_T\tilde{\Delta}_T^2}{32e^{\beta}T} \geqslant \frac{V_T\tilde{\Delta}_T}{32e^{\beta}},\end{aligned}$$

where the first four inequalities follow from arguments given in Step 3 in the proof of Theorem 2, (a) holds by (25), and (b) holds by $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$. Given the selection of $\tilde{\Delta}_T$, one has

$$\begin{aligned}\mathscr{R}^{\pi}_{\phi^{(1)}}(\mathscr{V}'_s, T) &\geqslant \frac{V_T}{32e^{\beta}} \cdot \left\lfloor \frac{1}{\sqrt{2\tilde{C}}} \cdot \sqrt{\frac{T}{V_T}} \right\rfloor \\ &\geqslant \frac{V_T}{32e^{\beta}} \cdot \left(\frac{\sqrt{T} - \sqrt{2\tilde{C}V_T}}{\sqrt{2\tilde{C}V_T}} \right) \\ &\geqslant \frac{1}{64e^{\beta}\sqrt{2\tilde{C}}} \cdot \sqrt{V_T T},\end{aligned}$$

where the last inequality holds if $T \geqslant 8\tilde{C}V_T$. If $T < 8\tilde{C}V_T$, by Proposition 1, there exists a constant C such that $\mathscr{R}^{\pi}_{\phi^{(1)}}(\mathscr{V}_s, T) \geqslant CT \geqslant C\sqrt{V_T T}$. Then, recalling that $\mathscr{V}'_s \subseteq \mathscr{V}_s$, we have established that

$$\mathscr{R}^{\pi}_{\phi^{(1)}}(\mathscr{V}_s, T) \geqslant \mathscr{R}^{\pi}_{\phi^{(1)}}(\mathscr{V}'_s, T) \geqslant \frac{1}{64e^{\beta}\sqrt{2\tilde{C}}} \cdot \sqrt{V_T T}.$$

This concludes the proof. □

Proof of Theorem 5. Part 1. Fix $T \geqslant 1$ and $1 \leqslant V_T \leqslant T$. For any $\Delta_T \in \{1, \ldots, T\}$, consider the EGS algorithm $\mathscr{A}$ given in §5.2 with $a_t = 2/Ht$ and $\delta_t = h_t = a_t^{1/4}$ for $t = 1, \ldots, \Delta_T$, and let π be the policy defined by the restarting procedure with subroutine $\mathscr{A}$ and batch size Δ_T. By Lemma 1 (see Appendix C), we have

$$\mathscr{G}^{\mathscr{A}}_{\phi}(\mathscr{F}_s, \Delta_T) \leqslant C_1 \cdot \sqrt{\Delta_T} \tag{26}$$

with $C_1 = 2G + (G^2 + \sigma^2 + H)d^{3/2}/\sqrt{2H}$. Therefore, by Proposition 2,

$$\begin{aligned}\mathscr{R}^{\pi}_{\phi^{(0)}}(\mathscr{V}_s, T) &\leqslant \left(\frac{T}{\Delta_T} + 1\right) \cdot \mathscr{G}^{\mathscr{A}}_{\phi^{(0)}}(\mathscr{F}, \Delta_T) + 2V_T\Delta_T \\ &\overset{(a)}{\leqslant} C_1 \cdot \frac{T}{\sqrt{\Delta_T}} + C_1 \cdot \sqrt{\Delta_T} + 2V_T\Delta_T,\end{aligned}$$

where (a) holds by (26). By selecting $\Delta_T = \lceil (T/V_T)^{2/3} \rceil$, one obtains

$$\begin{aligned}\mathscr{R}^{\pi}_{\phi^{(0)}}(\mathscr{V}_s, T) &\leqslant C_1 \cdot \frac{T}{(T/V_T)^{1/3}} \\ &\quad + C_1 \cdot \left(\left(\frac{T}{V_T}\right)^{1/3} + 1 \right) \\ &\quad + 2V_T \left(\left(\frac{T}{V_T}\right)^{2/3} + 1 \right) \\ &\overset{(b)}{\leqslant} (C_1 + 4)V_T^{1/3}T^{2/3} + C_1 \cdot \left(\frac{T}{V_T}\right)^{1/3} + C_1 \\ &\overset{(c)}{\leqslant} (3C_1 + 4)V_T^{1/3}T^{2/3},\end{aligned}$$

where (b) and (c) hold since $1 \leqslant V_T \leqslant T$.

