Sequential Optimization in Changing Environments: Theory and Application to Online Content Recommendation Services

Yonatan Gur

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Sequential Optimization in Changing Environments: Theory and Application to Online Content Recommendation Services

Yonatan Gur

Recent technological developments allow the online collection of valuable information that can be efficiently used to optimize decisions "on the fly" and at a low cost. These advances have greatly influenced the decision-making process in various areas of operations management, including pricing, inventory, and retail management. In this thesis we study methodological as well as practical aspects arising in online sequential optimization in the presence of such real-time information streams. On the methodological front, we study aspects of sequential optimization in the presence of temporal changes, such as designing decision making policies that adopt to temporal changes in the underlying environment (that drives performance) when only partial information about this changing environment is available, and quantifying the added complexity in sequential decision making problems when temporal changes are introduced. On the applied front, we study practical aspects associated with a class of online services that focus on creating customized recommendations (e.g., Amazon, Netflix). In particular, we focus on *online content recommendations*, a new class of online services that allows publishers to direct readers from articles they are currently reading to other web-based content they may be interested in, by means of links attached to said article.

In the first part of the thesis we consider a non-stationary variant of a sequential stochastic optimization problem, where 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. As a yardstick to quantify performance in non-stationary settings we propose a regret measure relative to a *dynamic oracle* benchmark.

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: adversarial online convex optimization; and 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.

In the second part of the thesis we consider another core stochastic optimization problem couched in a multi-armed bandit (MAB) setting. We develop a MAB formulation that allows for a broad range of temporal uncertainties in the rewards, characterize the (regret) complexity of this class of MAB problems by establishing a direct link between the extent of allowable reward "variation" and the minimal achievable worst-case regret, and provide an optimal policy that achieves that performance. Similarly to the first part of the thesis, our analysis draws concrete connections between two strands of literature: the adversarial and the stochastic MAB frameworks.

The third part of the thesis studies applied optimization aspects arising in *online content recommendations*, that allow web-based publishers to direct readers from articles they are currently reading to other web-based content. We study the content recommendation problem and its unique dynamic features from both theoretical as well as practical perspectives. Using a large data set of browsing history at major media sites, we develop a representation of content along two key dimensions: *clickability*, the likelihood to *click to* an article when it is recommended; and *engageability*, the likelihood to *click from* an article when it hosts a recommendation. Based on this representation, we propose a class of user path-focused heuristics, whose purpose is to simultaneously ensure a high instantaneous probability of clicking recommended articles, while also optimizing engagement along the future path. We rigorously quantify the performance of these heuristics and validate their impact through a live experiment. The third part of the thesis is based on a collaboration with a leading provider of content recommendations to online publishers.

Table of Contents

Acknowledgments

I wish to wholeheartedly thank:

Assaf and Omar, for teaching, guiding, believing, and leading by example;

for showing me:
how to choose my battles (and how to avoid unchosen ones),
how to ask the right questions (and how to answer the wrong ones),
and how to tell the hot from the not;

and for leaving the empty space I grew into.

Nuphar, for being my outer-self, my alter-ego, my out-of-body experience;

and for willing to be an academician's wife - I really hope it goes well for you.

Chapter 1

Introduction

1.1 Sequential optimization in changing environments

In the presence of uncertainty and partial feedback, an agent that faces a sequence of decisions needs to judiciously use information collected from past observations when trying to optimize future actions. This fundamental paradigm is present in a variety of applications in dynamic pricing, inventory control, retail management, and assortment selection: an online retailer that launches a new product needs to set a price to maximize profits but does not know the demand curve; that retailer may also need to select an assortments of products to suggest an arriving customer, but does not know the preferences of that customer over available products; other web-based companies may try to suggest articles, music, or videos to individual consumers whose tastes are a-priori not known; as well as many other instances. In all the above examples decisions can be adjusted on a weekly, daily or hourly basis (if not more frequently), and the history of observations may be used to optimize current and future performance. Two widely studied paradigms that capture sequential decision-making in the presence of uncertainty and partial feedback are the Stochastic approximation (SA) formulation that is typically applied when the available action set is continuous (such as in dynamic pricing problems), and the Multi-armed bandit (MAB) formulation, typically applied when that action set is discrete (such as in assortment selection). While in many application domains (such as the ones noted above) temporal structural changes may be an intrinsic characteristic of the problem, these potential changes are largely not dealt with in the traditional SA and (stochastic) MAB literature streams.

Stochastic approximation. In the prototypical setting of sequential stochastic optimization, a decision maker selects at each epoch $t \in \{1, \dots T\}$ a point X_t that belongs (typically) to some convex compact action set $\mathcal{X} \subset \mathbb{R}^d$, 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 the gradient of the cost. When the cost function is assumed to be strongly convex, a typical objective is to minimize the mean-squared-error, $\mathbb{E}\|X_T - x^*\|^2$, where x^* denotes the minimizer of $f(\cdot)$ in $\mathcal{X}$. 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 both 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, Operation 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, how to model temporal changes in a manner that is "rich" enough to capture a broad set of scenarios while still being mathematically tractable, and what is the performance that can be achieved in such settings in comparison to the stationary SA environment. Chapter 2 of this thesis 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 $\{f_t : t = 1, \dots, T\}$; like the stationary setting, in every epoch

$t = 1, \ldots, T$ the decision maker selects a point $X_t \in \mathcal{X}$ (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 chapter 2 we consider two canonical feedback structures alluded to earlier, namely, noisy access to the function value $f(X_t)$, and noisy access to the gradient $\nabla f(X_t)$. Let $\{x_t^* : t = 1, \ldots, T\}$ 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, $\sum_{t=1}^{T} \mathbb{E}\left[f_t(X_t) - f_t(x_t^*)\right]$ 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 that 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, in chapter 2 we introduce the concept of a *temporal uncertainty set* $\mathcal{V}$, which is driven by a *variation budget* V_T:

$$\mathcal{V} := \{\{f_1, \ldots, f_T\} \ : \ \mathrm{Var}(f_1, \ldots, f_T) \leq V_T\}.$$

The precise definition of the variation functional $\mathrm{Var}(\cdot)$ will be given in chapter 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 chapter 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 dynamic oracle:

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

[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 non-anticipating 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.

In this set up, a policy π is chosen and then nature (playing the role of the adversary) selects the sequence of functions $f := \{f_t\}_{t=1,\dots,T} \in \mathcal{V}$ 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,

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

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 $\mathcal{R}_{\phi}^{*}(\mathcal{V}, T)$ 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 $\mathcal{V}$. 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.

A discrete action set. A widely studied paradigm that captures the tension between the acquisition cost of new information (*exploration*) that may be used to improve future decisions and rewards, and the generation of instantaneous rewards based on the existing information (*exploitation*) is that of multi armed bandits (MAB), originally proposed in the context of drug testing by Thompson (1933), and placed in a general setting by Robbins (1952). The original setting has a gambler choosing among K slot machines at each round of play, and upon that selection observing a reward realization. In this classical formulation the rewards are assumed to be independent and identically distributed according to an unknown distribution that characterizes each machine. The objective is to maximize the expected sum of (possibly discounted) rewards received over a given (possibly infinite) time horizon.

Since the set of MAB instances in which one can identify the optimal policy is extremely limited, a typical yardstick to measure performance of a candidate policy is to compare it to a benchmark: an *oracle* that at each time instant selects the arm that maximizes expected reward. The difference between the performance of the policy and that of the oracle is called the *regret*. When the growth of the regret as a function of the horizon T is *sub-linear*, the policy is *long-run*

average optimal: its long run average performance converges to that of the oracle. Hence the first order objective is to develop policies with this characteristic. The precise rate of growth of the regret as a function of T provides a refined measure of policy performance. Lai and Robbins (1985) is the first paper that provides a sharp characterization of the regret growth rate in the context of the traditional setting (stationary random rewards) that is often referred to as the *stochastic* MAB problem. Most of the literature has followed this path with the objective of designing policies (often referred to as *rate optimal* policies) that exhibit the "slowest possible" rate of growth in the regret.

In chapter 3, following the meta-principle introduced in chapter 2, we show that in a broad class of stochastic non-stationary MAB problems one may achieve rate optimal performance by adapting policies from the adversarial MAB setting. Interestingly, we show that one may obtain a rate optimal performance with respect to all three parameters that characterize non-stationary stochastic MAB settings: not only the horizon length and the variation in the rewards, but also the *number of available arms*.

1.2 Online content recommendation services

Diversity and sheer number of content sites on the world wide web has been increasing at an extraordinary rate over the past years. One of the great technological challenges, and a major achievement of search portals, is the ability to successfully navigate users through this complex forest of information to their desired content. However, search is just one route for users to seek content, and one that is mostly relevant when users have a fairly good idea of what they are searching for. Recent years have witnessed the emergence of dynamically customized *content recommendations*, a new class of online services that complement search and allows publishers to direct users from articles they are currently reading to other web-based content they may be interested in consuming. Chapter 4 focuses on performance assessment and optimization for this new class of services.

Brief overview of the service. When a reader arrives to an online article (for example, from the publisher's front page), a customized recommendation is generated at the bottom of the article (Figure 1.1 depicts such an example). The recommendation typically contains 3 to 12 links

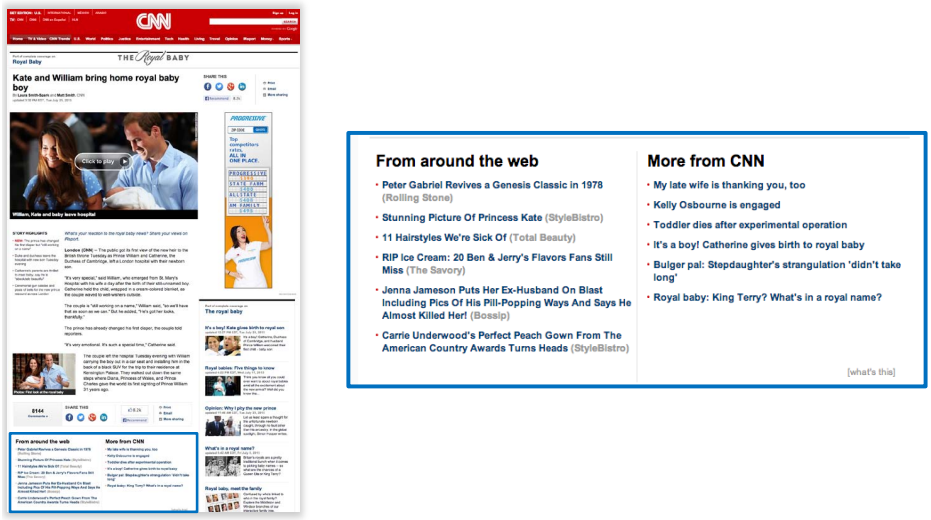


Figure 1.1: **Online content recommendation.** (*Left*) The position of the recommendation, at the bottom of a CNN article. (*Right*) The enlarged recommendation, containing links to recommended articles. The right side of this recommendation contains internal links to other articles on CNN's website, or CNN owned blogs. The left side of the recommendations contains external (sponsored) links to articles from other media sites.

that point readers to recommended articles. The links specify the title of the recommended article. By clicking on one of these links the reader is sent to the recommended article, at the bottom of which a new recommendation is provided, etc. These recommendations are typically generated by a service provider (not the media site), and recommended articles may be internal (organic), leading readers to other articles published in the host media site, or external (sponsored), in general leading readers to other publishers. While internal recommendations are typically given as a service to the host publisher, external links are sponsored (either by the site on the receiving end of the recommendation, or by a third party that promotes the content) based on a fee-per-click, which is split between the service provider and the publisher that hosts the recommendation. This simple revenue share business model is predicated on the service's success in matching users with customized content that is relevant for them at the time of their visit. The dynamic matching problem between users and content lies at the heart of both the service provider's and online publishers' revenue maximization problems, and determines the value of the service to the publishers and their readers.

At a high level, the process of matching a reader with a bundle of recommended articles takes

the following form. When a reader arrives to an article, a request for a recommendation is sent by the host publisher. This request may include some information regarding the host article as well as the reader. The service provider also has access to a database of feasible articles with information such as topic classification, publish date, and click history. The available information is processed by several competing and complementary algorithms that analyze different aspects of it: the contextual connection between the host article and the recommendation candidates; the reading behavior and patterns associated with articles; and additional information such as the popularity of articles (and general traffic trends in the content network). These inputs are combined to generate a content recommendation.

Salient features. While the problem of recommending articles to readers shares similar features with the ones faced by more traditional product recommendation services (such as Amazon or Netflix), it has several unique characteristics. Such features include the rate at which new "products" (articles) are added to the system (roughly 1M daily), the typical short shelf life of many articles (in many cases these may lose relevancy in a matter of hours/days after publication), as well as rapid fluctuations of interest levels associated with different topics, driven by the evolving trends and buzz in the content world. These salient features introduce challenges that go beyond the traditional product recommendation problem (e.g., the need to base recommendations on dynamic and relatively limited information).

A key feature defining the content recommendation service, is that it stimulates *ongoing user engagement* in each interaction. While many online services are terminated after a single click, the content recommendation service is dynamic, as each successful recommendation leads to a new opportunity for interaction: following the first click, the user arrives to a new article, at the bottom of which a new recommendation is generated, and so on. Thus, content recommendations often serve as a navigation tool for readers, inducing a chain of discovered articles. In such an environment, a central question is how to measure (and optimize) the performance of the recommendation service?

The key performance indicator that is currently used to evaluate various articles as candidates for recommendation is the click through rate (CTR): the number of times a link to an article was clicked, divided by the number of times the link was shown. The CTR performance indicator

is adopted by many online services that have the objective of generating a *single* click per user. Under such a myopic approach, optimization techniques typically focus on integrating the probability to click on a recommendation with the potential revenue generated by a click (see Jansen and Mullen (2008) and Feng et al. (2007) for an overview). Following this common approach, content recommendation algorithms used in current practice are designed to maximize the instantaneous CTR (or alternatively, the instantaneous revenue) of the generated recommendation. While high CTR signals that the service is frequently being used and that revenue is generated by the service provider, CTR also has an important limitation: it measures the probability to click at the current step, but does not account for interactions that may come after the click, and in particular, *future* clicks along the potential visit path of the reader.

1.3 Overview of main contributions

1.3.1 Non-stationary stochastic optimization

The main results and key qualitative insights of Chapter 2 can be summarized as follows.

Necessary and sufficient conditions for sublinear regret. We first show that if the variation budget V_T is *linear* in T, then sublinear regret *cannot* be achieved by any admissible policy, and conversely, 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.

Complexity characterization. We prove a sequence of results that characterizes the order of the minimax regret for both 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:

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

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

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. The table below summarizes our main findings. Note that even in

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	$\sqrt{T}$	$V_T^{1/3}T^{2/3}$

Table 1.1: **The price of non-stationarity.** 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.

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 chapter 2.)

A meta principle for constructing optimal policies. One of the key insights we wish to communicate in chapter 2 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 *adversarial* and the *stochastic* frameworks. The former, which in our context is often refereed to as online convex optimization (OCO), allows nature to choose 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. While 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 stochastic non-stationary 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 "meta principle":

> If a policy has "good" performance relative 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 adversarial 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 while policies that admit these properties have, by and large, been identified in the online convex optimization literature[2], to the best of our knowledge there are no counterparts to date in a non-stationary stochastic setting.

1.3.2 Multi-armed bandit problems with non-stationary rewards

At a high level, the main contribution of chapter 3 lies in fully characterizing the (regret) complexity of a broad class of MAB problems with non-stationary reward structure by establishing a direct link between the extent of reward "variation" and the minimal achievable worst-case regret. More specifically, the contributions of chapter 3 are along three dimensions.

Modeling a broad class of MAB problems with non-stationary rewards. We formulate a class of non-stationary reward structures that is quite general, and hence can be used to realistically capture a variety of real-world type phenomena, yet remain mathematically tractable. The main constraint that we impose on the evolution of the mean rewards is that their variation over the relevant time horizon is bounded by a variation budget V_T. This limits the power of nature compared to the adversarial setup discussed above where rewards can be picked to maximally damage the policy at each instance within $\{1, \ldots, T\}$. Nevertheless, this constraint still allows for a very rich class of temporal changes. This class extends most of the treatment in the

[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.

non-stationary stochastic MAB literature which mainly focuses on a finite (known) number of changes in the mean reward values, see, e.g., Garivier and Moulines (2011) and references therein (see also Auer, Cesa-Bianchi, Freund and Schapire (2002) in the adversarial context), and is consistent with more extreme settings, such as the one treated in Slivkins and Upfal (2008) where reward distributions evolve according to a Brownian motion and hence the regret is linear in T. (We will explain these connections in more detail in chapter 3.)

Characterizing complexity and designing a near-optimal policy. For the class of non-stationary reward distributions described above, we establish lower bounds on the performance of *any* non-anticipating policy relative to the *dynamic* oracle, and show that these bounds can be achieved, uniformly over the class of admissible reward distributions, by a suitable policy construction. The term "achieved" is meant in the sense of the order of the regret as a function of the time horizon T, the variation budget V_T, and the number of arms K. Thus, up to a logarithmic scale of the number of arms our policies are shown to be minimax optimal. The regret is sublinear and is of the order of $(KV_T)^{1/3}\,T^{2/3}$. Auer et al. (2002), in the adversarial setting, and Garivier and Moulines (2011) in the stochastic setting, considered non-stationary rewards where the identity of the best arm can change a *finite* number of times; the regret in these instances (relative to a dynamic oracle) is shown to be of order $\sqrt{T}$. Our analysis complements these results by treating a broader and more flexible class of temporal changes in the reward distributions, yet still establishing optimality results and showing that sublinear regret is achievable. When V_T increases with the time horizon T, our results provide a spectrum of minimax regret performance between order $T^{2/3}$ (when V_T is a constant independent of T) and order T (when V_T grows linearly with T), and by that, map the allowed variation to the best achievable performance.

Identifying and optimizing salient tradeoffs. With the analysis described above we shed light on the exploration-exploitation trade off that is a characteristic of the non-stationary reward setting, and the change in this trade off compared to the stationary setting. In particular, our results highlight the tension that exists between the need to "remember" and "forget." This is characteristic of several algorithms that have been developed in the adversarial MAB literature, e.g., the family of exponential weight methods such as EXP3, EXP3.S and the like; see, e.g., Auer, Cesa-Bianchi, Freund and Schapire (2002), and Cesa-Bianchi and Lugosi (2006). In a nutshell, the fewer past observations one retains, the larger the stochastic error associated with one's estimates

of the mean rewards, while at the same time the more past observations are used, the higher the risk of these being biased. One interesting observation, that is formalized as one of our main theorems, is that an optimal policy in the sense of performance relative to a static oracle in the adversarial setting can be used to construct a policy that achieves optimal performance relative to the more ambitious dynamic oracle that we employ in our setting. We leverage this to show that the EXP3 type algorithms can be properly customized to our stochastic non-stationary MAB setting and yield rate optimal performance.

1.3.3 Optimization in online content recommendation services

Chapter 4 studies theoretical properties and practical real-time optimization of the content recommendation service. We develop a predictive analytics model of clicks that enables to identify click drivers along the path of readers, which in turn gives rise to concrete and implementable insights that lead to recommendations that account for the future path of readers. Furthermore, we conduct a controlled experiment to validate the value of the proposed prescription. In more detail, the contribution of chapter 4 can be described along the following four components.

Diagnostic. We formulate the optimal content recommendation problem and show that it is NP-hard. We then formalize the myopic heuristic that is used in practice, and whose objective is to maximize CTR, namely, maximizing the probability to click on the current recommendation. We establish that the gap between the performance of optimal recommendations and that of the myopic heuristic may be arbitrarily large. In that sense, theoretically, myopic recommendations may have poor performance. Analyzing the data, we provide empirical evidence that indeed there might be significant room for improvement over the myopic heuristic of maximizing CTR.

Introducing and validating the notion of engageability. We analyze the click behavior of users by introducing and estimating a choice model. In particular, we model the characteristics of the articles and those of the displayed recommendation box that impact the "content path" of a reader within the recommendation network. We calibrate this model based on a large data set, in a manner that accounts for the evolution of articles' relevancy over time. Based on our model, we develop a representation of content along two key dimensions: (1) *clickability*, the likelihood to *click to* an article when it is recommended; and (2) *engageability*, the likelihood to *click from* an article when it hosts a recommendation; the full meaning of this terminology will become

apparent in what follows. Our suggested "space of articles" is compact, but captures a key new dimension (engageability) and is therefore significantly richer than the one adopted by current practice (which, as we explain later, may be interpreted as focusing on clickability alone). This new space quantifies both the likelihood to click on each candidate article, and the likelihood to continue using the service in the next step, if this article is indeed clicked and becomes the host of a recommendation.

Leveraging engageability. Based on the aforementioned content space representation, we propose an efficient one-step look-ahead heuristic that balances clickability and engageability. We then demonstrate that by accounting for engageability, this heuristic yields performance that is close to the one of the optimal (and computationally intractable) recommendation policy.

Validating key ideas through a live controlled experiment. We study the implementation of a new class of one-step look-ahead recommendation policies, balancing clickability and engageability using proxies that are observed in real time throughout the recommendation process, without increasing the complexity of the existing practice. Together with our industry partner, we design and implement a controlled experiment that measures the impact of lookahead recommendations (that are based on the above representation) compared to myopic ones, validating the potential of the proposed approach.

1.4 Related Literature

Stochastic approximation. The use of the cumulative performance criterion and regret, while mostly absent from the traditional SA stream of literature, has been adapted in several occasions. 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 $\log T$.

While temporal changes in the cost function are largely not dealt with in the traditional stationary SA literature (see Kushner and Yin (2003), chapter 3 for some 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*. As 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 sub-linearity 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* 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 2005, Hazan et al. 2007), or other feedback structures such as multi-point 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 study 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. 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).

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). Several papers study dynamic pricing where the demand function is unknown, and noisy cost observations are obtained at each step (see, e.g., Broder and Rusmevichientong (2012), den Boer and Zwart (2014), and Harisson et al. (2014)). Other studies consider a problem of inventory control with censored demand, where noisy observations of the gradient can be obtained in each step (see Huh and Rusmevichientong (2009), Besbes and Muharremoglu (2013), and references therein). Other applications arise in queueing networks, wireless communications, and manufacturing systems, among other areas (see Kushner and Yin (2003) for an overview).

Most of the studies in the literature focus on a setting in which the underlying environment (while unknown) is stationary. While several papers have considered settings where changes in the governing 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 study suggests a general framework to study problems such as the ones mentioned above,

while assuming a broad array of changes in the underlying environment. By introducing and characterizing the regret with respect to a dynamic oracle we map the extent of environmental changes to the best achievable performance, and provide a general approach of designing rate-optimal policies.

Multi-armed bandits. Since their inception, MAB problems with various modifications have been studied extensively in Statistics, Economics, Operations Research, and Computer Science, and are used to model a plethora of dynamic optimization problems under uncertainty; examples include clinical trials (Zelen 1969), strategic pricing (Bergemann and Valimaki 1996), investment in innovation (Bergemann and Hege 2005), packet routing (Awerbuch and Kleinberg 2004), on-line auctions (Kleinberg and Leighton 2003), assortment selection (Caro and Gallien 2007*a*), and on-line advertising (Pandey et al. 2007), to name but a few. For overviews and further references cf. the monographs by Berry and Fristedt (1985), Gittins (1989) for Bayesian / dynamic programming formulations, and Cesa-Bianchi and Lugosi (2006) that covers recent advances in the machine learning literature and the so-called adversarial setting.

While temporal changes in the structure of the reward distribution are ignored in the traditional stochastic MAB formulation, there have been several attempts to extend that framework. The origin of this line of work can be traced back to Gittins and Jones (1974) who considered a case where only the state of the chosen arm can change, giving rise to a rich line of work (see, e.g., Gittins 1979, and Whittle 1981). In particular, Whittle (1988) introduced the term *restless bandits*; a model in which the states (associated with the reward distributions) of the arms change in each step according to an arbitrary, yet known, stochastic process. Considered a notoriously hard class of problems (cf. Papadimitriou and Tsitsiklis 1994), this line of work has led to various approximation approaches, see, e.g., Bertsimas and Nino-Mora (2000), and relaxations, see, e.g., Guha and Munagala (2007) and references therein.

Departure from the stationarity assumption that has dominated much of the MAB literature raises fundamental questions as to how one should model temporal uncertainty in rewards, and how to benchmark performance of candidate policies. One extreme view, is to allow the reward realizations of arms to be selected at any point in time by an *adversary*. These ideas have their origins in game theory with the work of Blackwell (1956) and Hannan (1957), and have since seen

significant development; Foster and Vohra (1999) and Cesa-Bianchi and Lugosi (2006) provide reviews of this line of research. Within this so called *adversarial* formulation, the efficacy of a policy over a given time horizon T is often measured relative to a benchmark which is defined by the single best action one could have taken in hindsight (after seeing all reward realizations). The single best action benchmark represents a *static* oracle, as it is constrained to a single (static) action. For obvious reasons, this static oracle can perform quite poorly relative to a "dynamic oracle" that follows the optimal *dynamic* sequence of actions, as the latter optimizes the (expected) reward at each time instant over all possible actions.[3] Thus, a potential limitation of the adversarial framework is that even if a policy has a "small" regret relative to a static oracle, there is no guarantee with regard to its performance relative to the dynamic oracle.