Part 2. The proof of this part of the theorem follows the steps and uses notation introduced in the proof of Theorem 2. The different feedback structure affects the bound on the KL divergence, and the selected value of $\tilde{\Delta}_T$ in Step 2, as well as the resulting regret analysis in Step 3. Details are given below.

Step 1. We define a class $\mathscr{V}'_s$ as it is defined in the proof of Theorem 4, using the quadratic functions f^1 and f^2 that are given in (23), and the partition of $\mathscr{T}$ to batches in (8). Again, selecting $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$, we have $\mathscr{V}'_s \subset \mathscr{V}_s$.

Step 2. Fix some policy $\pi \in \mathscr{P}_{\phi^{(0)}}$. At each $t \in \mathscr{T}_j$, $j = 1, \ldots, m$, the decision maker selects $X_t \in \mathscr{X}$ and observes a noisy feedback $\phi^{(0)}_t(X_t, f^k)$. For any $f \in \mathscr{F}$, $\tau \geqslant 1$, $A \subset \mathbb{R}^{\tau}$ and $B \subset \mathscr{U}$, denote $\mathbb{P}^{\pi,\tau}_f(A, B) := \mathbb{P}_f\{\{\phi^{(0)}_t(X_t, f_t)\}_{t=1}^{\tau} \in A, U \in B\}$. In this part of the proof, we use the following counterpart of Lemma A.1 for the case of noisy cost feedback structure.

Lemma A.3. (Bound on KL Divergence for Noisy Cost Observations). *Consider the feedback structure $\phi = \phi^{(0)}$ and let Assumption 2 hold. Then, for any $\tau \geqslant 1$ and f, $g \in \mathscr{F}$,*

$$\mathscr{K}(\mathbb{P}^{\pi,\tau}_f \,\|\, \mathbb{P}^{\pi,\tau}_g) \leqslant \tilde{C}\,\mathbb{E}^{\pi}_f\left[\sum_{t=1}^{\tau}(f_t(X_t) - g_t(X_t))^2\right],$$

where $\tilde{C}$ is the constant that appears in the second part of Assumption 2.

The proof of the lemma appears in Appendix B. We next bound $\mathscr{K}(\mathbb{P}^{\pi,|\mathscr{T}_j|}_{f^1} \,\|\, \mathbb{P}^{\pi,|\mathscr{T}_j|}_{f^2})$ throughout an arbitrary batch $\mathscr{T}_j$, $j \in \{1, \ldots, m\}$, for a given batch size $\tilde{\Delta}_T$. Define

$$\begin{aligned}R^{\pi}_j &= \frac{1}{2}\,\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathscr{T}_j}(f^1(X_t) - f^1(x^*_1))\right] \\ &\quad + \frac{1}{2}\,\mathbb{E}^{\pi}_{f^2}\left[\sum_{t\in\mathscr{T}_j}(f^2(X_t) - f^2(x^*_2))\right].\end{aligned}$$

Then, one has

$$\begin{aligned}\mathscr{K}(\mathbb{P}^{\pi,|\mathscr{T}_j|}_{f^1} \,\|\, \mathbb{P}^{\pi,|\mathscr{T}_j|}_{f^2}) &\overset{(a)}{\leqslant} \tilde{C}\,\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathscr{T}_j}(f^1(X_t) - f^2(X_t))^2\right] \\ &= \tilde{C}\,\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathscr{T}_j}\left(\delta X_t - \frac{\delta}{2}\right)^2\right] \\ &= \tilde{C}\,\mathbb{E}^{\pi}_{f^1}\left[\delta^2\sum_{t\in\mathscr{T}_j}(X_t - x^*_1)^2\right] \\ &\overset{(b)}{=} 2\tilde{C}\delta^2\,\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathscr{T}_j}(f^1(X_t) - f^1(x^*_1))\right] \\ &\overset{(c)}{\leqslant} \frac{8\tilde{C}\tilde{\Delta}_T V_T}{T} \cdot R^{\pi}_j,\end{aligned} \tag{27}$$

where (a) follows from Lemma A.3, (b) holds since

$$\begin{aligned}f^1(x) - f^1(x^*_1) &= \nabla f^1(x^*_1)(x - x^*_1) + \frac{1}{2} \cdot \nabla f^1(x^*_1)(x - x^*_1)^2 \\ &= \frac{1}{2}(x - x^*_1)^2\end{aligned}$$