Online content recommendation services. At the technical level, the service provider's main problem is to dynamically select a set of recommended links for each reader. This has some similarities to the assortment planning problem studied in the operations management literature under various settings and demand models (see Kök et al. (2009) for a comprehensive review). When assortment selection is dynamic, Caro and Gallien (2007*b*) have studied the tradeoff between exploration and exploitation (when demand is unknown); see also Rusmevichientong et al. (2010), Alptekinoglu et al. (2012), and Saure and Zeevi (2013). A paper that studies dynamic assortment selection in an environment that is closer to the one of content recommendations is that of Caro et al. (2013), that considers a problem in which the attractiveness of products decay with time once they are introduced in the selected assortment. In their formulation, one needs to decide in advance the timing at which different products are introduced in the selected assortment, when each product can be introduced only once, and there are no inventory or capacity constraints.

The current study also relates to studies that focus on performance metrics and heuristics in online services (see, e.g., Kumar et al. (2006) and Araman and Fridgeirsdottir (2011) in the context of online advertising); the main distinction is driven by the dynamic nature that governs the content recommendation service, and thus, as we will see, calls for performance metrics (and

[3]Under non-stationary reward structure it is immediate that the single best action may be sub-optimal in a large number of decision epochs, and the gap between the performance of the static and the dynamic oracles can grow linearly with T.

appropriate heuristics) that account for the future path of users. In that respect, our study also relates to papers that study operational challenges of using path data to model and analyze consumers' behavior in various markets, such as retail, e-commerce, and advertising; for an overview cf. the survey by Hui et al. (2009).

An active stream of literature has been studying recommender systems, focusing on the tactical aspects that concern modeling and establishing connections between users and products, as well as implementing practical algorithms based on these connections (see the book by Ricci et al. (2011) and the survey by Adomavicius and Tuzhilin (2005) for an overview). A typical perspective that is taken in this rich line of work is that of the *consumer*, focusing on the main objective of maximizing the probability to click on a recommendation. Common approaches that are used for this purpose are nearest neighbor methods, relevance feedback methods, probabilistic (non-parametric or Bayesian) learning methods, and linear classification methods (see Pazzani and Billsus (2007) and references therein). Another common class of algorithms focuses on collaborative filtering; see the survey by Su and Khoshgoftaar (2009) and references therein, as well as the industry report by Linden et al. (2003) on Amazon's item-based collaborative filtering approach. The current study does not focus on these tactical questions, but rather on the higher level principles that guide the design of such algorithms when one accounts for the path of a user. By doing so, to the best of our knowledge the current paper is the first to focus on the perspective of the *recommender system* (the service provider), in a context of a multi-step service in which the system's objective is not necessarily aligned with that of the consumer.

1.5 Conclusions

In this thesis we study methodological as well as practical aspects arising in online sequential optimization in the presence of online partial feedback and a changing environment. On the methodological front, we study aspects of sequential optimization in the presence of temporal changes, such as designing decision making policies that adopt to temporal changes in the underlying environment when only partial feedback is available. In doing so we focus on two widely studied paradigms of sequential optimization: the stochastic approximation (SA) formulation, and the multi-armed bandit (MAB) formulation, when couched in a non-stationary setting.

In the first part of the thesis we consider a non-stationary variant of the SA problem, where the underlying cost functions may change along the horizon. In the second part of the thesis we consider a multi-armed bandit (MAB) formulation that allows for a broad range of temporal uncertainties in the rewards. Both of these sequential optimization settings, that are widely applied in the Operations Research, Economics, Statistics, and Computer Science literature. In both the SA and the MAB settings we establish tight bounds on the regret relative to the dynamic oracle, characterizing the complexity of these classes of problems in terms of the best achievable performance. These bounds maps the extent of allowable "variation" to the best achievable performance. Our analysis quantifies the "price of non-stationarity": the added complexity embedded in a temporally changing environment versus a stationary one. Our analysis also suggests key ingredients in polices that are designed to "perform well" in non-stationary environments, such as the the balance of "remembering and forgetting", captured by the restarting property of our suggested near-optimal policies. Our study draws a strong and concrete connection between rather disparate strands of literature: connecting the adversarial online convex optimization literature stream with that of the more traditional stochastic approximation paradigm; and connecting the adversarial MAB framework with the stochastic MAB one. These connections are the key in designing "well performing" policies in stochastic, non-stationary environments, by leveraging the structure of optimal policies in adversarial settings.

On the applied front, in the third part of the thesis we study practical aspects arising in *online content recommendations*, a new class of online services that allows web-based publishers to direct readers from articles they are currently reading to other web-based content. We study the dynamic optimization problem faced by the service provider, focusing on the salient features of that problem: the short time frames in which decisions are taken, the short shelf life of products, and the path-based structure of the service. Using a large data set of browsing history at major media sites, we develop a representation of content along two key dimensions: *clickability*, the likelihood to *click to* an article when it is recommended; and *engageability*, the likelihood to *click from* an article when it hosts a recommendation. Based on this representation, we propose a class of user path-focused heuristics, and validate their impact through theoretical bounds, simulation, and a live experiment.

All together, our thesis provide both theoretical as well as practical aspects that are faced in rapidly emerging application domains, such as online dynamic pricing and online assortment selection. On the methodological level our formulation allows significant departure from stationary assumptions that have governed most of stochastic optimization models, by introducing the *variation budget* and the *dynamic oracle*. On the applied level, collaborating with a major supplier of online content recommendations to web-based publishers, we were able to complete a cycle, from identifying a performance gap in current practice, through model and problem formulation, empirical analysis that lead to the design of improved heuristics, that in turn were validated by theoretical bounds and a simulation, and finally, an implementation study and a validation through a controlled experiment.

Chapter 2

Non-stationary Stochastic Optimization

The material presented in this chapter is based on Besbes, Gur and Zeevi (2014*a*).

In this chapter we consider a non-stationary variant of a sequential stochastic optimization problem, where the underlying cost functions may change along the horizon. §2.1 contains the problem formulation, where we propose a measure, termed *variation budget*, that controls the extent of said change, and in the following sections we study how restrictions on this budget impact achievable performance. In §2.2 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 principle connecting two rather disparate strands of literature: adversarial online convex optimization; and 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. §2.3 and §2.4 present the main rate optimality results for the convex and strongly convex settings, respectively. The tight bounds on the minimax regret that are established in §2.3 and §2.4 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. Finally, §2.5 presents concluding remarks. Proofs can be found in Appendix A.

2.1 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 $\mathcal{X}$ be a convex, compact, non-empty *action set*, and $\mathcal{T} = \{1, \dots, T\}$ be the sequence of decision epochs. Let $\mathcal{F}$ be a class of sequences $f := \{f_t : t = 1, \dots, T\}$ of convex cost functions from $\mathcal{X}$ into $\mathbb{R}$, that submit to the following two conditions:

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

$$|f_t(x)| \leq G, \qquad \|\nabla f_t(x)\| \leq G. \tag{2.1}$$

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

$$\Big\{x \in \mathbb{R}^d \ : \ \|x - x_t^*\| \leq \nu\Big\} \subset \mathcal{X} \qquad \text{for all } t \in \mathcal{T}, \tag{2.2}$$

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

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

For all $x \in \mathcal{X}$ and f_t, $t \in \{1, \dots, T\}$, 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 $\mathcal{F}$ and $\mathcal{X}$.

Example 2.1. (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 $(\mathbb{U},\mathcal{U},\mathbf{P}_u)$. Let $\pi_1 : \mathbb{U} \to \mathbb{R}^d$ and $\pi_t : \mathbb{R}^{(t-1)k} \times \mathbb{U} \to \mathbb{R}^d$ 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 $\{\pi_t : t = 1, \ldots, T\}$ together with the distribution $\mathbf{P}_u$ define the class of admissible policies with respect to feedback ϕ. We denote this class by $\mathcal{P}_\phi$. We further denote by $\{\mathcal{H}_t,\ t = 1, \ldots, T\}$ the *filtration* associated with a policy $\pi \in \mathcal{P}_\phi$, such that $\mathcal{H}_1 = \sigma(U)$ and $\mathcal{H}_t = \sigma\left(\{\phi_j(X_j, f_j)\}_{j=1}^{t-1}, U\right)$ for all $t \in \{2, 3, \ldots\}$. Note that policies in $\mathcal{P}_\phi$ are non-anticipating, 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 $\mathcal{F}$ 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:

$$\mathrm{Var}(f_1, \ldots, f_T) \ := \ \sum_{t=2}^{T} \|f_t - f_{t-1}\|, \tag{2.3}$$

where for any bounded functions g and h from $\mathcal{X}$ into $\mathbb{R}$ we denote $\|g-h\| := \sup_{x\in\mathcal{X}} |g(x) - h(x)|$. Let $\{V_t : t = 1, 2, \ldots\}$ be a non-decreasing sequence of real numbers such that $V_t \leq t$ for all t, $V_1 = 0$, and for normalization purposes set $V_2 \geq 1$. We refer to V_T as the *variation budget* over $\mathcal{T}$. 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 $\{1, \ldots, T\}$:

$$\mathcal{V} = \left\{ \{f_1, \ldots, f_T\} \subset \mathcal{F} \ : \ \sum_{t=2}^{T} \|f_t - f_{t-1}\| \leq V_T \right\}. \tag{2.4}$$

While 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 non-constant rate of change. Two possible variations instances are illustrated in Figure 3.1.

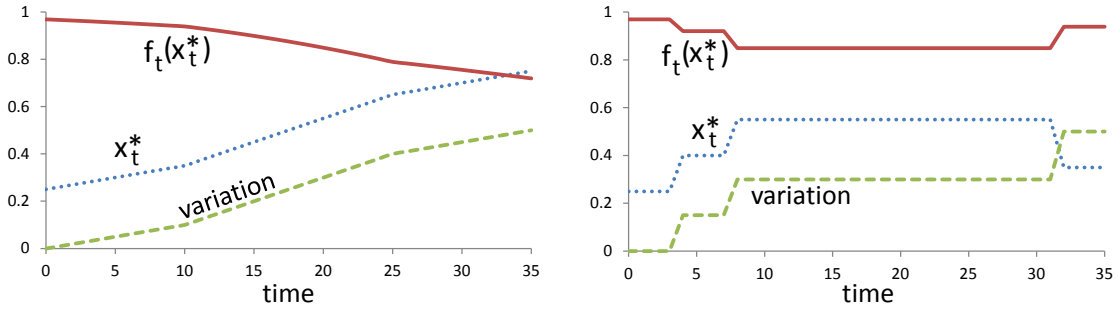


Figure 2.1: **Variation instances within a temporal uncertainty set.** Assume a quadratic cost of the form $f_t(x) = \frac{1}{2}x^2 - b_t x + 1$. The change in the minimizer $x_t^* = b_t$, the optimal performance $f_t(x_t^*) = 1 - \frac{1}{2}b_t^2$, and the variation measured by (2.3), is illustrated for cases characterized by continuous changes (left), and "jump" changes (right) in b_t. In both instances the variation budget is $V_T = 1/2$.

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

$$\mathcal{R}_\phi^\pi(\mathcal{V}, T) = \sup_{f \in \mathcal{V}} \left\{ \mathbb{E}^\pi \left[\sum_{t=1}^{T} f_t(X_t) \right] - \sum_{t=1}^{T} f_t(x_t^*) \right\}, \tag{2.5}$$

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 $f \in \mathcal{V}$ 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[1]. Recall that a policy π is said to have *sublinear* regret if $\mathcal{R}_\phi^\pi(\mathcal{V}, T) = o(T)$, 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 $\pi \in \mathcal{P}_\phi$, is given by:

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

We refer to a policy π as *rate optimal* if there exists a constant $\bar{C} \geq 1$, independent of V_T and T, such that for any $T \geq 1$,

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

Such policies achieve the lowest possible growth rate of regret.

[1]In particular, while 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.

Contrasting with the adversarial online convex optimization paradigm. An OCO problem consists of a convex set $\mathcal{X} \subset \mathbb{R}^d$ and an a-priori unknown sequence $f = \{f_1, \ldots, f_T\} \in \mathcal{F}$ of convex cost functions. At any epoch t the decision maker selects a point $X_t \in \mathcal{X}$, 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 $\mathcal{A}$:

$$\mathcal{G}_{\phi}^{\mathcal{A}}(\mathcal{F}, T) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}^{\pi} \left[\sum_{t=1}^{T} f_t(X_t) \right] - \min_{x \in \mathcal{X}} \left\{ \sum_{t=1}^{T} f_t(x) \right\} \right\}, \tag{2.6}$$

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)[2]. Interchanging the sum and $\min\{\cdot\}$ operators in the right-hand-side of (2.6) we obtain the definition of regret in the non-stationary stochastic setting, as in (2.5). 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 (2.6).

Example 2.2. (Contrasting the static and dynamic oracles) Assume an action set $\mathcal{X} = [-1, 2]$, and variation budget $V_T = 1$. Set

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

for any $x \in \mathcal{X}$. Then, the single best action is sub-optimal at any decision epoch, and

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

[2]We note that most results in the OCO literature allow sequences that can adjust the cost function adversarially at each epoch. For the sake of consistency with the definition of (2.5), in the above regret measure nature commits to a sequence of functions in advance.

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 online convex optimization context can in fact be adapted to perform well in the non-stationary stochastic setting laid out in this chapter.

2.2 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 2.1. (Linear variation budget implies linear regret) *Assume a feedback structure $\phi \in \left\{\phi^{(0)}, \phi^{(1)}\right\}$. If there exists a positive constant C_1 such that $V_T \geq C_1 T$ for any $T \geq 1$, then there exists a positive constant C_2, such that for any admissible policy $\pi \in \mathcal{P}_\phi$,*

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

The proposition states that whenever the variation budget is at least of order T, *any* policy which 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.

A class of candidate policies. We introduce a class of policies that leverages existing algorithms designed for fully adversarial environments. We denote by $\mathcal{A}$ an online optimization algorithm that given a feedback structure ϕ achieves a regret $\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}, T)$ (see (2.6)) with respect to the *static* benchmark of the single best action. Consider the following generic "restarting" procedure, which takes as input $\mathcal{A}$ and a batch size Δ_T, with $1 \leq \Delta_T \leq T$, and consists of restarting $\mathcal{A}$ every Δ_T periods. To formalize this idea we first refine our definition of history-adapted policy and the actions it generates. Given a feedback ϕ and a restarting epoch $\tau \geq 1$, we

define the history at time $t \geq \tau + 1$ to be:

$$\mathcal{H}_{\tau,t} = \begin{cases} \sigma(U) & \text{if } t = \tau + 1, \\ \sigma\left(\{\phi_j(X_j, f_j)\}_{j=\tau+1}^{t-1}, U\right) & \text{if } t > \tau + 1. \end{cases} \tag{2.7}$$

Then, for any t we have that X_t is $\mathcal{H}_{\tau,t}$-measurable. In particular $X_{\tau+1} = \mathcal{A}_1(U)$, $X_t = \mathcal{A}_{t-\tau}(\mathcal{H}_{\tau,t})$ for $t > \tau + 1$, and the sequence of measurable mappings $\mathcal{A}_t$, $t = 1, 2, \ldots$ is prescribed by the algorithm $\mathcal{A}$. The following procedure restarts $\mathcal{A}$ 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 $\mathcal{A}$, and a batch size Δ_T.

1. Set $j = 1$
2. Repeat while $j \leq \lceil T/\Delta_T \rceil$:
   (a) Set $\tau = (j-1)\Delta_T$
   (b) For any $t = \tau + 1, \ldots, \min\{T, \tau + \Delta_T\}$, select the action $X_t = \mathcal{A}_{t-\tau}(\mathcal{H}_{\tau,t})$
   (c) Set $j = j + 1$, and return to step 2.

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

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

Theorem 2.1. (Long-run-average optimality) *Set a feedback structure $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$. Let $\mathcal{A}$ be an OCO algorithm with $\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}, T) = o(T)$. Let π be the policy defined by the restarting procedure that uses $\mathcal{A}$ 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 $T \to \infty$,*

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

In other words, the theorem establishes the following meta-principle: 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 2.1 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 $\left(\mathcal{F}, \phi^{(0)}\right)$ and $\left(\mathcal{F}, \phi^{(1)}\right)$ an online gradient descent 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 2.1 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 $\mathcal{A}$ with respect to the single best action in the adversarial setting.

Proposition 2.2. (Connecting performance in OCO and non-stationary SA) *Set* $\phi \in \left\{\phi^{(0)}, \phi^{(1)}\right\}$. *Let* π *be the policy defined by the restarting procedure that uses* $\mathcal{A}$ *as a subroutine, with batch size* Δ_T. *Then, for any* $T \geq 1$,

$$\mathcal{R}_{\phi}^{\pi}(\mathcal{V}, T) \;\leq\; \left\lceil \frac{T}{\Delta_T} \right\rceil \cdot \mathcal{G}_{\phi}^{\mathcal{A}}(\mathcal{F}, \Delta_T) \;+\; 2\Delta_T V_T. \tag{2.8}$$

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 $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size Δ_T each (except, possibly the last batch):

$$\mathcal{T}_j = \{t \;:\; (j-1)\Delta_T + 1 \leq t \leq \min\{j\Delta_T, T\}\}, \text{ for all } \; j = 1, \ldots, m, \tag{2.9}$$

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

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

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 (2.8). 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. While 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. Theorem 2.1 directly follows. □

Remark (Alternative forms of feedback) The principle laid out in Theorem 2.1 can also be derived for other forms of feedback using Proposition 2.2. For example, the proof of Theorem 2.1 holds for settings with richer feedback structures, such as noiseless access to the full cost function (Zinkevich 2003), or a multi-point access (Agarwal et al. 2010).

2.3 Rate Optimality: The General Convex Case

A natural question arising from the analysis of §2.2 is whether the restarting procedure introduced there enables to carry over the property of rate optimality from the adversarial environment to the non-stationary stochastic environment. 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).

Subroutine OCO algorithm. As a subroutine algorithm, we will use an adaptation of the online gradient descent (OGD) algorithm introduced by Zinkevich (2003):

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

1. Select some $X_1 = x_1 \in \mathcal{X}$
2. For any $t = 1, \ldots, T-1$, set

$$X_{t+1} \;=\; P_{\mathcal{X}}\left(X_t - \eta_{t+1}\phi_t^{(1)}(X_t, f_t)\right),$$

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 $\{\mathcal{A}_{t-\tau}\}$, $t \geq \tau+1$, as follows:

$$\mathcal{A}_{t-\tau}\left(\mathcal{H}_{\tau,t}\right) = \begin{cases} x_1 & \text{if } t = \tau+1 \\ P_{\mathcal{X}}\left(X_{t-1} - \eta_{t-\tau}\phi_{t-1}^{(1)}\right) & \text{if } t > \tau+1, \end{cases}$$

for any epoch $t \geq \tau+1$, where $\mathcal{H}_{\tau,t}$ is defined in (2.7). For the structure $\left(\mathcal{F}, \phi^{(1)}\right)$ of convex cost functions and noisy gradient access, Flaxman et al. (2005) consider the OGD algorithm with the selection $\eta_t = r/G\sqrt{T}$, $t = 2, \ldots, T$, where r denotes the radius of the action set:

$$r \;=\; \inf\left\{y > 0 :\; \mathcal{X} \subseteq \mathbf{B}_y(x) \text{ for some } x \in \mathbb{R}^d\right\},$$

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 A.7 (given in Appendix A.2) that under Assumption 2.1 (a structural assumption on the feedback, given later in this section), this performance is rate optimal in the adversarial OCO setting.

Performance analysis. We first consider the performance of the OGD algorithm *without restarting*, relative to the dynamic benchmark. The following illustrates that this algorithm will yield linear regret for a broad set of variation budgets.

Example 2.3. (Failure of OGD without restarting) Consider a partition of the horizon $\mathcal{T}$ into batches $\mathcal{T}_1, \ldots \mathcal{T}_m$ according to (3.2), with each batch of size Δ_T. Consider the following cost functions:

$$g_1(x) = (x-\alpha)^2, \qquad g_2(x) = x^2; \qquad x \in [-1, 3]\,.$$

Assume that nature selects the cost function to be $g_1(\cdot)$ in the even batches and $g_2(\cdot)$ in the odd batches. Assume that at every epoch t, after selecting an action $x_t \in \mathcal{X}$, a *noiseless* access to the gradient of the cost function at point x_t is granted, that is, $\phi_t^{(1)}(x, f_t) = f_t'(x)$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$. Assume that the decision maker is applying the OGD algorithm with a sequence of step sizes $\{\eta_t\}_{t=2}^T$, and $x_1 = 1$. We consider two classes of step size sequences that have been shown to be rate optimal in two OCO settings (see Flaxman et al. (2005), and Hazan et al. (2007)).

1. Suppose $\eta_t = \eta = C/\sqrt{T}$. Then, selecting a batch size Δ_T of order $\sqrt{T}$, and $\alpha = 1 + (1+2\eta)^{\Delta_T}$, the variation budget V_T is at most of order $\sqrt{T}$, and there is a constant C_1 such that $\mathcal{R}_\phi^\pi(\mathcal{V}, T) \geq\; C_1 T$.

2. Suppose that $\eta_t = C/t$. Then, selecting a batch size Δ_T of order T, and $\alpha = 1$, the variation budget V_T is a fixed constant, and there is a constant C_2 such that $\mathcal{R}_\phi^\pi(\mathcal{V}, T) \geq C_2 T$. $\square$

In both of the cases that are described in the example, we analyze the trajectory of deterministic actions $\{x_t\}_{t=1}^T$ that is generated by the OGD algorithm, and show that there is a fraction of the horizon in which the action x_t is not "close" to the minimizer x_t^*, and therefore linear regret is incurred. At a high level, this example illustrates that, not surprisingly, OGD-type policies with classical step size selections do not perform well in non-stationary environments.

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

Theorem 2.2. (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 $\Delta_T = \left\lceil (T/V_T)^{2/3} \right\rceil$, and the OGD algorithm parameterized by $\eta_t = \frac{r}{G\sqrt{\Delta_T}}$, $t = 2, \ldots, \Delta_T$ as a subroutine. Then, there is some finite constant $\bar{C}$, independent of T and V_T, such that for all $T \geq 2$:*

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

Recalling the connection between the regret in the adversarial setting and the one in the non-stationary SA setting (Proposition 2.2), 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 $\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F}, \Delta_T)$ of order $\sqrt{\Delta_T}$, and the result follows by balancing the terms in (2.8) 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. On the other hand, 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.

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

Assumption 2.1. (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, $t \geq 1$, *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* $a \in \mathbb{R}^d$, $\int \log\left(\frac{dG(y)}{dG(y+a)}\right) dG(y) \leq \tilde{C}\,\|a\|^2$.

Remark. 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: $\mathbb{P}\left(\phi_t^{(1)}(x, f_t) \in A\right) > 0$ 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 2.1 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 2.3. (Lower bound on achievable performance) *Let Assumption 2.1 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* $T \geq 1$*:*

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

The result above, together with Theorem 2.2, implies that the performance of restarted OGD (provided in Theorem 2.2) is rate optimal, and the minimax regret under structure $\left(\mathcal{V}, \phi^{(1)}\right)$ 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 $0 \leq \alpha \leq 1$, 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 2.1).

Key ideas in the proof of Theorem 2.3. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on a probability space $\mathcal{Y}$, let

$$\mathcal{K}\left(\mathbb{P}\|\mathbb{Q}\right) \;=\; \mathbb{E}\left[\log\left(\frac{d\mathbb{P}\left\{Y\right\}}{d\mathbb{Q}\left\{Y\right\}}\right)\right], \tag{2.10}$$

where $\mathbb{E}\left[\cdot\right]$ is the expectation with respect to $\mathbb{P}$, and Y is a random variable defined over $\mathcal{Y}$. This quantity is is known as the Kullback-Leibler 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: (a) any drawn sequence must maintain the variation constraint; and (b) the functions are chosen to be "close" enough while 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 the batches, and hence any admissible policy trying to identify the current cost function in a given batch can only do so with a strictly positive error probability.)

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 $T^{3/4}$. 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 2.2, 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. While 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 environment by applying the restarting procedure. The next section further supports this conjecture by examining the case of strongly convex cost functions.

2.4 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

$$H\mathbf{I}_d \preceq \nabla^2 f_t(x) \preceq G\mathbf{I}_d \quad \text{for all } x \in \mathcal{X}, \text{ and all } t \in \mathcal{T}, \tag{2.11}$$

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 semi-definite, and $\nabla^2 f(x)$ denotes the Hessian of $f(\cdot)$, evaluated at point $x \in \mathcal{X}$.

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:

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

Using the above, we measure variation by:

$$\mathrm{Var}_s(f_1, \ldots, f_T) := \sum_{t=2}^{T} \sup_{x \in \mathcal{X}^*} |f_t(x) - f_{t-1}(x)| \,. \tag{2.12}$$

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

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

We note that the proof of Proposition 2.2 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.