for any $x \in \mathscr{X}$, and (c) holds since $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$, and $R^{\pi}_j \geqslant \frac{1}{2}\,\mathbb{E}^{\pi}_{f^1}[\sum_{t\in\mathscr{T}_j}(f^1(X_t) - f^1(x^*_1))]$. Thus, for any $x_0 \in \mathscr{X}$, using Lemma A.2 with $\varphi_t = \{X_t > x_0\}$, we have

$$\begin{aligned}&\max\{\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\}, \mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\}\} \\ &\geqslant \frac{1}{4}\exp\left\{-\frac{8\tilde{C}\tilde{\Delta}_T V_T}{T} \cdot R^{\pi}_j\right\} \quad \text{for all } t \in \mathscr{T}_j,\ 1 \leqslant j \leqslant m.\end{aligned} \tag{28}$$

Step 3. Set $x_0 = \frac{1}{2}(x_1^* + x_2^*) = 1/2 + \delta/4$. Let $\tilde{f}$ be the random sequence of functions that is described in Step 3 in the proof of Theorem 4. Taking expectation over $\tilde{f}$, one has

$$\begin{aligned}\mathscr{R}^{\pi}_{\phi^{(0)}}(\mathscr{V}'_s, T) &\geqslant \sum_{j=1}^{m}\left(\frac{1}{2}\cdot\mathbb{E}^{\pi}_{f^1}\left[\sum_{t\in\mathscr{T}_j}(f^1(X_t)-f^1(x_1^*))\right]\right.\\ &\qquad\left.+\frac{1}{2}\cdot\mathbb{E}^{\pi}_{f^2}\left[\sum_{t\in\mathscr{T}_j}(f^2(X_t)-f^2(x_2^*))\right]\right)\\ &=: \sum_{j=1}^{m} R_j^{\pi}.\end{aligned}$$

In addition, for each $1 \leqslant j \leqslant m$, one has

$$\begin{aligned}R_j^{\pi} &\geqslant \frac{1}{2}\left(\sum_{t\in\mathscr{T}_j}(f^1(x_0)-f^1(x_1^*))\right)\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\}\\ &\qquad+\sum_{t\in\mathscr{T}_j}(f^2(x_0)-f^2(x_2^*))\,\mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\}\Big)\\ &\geqslant \frac{\delta^2}{16}\sum_{t\in\mathscr{T}_j}(\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\}+\mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\})\\ &\geqslant \frac{\delta^2}{16}\sum_{t\in\mathscr{T}_j}\max\{\mathbb{P}^{\pi}_{f^1}\{X_t > x_0\},\,\mathbb{P}^{\pi}_{f^2}\{X_t \leqslant x_0\}\}\\ &\overset{(a)}{\geqslant} \frac{\delta^2}{16}\sum_{t\in\mathscr{T}_j}\frac{1}{4}\exp\left\{-\frac{8\tilde{C}\tilde{\Delta}_T V_T}{T}\cdot R_j^{\pi}\right\}\\ &= \frac{\delta^2\tilde{\Delta}_T}{64}\exp\left\{-\frac{8\tilde{C}\tilde{\Delta}_T V_T}{T}\cdot R_j^{\pi}\right\}\\ &\overset{(b)}{=} \frac{\tilde{\Delta}_T^2 V_T}{32T}\exp\left\{-\frac{8\tilde{C}\tilde{\Delta}_T V_T}{T}\cdot R_j^{\pi}\right\},\end{aligned}$$

where the first three inequalities follow arguments given in Step 3 in the proof of Theorem 3, (a) holds by (28), and (b) holds by $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$. Assume that $\sqrt{\tilde{C}}\cdot V_T \leqslant 2T$. Then, taking $\tilde{\Delta}_T = \lceil (4/\tilde{C})^{1/3}(T/V_T)^{2/3}\rceil$, one has:

$$\begin{aligned}R_j^{\pi} &\geqslant \frac{1}{32}\cdot\left(\frac{4}{\tilde{C}}\right)^{2/3}\left(\frac{T}{V_T}\right)^{1/3}\\ &\qquad\cdot\exp\left\{-\frac{8\tilde{C}V_T}{T}\cdot\left(\left(\frac{4}{\tilde{C}}\right)^{1/3}\left(\frac{T}{V_T}\right)^{2/3}+1\right)\cdot R_j^{\pi}\right\}\\ &\geqslant \frac{1}{32}\cdot\left(\frac{4}{\tilde{C}}\right)^{2/3}\left(\frac{T}{V_T}\right)^{1/3}\exp\left\{-16\tilde{C}^{2/3}\cdot 4^{1/3}\cdot\left(\frac{V_T}{T}\right)^{1/3}R_j^{\pi}\right\},\end{aligned}$$