2.4.1 Noisy access to the gradient

For the strongly convex function 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 adversarial OCO framework a regret guarantee of order $\log T$ (with respect to the

single best action benchmark). For completeness, we provide in Appendix A.2 a simple adaptation of this result to the case of *noisy* gradient access. 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.[3]

Theorem 2.4. (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* $\Delta_T = \left\lceil \sqrt{T \log T / V_T} \right\rceil$*, 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* $T \geq 2$*:*

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

2. *Let Assumption 2.1 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* $T \geq 1$*:*

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

Up to a logarithmic term, Theorem 2.4 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. In §2.5 we show that one may achieve a performance of $O(\sqrt{V_T T})$ through a slightly modified procedure, and hence the *minimax regret* under structure $\left(\mathcal{F}_s, \phi^{(1)}\right)$ is:

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

Theorem 2.4 further validates the "meta-principle" 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).

[3]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.

The first part of Theorem 2.4 is derived directly from Proposition 2.2, by plugging in a bound on $\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}_s, \Delta_T)$ of order $\log T$ (given by Lemma A.5 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 2.3, adjusting for strongly convex cost functions.

2.4.2 Noisy access to the cost

We now consider the structure $\left(\mathcal{V}_s, \phi^{(0)}\right)$, in which the cost functions are strongly convex and the decision maker has noisy access to the cost. In order 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 $\left(\mathcal{F}_s, \phi^{(0)}\right)$.

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

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

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

For $k = 1, ..., d$, let $e^{(k)}$ denote the unit vector with 1 at the k^{th} coordinate. The *estimated gradient step* (EGS) algorithm is defined through three sequences of real numbers $\{h_t\}$, $\{a_t\}$, and $\{\delta_t\}$, where[4] $\nu \geq \delta_t \geq h_t$ for all $t \in \mathcal{T}$:

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

---

[4]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 A.4 remaining unchanged.

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

$$\mathcal{A}_{t-\tau}\left(\mathcal{H}_{\tau,t}\right) = \begin{cases} \text{some } Z_1 & \text{if } t = \tau + 1 \\ Z_{t-\tau} + h_{t-\tau}\psi_{t-\tau-1} & \text{if } t > \tau + 1. \end{cases}$$

Note that $\mathbb{E}[\hat{\nabla}_h f_t(Z_t)|X_t] = \nabla f_t(Z_t)$ (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 A.4 (Appendix A.2) 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 $\left(\mathcal{F}_s, \phi^{(0)}\right)$. For completeness, we establish in Lemma A.6 (Appendix A.2), that under Assumption 2.2 (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 2.1 for the case of cost feedback, that will be used in deriving a lower bound on the regret.

Assumption 2.2. (Cost feedback structure)

1. *$\phi_t^{(0)}(x, f_t) = f_t(x) + \varepsilon_t$ for any $f \in \mathcal{F}$, $x \in \mathcal{X}$, and $t \in \mathcal{T}$, where ε_t, $t \geq 1$, 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 $a \in \mathbb{R}$, $\int \log\left(\frac{dG(y)}{dG(y+a)}\right) dG(y) \leq \tilde{C} \cdot a^2$.*

Theorem 2.5. (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 $\Delta_T = \left\lceil (T/V_T)^{2/3} \right\rceil$. Then, there exists a finite constant $\bar{C} > 0$, independent of T and V_T, such that for all $T \geq 2$:*

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

2. *Let Assumption 2.2 hold. Then, there exists a constant $C > 0$, independent of T and V_T, such that for any policy $\pi \in \mathcal{P}_{\phi^{(0)}}$ and for all $T \geq 1$:*

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

Theorem 2.5 again establishes the ability to "port over" rate optimality from the adversarial OCO setting to the non-stationary stochastic setting, this time under structure $\left(\mathcal{F}_s, \phi^{(0)}\right)$. The theorem establishes a characterization of the minimax regret under structure $\left(\mathcal{V}_s, \phi^{(0)}\right)$:

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

2.5 Concluding Remarks

Batching versus continuous updating. While the restarting procedure (together with suitable balancing of the batch size) can be used as a template for deriving "good" policies in the non-stationary SA setting, it is important to note that there are alternative paths to achieving this goal. One of them relies on directly re-tuning the parameters of the OCO algorithm. To demonstrate this idea we show that the OGD algorithm can be re-tuned to achieve rate optimal regret in a non-stationary stochastic setting under structure $\left(\mathcal{F}_s, \phi^{(1)}\right)$, matching the lower bound given in Theorem 2.4 (part 2).

Theorem 2.6. *Consider the feedback structure $\phi = \phi^{(1)}$, and let π the OGD algorithm with $\eta_t = \sqrt{V_T/T}$, $t = 2, \ldots, T$. Then, there exists a finite constant $\bar{C}$, independent of T and V_T, such that for all $T \geq 2$:*

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

The key to tuning the OGD algorithm so that it achieves rate optimal performance in the non-stationary SA setting is a suitable adjustment of the step size sequence as a function of the variation budget V_T: intuitively, the larger the variation is (relative to the horizon length T), the larger the step sizes that are required in order to "keep up" with the changing environment.

On the transition from stationary to non-stationary settings. Throughout this chapter we address "significant" variation in the cost function, and for the sake of concreteness assume $V_T \geq 1$. 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. The following table summarizes the minimax regret rates for a variation budget of the form $V_T = T^\alpha$, and documents the critical variation scales in different settings.

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

Table 2.1: **Critical variation scales.** The growth rates of the minimax regret in different settings for $V_T = T^\alpha$ (where $\alpha \leq 1$) and the variation scales that separate the stationary / non-stationary regimes.

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.

Adapting to an unknown variation budget. The policies introduced in the current chapter rely on prior knowledge of the variation budget 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 study.

Chapter 3

Multi-Armed-Bandit Problems with Non-stationary Rewards

The material presented in this chapter is based on Besbes, Gur and Zeevi (2014*b*).

In this chapter we consider a Multi-armed bandit (MAB) formulation which allows for a broad range of temporal uncertainties in the rewards, while also being mathematically tractable. §3.1 introduces the basic formulation of stochastic non-stationary MAB with a variation budget. In §3.2 we provide a lower bound on the regret that any admissible policy must incur relative to a dynamic oracle. §3.3 introduces a policy that achieves that lower bound. Together, these results fully characterize the regret complexity of this class of MAB problems, establishing a direct link between the extent of allowable reward "variation" and the minimal achievable worst-case regret. The analysis in this chapter draws concrete connections between two rather disparate strands of literature: the adversarial and the stochastic MAB frameworks. §3.4 briefly discusses some concluding remarks. Proofs can be found in Appendix B.

3.1 Problem Formulation

Let $\mathcal{K} = \{1, \ldots, K\}$ be a set of arms. Let $\mathcal{T} = \{1, 2, \ldots, T\}$ denote the sequence of decision epochs faced by the decision maker. At any epoch $t \in \mathcal{T}$, a decision-maker pulls one of the K arms. When pulling arm $k \in \mathcal{K}$ at epoch $t \in \mathcal{T}$, a reward $X_t^k \in [0, 1]$ is obtained, where X_t^k is a

random variable with expectation

$$\mu_t^k \;=\; \mathbb{E}\left[X_t^k\right].$$

We denote the best possible expected reward at decision epoch t by μ_t^*, i.e.,

$$\mu_t^* \;=\; \max_{k\in\mathcal{K}}\left\{\mu_t^k\right\}.$$

Changes in the expected rewards of the arms. We assume the expected reward of each arm μ_t^k may change at any decision point. We denote by μ^k the sequence of expected rewards of arm k: $\mu^k = \left\{\mu_t^k\right\}_{t=1}^T$. In addition, we denote by μ the sequence of vectors of all K expected rewards: $\mu = \left\{\mu^k\right\}_{k=1}^K$. We assume that the expected reward of each arm can change an arbitrary number of times, but bound the total variation of the expected rewards:

$$\sum_{t=1}^{T-1}\sup_{k\in\mathcal{K}}\left|\mu_t^k-\mu_{t+1}^k\right|. \tag{3.1}$$

Let $\{V_t : t = 1, 2, \ldots\}$ be a non-decreasing sequence of positive real numbers such that $V_1 = 0$, $KV_t \leq t$ for all t, and for normalization purposes set $V_2 = 2\cdot K^{-1}$. We refer to V_T as the *variation budget* over $\mathcal{T}$. We define the corresponding *temporal uncertainty set*, as the set of reward vector sequences that are subject to the variation budget V_T over the set of decision epochs $\{1,\ldots,T\}$:

$$\mathcal{V} \;=\; \left\{\mu\in[0,1]^{K\times T} \;:\; \sum_{t=1}^{T-1}\sup_{k\in\mathcal{K}}\left|\mu_t^k-\mu_{t+1}^k\right|\leq V_T\right\}.$$

The variation budget captures the constraint imposed on the non-stationary environment faced by the decision-maker. While limiting the possible evolution in the environment, it allows for many different forms in which the expected rewards may change: continuously, in discrete shocks, and of a changing rate (for illustration, Figure 3.1 depicts two different variation patterns that correspond to the same variation budget). In general, the variation budget V_T is designed to depend on the number of pulls T.

Admissible policies, performance, and regret. Let U be a random variable defined over a probability space $(\mathbb{U},\mathcal{U},\mathbf{P}_u)$. Let $\pi_1 : \mathbb{U}\to\mathcal{K}$ and $\pi_t : [0,1]^{t-1}\times\mathbb{U}\to\mathcal{K}$ for $t = 2,3,\ldots$ be measurable functions. With some abuse of notation we denote by $\pi_t\in\mathcal{K}$ the action at time t, that is given by

$$\pi_t \;=\; \begin{cases} \pi_1(U) & t=1,\\ \pi_t\left(X_{t-1}^\pi,\ldots,X_1^\pi,U\right) & t=2,3,\ldots,\end{cases}$$

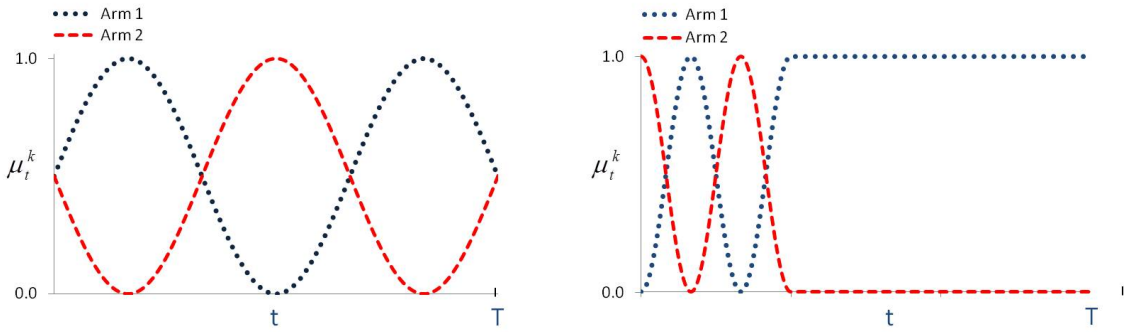


Figure 3.1: Two instances of variation in the expected rewards of two arms: (*Left*) Continuous variation in which a fixed variation budget (that equals 3) is spread over the whole horizon. (*Right*) "Compressed" instance in which the same variation budget is "spent" in the first third of the horizon.

The mappings $\{\pi_t : t = 1, \ldots, T\}$ together with the distribution $\mathbf{P}_u$ define the class of admissible policies. We denote this class by $\mathcal{P}$. We further denote by $\{\mathcal{H}_t, t = 1, \ldots, T\}$ the filtration associated with a policy $\pi \in \mathcal{P}$, such that $\mathcal{H}_1 = \sigma(U)$ and $\mathcal{H}_t = \sigma\left(\left\{X_j^\pi\right\}_{j=1}^{t-1}, U\right)$ for all $t \in \{2, 3, \ldots\}$. Note that policies in $\mathcal{P}$ are non-anticipating, i.e., depend only on the past history of actions and observations, and allow for randomized strategies via their dependence on U.

We define the *regret* under policy $\pi \in \mathcal{P}$ compared to a *dynamic* oracle as the worst-case difference between the expected performance of pulling at each epoch t the arm which has the highest expected reward at epoch t (the dynamic oracle performance) and the expected performance under policy π:

$$\mathcal{R}^\pi(\mathcal{V}, T) \;=\; \sup_{\mu \in \mathcal{V}} \left\{ \sum_{t=1}^{T} \mu_t^* - \mathbb{E}^\pi \left[\sum_{t=1}^{T} \mu_t^\pi \right] \right\},$$

where the expectation $\mathbb{E}^\pi[\cdot]$ is taken with respect to the noisy rewards, as well as to the policy's actions. In addition, we denote by $\mathcal{R}^*(\mathcal{V}, T)$ the minimal worst-case regret that can be guaranteed by an admissible policy $\pi \in \mathcal{P}$:

$$\mathcal{R}^*(\mathcal{V}, T) \;=\; \inf_{\pi \in \mathcal{P}} \mathcal{R}^\pi(\mathcal{V}, T).$$

$\mathcal{R}^*(\mathcal{V}, T)$ is the best achievable performance. In the following sections we study the magnitude of $\mathcal{R}^*(\mathcal{V}, T)$. We analyze the magnitude of this quantity by establishing upper and lower bounds; in these bounds we refer to a constant C as *absolute* if it is independent of K, V_T, and T.

3.2 Lower bound on the best achievable performance

We next provide a lower bound on the the best achievable performance.

Theorem 3.1. *Assume that rewards have a Bernoulli distribution. Then, there is some absolute constant $C > 0$ such that for any policy $\pi \in \mathcal{P}$ and for any $T \geq 1$, $K \geq 2$ and $V_T \in \left[K^{-1}, K^{-1}T\right]$,*

$$\mathcal{R}^{\pi}(\mathcal{V}, T) \;\geq\; C\,(KV_T)^{1/3}\,T^{2/3}.$$

We note that when reward distributions are stationary, there are known policies such as UCB1 and ε-greedy (Auer, Cesa-Bianchi and Fischer 2002) that achieve regret of order $\sqrt{T}$ in the stochastic setup. When the environment is non-stationary and the reward structure is defined by the class $\mathcal{V}$, then no policy may achieve such a performance and the best performance must incur a regret of at least order $T^{2/3}$. This additional complexity embedded in the stochastic non-stationary MAB problem compared to the stationary one will be further discussed in §3.4.

Remark 3.1. (Growing variation budget) Theorem 3.1 holds when V_T is increasing with T. In particular, when the variation budget is linear in T, the regret grows linearly and long run average optimality is not achievable. This also implies the observation of Slivkins and Upfal (2008) about linear regret in an instance in which expected rewards evolve according to a Brownian motion.

The driver of the change in the best achievable performance (relative to the one established in a stationary environment) is the optimal exploration-exploitation balance. Beyond the tension between exploring different arms and capitalizing on the information already collected, captured by the "classical" exploration-exploitation trade-off, a second tradeoff is introduced by the non-stationary environment, between "remembering" and "forgetting": estimating the expected rewards is done based on past observations of rewards. While keeping track of more observations may decrease the variance of mean rewards estimates, the non-stationary environment implies that "old" information is potentially less relevant and creates a bias that stems from possible changes in the underlying rewards. The changing rewards give incentive to dismiss old information, which in turn encourages enhanced exploration. The proof of Theorem 3.1 emphasizes these two tradeoffs and their impact on achievable performance. At a high level the proof of Theorem 3.1 builds on ideas of identifying a worst-case "strategy" of nature (e.g., Auer, Cesa-Bianchi,

Freund and Schapire 2002, proof of Theorem 5.1) adapting them to our setting. While the proof is deferred to the appendix, we next describe the key ideas.

Selecting a subset of feasible reward paths. We define a subset of vector sequences $\mathcal{V}' \subset \mathcal{V}$ and show that when μ is drawn randomly from $\mathcal{V}'$, any admissible policy must incur regret of order $(KV_T)^{1/3}T^{2/3}$. We define a partition of the decision horizon $\mathcal{T}$ into batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size $\tilde{\Delta}_T$ each (except, possibly the last batch):

$$\mathcal{T}_j \;=\; \left\{t \;:\; (j-1)\tilde{\Delta}_T + 1 \leq t \leq \min\left\{j\tilde{\Delta}_T, T\right\}\right\}, \quad \text{for all } j = 1, \ldots, m, \tag{3.2}$$

where $m = \lceil T/\tilde{\Delta}_T \rceil$ is the number of batches. In $\mathcal{V}'$, in every batch there is exactly one "good" arm with expected reward $1/2 + \varepsilon$ for some $0 < \varepsilon \leq 1/4$, and all the other arms have expected reward $1/2$. The "good" arm is drawn independently in the beginning of each batch according to a discrete uniform distribution over $\{1, \ldots, K\}$. Thus, the identity of the "good" arm can change only between batches. See Figure 3.2 for a description and a numeric example of possible realizations of a sequence μ that is randomly drawn from $\mathcal{V}'$. Since there are m batches we obtain

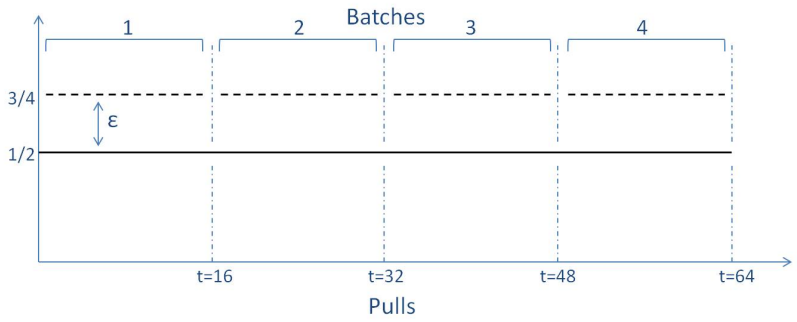


Figure 3.2: (Drawing a sequence from $\mathcal{V}'$.) A numerical example of possible realizations of expected rewards. Here $T = 64$, and we have 4 decision batches, each contains 16 pulls. We have K^4 possible realizations of reward sequences. In every batch one arm is randomly and independently drawn to have an expected reward of $1/2 + \varepsilon$, where in this example $\varepsilon = 1/4$. This example corresponds to a variation budget of $V_T = \varepsilon\tilde{\Delta}_T = 1$.

a set $\mathcal{V}'$ of K^m possible, equally probable realizations of μ. By selecting ε such that $\varepsilon T/\tilde{\Delta}_T \leq V_T$, any $\mu \in \mathcal{V}'$ is composed of expected reward sequences with a variation of at most V_T, and therefore $\mathcal{V}' \subset \mathcal{V}$. Given the draws under which expected reward sequences are generated, nature prevents any accumulation of information from one batch to another, since at the beginning of each batch a new "good" arm is drawn independently of the history.

Countering possible policies. For the sake of simplicity, the discussion in this paragraph assumes a variation budget that is fixed and independent of T (the proof of the theorem details the more general treatment for a variation budget that depends on T). The proof of Theorem 3.1 establishes that under the setting described above, if $\varepsilon \approx 1/\sqrt{\tilde{\Delta}_T}$ no admissible policy can identify the "good" arm with high probability within a batch. Since there are $\tilde{\Delta}_T$ epochs in each batch, the regret that any policy must incur along a batch is of order $\tilde{\Delta}_T \cdot \varepsilon \approx \sqrt{\tilde{\Delta}_T}$, which yields a regret of order $\sqrt{\tilde{\Delta}_T} \cdot T/\tilde{\Delta}_T \approx T/\sqrt{\tilde{\Delta}_T}$ throughout the whole horizon. Selecting the smallest feasible $\tilde{\Delta}_T$ such that the variation budget constraint is satisfied leads to $\tilde{\Delta}_T \approx T^{2/3}$, yielding a regret of order $T^{2/3}$ throughout the horizon.

3.3 A near-optimal policy

In this section we apply the ideas underlying the lower bound in Theorem 3.1 to develop a rate optimal policy for the non-stationary MAB problem with a variation budget. Consider the following policy:

Rexp3. Inputs: a positive number γ, and a batch size Δ_T.

1. Set batch index $j = 1$
2. Repeat while $j \leq \lceil T/\Delta_T \rceil$:
   (a) Set $\tau = (j-1)\,\Delta_T$
   (b) Initialization: for any $k \in \mathcal{K}$ set $w_t^k = 1$
   (c) Repeat for $t = \tau+1, \ldots, \min\{T, \tau+\Delta_T\}$:
   - For each $k \in \mathcal{K}$, set
   $$p_t^k \;=\; (1-\gamma)\,\frac{w_t^k}{\sum_{k'=1}^K w_t^{k'}} + \frac{\gamma}{K}$$
   - Draw an arm k' from $\mathcal{K}$ according to the distribution $\left\{p_t^k\right\}_{k=1}^K$
   - Receive a reward $X_t^{k'}$
   - For arm k' set $\hat{X}_t^{k'} = X_t^{k'}/p_t^{k'}$, and for any $k \neq k'$ set $\hat{X}_t^k = 0$. For all $k \in \mathcal{K}$ update:
   $$w_{t+1}^k \;=\; w_t^k \exp\left\{\frac{\gamma \hat{X}_t^k}{K}\right\}$$
   (d) Set $j = j+1$, and return to the beginning of step 2

Clearly $\pi \in \mathcal{P}$. The Rexp3 policy uses Exp3, a policy introduced by Freund and Schapire (1997) for solving a worst-case sequential allocation problem, as a subroutine, restarting it every Δ_T epochs.

Theorem 3.2. *Let π be the Rexp3 policy with a batch size $\Delta_T = \left\lceil (K \log K)^{1/3} (T/V_T)^{2/3} \right\rceil$ and with $\gamma = \min\left\{1\ ,\ \sqrt{\frac{K \log K}{(e-1)\Delta_T}}\right\}$. Then, there is some absolute constant $\bar{C}$ such that for every $T \geq 1$, $K \geq 2$, and $V_T \in \left[K^{-1}, K^{-1}T\right]$:*

$$\mathcal{R}^{\pi}(\mathcal{V}, T) \ \leq\ \bar{C}\,(K \log K \cdot V_T)^{1/3}\, T^{2/3}.$$

Theorem 3.2 is obtained by establishing a connection between the regret relative to the single best action in the adversarial setting, and the regret with respect to the dynamic oracle in non-stationary stochastic setting with variation budget. Several classes of policies, such as exponential-weight policies (including Exp3) and polynomial-weight policies, have been shown to achieve regret of order $\sqrt{T}$ with respect to the single best action in the adversarial setting (see Auer, Cesa-Bianchi, Freund and Schapire (2002) and chapter 6 of Cesa-Bianchi and Lugosi (2006) for a review). While in general these policies tend to perform well numerically, there is no guarantee for its performance with respect to the dynamic oracle studied here (see also Hartland et al. (2006) for a study of the empirical performance of one class of algorithms), since the single best action itself may incur linear (with respect to T) regret relative to the dynamic oracle. The proof of Theorem 3.2 shows that *any* policy that achieves regret of order $\sqrt{T}$ with respect to the single best action in the adversarial setting, can be used as a subroutine to obtain near-optimal performance with respect to the dynamic oracle in our setting.

Rexp3 emphasizes the two tradeoffs discussed in the previous section. The first tradeoff, information acquisition versus capitalizing on existing information, is captured by the subroutine policy Exp3. In fact, any policy that achieves a good performance compared to a single best action benchmark in the adversarial setting must balance exploration and exploitation, and therefore the loss incurred by experimenting on sub-optimal arms is indeed balanced with the gain of better estimation of expected rewards. The second tradeoff, "remembering" versus "forgetting," is captured by restarting Exp3 and forgetting any acquired information every Δ_T pulls. Thus, old information that may slow down the adaptation to the changing environment is being discarded.

Taking Theorem 3.1 and Theorem 3.2 together, we have characterized the minimax regret (up to a multiplicative factor, logarithmic in the number of arms) in a full spectrum of variations V_T:

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

Hence, we have quantified the impact of the extent of change in the environment on the best achievable performance in this broad class of problems. For example, for the case in which $V_T = C \cdot T^{\beta}$, for some absolute constant C and $0 \leq \beta < 1$ the best achievable regret is of order $T^{(2+\beta)/3}$.

3.3.1 Numerical Results

We illustrate the upper bound on the regret by a numerical experiment that measures the average regret that is incurred by Rexp3, in the presence of changing environments.

Setup. We consider instances where two arms are available: $\mathcal{K} = \{1, 2\}$. The reward X_t^k associated with arm k at epoch t has a Bernoulli distribution with a changing expectation μ_t^k:

$$X_t^k = \begin{cases} 1 & \text{w.p. } \mu_t^k \\ 0 & \text{w.p. } 1 - \mu_t^k \end{cases}$$

for all $t = 1, \ldots, T$, and for any pulled arm $k \in \mathcal{K}$. The evolution patterns of μ_t^k, $k \in \mathcal{K}$ will be specified below. At each epoch $t \in \mathcal{T}$ the policy selects an arm $k \in \mathcal{K}$. Then, the binary rewards are generated, and X_t^k is observed. The pointwise regret that is incurred at epoch t is $X_t^k - X_t^{k_t^*}$, where $k_t^* = \arg\max_{k \in \mathcal{K}} \mu_t^k$. We note that while the pointwise regret at epoch t is not necessarily positive, its expectation is. Summing over the whole horizon and replicating 20,000 times for each instance of changing rewards, the average regret approximates the expected regret compared to the dynamic oracle.

First stage (Fixed variation, different time horizons). The objective of the first part of the simulation is to measure the growth rate of the average regret incurred by the policy, as a function of the horizon length, under a fixed variation budget. We use two basic instances. In the first instance (displayed on the left side of Figure 3.1) the expected rewards are sinusoidal:

$$\mu_t^1 = \frac{1}{2} + \frac{1}{2} \sin\left(\frac{V_T \pi t}{T}\right), \qquad \mu_t^2 = \frac{1}{2} + \frac{1}{2} \sin\left(\frac{V_T \pi t}{T} + \pi\right)$$

for all $t = 1, \ldots, T$. In the second instance (depicted on the right side of Figure 3.1) similar sinusoidal evolution of the expected reward is "compressed" into the first third of the horizon, where in the rest of the horizon the expected rewards remain fixed:

$$\mu_t^1 = \begin{cases} \frac{1}{2} + \frac{1}{2}\sin\left(\frac{3V_T \pi t}{T} + \frac{\pi}{2}\right) & \text{if } t < \frac{T}{3} \\ 0 & \text{otherwise} \end{cases} \qquad \mu_t^2 = \begin{cases} \frac{1}{2} + \frac{1}{2}\sin\left(\frac{3V_T \pi t}{T} - \frac{\pi}{2}\right) & \text{if } t < \frac{T}{3} \\ 1 & \text{otherwise} \end{cases}$$

for all $t = 1, \ldots, T$. Both instances describe different changing environments under the same (fixed) variation budget $V_T = 3$. While in the first instance the variation budget is spent throughout the whole horizon, in the second one the same variation budget is spent only over the first third of the horizon. For different values of T (between 3000 and 40000) and for both variation instances we estimated the regret through 20,000 replications (the average performance trajectory of Rexp3 for $T = 5000$ is depicted in the upper-left and upper-right plots of Figure 3.3).

Discussion of the first stage. The first part of the simulation illustrates the decision process of the policy, as well as the order $T^{2/3}$ growth rate of the regret. The upper parts of Figure 3.3 describe the performance trajectory of the policy. One may observe that the policy identifies the arm with the higher expected rewards, and selects it with higher probability. The Rexp3 policy adjusts to changes in the expected rewards and updates the probabilities of selecting each arm according to the received rewards. While the policy adapts quickly to the changes in the expected rewards (and in the identity of the "better" arm), it keeps experimenting with the sub-optimal arm (the policy's trajectory doesn't reach the one of the dynamic oracle). The Rexp3 policy balances the remembering-forgetting tradeoff using the restarting points, occurring every Δ_T epochs. The exploration-exploitation tradeoff is balanced throughout each batch by the subroutine policy Exp3. While Exp3 explores at an order of $\sqrt{\Delta_T}$ epochs in each batch, restarting it every Δ_T (V_T is fixed, therefore one has an order of $T^{1/3}$ batches, each batch with an order of $T^{2/3}$ epochs) yields an exploration rate of order $T^{2/3}$.

The lower-left and lower-right parts in Figure 3.3 show plots of the natural logarithm of the averaged regret as a function of the natural logarithm of the the horizon length. All the standard errors of the data points in these log-log plots are lower than 0.004. These plots detail the linear dependence between the natural logarithm of the averaged regret, and the natural logarithm of T. In both cases the slope of the linear fit for increasing values of T supports the $T^{2/3}$ dependence of the minimax regret.

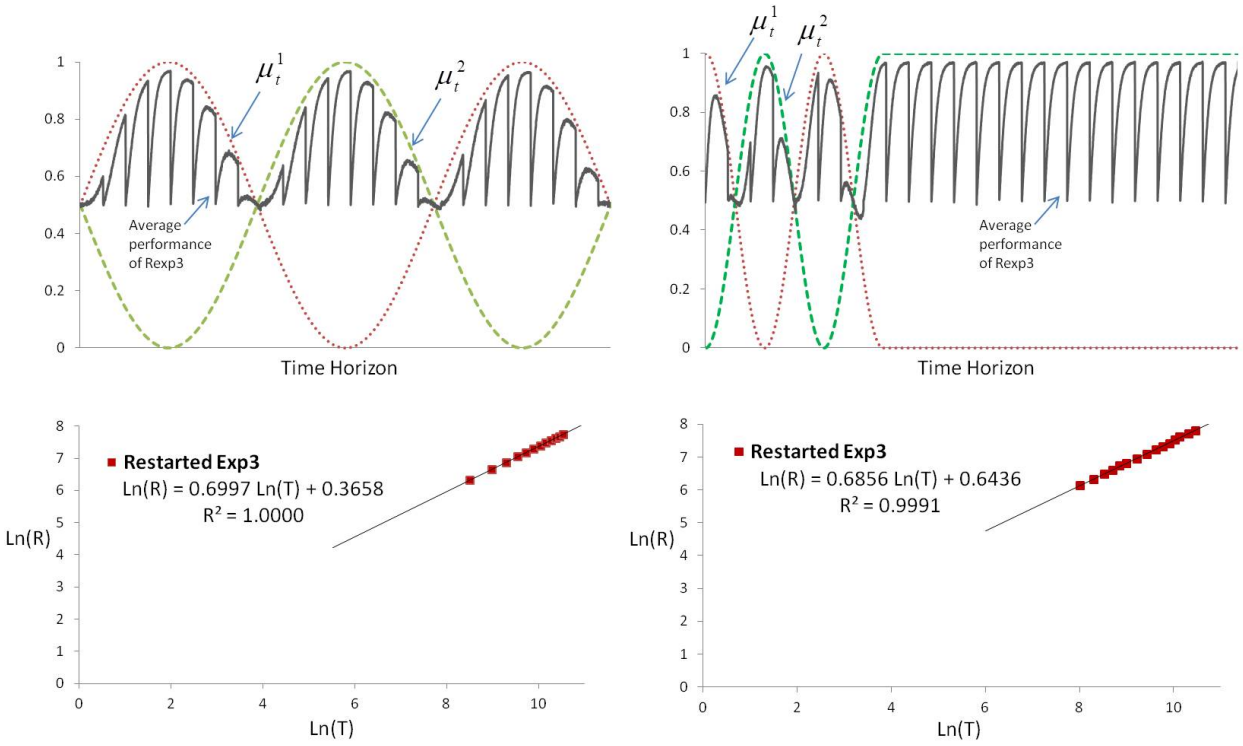
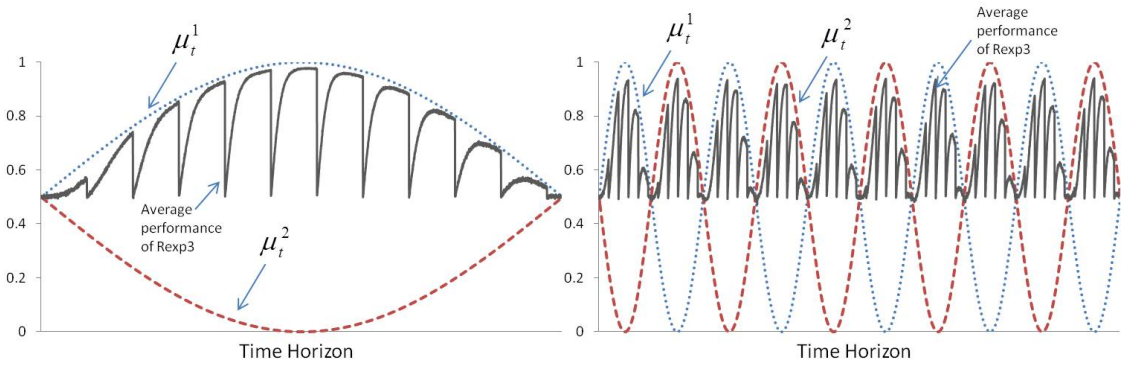


Figure 3.3: Numerical simulation of the performance of Rexp3, in two complementary instances: (*Upper left*) The average performance trajectory in the presence of sinusoidal expected rewards, with a fixed variation budget $V_T = 3$. (*Upper right*) The average performance trajectory under an instance in which the same variation budget is "spent" only over the first third of the horizon. In both of the instances the average performance trajectory of the policy is generated along $T = 5,000$ epochs. (*Bottom*) Log-log plots of the averaged regret as a function of the horizon length T.

Second stage (Increasing the variation). The objective of the second part of the simulation is to measure how the growth rate of the averaged regret (as a function of T) established in the first part changes when the variation increases. For this purpose we used a variation budget of the form $V_T = 3T^{\beta}$. Using first instance of sinusoidal variation, we repeated the first step for different values of β between 0 (implying a constant variation, that was simulated at the first stage) and 1 (implying linear variation). The upper plots of Figure 3.4 depicts the average performance trajectories of the Rexp3 policy under different variation budgets. The different slopes, representing different growth rate of the regret for different values of β appear in the table and the plot, at the bottom of Figure 3.4.



β Value	Estimated Slope
0.0	0.6997
0.1	0.7558
0.2	0.7915
0.3	0.8421
0.4	0.8801
0.5	0.9210
0.6	0.9520
0.7	0.9813
0.8	0.9942
0.9	1.0036

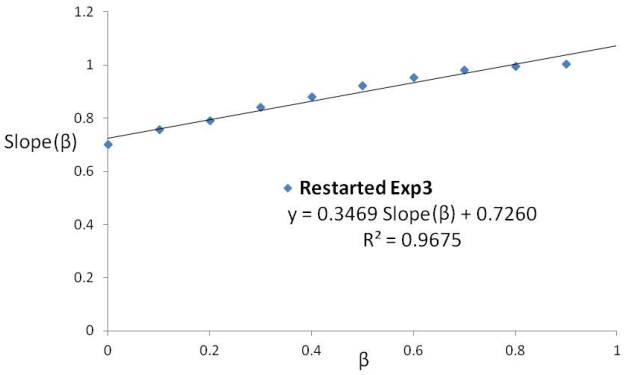


Figure 3.4: Variation and performance: (*Upper left*) The averaged performance trajectory for $V_T = 1$, and $T = 5000$. (*Upper right*) The averaged performance trajectory for $V_T = 10$, and $T = 5000$. (*Bottom*) The slope of the linear fit between the data points of Table 3.1 imply the growth rate $V_T^{1/3}$.

Discussion of the second stage. The second part of the simulation illustrates the way variation affects the policy decision process and the minimax regret. Since Δ_T is of order $(T/V_T)^{2/3}$, holding T fixed and increasing V_T affects the decision process and in particular the batch size of the policy. This is illustrated at the top plots of Figure 3.4. The slopes that were estimated for each β value (in the variation structure $V_T = 3T^\beta$) ranging from 0.1 to 0.9 describing the linear log-log dependencies (the case of $\beta = 0.0$ is already depicted at the bottom-left plot in Figure 3.3) are summarized in Table 3.1. The bottom part of Figure 3.4 show the slope of the linear fit between the data points of Table 3.1, illustrates the growth rate of the regret when the variation (as a function of T) increases, supports the $V_T^{1/3}$ dependence of the minimax regret, and emphasizes the full spectrum of minimax regret rates (of order $V_T^{1/3}T^{2/3}$) that are obtained for different variation levels.

β value	Estimated slope
0.0	0.6997
0.1	0.7558
0.2	0.7915
0.3	0.8421
0.4	0.8801
0.5	0.9210
0.6	0.9519
0.7	0.9813
0.8	0.9942
0.9	1.0036

Table 3.1: **Estimated slopes for growing variation budgets.** The estimated log-log slopes obtained for different β values in the variation structure $V_T = 3T^{\beta}$.

3.4 Concluding remarks

A continuous near-optimal policy. To achieve rate-optimal regret rate one may use the restarting procedure with any policy which is rate optimal in the adversarial setting relative to the static oracle as a subroutine. Nevertheless, it is notable that one may adopt rate optimal policies from the adversarial setting to obtain near optimal regret rate in a continuous fashion (without restarting). To illustrate this, we use the Exp3.S policy, provided in Auer et al. (2002).

Policy Exp3.S. Inputs: positive numbers γ, α.

1. Initialization: for any $k \in \mathcal{K}$ set $w_1^k = 1$
2. Loop: for each $t = 1, 2, \ldots$
   - Set
   $$p_t^k = (1-\gamma)\frac{w_t^k}{\sum_{k'=1}^K w_t^{k'}} + \frac{\gamma}{K} \quad \text{for all } k \in \mathcal{K}$$
   - Draw an arm k' from $\mathcal{K}$ according to the distribution $\left\{p_t^k\right\}_{k=1}^K$
   - Receive a reward $X_t^{k'}$

- For the drawn arm k', set $\hat{X}_t^{k'} = X_t^{k'}/p_t^{k'}$, and for any $k \neq k'$ set $\hat{X}_t^k = 0$. Then, for all $k \in \mathcal{K}$ update:

$$w_{t+1}^k = w_t^k \exp\left\{\frac{\gamma \hat{X}_t^k}{K}\right\} + \frac{e\alpha}{K}\sum_{k=1}^{K} w_i^k$$

3. Repeat (2) until there are no more pulls

Exp3.S is itself an adaptation of the Exp3 policy. Given a finite number S of times in which the identity of the best arm changes, this adaptation allows it, using the right tuning parameters ($\alpha \sim 1/T$, $\gamma \sim \sqrt{SK/T}$), to achieve regret of order $S\sqrt{KT\log(KT)}$ compared to a dynamic benchmark (Theorem 8.1 in Auer, Cesa-Bianchi, Freund and Schapire 2002). Nevertheless, this policy can be further adopted to achieve a near optimal performance compared to the dynamic oracle in our setting.

Theorem 3.3. *Let π be the Exp3.S policy with $\alpha = \frac{1}{T}$, and $\gamma_1 = \min\left\{1\ ,\ \sqrt[3]{\frac{2V_T K\log(KT)}{(e-1)^2 T}}\right\}$. Then, there exists some absolute constant $\bar{C}$ such that for every $T \geq 1$, $K \geq 2$, and $TK^{-1} \geq V_T \geq K^{-1}$:*

$$\mathcal{R}^{\pi}(\mathcal{V}, T) \leq \bar{C}\left(V_T K \log(KT)\right)^{1/3} \cdot T^{2/3}.$$

When Exp3.S is tuned by α and γ_1, it achieves the minimax regret rate up to a logarithmic factor. Nevertheless, simulating the policy's performance in several instances did not show any observable difference in the growth rate of the regret compared to the restarting procedure, with the Exp3 as a subroutine. Figure 3.5 shows the average performance trajectory of the tuned Exp3.S under the variation instances that where used in the first stage of the simulation described in §3.3.1.

Nevertheless, while the restarting procedure can be used as a "black box" mechanism to adopt policies from the adversarial setting, and requires no knowledge of the policy other than the regret rate it guarantees compared to the single best action, a continuous (epoch-by-epoch) adoption of a policy is done by changing the policy, its parametric values, or both. Therefore, it requires technical knowledge about the policy that is not required by the restarting procedure.

Knowledge of problem parameters. We have characterized the minimax regret for different non-stationary MAB environments as a functions of the number of arms K, the variation budget V_T, and the horizon length T. In that respect, the tuning parameter γ_0 (of Exp3) used by in

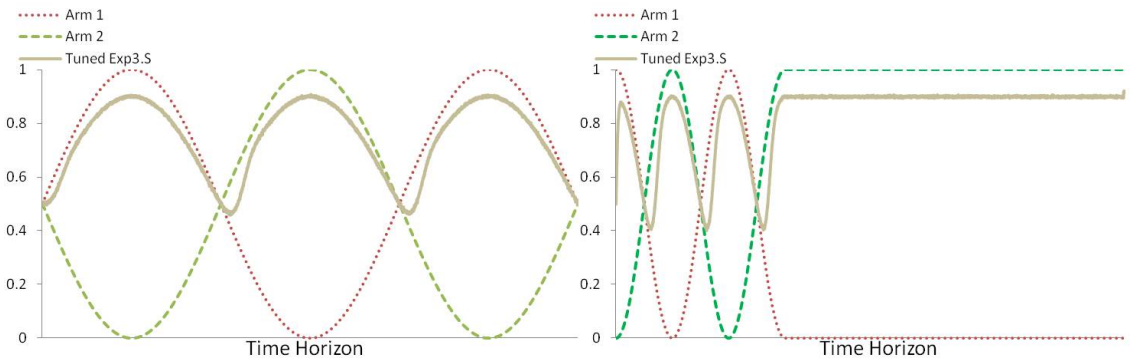


Figure 3.5: Average performance trajectories of Exp3.S(α, γ_1). (*Left*) Time-homogenous variation instance. (*Right*) Time-heterogenous variation instance. In both instance $T = 5000$, and $V_T = 3$.

Theorem 3.2, and the parameters α and γ_1 (of Exp3.S) used by in Theorem 3.3 require knowledge of T, K and V_T. While K is typically known, the number of pulls T and the variation budget V_T may be unknown. It is in general possible to adjust for the lack of knowledge of T by a classical "doubling trick" (The proof of Theorem 3.3 ends with a procedure that uses Exp3.S as a subroutine to achieve the same order of regret when T is unknown). However, V_T, and specifically the dependence of V_T in T needs to be known. One way to estimate V_T from historical data, given $\tilde{T}$ data points for each arm (when such historical data about the rewards generated by all arms is available), is to assume the structure $V_T = T^\beta$ and regressing $\log V_T$ on $\log \tilde{T}$ to recover an estimator $\hat{\beta}$ from the regression slope. Nevertheless, it remains an open problem to design a policy that can adjust to the extent of variation in an online fashion.

Contrasting with traditional (stationary) MAB problems. The tight bounds that were established on the minimax regret in our stochastic non-stationary MAB problem allows one to quantify the "price of non-stationarity," which mathematically captures the added complexity embedded in changing rewards versus stationary ones. While Theorem 3.1 and Theorem 3.2 together characterize minimax regret of order $V_T^{1/3}T^{2/3}$, the characterized minimax regret in the stationary stochastic setting is of order $\log T$ in the case where rewards are guaranteed to be "well separated" one from the other, and of order $\sqrt{T}$ when expected rewards can be arbitrarily close to each other (see Lai and Robbins (1985) and Auer, Cesa-Bianchi and Fischer (2002) for more details). Contrasting the different regret growth rates quantifies the "price," in terms of best

achievable performance, of non-stationary rewards compared to stationary ones, as a function of the variation that is allowed in the non-stationary case. Clearly, this comparison shows that additional complexity is introduced even when the allowed variation is fixed and independent of the horizon length.

Contrasting with other non-stationary MAB instances. The class of MAB problems with non-stationary rewards that is formulated in the current chapter extends other MAB formulations that allow rewards to change in a more structured manner. We already discussed in Remark 3.1 the consistency of our results (in the case where the variation budget grows linearly with the time horizon) with the setting treated in Slivkins and Upfal (2008) where reward evolve according to a Brownian motion and hence the regret is linear in T. Two other representative studies are those of Garivier and Moulines (2011), that study a stochastic MAB problems in which expected rewards may change a finite number of times, and Auer, Cesa-Bianchi, Freund and Schapire (2002) that formulate an adversarial MAB problem in which the identity of the best arm may change a finite number of times. Both studies suggest policies that, utilizing the prior knowledge that the number of changes must be finite, achieve regret of order $\sqrt{T}$ relative to the best sequence of actions. However, the performance of these policies can deteriorate to regret that is linear in T when the number of changes is allowed to depend on T. When there is a finite variation (V_T is fixed and independent of T) but not necessarily a finite number of changes, we establish that the best achievable performance deteriorate to regret of order $T^{2/3}$. In that respect, it is not surprising that the "hard case" used to establish the lower bound in Theorem 3.1 describes a nature's strategy that allocates the allowed variation over a large (increasing function of T) number of changes in the expected rewards.

Estimation in a changing environment. Our work demonstrates the effect a changing environment has on the exploration-exploitation balance, and on the incurred regret. When estimating vector of fixed expected rewards by T noisy observations that are iid, the calculated estimators have a stochastic error term (a) of order $1/\sqrt{T}$, that stems from estimating with noisy observations. The way in which exploration affects the quality of the estimators is clear: the longer we experiment the smaller this stochastic term turns to be, and the better our estimator gets (See Figure 3.6). However, when the true values of the expected rewards evolve, and observa-

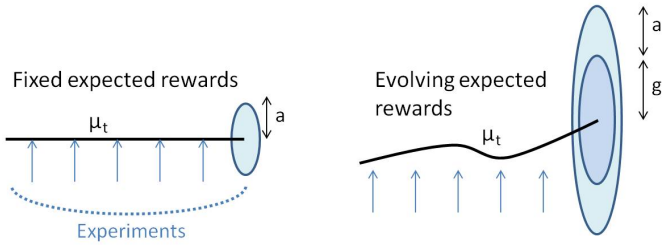


Figure 3.6: (*Left*) Estimating a fixed expected rewards: stochastic error term a which is decreasing with T the number of observations. (*Right*) Estimating evolving expected rewards: stochastic error term a which is decreasing with T and deterministic error term g which is increasing with T

.

tions are not identically distributed anymore. To the stochastic error term (a) we have to add a deterministic error term (g) that of order V_T that stems from the dynamic environment, and reflects the possible way in which expected rewards may change. In addition to the introduced "remembering" versus "forgetting" tradeoff, the exploration-exploitation balance may be affected as well. The tension between these two error terms is illustrated on the right side of Figure 3.6: Intuitively, focusing on exploration, the decision-maker would like to minimize the surface of the larger ellipsoid, considering both (a) and (g).

Chapter 4

Optimization in Online Content Recommendation Services

> The material presented in this chapter is based on Besbes, Gur and Zeevi (2014*c*). It is based on a collaboration with Outbrain, a leading provider of customized content recommendations to online publishers. Outbrain's recommendations appear in over 100,000 media sites, including many premium online publishers. These recommendations exhibit extraordinary exposure, and millions of articles are read on a daily basis via Outbrain's recommendations. The research in this chapter is based on a large data set (of 15 tera-bytes), including billions of recommendations generated for various online publishers, as well as data from a controlled experiment.

This chapter studies *online content recommendations*, a new class of online services allows publishers to direct readers from articles they are currently reading to other web-based content they may be interested in. In §4.1 we formulate and provide a diagnostic of the content recommendation problem. In §4.2, using a large data set of browsing history at major media sites, we develop a representation of content along two key dimensions: *clickability*, the likelihood to *click to* an article when it is recommended; and *engageability*, the likelihood to *click from* an article when it hosts a recommendation. Based on this representation, in §4.3 we propose a class of user path-focused heuristics, whose purpose is to simultaneously ensure a high instantaneous probability of clicking recommended articles, while also optimizing engagement along the future path. Using a

simulation study as well as supporting theoretical bounds, we rigorously quantify the gap between performance of the optimal recommendation policy and the one of myopic policies that are used in practice, and estimate the fraction of this gap that may be captured by our one-step look-ahead heuristic. To validate the impact of our proposed heuristics, we study in §4.4 an implementation (a controlled "live" experiment) of a practical class of one-step look-ahead recommendations, and studies its impact relative to current practice. In §4.5 we provide some concluding remarks. Auxiliary results as well as details on the estimation procedure can be found in Appendix C.

4.1 The content recommendation problem

The content recommendation problem (CRP) is faced by the recommendation service provider when a reader arrives to some initial (landing) article, typically by clicking on a link placed on the front page of the publisher. Then, the provider needs to plan a schedule of T recommendations (each recommendation being an assortment of links) to show the reader along the stages of her visit, according to the path the reader takes by clicking on recommended links. The reader can terminate the service at any stage, by leaving the current article she reads through any mean other than clicking on a content recommendation (e.g., closing the window, clicking on a link which is not a content recommendation, or typing a new URL). The objective of the provider is to plan a schedule of recommendations to maximize the value generated by clicks along the path of the reader before she terminates the services.

A model for recommending content. The CRP is formalized as follows. Let $1, \ldots, T$ be the horizon of the CRP throughout a visit of a single reader. We denote by ℓ the number of links that are introduced in each recommendation. We denote by x_{t-1} the article that hosts the recommendation at epoch t (for example, x_0 denotes the article that hosts the recommendation at $t = 1$; the article that the reader starts her journey from). We denote by $\mathcal{X}_t$ the set of articles that are available to be recommended at epoch t. $\mathcal{X}_0$ is the initial set of available articles, and we assume this set is updated at each epoch by $\mathcal{X}_t = \mathcal{X}_{t-1} \setminus \{x_{t-1}\}$ (for example, at $t = 1$ all the articles that are initially available can be recommended, except for x_0, that hosts the first recommendation). We assume $T \leq |\mathcal{X}_0| - \ell$.

We denote by $\mathcal{U}$ the set of reader (user) types. We denote by u_0 the initial type of the reader. This type may account for various ways by which a reader can be characterized, such as geographical location, as well as her reading and clicking history. We assume the type of the reader to be updated at each epoch according to $u_t = f_t(u_{t-1}, x_{t-1})$. This update may account for articles being read, as well as for epoch-dependent effects such as fatigue (for example, u_1, the type at $t = 1$, may account for the initial type u_0, the initial article x_0, and the fact that the reader sees the recommendation after she already read one article). We do not specify here a concrete structure of the functions $f_t(\cdot, \cdot)$; a special case of this update rule will be introduced and used in §4.2.1.