where the last inequality follows from $\sqrt{\tilde{C}}\cdot V_T \leqslant 2T$. Then, for $\beta = 16(4\tilde{C}^2\cdot V_T/T)^{1/3}$, one has

$$\beta R_j^{\pi} \geqslant \frac{32T}{\tilde{\Delta}_T^2 V_T} \geqslant \exp\{-\beta R_j^{\pi}\}. \tag{29}$$

Let y_0 be the unique solution to the equation $y = \exp\{-y\}$. Then, (29) implies $\beta R_j^{\pi} \geqslant y_0$. In particular, since $y_0 > 1/2$ this implies $R_j^{\pi} \geqslant 1/(2\beta) = 1/(32(2\tilde{C})^{2/3})(T/V_T)^{1/3}$ for all $1 \leqslant j \leqslant m$. Hence

$$\begin{aligned}\mathscr{R}^{\pi}_{\phi^{(0)}}(\mathscr{V}_s, T) &\geqslant \sum_{j=1}^{m} R_j^{\pi} \geqslant \frac{T}{\tilde{\Delta}_T}\cdot\frac{1}{32(2\tilde{C})^{2/3}}\left(\frac{T}{V_T}\right)^{1/3}\\ &\overset{(a)}{\geqslant} \frac{1}{64\cdot 2^{4/3}\tilde{C}^{1/3}}\cdot V_T^{1/3}T^{2/3},\end{aligned}$$

where (a) holds if $\sqrt{\tilde{C}}\cdot V_T \leqslant 2T$. If $\sqrt{\tilde{C}}\cdot V_T > 2T$, by Proposition 1 there is a constant C such that $\mathscr{R}^{\pi}_{\phi^{(0)}}(\mathscr{V}_s, T) \geqslant CT \geqslant CV_T^{1/3}T^{2/3}$; the last inequality holds by $T \geqslant V_T$. This concludes the proof. □

Endnotes

1. A more precise definition of an admissible policy will be advanced in the next section, but roughly speaking, we restrict attention to policies that are nonanticipating and adapted to past actions and observed feedback signals, allowing for auxiliary randomization; hence the expectation above is taken with respect to any randomness in the feedback, as well as in the policy's actions.
2. For the sake of completeness, to establish the connection between the adversarial and the stochastic literature streams, we adapt, where needed, results in the former setting to the case of noisy feedback.
3. In particular, for the sake of simplicity and concreteness, we use the above notation, our analysis applies to the case of sequences in which in every step, only the next cost function is selected, in a fully adversarial manner that takes into account the realized trajectory of the policy and is subjected only to the bounded variation constraint.
4. OCO settings typically allow sequences of cost functions that can adjust adversarially at each epoch. For the sake of consistency with the definition of (5), in the above regret measure, nature commits to a sequence of functions in advance.
5. Considering the special case of linear cost functions, there are known policies that guarantee regret of order $\sqrt{T}$ relative to the single best action in adversarial settings (see, e.g., Kalai and Vempala 2003 for full access to the function, and McMahan and Blum 2004 for point feedback). Using an adaptation of such policy as a subroutine of the restarting procedure would guarantee regret of order $V_T^{1/3}T^{2/3}$ relative to the dynamic oracle in our setting; a matching lower bound may be obtained by a rather straightforward adaptation of the proof of Theorem 2.
6. In fact, Hazan and Kale (2011) show that even in a stationary stochastic setting with strongly convex cost function and a class of unbiased gradient access, any policy must incur regret of at least order $\log T$ compared to a static benchmark.
7. For any t such that $\nu < \delta_t$, one may use the numbers $h'_t = \delta'_t = \min\{\nu, \delta_t\}$ instead, with the rate optimality obtained in Lemma 1 remaining unchanged.

References

Agarwal A, Dekel O, Xiao L (2010) Optimal algorithms for online convex optimization with multi-point bandit feedback. *Proc. 23rd Ann. Conf. Learn. Theory, COLT '10* (Omnipress, Madison, WI), 28–40.

Agarwal A, Foster DP, Hsu D, Kakade SM, Rakhlin A (2013) Stochastic convex optimization with bandit feedback. *SIAM J. Optim.* 23:213–240.