A recommendation assortment is an ordered list of ℓ links to articles that are available for recommendation. We denote by $\mathcal{A}^\ell(\mathcal{X}_t)$ the set of all possible assortments at epoch t. At each epoch $t = 1, \ldots, T$ the recommendation provider selects a recommendation assortment $A_t \in \mathcal{A}^\ell(\mathcal{X}_t)$ to present the reader with. For a given user type u, a host article x and a recommendation assortment A, we denote by $\mathbb{P}_{u,x,y}(A)$ the click probability to any article $y \in A$. With some abuse of notation we sometimes denote assortments as sets of links, and note that the probability to click on a link that belongs to an assortment depends on all the links in the assortment as well as on the way they are ordered.[1] Finally, we denote by $w(x)$ the value (for the service provider) generated by a click on article x.

The structure described above assumes Markovian dynamics, that are used in the following. Given an initial reader type u_0, an initial set of articles $\mathcal{X}_0$, and a host article x_0, the CRP of maximizing the value generated by clicks throughout the visit is defined by the following Bellman equations:[2]

$$V_t^*(u_t, \mathcal{X}_t, x_{t-1}) \;=\; \max_{A \in \mathcal{A}^\ell(\mathcal{X}_t)} \left\{ \sum_{x_t \in A} \mathbb{P}_{u_t, x_{t-1}, x_t}(A) \left(w(x_t) + V_{t+1}^*(u_{t+1}, \mathcal{X}_{t+1}, x_t) \right) \right\}, \tag{4.1}$$

[1] Therefore, $y \in A$ and $y \in A'$ does *not* imply $\mathbb{P}_{u,x,y}(A) = \mathbb{P}_{u,x,y}(A')$. Similarly, A and A' containing the same articles does *not* imply $\sum_{y \in A} \mathbb{P}_{u,x,y}(A) = \sum_{y \in A'} \mathbb{P}_{u,x,y}(A')$, as articles may be ordered differently.

[2] We assume that the value of clicking on each article is known to the provider. This value can represent actual revenue (in the case of sponsored links), or tangible value (in the case of organic links that drive publishers to partner with the provider). While in practice there may be constrains on the number of organic/sponsored links, in our model we only limit the overall number of links in each assortment.

for $t = 1, \ldots, T$, where $V^*_{T+1}(u_{T+1}, \mathcal{X}_{T+1}, x_T) = 0$ for all $u_{T+1}, \mathcal{X}_{T+1}$, and x_T. Since the CRP accounts for the future path of readers, the computational complexity that is associated with finding its optimal solution increases rapidly when the set of available articles gets large.

Theoretical observation 1. *The content recommendation problem defined by (4.1) is NP-hard.*

For further details see Proposition C.1 in Appendix B.1; we establish that the Hamiltonian path problem, a known NP-hard problem (Gary and Johnson 1979), can be reduced to a special case of the CRP, and therefore, even when the click probabilities between hosting articles and recommended articles are known for each arriving reader, the CRP is NP-hard.[3] Given the large amount of available articles, and the high volume of reader arrivals, this result implies that it is impractical for the service provider to look for an optimal solution for the CRP for each arriving reader. This motivates the introduction of customized recommendation algorithms that, although lacking performance guarantees for arbitrary problem instances, perform well empirically given the special structure of the problem at hand.

The myopic heuristic. One class of such algorithms is the one used in current practice, with the objective of recommending at each epoch t (until the reader terminates the service) an assortment of links that maximizes the instantaneous performance in the current step, without accounting for the future path of the reader. We refer to this approach as the *myopic* content recommendation problem (MCRP), and formally define the value associated with it by:

$$V^m_t(u_t, \mathcal{X}_t, x_{t-1}) \;=\; \sum_{x_t \in A^m_t} \mathbb{P}_{u_t, x_{t-1}, x_t}(A^m_t)\left(w(x_t) + V^m_t(u_{t+1}, \mathcal{X}_{t+1}, x_t)\right); \quad t = 1, \ldots, T, \qquad (4.2)$$

where

$$A^m_t \in \underset{A \in \mathcal{A}^\ell(\mathcal{X}_t)}{\arg\max} \left\{ \sum_{x_t \in A} \mathbb{P}_{u_t, x_{t-1}, x_t}(A) w(x_t) \right\}; \quad t = 1, \ldots, T,$$

and where $V^m_{T+1}(u_{T+1}, \mathcal{X}_{T+1}, x_T) = 0$ for all u_{T+1}, $\mathcal{X}_{T+1}$, and x_T. The MCRP can be solved at each epoch separately, based on the current host article, reader type, and set of available articles (where the host article is the one that was clicked at the previous epoch).

[3]Various relaxation methods as well as approximation algorithms that have been suggested in order to deal with the intractability of the Hamiltonian path problem appear in Uehara and Uno (2005), and Karger et al. (1997).

The sub-optimality of myopic recommendations. While recommending articles myopically is a practical approach that is currently implemented in content recommendation services, simple problem instances reveal that myopic recommendations may generate poor performance compared to the optimal schedule of recommendations. In one such instance that is depicted in Figure 4.1, myopic recommendations generate only two thirds of the expected clicks generated by optimal recommendations. While Figure 4.1 provides a single instance in which there is a

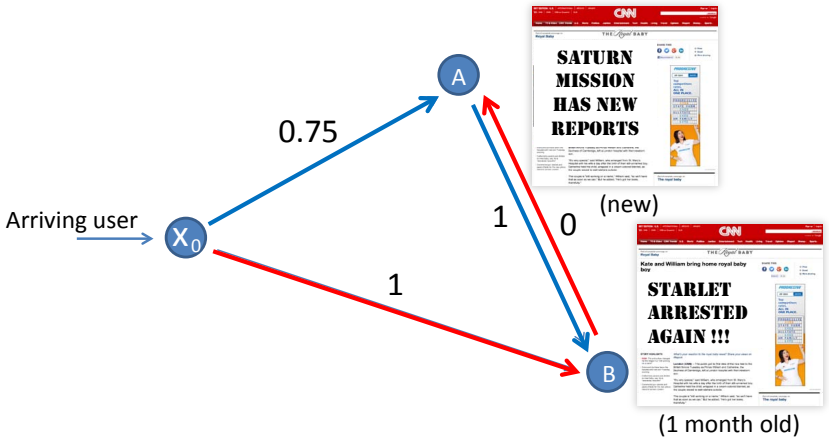


Figure 4.1: **Sub-optimality of myopic recommendations.** A content recommendation instance, with $\ell = 1$ (single-link assortments), $T = 2$, $\mathcal{X}_0 = \{x_0, A, B\}$, and uniform click values. x_0 is the initial host. The click probabilities (accounting for evolution of user type and available article set) illustrate a scenario where article B has attractive title but irrelevant content that drives users to terminate the service. Myopic schedule first recommends B and then A, generating a single click. An optimal schedule first recommends A and then B, generating $0.75 + 0.75 \times 1 = 1.5$ expected clicks.

significant gap between the performance of myopic recommendations and that of optimal recommendations, such a performance gap appears in many simple instances. Moreover, theoretically, this performance gap can be very large.

Theoretical observation 2. *The performance gap between myopic recommendations and optimal recommendations can be arbitrarily large when the set of articles is large.*

For a precise statement and details see Proposition C.2 in Appendix B.1.

Empirical insights. While a descriptive discussion on the available data is deferred to §4.2, we wish to bring forward at this point a preliminary empirical observation that supports the existence of such a performance gap in practice, and to verify our premise that the content recommendation service is more than a one-click service. We construct the visit paths of readers, from arrival to some host article through a sequence of clicks (if such clicks take place) on internal links. The distribution of clicks along visit steps in two representative media sites is shown on the left part of Figure 4.2. We observe that a significant portion of the service is provided *along* the

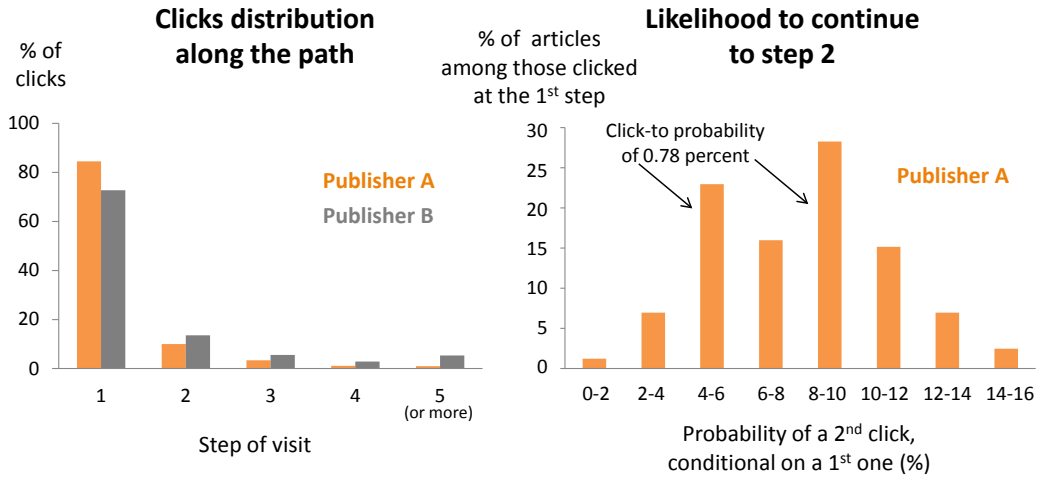


Figure 4.2: **Aggregate analysis of clicks along the visit path.** (*Left*) The distribution of clicks along visit steps in two representative media sites (A and B). (*Right*) The distribution of the probability to click again among articles to which readers arrived after their first click (in media site A).

visit path: between 15% and 30% of clicks were generated in advanced stages of the visit, namely *after* the first click (this range is representative of other publishers as well).[4] Next, consider the set of links that were clicked in the first step. The right part of Figure 4.2 depicts the distribution of the probability to click on a recommendation again, *from* articles to which readers arrived after clicking at the first step. While this conditional probability is relatively high for some articles, it is relatively low for others. This points out that the recommendation that is selected at the first step clearly affects clicks generated along the future path of readers. Moreover, we observe that

[4]It is important to note that the portion of clicks along the path is significant even though the future path is not accounted by the first recommendation. One can only expect this portion of the business volume to grow if recommendations account for the future path of readers.

the average CTR *to* articles along this distribution is similar, ranging from 0.7 to 0.85 percent (in particular, the probability to click to articles at both the third and the fifth columns is ~ 0.78 percent). This suggests that myopic recommendations might be "leaving clicks on the table."

This observation leads to the following question: are some articles more "engaging" than others, in the sense that readers are more likely to continue the service after clicking to these articles? In the next section we will see that "engageability" is indeed an important characteristic of articles, and is a significant click driver along the path of readers.

4.2 Identifying click drivers along a visit

In this section we estimate a choice model that aims to capture click drivers using a large data set, in a manner that accounts for potential changes in articles' features along time. Our model leads to a new representation of the value of articles along two key dimensions: clickability and engageability. We begin by describing our data set.

Available (and unavailable) data. Our database includes access to over 5 billion internal recommendations that were shown on various media sites, including anonymous information about articles, readers, recommendations, as well as observed click/no-click feedback. Every article that was visited or recommended in the database has a unique id. This id is linked to the publish date of the article, and the main topic of the article, which is classified into 84 topic categories (for example, representative categories include "sports: tennis," "entertainment: celebrities," and "health: fitness"). Every event of a reader arriving to an article is associated with a unique recommendation id, reader id, and host article id. Each recommendation id is linked to:

- the list of internal articles that were recommended (ordered by position),
- the time at which the recommendation was created,
- the time spent by the reader on the host article before leaving it (for some media sites),
- the recommendation id that the reader clicked on to arrive to the current page (if the reader arrived by clicking on an internal Outbrain recommendation).

Our data does not include information about additional article features such as length, appearance of figures/pictures, links presented in the article, or display ads. We also do not have access to

the sponsored links that were shown, nor to clicks on them.

Main click drivers along the path. As the recommendation service aims to suggest attractive links, one of the important parameters on which recommendation algorithms focus is the id of recommended articles. Other potential drivers include the position of links within recommended assortments, the topics of candidate articles, and the extent to which a user is familiar with the service.[5] These elements are typically taken into account throughout the recommendation process. In what follows we add a new click driver that has been overlooked so far: the id of the article that *hosts* the recommendation. While content recommendations aim to match readers with attractive links, the id of the host article describes the environment in which this matching is taking place (recommendations are placed at the bottom of host articles), and therefore potentially impacts the likelihood to click on a recommendation.

4.2.1 Choice model

To capture main key drivers, we propose to estimate a multinomial logit model. Given a reader type $u \in \mathcal{U}$ and an assortment A that is placed at the bottom of a host article x, we define:

$$\mathbb{P}_{u,x,y}(A) \;=\; \begin{cases} \frac{\phi_{u,x,y}(A)}{1+\sum_{y' \in A} \phi_{u,x,y'}(A)} & \text{if } y \text{ appears in } A \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Whenever y appears in A, we define:

$$\phi_{u,x,y}(A) \;=\; \exp\left\{\alpha + \beta_x + \gamma_y + \mu_{x,y} + \theta_u + \lambda_{p(y,A)}\right\}, \tag{4.4}$$

where $p : (y, A) \to \{0, \ldots, 5\}$ denotes the *position* of article y in the assortment ($p(y, A) = 0$ implies that y is placed at the highest position in A). Using the model above we aim to show that there is potential value in accounting for the host effect (β), in addition to the link effect (γ) on which current recommendations focus. As we further discuss in §5, this model is selected with the prospect of implementing practical algorithms that account for the host effect (in addition to the link effect) of candidate articles via proxies that are observable in real time throughout

[5]Due to technical reasons (for example, the content recommendation service is not subscription-based), information on the preferences of readers is typically limited (and in particular, does not appear in our data set).

the recommendation process. In what follows we focus on the pair of parameters γ and β that describe each article, but first we briefly discuss the control parameters (a thorough description of these parameters is given in Appendix C.2).

$\lambda_1 \ldots, \lambda_5$ are position parameters that capture the effect of different positions links may take within the recommended assortment, compared to the highest position. θ_u is a dummy variable that separates readers that are familiar with internal content recommendations from users that are not.[6] $\mu_{x,y}$ is the contextual relation between the host article and the recommended ones. We use a single parameter that flags cases in which the recommended article directly relates to the topic discussed in the host article.[7]

Host effect (engageability). The parameter β is associated with the likelihood to *click from* an article whenever it hosts a recommendation, and is driven by the actual content of the article. We refer to β as the *engageability* of an article. Under our model, the engageability of an article may account for two potentially different effects. The first one is "homogeneous" with respect to all recommended links that may be placed at the bottom of it. Intuitively, when an article is well-written, interesting, relevant, and of reasonable length, the reader is more likely to read through it, arrive to the recommendation at the bottom of it in an engaging mood, as well as more likely to click on a recommendation. On the other hand, when content is poor or irrelevant, a reader is more likely to terminate the service rather than scrolling down reading the article, and therefore she is less likely to see the recommendation and click on it. Engageability of an article may also capture in an aggregate manner an effect which is "heterogeneous" with respect to different potential links: the extent to which it encourages readers to continue and read specific other articles that may be recommended.[8] We note that the engageability of a given article may

[6] As described in Appendix C.2, we use the first 10 days of the data to identify experienced readers. Then, during the 30 days over which the model was estimated we update reader types from "inexperienced" to "experienced" once they click on an internal recommendation. This update rule is a special case of the one given in §4.1: $u_0 \in \{u_{exp}, u_{inexp}\}$ is set according to whether or not the reader has clicked a link in the first 10 days. Whenever $u_0 = u_{exp}$ one has $u_t = u_0$ for all t, and whenever $u_0 = u_{inexp}$ one has $u_1 = u_0$ and $u_t = u_{exp}$ for all $t \geq 2$.

[7] Approaches such as a matrix that describes relations between all combinations of article or topics are impractical for estimation, due to the limited number of observations, as well as the dynamic nature of these connections (driven by the introduction of new articles and the aging of old ones) that necessitates estimating them repeatedly.

[8] Theoretically, such connections between articles may potentially be separated from the first, "homogeneous"

change with time, along with the relevancy of its content.

Link effect (clickability). The parameter γ is associated with the likelihood to *click to* an article whenever it is recommended, and is driven by the title of the article (which is typically the only information given on the link itself). We refer to γ as the *clickability* of an article. Like engageability, the clickability of articles may change with time.

Model limitations. As discussed, our data does not include access to factors that may affect the likelihood to click on internal recommendations. These would be crucial if our objective would be to quantify the magnitude of the different effects. Instead, we aim to identify main click drivers (and in particular, the impact of engageability) by testing *out-of-sample* the ability to use the estimated model parameters to predict which assortments are eventually clicked. By doing so, in §4.2.3 we quantify the *predictive power* of the model, and by comparing it to the ones of alternative models we validate that accounting for engageability is key to maximizing the number of clicks along a visit path.

Estimation process. A description of the estimation process is given in Appendix C.2. The model was estimated using a database that includes 40 days of internal recommendations presented on articles of one media site. Since clickability and engageability of articles may be time-varying, the model was estimated independently over 360 batches of two hours. In each such batch approximately 500 different articles hosted recommendations, and a similar number of articles were recommended (approximately 90 percent of host articles were also recommended in the same batch, and vice versa). Along each batch approximately 1,000 parameters where estimated (including the control parameters). Estimation in each batch was based on approximately 2M recommendations (of 6 links each) and 100,000 clicks.

4.2.2 Content representation

We use the clickability and engageability estimates to represent articles in a two-dimensional content space. Figure 4.3 depicts the representation of articles in that space. The dimensions of

engageability effect, using a complex description of contextual relations between articles/topics, but such approaches significantly increase the number of estimated parameters and are impractical. In this study we do not aim to separate between the two, but rather focus on the value of recommending articles with higher engageability, in the sense that lead to future clicks, independently of the underlying reason for these clicks.

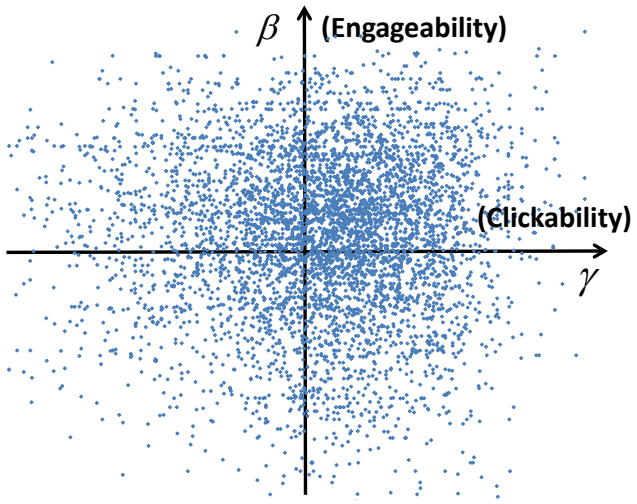


Figure 4.3: **Articles in the content space.** Every article is positioned in the content space according to its two estimates, β (engageability) and γ (clickability). The 5012 articles that appear in the plot have at least 500 appearances as a host and as a recommended link during the estimation segment. The estimated values in the figure range from -3 to 3 along both axes.

the content space have meaningful interpretation for the service provider, when examining articles as candidates for recommendations: it captures not only the likelihood to click on an article when it is recommended, but also the likelihood of readers to continue using the service if this article is indeed clicked, and thus hosts the recommendation in the following step.

One clear observation from Figure 4.3, is that engageability and clickability (and intuitively, their main drivers: the title attractiveness and the actual content) are content features that represent fundamentally different click drivers. In fact, the correlation between the two characteristics is 0.03. A potential benefit of our content representation (compared with the current practice, which focuses only on clickability/CTR), is that it allows one to differentiate between articles that have similar clickability. In particular, one may use this framework to tune recommendation algorithms to select articles that have not only high clickability (generating high instantaneous CTR), but also high engageability (guaranteeing high CTR in the next step). We note that the space of articles also allows one to study the dynamics of articles' relevancy from the time they are published and on. In §6 we further discuss this "aging process" of articles, and the way it can be tracked in terms of clickability and engageability.

We turn to specify representative classes of articles (illustrated in Figure 4.4). We refer to

articles with high clickability and high engageability as "good articles": readers are likely to click on them, and are also likely to click from them and continue the service. On the other hand, the class of "bad articles" is characterized by low clickability and low engageability. We refer to

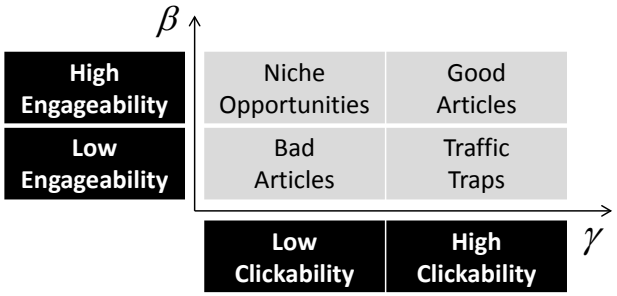


Figure 4.4: **The content matrix.**

articles with high clickability but low engageability as "traffic traps": these articles attract a lot of readers, but these readers tend to terminate the service upon arrival. Unlike bad articles, that are naturally screened out of the recommendation pool due to their low clickability, traffic traps represent a threat to the service provider: since they are being clicked on, algorithms that focus on clickability keep recommending them despite potentially poor content that harms the service performance in the short term (along the path of readers) and in the long term (decreasing the value of the service in the eyes of readers).

Finally, we refer to articles with low clickability and high engageability as "niche opportunities". Readers do not tend to click on these articles, but those who do click on them tend to continue the service afterwards. These articles often deal with more "professional" topics such as architecture, arts, fitness, and health. Interestingly, we observe that these articles tend to stay relevant (and maintain high engageability) much longer than other articles, and therefore there is a long-term opportunity in identifying them and recommending them to appropriate readers. With that in mind, the engageability dimension suggests a practical approach to separate traffic traps from good articles and niche opportunities from bad articles.

4.2.3 Validating the notion of engageability

In-sample testing. In each estimation batch we perform a likelihood ratio test with the null hypothesis being that click observations follow a link-focused model, that is a special case of the

model we describe in §3.1. The link-focused model follows the one in (4.3) with $\phi_{u,x,y}(A)$ being:

$$\phi^{lf}_{u,x,y}(A) \;=\; \exp\left\{\alpha + \gamma_y + \mu_{x,y} + \theta_u + \lambda_{p(y,A)}\right\}, \tag{4.5}$$

where the control parameters $\mu_{x,y}$, θ_u, and $\lambda_{p(y,A)}$ are defined as in §3.1. In the link-focused model engageability is always constrained to be zero. For each two-hour batch we measure

$$R = -2\ln\left[\frac{\text{likelihood for link-focused model}}{\text{likelihood for full model}}\right],$$

which is approximately distributed according to a chi-squared distribution with the number of degrees of freedom being the number of engageability parameters (which is the number of articles, roughly 500 in each batch). The obtained p-values of the various batches were all smaller than 0.05, implying the significance of host engageability. We next turn to establish a *stronger* notion of validation through out-of-sample testing and predictive analysis.

Out-of-sample testing. We use each set of estimates generated over a batch to predict click/no-click outcomes for impressions in the following batch. We test the ability to predict a click on the whole recommendation, rather than on a specific link, focusing only on impressions in which all the recommended articles were estimated in the following batch. The procedure of testing the predictive power of the model is as follows.

Testing procedure. Input: $\delta \in [0,1]$

1. For each 2-hour data batch $1 \le j \le 359$:
   (a) Estimate model parameters according to §4.2.1, using the data set of segment j.
   (b) Test predictive power in the *following* 2-hour batch: for any recommended assortment A in batch $j+1$, calculate the assortment click probability according to the estimates of batch j:
   $$\mathbb{P}_{u,x}(A) \;=\; \sum_{y\in A} \mathbb{P}_{u,x,y}(A),$$
   where $\mathbb{P}_{u,x,y}(A)$ is defined according to (4.3) and $\phi_{u,x,y}(A)$ according to (4.4). Then, classify:
   $$C_\delta(A) = \begin{cases} 1 & \text{if } \mathbb{P}_{u,x}(A) \ge \delta \\ 0 & \text{if } \mathbb{P}_{u,x}(A) < \delta \end{cases}$$
2. Using the click/no-click reader's feedback, calculate throughout the entire data horizon:

(a) the false positive rate:

$$R_\delta^{fp} \;=\; \frac{\#\left\{A:\ \text{not clicked}, C_\delta(A)=1\right\}}{\#\left\{A:\ \text{not clicked}\right\}}$$

(b) the true positive rate:

$$R_\delta^{tp} \;=\; \frac{\#\left\{A:\ \text{clicked}, C_\delta(A)=1\right\}}{\#\left\{A:\ \text{clicked}\right\}}$$

Benchmarks. We compare the predictive power of the model to those calculated for the following benchmark models.

1. *Random click probabilities.* The first one is a random classifier, in which $\mathbb{P}_{u,x}(A)$ is an independent uniform distribution over $[0,1]$.

2. *Link-focused model.* We estimated the model in (4.3) with $\phi_{u,x,y}(A)$ defined by:

$$\phi_{u,x,y}^{lf}(A) \;=\; \exp\left\{\alpha+\gamma_y+\mu_{x,y}+\theta_u+\lambda_{p(y,A)}\right\}, \tag{4.6}$$

   where the control parameters $\mu_{x,y}$, θ_u, and $\lambda_{p(y,A)}$ are defined as in §3.1.

3. *Host-focused model.* We estimated the model in (4.3) with $\phi_{u,x,y}(A)$ defined by:

$$\phi_{u,x,y}^{hf}(A) \;=\; \phi_{u,x}^{hf}(A) \;=\; \exp\left\{\alpha+\beta_x+\theta_u\right\}, \tag{4.7}$$

   where θ_u is defined as in §4.2.1. The host-focused model accounts only for the engageability of the host, and the experience level of readers.

We repeat the above testing procedure for our model as well as for the three benchmarks for various values of $\delta \in [0,1]$. To put the predictive power of our model in perspective, we compared with the three benchmarks in the receiver operating characteristic (ROC) space, in which the true positive rate is presented as a function of the false positive rate (for a spectrum of δ values).

Predictive power. Figure 4.5 details the ROC curve of our model, compared to that of the link-focused and the host-focused benchmarks, as well as the random classification diagonal. The large gap between the ROC curve of the full model and the one of the link-focused model implies the decisiveness of the host effect in generating a successful recommendation assortment. The importance of the host is also implied by a relatively small gap between the ROC curve of the full model and that of the host-focused model. Indeed, the predictive power of a model that accounts

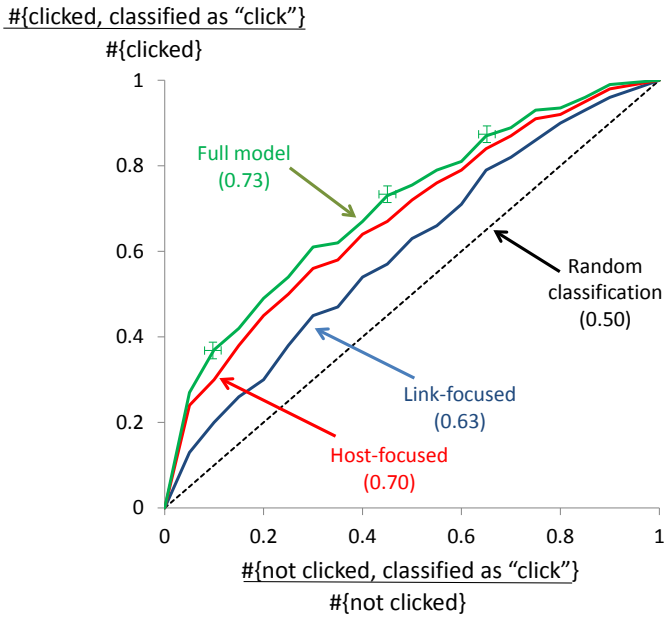


Figure 4.5: **Quantifying predictive power in the ROC space**. The plot shows the ROC curve generated by our model, together with 3 benchmarks: the link-focused model, the host-focused model, and the random classification model. The area under each curve (AUC) appears in parentheses. All standard errors (with respect to both axes) are smaller than 0.02; three illustrative examples are depicted in the "full model" curve.

only for the host effect, as well as the level of the reader's familiarity with the recommendation service, significantly exceeded that of the much richer link-focused model, that does not account for host engageability. Comparing the predictive power of the host-focused model with that of the link-focused model indicates, among other things, that while elements such as clickability and position are controlled for and taken into account by the service provider in the recommendation process (and thus a model that does not account for such elements may successfully predict click behavior), the host engageability is not taken into account by the current process.

Discussion on potential over-fitting. Since in each two-hour batch we estimate approximately 1,000 parameters, a natural issue to be concerned of is the one of over-fitting. To verify that this is not the case, we tested the predictive power of the full model in sample as well (that is, tested each set of estimators along the batch over which these estimators were produced. The in sample ROC of the full model is similar to the one generated by the out-of-sample test. The area under this curve (AUC) is 0.78. This gap between the in-sample AUC and the out-of-sample

AUC (that would be much larger in the case of significant over-fitting) should be analyzed by accounting for three factors: first, the in-sample predictive power of any model is expected to be higher than an out-of-sample one; second, clickability and engageability are time varying, and therefore predictive power may be lost as time goes by; finally, there is some potential over-fitting caused by articles that appeared only a few times during an estimation batch. However, since typically these articles rarely appear in the following test batch, these articles have a limited impact on the estimation of the control parameters, and on the predictive power of the model.

4.3 Leveraging engageability

Having established the importance of engageability in predicting click behavior, we next turn to leverage engageability for the purpose sof recommending articles. We propose a heuristic that accounts for one step forward in a reader's path when creating each recommendation. We asses the impact of these one-step look-ahead recommendations, compared to the optimal schedule of recommendations as well as to the myopic schedule of recommendations.

One-step look-ahead heuristic. We suggest recommending articles with the objective of solving the *one-step look-ahead* recommendation problem, defined by the following equations:

$$V_t^{one}(u_t, \mathcal{X}_t, x_{t-1}) \;=\; \sum_{x_t \in A_t^{one}} \mathbb{P}_{u_t, x_{t-1}, x_t}(A_t^{one}) \left(w(x_t) + V_{t+1}^{one}(u_{t+1}, \mathcal{X}_{t+1}, x_t) \right), \tag{4.8}$$

for $t = 1, \ldots, T-1$, where

$$A_t^{one} \in \operatorname*{arg\,max}_{A \in \mathcal{A}^\ell(\mathcal{X}_t)} \left\{ \sum_{x_t \in A} \mathbb{P}_{u_t, x_{t-1}, x_t}(A) \left(w(x_t) + \max_{A' \in \mathcal{A}^\ell(\mathcal{X}_{t+1})} \left\{ \sum_{x_{t+1} \in A'} \mathbb{P}_{u_{t+1}, x_t, x_{t+1}}(A') w(x_{t+1}) \right\} \right) \right\},$$

for $t = 1, \ldots, T-1$, where $V_T^{one}(u_T, \mathcal{X}_T, x_{T-1}) = V_T^m(u_T, \mathcal{X}_T, x_{T-1})$ for all u_T, $\mathcal{X}_T$, and x_{T-1}, that is, in the last time slot one-step lookahead recommendations are simply myopic.

4.3.1 Simulation

To estimate the relation between the optimal, myopic, and one-step look-ahead performances based on our data, we conduct the following simulation based on our model estimates.

Setup. Our estimates include approximately 500 articles in each of the 360 two-hours estimation batches. We assume $\ell = 5$, that is, each assortment contains exactly five links. For each

of the batches, we simulated the performance of different recommendation approaches using the following procedure:

Simulation procedure. Inputs: $k \in \{5, 10, 25, 50, 75, 100\}$, and a reader type $u_0 \in \{u_{exp}, u_{inexp}\}$

1. Set batch index $j = 1$
2. Repeat from $\tau = 1$ to $\tau = 1,000$:
   - Construct $\mathcal{X}_0$ by randomly drawing k articles (uniformly) out of those appear in batch j (with the corresponding estimates).
   - Assign a uniform price, $w(\cdot) = 1$, for any article.
   - From the set of available article draw randomly one article (uniformly) to be the landing article x_0.
   - Set $T = k - 1$. For all $t = 1, \ldots, T$ follow the update rules: $\mathcal{X}_t = \mathcal{X}_{t-1} \setminus \{x_{t-1}\}$; $u_t = u_{inexp}$ if $u_0 = u_{inexp}$ and $t \leq 1$, otherwise $u_t = u_{exp}$. Based on the model estimation output in batch j, calculate recommendation schedules that solve: the content recommendation problem[9] (4.1), the myopic content recommendation problem (4.2), and the one-step lookahead content recommendation problem (4.8), obtaining:

$$V^*_{j,\tau} = V^*_1(u_1, \mathcal{X}_1, x_0); \qquad V^m_{j,\tau} = V^m_1(u_1, \mathcal{X}_1, x_0); \qquad V^{one}_{j,\tau} = V^{one}_1(u_1, \mathcal{X}_1, x_0).$$

3. Update batch index $j \to j + 1$, and while $j \leq 360$ go back to step 2.
4. Calculate average clicks-per-visit performances:

$$\bar{V}^* = \sum_{j=1}^{360} \sum_{\tau=1}^{1,000} V^*_{j,\tau}; \qquad \bar{V}^m = \sum_{j=1}^{360} \sum_{\tau=1}^{1,000} V^m_{j,\tau}; \qquad \bar{V}^{one} = \sum_{j=1}^{360} \sum_{\tau=1}^{1,000} V^{one}_{j,\tau}.$$

We repeated the simulation procedure for combinations of $k \in \{5, 10, 25, 50, 75, 100\}$ and $u_0 \in \{u_{exp}, u_{inexp}\}$. The average performances for different numbers of available articles are depicted in Figure 4.6.

Discussion. Figure 4.6 shows that while optimal recommendations that account for the whole future path of readers may generate an increase of approximately 50 percent in clicks per

[9]The optimal recommendation schedule was determined by exhaustively comparing all the possible recommendation schedules. To reduce the computation time we used the monotonicity of the value function in (4.1) with respect to the engageability and clickability values of recommended articles to dismiss suboptimal schedules.

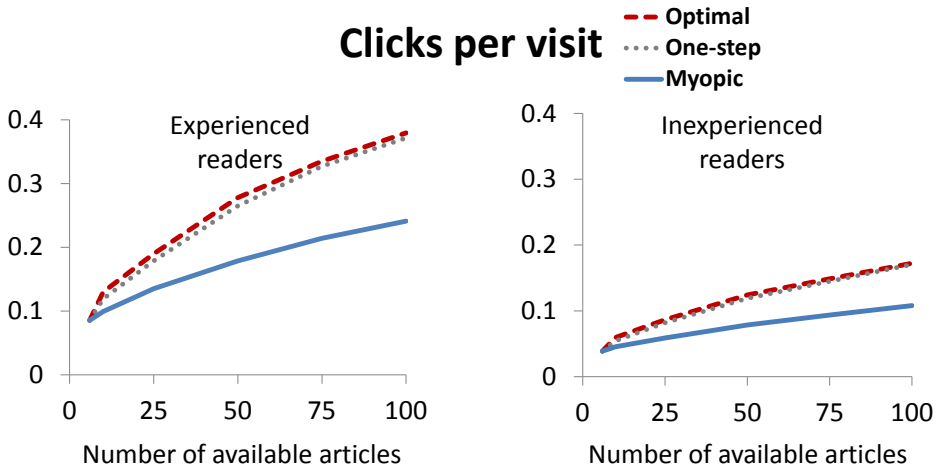


Figure 4.6: **The near-optimality of one-step look-ahead recommendations.** (*Left*) The average performance, in clicks-per-visit, of optimal, one-step lookahead, and myopic recommendations, for readers that recently clicked on internal recommendations. (*Right*) The average performances for readers that did not click recently on an internal recommendation.

visit compared to myopic recommendations, the major part of this performance gap may be captured by one-step lookahead recommendations. While readers that are familiar with internal recommendations and have clicked on those before tend to generate approximately twice as many clicks per visit compared to readers that did not click on internal recommendations recently, the significant impact of one-step look-ahead recommendations is robust over both "experienced" as well as "inexperienced" readers. The near optimality of one-step look-ahead recommendation can be backed up by theoretical bounds.

Theoretical observation 3. *Under mild structural assumptions the performance gap between the one-step look-ahead policy and that of the optimal recommendation policy is suitably small.*

The formal result, which quantifies the structural assumptions as well as the notion of "small" for the performance gap, is given in Proposition C.3 in Appendix B.1. At a high level, we show that whenever there is a continuum of available articles (intuitively, when enough articles are available), the performance of one-step look-ahead recommendations can be guaranteed to approach the one of optimal recommendations when the (optimal) click probabilities are small, and when the efficient frontier set of available articles correspond to a mild tradeoff between engageability and clickability.

4.4 Implementation Study: A Controlled Experiment

The analysis presented in §4.3.1 implies that there might be significant value in departing from myopic recommendations towards recommendations that account for a single future step in the potential path of readers. In collaboration with Outbrain, we have designed an implementation of a simple class of one-step look-ahead recommendation policies, and tested the impact of such recommendations in cooperation with a publisher that has agreed to take part in a planned pilot experiment. The objective of the experiment is to measure the change in the performance (in clicks on internal recommendations per reader's visit) when accounting for the engageability of recommended articles, relative to the performance of current practice (that myopically accounts only for clickability). An important part of the implementation study that is described below was to adjust the approach that is described in §4.3 to fit the limitations of the operating recommendation system.

4.4.1 Methodology

An adjusted-myopic proxy for the one-step look-ahead heuristic. Finding a solution for the one-step look-ahead problem involves computational complexity of order $|\mathcal{X}|^2$, compared to order $|\mathcal{X}|$ that is required in order to find the best myopic recommendation. Since the set of available articles is typically very large, a first step towards implementation was to find a proxy for the one-step look-ahead policy that requires computational complexity of order $|\mathcal{X}|$, and that follows a procedure which is similar to the one currently in place.

Recommendation algorithms that are currently being used, at a high level, operate as index policies that assign grades to candidate articles. In general, the grades generated by algorithms on which we focus in the experiment do not account for a reader type or the contextual connection between the host article and the recommended article.[10] Moreover, once grades are assigned to candidate articles, the recommended assortment typically includes the articles with the highest grades, in a manner that does not account for position effects. Finally, since the clickability and the engageability of articles (the sequences of γ and β estimates) are obtained by an off-line

[10]While some classes of recommendation algorithm are based on collaborative filtering or similarity ideas that may account for a reader type as well as the context of candidate articles, these algorithms are not modified in the experiment.

estimation and currently are not available online, we use proxies that are collected and measured in an online fashion throughout the recommendation process. An intuitive proxy for probability to *click to* an article is the CTR of the article, defined by

$$\text{CTR}(x) \;=\; \frac{\#\left\{\text{clicks to } x\right\}}{\#\left\{\text{times } x \text{ is recommended }\right\}},$$

for any article x. The CTR of each article is calculated over some time window along which offerings and click observations are documented. We found the correlation between the values of $\mathbb{P}_{u_t,x_{t-1},x_t}(A)$, when constructed by our estimators (considering the recommended article (x_t), the host article (x_{t-1}), the reader type (u_t) and the whole assortment that was offered), and the values of $\text{CTR}(x_t)$ (of the recommended article, x_t) that were calculated based on observations documented in the same estimation batch to be 0.29. In a similar manner, a potential proxy for probability to *click from* an article is the exit-CTR of an article, defined by

$$\text{exit-CTR}(x) \;=\; \frac{\#\left\{\text{events of at least one additional page-view after reading } x\right\}}{\#\left\{\text{times } x \text{ was viewed}\right\}}.$$

The exit-CTR above accounts not only for clicks on organic links, but also for other events, such as clicks on links in the text of x that lead to an additional article in the same media site, or an additional article that was read at the same publisher shortly after visiting article x, for example, after a short visit in the front page of the media site. We found the correlation between the values of $\max_{A'\in\mathcal{A}^\ell(\mathcal{X}_{t+1})}\left\{\sum_{x_{t+1}\in A'}\mathbb{P}_{u_{t+1},x_t,x_{t+1}}(A')\right\}$, when constructed by our estimators (considering the recommended article (x_t), the host article (x_{t-1}), the reader type (u_t), the whole assortment that was offered, as well as the set of articles that were available for recommendation at the following step), and the values of exit-$\text{CTR}(x_t)$ (of the host article, x_t) that were calculated based on observations documented in the same estimation batch to be 0.25.

Based on these findings, and assuming a uniform article value $w(\cdot) = 1$, we suggest the following *adjusted-myopic* recommendation policy that recommends the ℓ articles with the highest index value:

$$\text{Index}(y) \;=\; \text{CTR}(y)\left[1+\text{exit-CTR}(y)\right].$$

Recalling the one-step look-ahead heuristic in (4.8), the adjusted myopic policy uses observable proxies of the elements of that heuristic to recommend articles based on a proxy of their one-step look-ahead value. This policy accounts for the potential future path of the reader upfront, without increasing the computational complexity of index policies that are currently used by the system.

Current practice: click-based policies. Whenever a reader arrives to an article the recommended list of links that appears at the bottom of the web-page consists of links that may be generated by different classes of algorithms, each of these may use different methods and inputs in the process of evaluating candidate articles. In the designed experiment (which is described below) we focus on an important class of algorithms that directly use observed CTR values of candidate articles. At a high level, the class of algorithms on which we focus operates as follows.

CTR-based recommendation procedure $\mathcal{P}$. Inputs: a set $\mathcal{X}$ of available articles, and a time window τ of recent observations.

1. For each candidate article $x \in \mathcal{X}$ calculate $\mathrm{CTR}(x)$ along a window of recent observations.
2. For each $x \in \mathcal{X}$ assign a weight
$$q(x) = \psi\left[\mathrm{CTR}(x)\right]$$
where $\psi : \mathbb{R} \to \mathbb{R}$ is some strictly increasing mapping.
3. For each $x \in \mathcal{X}$ assign a probability
$$p(x) = \frac{q(x)}{\sum_{x' \in \mathcal{X}} q(x')}.$$
4. Draw an article to recommend according to the distribution $\{p(x)\}_{x \in \mathcal{X}}$.

We note that the set $\mathcal{X}$ considers some system constraints (for example, the article that currently hosts the recommendation cannot be recommended). The class of algorithms that follow procedure $\mathcal{P}$ is typically used in approximately 30% of the organic links that are generated in each recommendation.

Accounting for the engageability of candidate articles. As an alternative to the procedure $\mathcal{P}$ we suggest a class of recommendation policies that account for the engageability of candidate articles.

A simple lookahead procedure $\tilde{\mathcal{P}}$. Input: a set $\mathcal{X}$ of available articles, and a time window of recent observations.

1. For each candidate article $x \in \mathcal{X}$ calculate $\mathrm{CTR}(x)$ and exit-$\mathrm{CTR}(x)$ along a window of recent observations.

2. For each $x \in \mathcal{X}$ assign a weight

$$\tilde{q}(x) = \psi\left[\text{CTR}(x) \cdot (1 + \text{exit-CTR}(x))\right],$$

where $\psi[\cdot]$ is the same mapping as in the procedure $\mathcal{P}$.

3. For each $x \in \mathcal{X}$ assign a probability

$$\tilde{p}(x) = \frac{\tilde{q}(x)}{\sum_{x' \in \mathcal{X}} \tilde{q}(x')}.$$

4. Draw an article to recommend according to the distribution $\{\tilde{p}(x)\}_{x \in \mathcal{X}}$.

4.4.2 Experiment Setup

In the experiment each reader is assigned either to a test group or to a control group based on their unique user id, a number that is uniquely matched with the reader and typically does not change over time. As a result, each reader is assigned to the same group (test or control) with each arrival throughout the entire time over which the experiment takes place. Whenever a reader arrives to an article, the number of links (out of the organic links) that are generated by the algorithm class described above is determined by a mechanism that is independent of the group the user belongs to. When the reader belonged to the control group, links were generated based on the procedure $\mathcal{P}$, that is, considering the CTR of candidate articles. When the reader belonged to the test group, links were generated based on the procedure $\tilde{\mathcal{P}}$, that is, considering both the CTR and the exit-CTR of candidate articles. The group to which a reader belongs did not impact the sponsored links that were offered.

The experiment focused on *active* readers that have just clicked on an organic recommended link (a special subset of experienced readers). A reader "entered" the experiment after the first click, and we do not differentiate with respect to the algorithm that generated the first clicked link. From that point, we tracked the path of the reader throughout organic recommendations that were generated by the described algorithm class, and compared the performance of that class of algorithms in the test group relative to the control group. In both test and control groups CTR and exit-CTR values were updated every 3 hours, based on observations documented in the previous 3 hours. The experiment took place over 56 consecutive hours in a single media site, beginning on midnight between Monday and Tuesday.

Performance indicators. We follow the number of consecutive clicks made by each active reader (after the first click) on links that were generated by the algorithm class on which we focus. When the reader clicks on a sponsored link or an organic link that was not generated by that class, or when the reader terminates the session in any way without clicking on a recommended link, the path of the reader ends. We partition the experiment period into 14 batches of four hours each. Along each batch we calculate, in both test and control groups, the average clicks per active reader's visit (not counting the first click after which the reader "entered" the experiment). We denote by $\nu_{control}(t)$ the average clicks per visit in the control group along batch t, and by $\nu_{test}(t)$ the average clicks per visit in the test group along batch t. We further denote by $r(t)$ the relative difference in performance in the test group compared to the control group in batch t:

$$r(t) \;=\; 100 \cdot \frac{\nu_{test}(t) - \nu_{control}(t)}{\nu_{control}(t)}.$$

4.4.3 Results

Throughout the experiment 58,116 visits of "active" readers were documented in the control group, and 13,155 visits were documented in the test group. The results represent a 7.7% improvement in clicks per visit in the test group compared to the control group. The volume of visits and the documented clicks per visit values in the two groups along different batches are depicted in Figure 4.7. The absolute and relative differences in clicks per visit appear in Table 4.1.

Discussion. Since the experiment took place with a publisher that is characterized by a relatively low volume of readers and since it took place over a relatively short time period, some of the differences in the performance are not statistically significant. Nevertheless, the results are encouraging: in most of the batches there was an improvement in the test group relative to the control group; in three batches (4, 11, and 12, in which the number of visits was relatively large) this performance improvement is statistically significant.

It is worthwhile to to note that these improvements are witnessed despite the fact that: *i)* only one class of algorithms is adjusted in the experiment; *ii)* the exit-CTR proxy accounts not only for clicks on Outbrain's links but also on for other events of future page views; and *iii)* the adjusted myopic policy may be fine-tuned to enhance performance. Other proxies that have higher correlation with elements of the one-step look-ahead heuristic may yield better performance. One

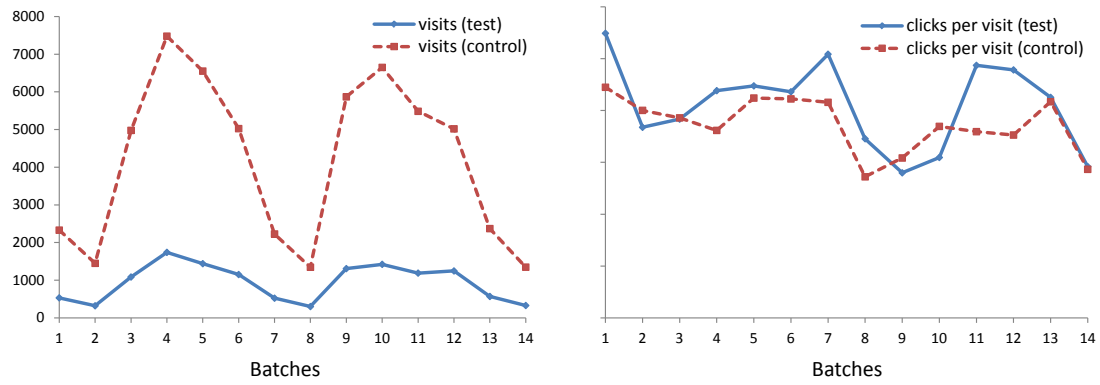


Figure 4.7: **Clicks per visit, test versus control.** Number of visits recorded along each 4-hour batch (left) and the average number of clicks per visit observed in each batch (right) in the test group and in the control one. Due to Non-disclosure agreement with Outbrain the clicks per visit units have been removed from the plot.

example, is the following measure of exit-CTR that accounts only for clicks on the recommendation that is hosted by the article:

$$\text{exit-CTR}'(x) \;=\; \frac{\#\left\{\text{clicks from } x \text{ when hosts a recommendation}\right\}}{\#\left\{\text{times } x \text{ hosts a recommendation along } \tau\right\}}.$$

We found the correlation between the values of $\max_{A' \in \mathcal{A}^{\ell}(\mathcal{X}_{t+1})} \left\{ \sum_{x_{t+1} \in A'} \mathbb{P}_{u_{t+1}, x_t, x_{t+1}}(A') \right\}$, and the values of the above exit-CTR(x_t) values that were calculated based on observations documented in the same estimation batch to be 0.36. The availability of a proxy with such correlation suggests that the results above represent a lower bound on the potential improvement.