Araman VF, Caldentey R (2011) Revenue management with incomplete demand information. Cochran JJ, Cox, Jr. LA, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ), 1–17.

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Math. Oper. Res.* 23(4):769–805.

Benveniste A, Priouret P, Metivier M (1990) *Adaptive Algorithms and Stochastic Approximations* (Springer, New York).

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53:464–501.

Besbes O, Muharremoglu A (2013) On implications of demand censoring in the newsvendor problem. *Management Sci.* 59(6):1407–1424.

Besbes O, Zeevi A (2011) On the minimax complexity of pricing in a changing environment. *Oper. Res.* 59(1):66–79.

Blackwell D (1956) An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6:1–8.

Broder J, Rusmevichientong P (2012) Dynamic pricing under a general parametric choice model. *Oper. Res.* 60(4):965–980.

Cesa-Bianchi N, Lugosi G (2006) *Prediction, Learning, and Games* (Cambridge University Press, Cambridge, UK).

Cope EW (2009) Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. Automatic Control* 54: 1243–1253.

den Boer A, Zwart B (2014) Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.* 60(3):770–783.

Flaxman AD, Kalai AT, McMahan HB (2005) Online convex optimization in the bandit setting: Gradient descent without gradient. *Proc. Sixteenth Ann. ACM-SIAM Sympos. Discrete Algorithms* (SIAM, Philadelphia), 385–394.

Hannan J (1957) Approximation to Bayes risk in repeated play. *Contubotions to the Theory of Games*, Vol. 3 (Princeton University Press, Princeton, NJ), 97–139.

Haykin SS (2001) *Kalman Filtering and Neural Networks* (John Wiley & Sons, New York).

Hazan E, Kale S (2010) Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine Learn.* 80:165–188.

Hazan E, Kale S (2011) Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. *J. Machine Learn. Res. Proceedings Track* 19:421–436.

Hazan E, Agarwal A, Kale S (2007) Logarithmic regret algorithms for online convex optimization. *Machine Learn.* 69:169–192.

Huh WT, Rusmevichientong P (2009) A non-parametric asymptotic analysis of inventory planning with censored demand. *Math. Oper. Res.* 34(1):103–123.

Kalai A, Vempala S (2003) Efficient algorithms for online decision problems. *Learning Theory and Kernel Machines* (Springer, Berlin), 26–40.

Kalman RE (1960) A new approach to linear filtering and prediction problems. *J. Fluids Engrg.* 81:35–45.

Keller G, Rady S (1999) Optimal experimentation in a changing environment. *Rev. Econom. Stud.* 66:475–507.

Keskin BN, Zeevi A (2014) Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.* 62(5):1142–1167.

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* 23:462–466.

Kleinberg RD (2004) Nearly tight bounds for the continuum-armed bandit problem. Saul L, Weiss Y, Bottou L, eds. *Adv. Neural Inform. Processing Systems 17* (MIT Press, Cambridge, MA), 697–704.

Kushner HJ, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications* (Springer, New York).

Lai TL (2003) Stochastic approximation. *Ann. Statist.* 31:391–406.

McMahan HB, Blum A (2004) Online geometric optimization in the bandit setting against an adaptive adversary. *Learning Theory* (Springer, Berlin), 109–123.

Nemirovski A, Yudin D (1983) *Problem Complexity and Method Efficiency in Optimization* (John Wiley & Sons, New York).

Robbins H, Monro S (1951) A stochastic approximatiom method. *Ann. Math. Statist.* 22:400–407.

Tsybakov AB (2008) *Introduction to Nonparametric Estimation* (Springer, New York).

Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *20th Internat. Conf. Machine Learn.* (AAAI, Palo Alto, CA), 928–936.

Omar Besbes is the Philip H. Geier, Jr. Associate Professor of Business at the Graduate School of Business, Columbia University. His research focuses on data-driven decision making and its applications in revenue management, operations, and service systems.

Yonatan Gur is an assistant professor of operations, information, and technology at the Graduate School of Business, Stanford University. His research interests include operations management, revenue management, and service systems, with a particular emphasis on data-driven, dynamic, online environments.

Assaf Zeevi is the Kravis Professor of Business at the Graduate School of Business, Columbia University. His research is broadly focused on the formulation and analysis of mathematical models of complex systems. His teaching centers on the areas of statistics and stochastic modeling. Assaf serves on several scientific advisory boards.