4.5 Concluding remarks

Engageability and quality. In this chapter we identify the notion of engageability and validate its significance as a click driver along the path of a reader. Intuitively, engageability may be driven by various article features, one of which is the quality of the content, as implied in §3. To examine the connection between engageability and quality we compared the β estimates of articles with the average time that was spent by readers in these articles during the batch of data in which these estimates were generated (average time spent is an independent and common measure of quality and user engagement in online services). While both engageability and time

Batch	Visits (control)	Visits (test)	$\nu_{test}(t) - \nu_{control}(t)$	$r(t)$
1	2329	532	0.052	23.4%
2	1449	321	−0.016	−11%
3	4977	1085	−0.001	0.6%
4	7487	1740	0.038*	21.2%*
5	6551	1439	0.012	5.5%
6	5024	1151	0.007	3.3%
7	2227	523	0.046	22.3%
8	1345	301	0.037	27.0%
9	5868	1308	−0.014	−9.3%
10	6649	1422	−0.029	−16.2%
11	5484	1189	0.063**	35.5%**
12	5018	1246	0.063**	35.6%**
13	2370	569	0.004	2.0%
14	1347	329	0.003	1.8%

** At confidence level $p < 0.05$

* At confidence level $p < 0.1$

Table 4.1: **Absolute and relative improvement, test compared to control.**

spent potentially indicate quality, these notions may capture different aspects of it. For example, while engageability may undervalue the quality of long and deep articles (by the end of which the reader may be unwilling to continue reading an additional article), time spent may undervalue the quality of artistic photos. Nevertheless, the correlation between the sequence of β estimates and the sequence of average time spent is 0.28; considering the noisy online environment this indeed provides further validation for the relation between the two. We note that the interpretation of engageability as content quality is given only for the sake of intuition (to establish such a relation one needs to begin with providing a proper definition of content quality, which is beyond the scope of the present study). Nevertheless, one potential way to define quality for a broad range of sequential services is through the likelihood of a user to continue using it at the next step.

Engageability and content aging. The space of articles developed in §3 also allows one to track the manner in which clickability and engageability of articles vary with time. Having estimated the model parameters separately every two hours allows one to track clickability and engageability of articles from the time they are published. We refer to the way these properties vary along time as *the aging process* of articles. Since most of the articles lose relevancy at a rapid pace from the time they are published, tracking the aging process of articles is crucial for the provider's ability to screen out articles that became non-relevant, and to keep recommending articles that maintain their relevancy in the long term. Indeed, tracking the varying clickability and engageability shows that most of the articles exhibit a decreasing trend in both dimensions from the time they are published, and until they are not recommended anymore, due to their declining clickability. However, some of the articles exhibit a decrease in engageability along time while maintaining high clickability. One potential interpretation of this observation is that these articles lose relevancy, but such loss is not reflected in the attractiveness of their title (this drives readers to click to it when it is recommended, but also to terminate the service rather then to click from it when it hosts a recommendation, due to poor user experience). Two instances that correspond to two representative aging processes appear in Figure 4.8.

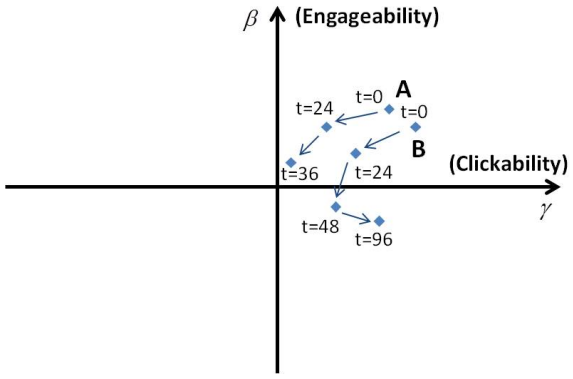


Figure 4.8: **Two instances of content aging processes.** Each point describes the estimated clickability and engageability in different ages (in hours since the time in which the articles were published). Article A exhibits decreasing clickability and engageability until it is screened out after 36 hours, and is not recommended later on. Article B exhibits decreasing engageability but maintains high clickability and as a result continues to be recommended.

Future research. We are currently in the process of designing, together with our industry collaborators, a second experiment that will take place throughout a longer period of time, with the collaboration of a larger media site (with a larger volume of readers). The objective of this experiment will be dual. On the one hand we aim to refine the analysis of the impact of accounting for exit-ctr in recommendations on the length of paths of users, by testing different combinations of CTR end exit-CTR. On the other hand, we aim to disentangle short time and longer impacts of such recommendations on the service performance: recommending articles with higher engageability may lead not only to more clicks in the short run but also to more frequent use of the service in the future.

Bibliography

Adomavicius, G. and Tuzhilin, A. (2005), 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', *Knowledge and Data Engineering, IEEE Transactions on* **17**, 734–749.

Agarwal, A., Dekel, O. and Xiao, L. (2010), 'Optimal algorithms for online convex optimization with multi-point bandit feedback', *In Proceedings of the 23rd Annual Conference on Learning Theory (COLT)* pp. 28–40.

Agarwal, A., Foster, D., Hsu, D., Kakade, S. and Rakhlin, A. (2013), 'Stochastic convex optimization with bandit feedback', *SIAM J. of Optim.* **23**, 213–240.

Alptekinoglu, A., Honhon, D. and Ulu, C. (2012), 'Learning consumer tastes through dynamic assortments', *Operations Research* **60**, 833–849.

Araman, V. and Fridgeirsdottir, K. (2011), 'A uniform allocation mechanism and cost-per-impression pricing for online advertising', *Working paper* .

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine Learning* **47**, 235–246.

Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R. E. (2002), 'The non-stochastic multi-armed bandit problem', *SIAM journal of computing* **32**, 48–77.

Awerbuch, B. and Kleinberg, R. D. (2004), 'Addaptive routing with end-to-end feedback: distributed learning and geometric approaches', *In Proceedings of the 36th ACM Symposiuim on Theory of Computing (STOC)* pp. 45–53.

Ben-Tal, A. and Nemirovski, A. (1998), 'Robust convex optimization', *Mathematics of Operations Research* **23**, 769–805.

Benveniste, A., Priouret, P. and Metivier, M. (1990), *Adaptive algorithms and stochastic approximations*, Springer-Verlag, New York.

Bergemann, D. and Hege, U. (2005), 'The financing of innovation: Learning and stopping', *RAND Journal of Economics* **36 (4)**, 719–752.

Bergemann, D. and Valimaki, J. (1996), 'Learning and strategic pricing', *Econometrica* **64**, 1125–1149.

Berry, D. A. and Fristedt, B. (1985), *Bandit problems: sequential allocation of experiments*, Chapman and Hall.

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

Bertsimas, D. and Nino-Mora, J. (2000), 'Restless bandits, linear programming relaxations, and primal dual index heuristic', *Operations Research* **48(1)**, 80–90.

Besbes, O., Gur, Y. and Zeevi, A. (2014*a*), 'Non-stationary stochastic optimization', *Working paper* .

Besbes, O., Gur, Y. and Zeevi, A. (2014*b*), 'Optimal exploration-exploitation in multi-armed-bandit problems with non-stationary rewards', *Working paper* .

Besbes, O., Gur, Y. and Zeevi, A. (2014*c*), 'Optimization in online content recommendation services: from clicks to engagement', *Working paper* .

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

Besbes, O. and Zeevi, A. (2011), 'On the minimax complexity of pricing in a changing environment', *Operations Research* **59**, 66–79.

Blackwell, D. (1956), 'An analog of the minimax theorem for vector payoffs', *Pacific Journal of Mathematics* **6**, 1–8.

Broder, J. and Rusmevichientong, P. (2012), 'Dynamic pricing under a general parametric choice model', *Operations Research* **60**, 965–980.

Caro, F. and Gallien, G. (2007*a*), 'Dynamic assortment with demand learning for seasonal consumer goods', *Management Science* **53**, 276–292.

Caro, F. and Gallien, J. (2007*b*), 'Dynamic assortment with demand learning for seasonal consumer goods', *Management Science* **53**, 276–292.

Caro, F., Martinez-de-Albeniz, V. and Rusmevichientong, P. (2013), 'The assortment packing problem: Multiperiod assortment planning for short-lived products', *Working paper* .

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

Cope, E. (2009), 'Regret and convergence bounds for a class of continuum-armed bandit problems', *IEEE Transactions on Automatic Control* **54**, 1243–1253.

den Boer, A. and Zwart, B. (2014), 'Simultaneously learning and optimizing using controlled variance pricing', *forthcoming in Management Science* .

Feng, J., Bhargava, H. K. and Pennock, D. M. (2007), 'Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms', *INFORMS Journal on Computing* **19**, 137–148.

Flaxman, A., Kalai, A. and McMahan, H. B. (2005), 'Online convex optimization in the bandit setting: Gradient descent without gradient', *Proc. 16th Annual ACM-SIAM Sympos. Discrete Algorithms, Vancouver, British Columbia, Canada* pp. 385–394.

Foster, D. P. and Vohra, R. (1999), 'Regret in the on-line decision problem', *Games and Economic Behaviour* **29**, 7–35.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *J. Comput. System Sci.* **55**, 119–139.

Garivier, A. and Moulines, E. (2011), On upper-confidence bound policies for switching bandit problems, *in* 'Algorithmic Learning Theory', Springer Berlin Heidelberg, pp. 174–188.

Gary, M. R. and Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, New York.

Gittins, J. C. (1979), 'Bandit processes and dynamic allocation indices (with discussion)', *Journal of the Royal Statistical Society, Series B* **41**, 148–177.

Gittins, J. C. (1989), *Multi-Armed Bandit Allocation Indices*, John Wiley and Sons.

Gittins, J. C. and Jones, D. M. (1974), *A dynamic allocation index for the sequential design of experiments*, North-Holland.

Guha, S. and Munagala, K. (2007), 'Approximation algorithms for partial-information based stochastic control with markovian rewards', *In 48th Annual IEEE Symposium on Fundations of Computer Science (FOCS)* pp. 483–493.

Hannan, J. (1957), *Approximation to bayes risk in repeated plays, Contributions to the Theory of Games, Volume 3*, Princeton University Press, Cambridge, UK.

Harisson, J., Keskin, B. and Zeevi, A. (2014), 'Dynamic pricing with an unknown linear demand model: Asymptotically optimal semi-myopic policies', *Working paper, Stanford University* .

Hartland, C., Gelly, S., Baskiotis, N., Teytaud, O. and Sebag, M. (2006), 'Multi-armed bandit, dynamic environments and meta-bandits', *NIPS-2006 workshop, Online trading between exploration and exploitation, Whistler, Canada* .

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

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

Hazan, E. and Kale, S. (2011), 'Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization', *Journal of Machine Learning Research - Proceedinds Track* **19**, 421–436.

Huh, W. and Rusmevichientong, P. (2009), 'A non-parametric asymptotic analysis of inventory planning with censored demand', *Mathematics of Operations Research* **34**, 103–123.

Hui, S., Fader, P. and Bradlow, E. (2009), 'Path data in marketing: An integrative framework and prospectus for model building', *Marketing Science* **28**, 320–335.

Jansen, B. J. and Mullen, T. (2008), 'Sponsored search: An overview of the concept, history, and technology', *International Journal of Electronic Business* **6**, 114–131.

Kalai, A. and Vempala, S. (2005), 'Efficient algorithms for online decision problems', *Journal of Computer and System Sciences* **71**, 291–307.

Karger, D., Motwani, R. and Ramkumar, G. D. S. (1997), 'On approximating the longest path in a graph', *Algorithmica* **18**, 82–98.

Keller, G. and Rady, S. (1999), 'Optimal experimentation in a changing environment', *The review of economic studies* **66**, 475–507.

Kiefer, J. and Wolfowitz, J. (1952), 'Stochastic estimation of the maximum of a regression function', *The Annuals of Mathematical Statistics* **23**, 462–466.

Kleinberg, R. D. and Leighton, T. (2003), 'The value of knowing a demand curve: Bounds on regret for online posted-price auctions', *In Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* pp. 594–605.

Kök, G. A., Fisher, M. L. and Vaidyanathan, R. (2009), *Assortment planning: Review of literature and industry practice. In Retail Supply Chain Management*, Springer US.

Kumar, S., Jacob, V. S. and Sriskandarajah, C. (2006), 'Scheduling advertisements on a web page to maximize revenue', *European journal of operational research* **173**, 1067–1089.

Kushner, H. and Yin, G. (2003), *Stochastic approximation and recursive algorithms and applications*, Springer-Verlag, New York.

Lai, T. (2003), 'Stochastic approximation', *The Annals of Statistics* **31**, 391–406.

Lai, T. L. and Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**, 4–22.

Linden, G., Smith, B. and York, J. (2003), 'Amazon.com recommendations: Item-to-item collaborative filtering', *Internet Computing, IEEE* **7**, 76–80.

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

Pandey, S., Agarwal, D., Charkrabarti, D. and Josifovski, V. (2007), 'Bandits for taxonomies: A model-based approach', *In SIAM International Conference on Data Mining* .

Papadimitriou, C. H. and Tsitsiklis, J. N. (1994), 'The complexity of optimal queueing network control', *In Structure in Complexity Theory Conference* pp. 318–322.

Pazzani, M. J. and Billsus, D. (2007), Content-based recommendation systems, *in* P. Brusilovsky, A. Kobsa and W. Nejdl, eds, 'The Adaptive Web', Springer-Varlag Berlin Heidelberg, pp. 325–341.

Ricci, F., Rokach, L. and Shapira, B. (2011), *Introduction to recommender systems handbook*, Springer US.

Robbins, H. (1952), 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society* **55**, 527–535.

Robbins, H. and Monro, S. (1951), 'A stochastic approximatiom method', *The Annuals of Mathematical Statistics* **22**, 400–407.

Rusmevichientong, P., Shen, Z. M. and Shmoys, D. B. (2010), 'Dynamic assortment optimization with a multinomial logit choice model and capacity constraint', *Operations Research* **58**, 1666–1680.

Saure, D. and Zeevi, A. (2013), 'Optimal dynamic assortment planning with demand learning', *Manufacturing & Service Operations Management* **15**, 387–404.

Slivkins, A. and Upfal, E. (2008), 'Adapting to a changing environment: The brownian restless bandits', *In Proceedings of the 21st Annual Conference on Learning Theory (COLT)* pp. 343–354.

Su, X. and Khoshgoftaar, T. M. (2009), 'A survey of collaborative filtering techniques', *Advances in artificial intelligence 2009: article number 4* .

Thompson, W. R. (1933), 'On the likelihood that one unknown probability exceeds another in view of the evidence of two samples', *Biometrika* **25**, 285–294.

Tsybakov, A. B. (2008), *Introduction to Nonparametric Estimation*, Springer.

Uehara, R. and Uno, Y. (2005), Efficient algorithms for the longest path problem, *in* 'Algorithms and computation', Springer Berlin Heidelberg, pp. 871–883.

Whittle, P. (1981), 'Arm acquiring bandits', *The Annals of Probability* **9**, 284–292.

Whittle, P. (1988), 'Restless bandits: Activity allocation in a changing world', *Journal of Applied Probability* **25A**, 287–298.

Zelen, M. (1969), 'Play the winner rule and the controlled clinical trials', *Journal of the American Statistical Association* **64**, 131–146.

Zinkevich, M. (2003), 'Online convex programming and generalized infinitesimal gradient ascent', *20th International Conference on Machine Learning* pp. 928–936.

Appendices

Appendix A

Appendix to Chapter 2

A.1 Proofs of main results

Proof of Proposition 2.1. The proof of the proposition is established in two steps. In the first step, we limit nature to a class of function sequences $\mathcal{V}'$ where in every epoch nature is limited to one of two specific cost functions, and show that $\mathcal{V}' \subset \mathcal{V}$. In the second step, we show that whenever $\phi \in \left\{\phi^{(0)}, \phi^{(1)}\right\}$, any admissible policy must incur regret of at least order T, even when nature is limited to the set $\mathcal{V}'$.

Step 1. Let $\mathcal{X} = [0,1]$ and fix $T \geq 1$. Let $V_T \in \{1, \ldots, T\}$ and assume that C_1 is a constant such that $V_T \geq C_1 T$. Let $C = \min\left\{C_1, \left(\frac{1}{2} - \nu\right)^2\right\}$ where ν appears in (2.2), and we assume $\nu < 1/2$. Consider the following two quadratic functions:

$$f^1(x) = x^2 - x + \frac{3}{4}, \qquad f^2(x) = x^2 - (1+2C)\,x + \frac{3}{4} + C.$$

Denoting $x_k^* = \arg\min_{x\in[0,1]} f^k(x)$, we have $x_1^* = \frac{1}{2}$, and $x_2^* = \frac{1}{2} + C$. Define $\mathcal{V}' = \left\{f\,; f_t \in \left\{f^1, f^2\right\}\ \forall t \in \mathcal{T}\right\}$. Then, for any sequence in $\mathcal{V}'$ the total functional variation is:

$$\sum_{t=2}^{T} \sup_{x\in\mathcal{X}} |f_t - f_{t-1}| \ \leq\ \sum_{t=2}^{T} \sup_{x\in\mathcal{X}} |2Cx - C| \ \leq\ CT \ \leq\ C_1 T \ \leq\ V_T.$$

For any sequence in $\mathcal{V}'$ the total functional variation (2.3) is bounded by V_T, and therefore $\mathcal{V}' \subset \mathcal{V}$.

Step 2. Fix $\phi \in \left\{\phi^{(0)}, \phi^{(1)}\right\}$, and let $\pi \in \mathcal{P}_\phi$. Let $\tilde{f}$ to be a random sequence in which in each epoch f_t is drawn according to a discrete uniform distribution over $\left\{f^1, f^2\right\}$ ($\tilde{f}_t$ is independent of $\mathcal{H}_t$ for any $t \in \mathcal{T}$). Any realization of $\tilde{f}$ is a sequence in $\mathcal{V}'$. In particular, taking expectation

over $\tilde{f}$, one has:

$$\begin{aligned}
\mathcal{R}_\phi^\pi(\mathcal{V}', T) &\geq \mathbb{E}^{\pi,\tilde{f}}\left[\sum_{t=1}^{T} \tilde{f}_t(X_t) - \sum_{t=1}^{T} \tilde{f}_t(x_t^*)\right] \\
&= \mathbb{E}^{\pi}\left[\sum_{t=1}^{T}\left(\frac{1}{2}\left(f^1(X_t) + f^2(X_t)\right) - \frac{1}{2}\left(f^1(x_1^*) + f^2(x_2^*)\right)\right)\right] \\
&\geq \sum_{t=1}^{T} \min_{x\in[0,1]}\left\{x^2 - (1+C)\,x + \frac{1}{4} + \frac{C}{2} + \frac{C^2}{2}\right\} \;=\; T\cdot\frac{C^2}{4},
\end{aligned}$$

where the minimum is obtained at $x^* = \frac{1+C}{2}$. Since $\mathcal{V}' \subseteq \mathcal{V}$, we have established that

$$\mathcal{R}_\phi^\pi(\mathcal{V}, T) \;\geq\; \mathcal{R}_\phi^\pi(\mathcal{V}', T) \;\geq\; \frac{C^2}{4}\cdot T,$$

which concludes the proof. □

Proof of Theorem 2.1. Fix $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$, and assume $V_T = o(T)$. Let $\mathcal{A}$ be a policy such that $\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}, T) = o(T)$, and let $\Delta_T \in \{1, \ldots, T\}$. Let π be the policy defined by the restarting procedure that uses $\mathcal{A}$ as a subroutine with batch size Δ_T. Then, by Proposition 2.2,

$$\frac{\mathcal{R}_\phi^\pi(\mathcal{V}, T)}{T} \;\leq\; \frac{\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}, \Delta_T)}{\Delta_T} + \frac{\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}, \Delta_T)}{T} + 2\Delta_T\cdot\frac{V_T}{T},$$

for any $1 \leq \Delta_T \leq T$. 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 2.2. Fix $\phi \in \{\phi^{(0)}, \phi^{(1)}\}$, $T \geq 1$, and $1 \leq V_T \leq T$. For $\Delta_T \in \{1, \ldots, T\}$, we break the horizon $\mathcal{T}$ into a sequence of batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size Δ_T each (except, possibly the last batch) according to (3.2). Fix $\mathcal{A} \in \mathcal{P}_\phi$, and let π be the policy defined by the restarting procedure that uses $\mathcal{A}$ as a subroutine with batch size Δ_T. Let $f \in \mathcal{V}$. We decompose the regret in the following way: $R^\pi(f, T) = \sum_{j=1}^m R_j^\pi$, where

$$\begin{aligned}
R_j^\pi &:= \mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}\left(f_t(X_t) - f_t(x_t^*)\right)\right] \\
&= \underbrace{\mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j} f_t(X_t)\right] - \min_{x\in\mathcal{X}}\left\{\sum_{t\in\mathcal{T}_j} f_t(x)\right\}}_{J_{1,j}} + \underbrace{\min_{x\in\mathcal{X}}\left\{\sum_{t\in\mathcal{T}_j} f_t(x)\right\} - \sum_{t\in\mathcal{T}_j} f_t(x_t^*)}_{J_{2,j}}. \qquad \text{(A.1)}
\end{aligned}$$

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 $\mathcal{F}$ (recall that $\mathcal{V} \subseteq \mathcal{F}$) and using the regret with respect to the single best action in the adversarial setting, one has:

$$J_{1,j} \;\leq\; \sup_{f\in\mathcal{F}} \left\{ \mathbb{E}^{\pi} \left[\sum_{t\in\mathcal{T}_j} f_t(X_t) \right] - \min_{x\in\mathcal{X}} \left\{ \sum_{t\in\mathcal{T}_j} f_t(x) \right\} \right\} \;\leq\; \mathcal{G}_{\phi}^{\mathcal{A}}\left(\mathcal{F}, \Delta_T\right), \tag{A.2}$$

where the last inequality holds using (2.6), and since in each batch decisions are dictated by $\mathcal{A}$, and since in each batch there are at most Δ_T epochs (recall that $\mathcal{G}_{\phi}^{\mathcal{A}}$ is non-decreasing in the number of epochs).

Step 2 (**Analysis of** $J_{2,j}$). Defining $f_0(x) = f_1(x)$, we denote by $V_j = \sum_{t\in\mathcal{T}_j} \|f_t - f_{t-1}\|$ the variation along batch $\mathcal{T}_j$. By the variation constraint (2.3), one has:

$$\sum_{j=1}^{m} V_j \;=\; \sum_{j=1}^{m} \sum_{t\in\mathcal{T}_j} \sup_{x\in\mathcal{X}} |f_t(x) - f_{t-1}(x)| \;\leq\; V_T. \tag{A.3}$$

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

$$\min_{x\in\mathcal{X}} \left\{ \sum_{t\in\mathcal{T}_j} f_t(x) \right\} - \sum_{t\in\mathcal{T}_j} f_t(x_t^*) \;\leq\; \sum_{t\in\mathcal{T}_j} \left(f_t(x_{\tilde{t}}^*) - f_t(x_t^*) \right) \;\leq\; \Delta_T \cdot \max_{t\in\mathcal{T}_j} \left\{ f_t(x_{\tilde{t}}^*) - f_t(x_t^*) \right\}. \tag{A.4}$$

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

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

where (a) and (b) follows from the fact that V_j is the maximal variation along batch $\mathcal{T}_j$. In particular, the above holds for $t = \tilde{t}$, contradicting the optimality of $x_{\tilde{t}}^*$ at epoch $\tilde{t}$. Therefore, one has from (A.4):

$$\min_{x\in\mathcal{X}} \left\{ \sum_{t\in\mathcal{T}_j} f_t(x) \right\} - \sum_{t\in\mathcal{T}_j} f_t(x_t^*) \;\leq\; 2\Delta_T V_j. \tag{A.5}$$

Step 3 (**Analysis of the regret over T periods**). Summing (B.6) over batches and using (A.3), one has

$$\sum_{j=1}^{m}\left(\min_{x\in\mathcal{X}}\left\{\sum_{t\in\mathcal{T}_j} f_t(x)\right\} - \sum_{t\in\mathcal{T}_j} f_t(x_t^*)\right) \;\le\; \sum_{j=1}^{m} 2\Delta_T V_j \;\le\; 2\Delta_T V_T. \tag{A.6}$$

Therefore, by the regret decomposition in (B.3), and following (A.2) and (A.6), one has:

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

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

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

This concludes the proof. □

Proof of Theorem 2.2. Fix $T\ge 1$, and $1\le V_T\le T$. For any $\Delta_T\in\{1,\dots,T\}$, let $\mathcal{A}$ be the OGD algorithm with $\eta_t=\eta=\frac{r}{G\sqrt{\Delta_T}}$ 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 relative to the single best action in the adversarial setting, and show (Flaxman et al. 2005, Lemma 3.1) that $\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F},\Delta_T)\le rG\sqrt{\Delta_T}$. Therefore, by Proposition 2.2,

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

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

$$\begin{aligned}
\mathcal{R}_{\phi^{(1)}}^{\pi}(\mathcal{V},T) &\le \frac{rG\cdot T}{(T/V_T)^{1/3}} + rG\left(\left(\frac{T}{V_T}\right)^{1/3}+1\right) + 2V_T\left(\left(\frac{T}{V_T}\right)^{2/3}+1\right)\\
&\overset{(a)}{\le} (rG+4)\cdot V_T^{1/3}T^{2/3} + rG\cdot\left(\frac{T}{V_T}\right)^{1/3} + rG\\
&\overset{(b)}{\le} (3rG+4)\cdot V_T^{1/3}T^{2/3},
\end{aligned} \tag{A.7}$$

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

Proof of Theorem 2.3. Fix $T \geq 1$ and $1 \leq V_T \leq T$. We will restrict nature to a specific class of function sequences $\mathcal{V}' \subset \mathcal{V}$. In any element of $\mathcal{V}'$ 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 $\mathcal{P}_{\phi^{(1)}}$ must incur regret of order $V_T^{1/3}T^{2/3}$.

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

$$f^1(x) = \begin{cases} \frac{1}{2} + \delta - 2\delta x + \left(x - \frac{1}{4}\right)^2 & x < \frac{1}{4} \\ \frac{1}{2} + \delta - 2\delta x & \frac{1}{4} \leq x \leq \frac{3}{4} \\ \frac{1}{2} + \delta - 2\delta x + \left(x - \frac{3}{4}\right)^2 & x > \frac{3}{4} \end{cases} \quad ; \quad f^2(x) = \begin{cases} \frac{1}{2} - \delta + 2\delta x + \left(x - \frac{1}{4}\right)^2 & x < \frac{1}{4} \\ \frac{1}{2} - \delta + 2\delta x & \frac{1}{4} \leq x \leq \frac{3}{4} \\ \frac{1}{2} - \delta + 2\delta x + \left(x - \frac{3}{4}\right)^2 & x > \frac{3}{4}, \end{cases} \tag{A.8}$$

for some $\delta > 0$ that will be specified shortly. Denoting $x_k^* = \arg\min_{x \in [0,1]} f^k(x)$, one has $x_1^* = \frac{3}{4} + \delta$, and $x_2^* = \frac{1}{4} - \delta$. 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 $\mathcal{X}$. For some $\tilde{\Delta}_T \in \{1, \ldots, T\}$ that will be specified below, define a partition of the horizon $\mathcal{T}$ to $m = \left\lceil T/\tilde{\Delta}_T \right\rceil$ batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size $\tilde{\Delta}_T$ each (except perhaps $\mathcal{T}_m$), according to (3.2). Define:

$$\mathcal{V}' \;=\; \left\{ f \;:\; f_t \in \left\{f^1, f^2\right\} \text{ and } f_t = f_{t+1} \text{ for } (j-1)\tilde{\Delta}_T + 1 \leq t \leq \min\left\{ j\tilde{\Delta}_T, T \right\} - 1,\; j = 1, \ldots, m \right\}. \tag{A.9}$$

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

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

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

Step 2 (Bounding the relative entropy within a batch). Fix any policy $\pi \in \mathcal{P}_{\phi^{(1)}}$. At each $t \in \mathcal{T}_j$, the decision maker selects $X_t \in \mathcal{X}$ and observes a noisy feedback $\phi_t^{(1)}(X_t, f_t)$. For any $f \in \mathcal{F}$: 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 $\tau \geq 1$, $A \subset \mathbb{R}^{d \times \tau}$ and $B \subset \mathcal{U}$, denote $\mathbb{P}_f^{\pi,\tau}(A, B) := \mathbb{P}_f^\pi \left\{ \left\{ \phi_t^{(1)}(X_t, f_t) \right\}_{t=1}^{\tau} \in A, U \in B \right\}$. In what follows we make use of the Kullback-Leibler divergence defined in (2.10).

Lemma A.1. (Bound on KL divergence for noisy gradient observations) *Consider the feedback structure $\phi = \phi^{(1)}$ and let Assumption 2.1 holds. Then, for any $\tau \geq 1$ and $f, g \in \mathcal{F}$:*

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

where $\tilde{C}$ is the constant that appears in the second part of Assumption 2.1.

The proof of Lemma A.1 is given later in the Appendix. 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 $\mathcal{K}(\mathbb{P}\|\mathbb{Q}) \leq \beta < \infty$. Then, for any $\mathcal{H}$-measurable real function $\varphi : \mathcal{H} \to \{0, 1\}$,*

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

Set $\tilde{\Delta}_T = \max\left\{\left\lfloor \left(\frac{1}{4\tilde{C}}\right)^{1/3}\left(\frac{T}{V_T}\right)^{2/3}\right\rfloor, 1\right\}$, (where $\tilde{C}$ is the constant that appears in part 2 of Assumption 2.1). We next show that for each batch $\mathcal{T}_j$, $\mathcal{K}\left(\mathbb{P}_{f^1}^{\pi,\tau} \| \mathbb{P}_{f^2}^{\pi,\tau}\right)$ is bounded for any $1 \leq \tau \leq |\mathcal{T}_j|$. Fix $j \in \{1, \ldots, m\}$. Then:

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

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 $V_T \leq T$. Set $\beta = \max\left\{1, 2\tilde{C}\right\}$. Since $\mathcal{K}\left(\mathbb{P}_{f^1}^{\pi,\tau} \| \mathbb{P}_{f^2}^{\pi,\tau}\right)$ is non-decreasing in τ throughout a batch, we deduce that $\mathcal{K}\left(\mathbb{P}_{f^1}^{\pi,\tau} \| \mathbb{P}_{f^2}^{\pi,\tau}\right)$ is bounded by β throughout each batch. Then, for any $x_0 \in \mathcal{X}$, using Lemma A.2 with $\varphi_t = \mathbb{1}\{X_t \leq x_0\}$, one has:

$$\max\left\{\mathbb{P}_{f^1}\left\{X_t \leq x_0\right\}, \mathbb{P}_{f^2}\left\{X_t > x_0\right\}\right\} \geq \frac{1}{4e^{\beta}} \qquad \text{for all } t \in \mathcal{T}. \tag{A.10}$$

Step 3 (A lower bound on the incurred regret for $f \in \mathcal{V}'$). Set $x_0 = \frac{1}{2}\left(x_1^* + x_2^*\right) = \frac{1}{2}$. 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 $\left\{f^1, f^2\right\}$, and applied throughout the whole batch. In particular, note that for any $1 \leq j \leq m$, 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:

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

where (a) holds since for any function $g : [0,1] \to \mathbb{R}^+$ and $x_0 \in [0,1]$ such that $g(x) \geq g(x_0)$ for all $x > x_0$, one has that $\mathbb{E}\left[g(X_t)\right] = \mathbb{E}\left[g(X_t)|X_t > x_0\right]\mathbb{P}\left\{X_t > x_0\right\} + \mathbb{E}\left[g(X_t)|X_t \leq x_0\right]\mathbb{P}\left\{X_t \leq x_0\right\} \geq g(x_0)\mathbb{P}\left\{X_t > x_0\right\}$ for any $t \in \mathcal{T}$, and similarly for any $x_0 \in [0,1]$ such that $g(x) \geq g(x_0)$ for all $x \leq x_0$, one obtains $\mathbb{E}\left[g(X_t)\right] \geq g(x_0)\mathbb{P}\left\{X_t \leq x_0\right\}$. In addition, (b) holds by (A.10) and (c) holds by $\delta = V_T\tilde{\Delta}_T/2T$. Suppose that $T \geq 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$. Applying the selected $\tilde{\Delta}_T$, one has:

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

where the last inequality follows from $T \geq 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$. If $T < 2^{5/2}\sqrt{\tilde{C}}\cdot V_T$, by Proposition 2.1

there exists a constant C such that $\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V},T) \geq C \cdot T \geq C \cdot V_T^{1/3} T^{2/3}$. Recalling that $\mathcal{V}' \subseteq \mathcal{V}$, we have:

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

This concludes the proof. □

Proof of Theorem 2.4. Part 1. We begin with the first part of the Theorem. Fix $T \geq 1$, and $1 \leq V_T \leq T$. 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 A.5 (see Appendix A.2), one has:

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}(\mathcal{F}_s,\Delta_T) \;\leq\; \frac{\left(G^2+\sigma^2\right)}{2H}\left(1+\log\Delta_T\right). \tag{A.11}$$

Therefore, by Proposition 2.2,

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

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

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

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

Part 2. We next prove the second part of the Theorem. The proof follows along similar steps as those described in the proof of Theorem 2.3, and uses the notation introduced in the latter. 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:

$$f^1(x) = x^2 - x + \frac{3}{4}, \qquad\qquad f^2(x) = x^2 - (1+\delta)\,x + \frac{3}{4} + \frac{\delta}{2} \tag{A.12}$$

for some small $\delta > 0$. Note that $x_1^* = \frac{1}{2}$, and $x_2^* = \frac{1+\delta}{2}$. 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 (3.2), where $\tilde{\Delta}_T$ will be specified below. Define the class $\mathcal{V}_s'$ according to (A.9), such that in every $f \in \mathcal{V}_s'$ the cost function is restricted to the set $\left\{f^1, f^2\right\}$, and cannot change throughout a batch. Note that all the sequences in $\mathcal{V}_s'$ consist of strongly convex functions (Condition (2.11) holds for any $H \leq 1$), with minimizers that are interior points in $\mathcal{X}$. Set $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$. Then, one has:

$$\sum_{t=2}^{T} \sup_{x\in\mathcal{X}^*} |f_t(x) - f_{t-1}(x)| \;\leq\; \sum_{j=2}^{m} \sup_{x\in\mathcal{X}^*} \left|f^1(x) - f^2(x)\right| \;\leq\; \frac{T}{\tilde{\Delta}_T} \cdot \frac{\delta^2}{2} \;=\; V_T,$$

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 $\tilde{\Delta}_T = \max\left\{\left\lfloor \frac{1}{\sqrt{2\tilde{C}}} \cdot \sqrt{\frac{T}{V_T}} \right\rfloor, 1\right\}$ ($\tilde{C}$ appears in part 2 of Assumption 2.1). Fix $j \in \{1, \ldots, m\}$. Then:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}_{f^1}^{\pi,|\mathcal{T}_j|} \| \mathbb{P}_{f^2}^{\pi,|\mathcal{T}_j|}\right) &\overset{(a)}{\leq} \tilde{C}\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j} \left(\nabla f^1(X_t) - \nabla f^2(X_t)\right)^2\right] \\
&\leq \tilde{C}\tilde{\Delta}_T\delta^2 \overset{(b)}{=} \frac{2\tilde{C}V_T\tilde{\Delta}_T^2}{T} \\
&\overset{(c)}{\leq} \max\left\{1, \frac{2\tilde{C}V_T}{T}\right\} \overset{(d)}{\leq} \max\left\{1, 2\tilde{C}\right\}, \qquad \text{(A.13)}
\end{aligned}$$

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 $V_T \leq T$. Set $\beta = \max\left\{1, 2\tilde{C}\right\}$. Then, for any $x_0 \in \mathcal{X}$, using Lemma A.2 with $\varphi_t = \mathbb{1}\{X_t > x_0\}$, one has:

$$\max\left\{\mathbb{P}_{f^1}\left\{X_t > x_0\right\}, \mathbb{P}_{f^2}\left\{X_t \leq x_0\right\}\right\} \geq \frac{1}{4e^{\beta}} \qquad \forall t \in \mathcal{T}. \qquad \text{(A.14)}$$

Step 3. Set $x_0 = \frac{1}{2}\left(x_1^* + x_2^*\right) = 1/2 + \delta/4$. 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:

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

where: the first four inequalities follow from arguments given in step 3 in the proof of Theorem 2.3; (a) holds by (A.14); and (b) holds by $\delta=\sqrt{2V_T\tilde{\Delta}_T/T}$. Given the selection of $\tilde{\Delta}_T$, one has:

$$\mathcal{R}^{\pi}_{\phi^{(1)}}\left(\mathcal{V}'_s, T\right) \;\geq\; \frac{V_T}{32e^{\beta}}\cdot\left\lfloor\frac{1}{\sqrt{2\tilde{C}}}\cdot\sqrt{\frac{T}{V_T}}\right\rfloor \;\geq\; \frac{V_T}{32e^{\beta}}\cdot\left(\frac{\sqrt{T}-\sqrt{2\tilde{C}V_T}}{\sqrt{2\tilde{C}V_T}}\right) \geq \frac{1}{64e^{\beta}\sqrt{2\tilde{C}}}\cdot\sqrt{V_TT},$$

where the last inequality holds if $T\geq 8\tilde{C}V_T$. If $T<8\tilde{C}V_T$, by Proposition 2.1 there exists a constant C such that $\mathcal{R}^{\pi}_{\phi^{(1)}}\left(\mathcal{V}_s,T\right)\geq CT\geq C\sqrt{V_TT}$. Then, recalling that $\mathcal{V}'_s\subseteq\mathcal{V}_s$, we have established that

$$\mathcal{R}^{\pi}_{\phi^{(1)}}\left(\mathcal{V}_s,T\right) \;\geq\; \mathcal{R}^{\pi}_{\phi^{(1)}}\left(\mathcal{V}'_s,T\right) \;\geq\; \frac{1}{64e^{\beta}\sqrt{2\tilde{C}}}\cdot\sqrt{V_TT}.$$

This concludes the proof. □

Proof of Theorem 2.5. Part 1. Fix $T\geq 1$, and $1\leq V_T\leq T$. For any $\Delta_T\in\{1,\ldots,T\}$, consider the EGS algorithm $\mathcal{A}$ 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 $\mathcal{A}$ and batch size Δ_T. By Lemma A.4 (see Appendix A.2), we have:

$$\mathcal{G}^{\mathcal{A}}_{\phi}(\mathcal{F}_s,\Delta_T)\leq C_1\cdot\sqrt{\Delta_T}, \tag{A.15}$$

with $C_1=2G+\left(G^2+\sigma^2+H\right)d^{3/2}/\sqrt{2H}$. Therefore, by Proposition 2.2,

$$\mathcal{R}^{\pi}_{\phi^{(0)}}(\mathcal{V}_s,T) \;\leq\; \left(\frac{T}{\Delta_T}+1\right)\cdot\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F},\Delta_T)+2V_T\Delta_T \overset{(a)}{\leq} C_1\cdot\frac{T}{\sqrt{\Delta_T}}+C_1\cdot\sqrt{\Delta_T}+2V_T\Delta_T,$$

where (a) holds by (A.15). By selecting $\Delta_T = \left\lceil (T/V_T)^{2/3} \right\rceil$, one obtains

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

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

Part 2. The proof of the second part of the theorem follows the steps described in the proof of Theorem 2.3, and uses notation introduced in the latter. 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. The details are given below.

Step 1. We define a class $\mathcal{V}'_s$ as it is defined in the proof of Theorem 2.4, using the quadratic functions f^1 and f^2 that are given in (A.12), and the partition of $\mathcal{T}$ to batches in (3.2). Again, selecting $\delta = \sqrt{2V_T \tilde{\Delta}_T / T}$, we have $\mathcal{V}'_s \subset \mathcal{V}_s$.

Step 2. Fix some policy $\pi \in \mathcal{P}_{\phi^{(0)}}$. At each $t \in \mathcal{T}_j$, $j = 1, \ldots, m$, the decision maker selects $X_t \in \mathcal{X}$ and observes a noisy feedback $\phi_t^{(0)}(X_t, f^k)$. For any $f \in \mathcal{F}$, $\tau \geq 1$, $A \subset \mathbb{R}^\tau$ and $B \subset \mathcal{U}$, denote $\mathbb{P}_f^{\pi,\tau}(A, B) := \mathbb{P}_f \left\{ \left\{ \phi_t^{(0)}(X_t, f_t) \right\}_{t=1}^{\tau} \in A, U \in B \right\}$. 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 2.2 holds. Then, for any $\tau \geq 1$ and $f, g \in \mathcal{F}$:*

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

where $\tilde{C}$ is the constant that appears in the second part of Assumption 2.2.

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

$$R_j^{\pi} \;=\; \frac{1}{2} \mathbb{E}_{f^1}^{\pi} \left[\sum_{t \in \mathcal{T}_j} \left(f^1(X_t) - f^1(x_1^*) \right) \right] + \frac{1}{2} \mathbb{E}_{f^2}^{\pi} \left[\sum_{t \in \mathcal{T}_j} \left(f^2(X_t) - f^2(x_2^*) \right) \right].$$

Then, one has:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}_{f^1}^{\pi,|\mathcal{T}_j|}\|\mathbb{P}_{f^2}^{\pi,|\mathcal{T}_j|}\right) &\overset{(a)}{\leq} \tilde{C}\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(f^1(X_t)-f^2(X_t)\right)^2\right] = \tilde{C}\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(\delta X_t-\frac{\delta}{2}\right)^2\right] \\
&= \tilde{C}\mathbb{E}_{f^1}^{\pi}\left[\delta^2\sum_{t\in\mathcal{T}_j}\left(X_t-x_1^*\right)^2\right] \overset{(b)}{=} 2\tilde{C}\delta^2\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(f^1(X_t)-f^1(x_1^*)\right)\right] \\
&\overset{(c)}{\leq} \frac{8\tilde{C}\tilde{\Delta}_T V_T}{T}\cdot R_j^{\pi} \qquad\qquad (A.16)
\end{aligned}$$

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

$$f^1(x)-f^1(x_1^*)=\nabla f^1(x_1^*)(x-x_1^*)+\frac{1}{2}\cdot\nabla f^1(x_1^*)(x-x_1^*)^2=\frac{1}{2}(x-x_1^*)^2,$$

for any $x\in\mathcal{X}$; and (c) holds since $\delta=\sqrt{2V_T\tilde{\Delta}_T/T}$, and $R_j^{\pi}\geq\frac{1}{2}\mathbb{E}_{f^1}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(f^1(X_t)-f^1(x_1^*)\right)\right]$. Thus, for any $x_0\in\mathcal{X}$, using Lemma A.2 with $\varphi_t=\mathbb{1}\{X_t>x_0\}$, we have:

$$\max\left\{\mathbb{P}_{f^1}^{\pi}\left\{X_t>x_0\right\},\mathbb{P}_{f^2}^{\pi}\left\{X_t\leq x_0\right\}\right\}\geq\frac{1}{4}\exp\left\{-\frac{8\tilde{C}\tilde{\Delta}_T V_T}{T}\cdot R_j^{\pi}\right\}\qquad\text{for all } t\in\mathcal{T}_j,\quad 1\leq j\leq m. \tag{A.17}$$

Step 3. Set $x_0=\frac{1}{2}\left(x_1^*+x_2^*\right)=1/2+\delta/4$. Let $\tilde{f}$ be the random sequence of functions that is described in Step 3 in the proof of Theorem 2.4. Taking expectation over $\tilde{f}$, one has:

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

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

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

where: the first three inequalities follow arguments given in step 3 in the proof of Theorem 3; (a) holds by (A.17); and (b) holds by $\delta = \sqrt{2V_T\tilde{\Delta}_T/T}$. Assume that $\sqrt{\tilde{C}} \cdot V_T \leq 2T$. Then, taking $\tilde{\Delta}_T = \left\lceil \left(\frac{4}{\tilde{C}}\right)^{1/3} \left(\frac{T}{V_T}\right)^{2/3} \right\rceil$, one has:

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

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

$$\beta R_j^\pi \;\geq\; \frac{32T}{\tilde{\Delta}_T^2 V_T} \;\geq\; \exp\left\{-\beta R_j^\pi\right\}. \tag{A.18}$$

Let y_0 be the unique solution to the equation $y = \exp\{-y\}$. Then, (A.18) implies $\beta R_j^\pi \geq y_0$. In particular, since $y_0 > 1/2$ this implies $R_j^\pi \geq 1/(2\beta) = \frac{1}{32\left(2\tilde{C}\right)^{2/3}} \left(\frac{T}{V_T}\right)^{1/3}$ for all $1 \leq j \leq m$. Hence:

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

where (a) holds if $\sqrt{\tilde{C}} \cdot V_T \leq 2T$. If $\sqrt{\tilde{C}} \cdot V_T > 2T$, by Proposition 2.1 there is a constant C such that $\mathcal{R}_{\phi^{(0)}}^\pi(\mathcal{V}_s, T) \geq CT \geq CV_T^{1/3}T^{2/3}$, where the last inequality holds by $T \geq V_T$. This concludes the proof. □

Proofs of Lemma A.1 and Lemma A.3. We start by proving Lemma A.1. Suppose that the feedback structure is $\phi = \phi^{(1)}$. In the proof we use the notation defined in §4 and in the proof of Theorem 3. For any $t \in \mathcal{T}$ denote $Y_t = \phi^{(1)}(X_t, \cdot)$, and denote by $y_t \in \mathbb{R}^d$ the realized feedback observation at epoch t. For convenience, for any $t \geq 1$ we further denote $y^t = (y_1, \ldots, y_t)$. Fix $\pi \in \mathcal{P}_\phi$. Letting $u \in \mathcal{U}$, we denote $x_1 = \pi_1(u)$, and $x_t := \pi_t\left(y^{t-1}, u\right)$ for $t \in \{2, \ldots, T\}$. For any $f \in \mathcal{F}$ and $\tau \geq 2$, one has:

$$\begin{aligned} d\mathbb{P}_f^{\pi,\tau}\left\{y^\tau, u\right\} &= d\mathbb{P}_f\left\{y^{\tau-1}, u\right\} d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\ &\overset{(a)}{=} d\mathbb{P}_f\left\{y_\tau \,|x_\tau\right\} d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\ &\overset{(b)}{=} dG\left(y_\tau - \nabla f(x_\tau)\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}, \end{aligned} \tag{A.19}$$

where: (a) holds since by the first part of Assumption 2.1 the feedback at epoch τ depends on the history only through $x_\tau = \pi_\tau\left(y^{\tau-1}, u\right)$; and (b) follows from the feedback structure given in the first part of Assumption 2.1. Fix $f, g \in \mathcal{F}$ and $\tau \geq 2$. One has:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}_f^{\pi,\tau} \| \mathbb{P}_g^{\pi,\tau}\right) &= \int_{u,y^\tau} \log\left(\frac{d\mathbb{P}_f^{\pi,\tau}\left\{y^\tau, u\right\}}{d\mathbb{P}_g^{\pi,\tau}\left\{y^\tau, u\right\}}\right) d\mathbb{P}_f^{\pi,\tau}\left\{y^\tau, u\right\} \\
&\overset{(a)}{=} \int_{u,y^\tau} \log\left(\frac{dG\left(y_\tau - \nabla f(x_\tau)\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}{dG\left(y_\tau - \nabla g(x_\tau)\right) d\mathbb{P}_g^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}\right) dG\left(y_\tau - \nabla f(x_\tau)\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}
\end{aligned}$$

where (a) holds by (A.19). We have that $\mathcal{K}\left(\mathbb{P}_f^{\pi,\tau} \| \mathbb{P}_g^{\pi,\tau}\right) = A_\tau + B_\tau$, where:

$$\begin{aligned}
A_\tau &:= \int_{u,y^\tau} \log\left(\frac{d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}{d\mathbb{P}_g^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}\right) dG\left(y_\tau - \nabla f(x_\tau)\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\
&= \int_{u,y^{\tau-1}} \log\left(\frac{d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}{d\mathbb{P}_g^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}\right) \left[\int_{y_\tau} dG\left(y_\tau - \nabla f(x_\tau)\right)\right] d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\
&= \int_{u,y^{\tau-1}} \log\left(\frac{d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}{d\mathbb{P}_g^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\}}\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \;=\; \mathcal{K}\left(\mathbb{P}_f^{\pi,\tau-1} \| \mathbb{P}_g^{\pi,\tau-1}\right),
\end{aligned}$$

and

$$\begin{aligned}
B_\tau &:= \int_{u,y^\tau} \log\left(\frac{dG\left(y_\tau - \nabla f(x_\tau)\right)}{dG\left(y_\tau - \nabla g(x_\tau)\right)}\right) dG\left(y_\tau - \nabla f(x_\tau)\right) d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\
&= \int_{u,y^{\tau-1}} \int_{y_\tau} \left[\log\left(\frac{dG\left(y_\tau - \nabla f(x_\tau)\right)}{dG\left(y_\tau - \nabla g(x_\tau)\right)}\right) dG\left(y_\tau - \nabla f(x_\tau)\right)\right] d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \\
&\overset{(b)}{\leq} \tilde{C} \int_{u,y^{\tau-1}} \left\|\nabla f_\tau(x_\tau) - g_\tau(x_\tau)\right\|^2 d\mathbb{P}_f^{\pi,\tau-1}\left\{y^{\tau-1}, u\right\} \;=\; \tilde{C}\mathbb{E}_f^\pi \left\|\nabla f_\tau(x_\tau) - g_\tau(x_\tau)\right\|^2,
\end{aligned}$$

where (b) follows the second part of Assumption 2.1. Repeating the above arguments, one has:

$$\mathcal{K}\left(\mathbb{P}_f^{\pi,\tau} \| \mathbb{P}_g^{\pi,\tau}\right) \;\leq\; \mathcal{K}\left(\mathbb{P}_f^{\pi,1} \| \mathbb{P}_g^{\pi,1}\right) + \tilde{C}\mathbb{E}_f^\pi \left[\sum_{t=2}^{\tau} \left\|\nabla f_t(x_t) - g_t(x_t)\right\|^2\right].$$

From the above it is also clear that:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}_f^{\pi,1} \| \mathbb{P}_g^{\pi,1}\right) &= \int_{u,y_1} \log\left(\frac{d\mathbb{P}_f^{\pi,1}\left\{y_1, u\right\}}{d\mathbb{P}_g^{\pi,1}\left\{y_1, u\right\}}\right) d\mathbb{P}_f^{\pi,1}\left\{y_1, u\right\} \\
&= \int_u \left[\int_{y_1} \log\left(\frac{dG\left(y_1 - \nabla f(x_1)\right)}{dG\left(y_1 - \nabla g(x_1)\right)}\right) dG\left(y_1 - \nabla f(x_1)\right)\right] d\mathbf{P}_u\left\{u\right\} \\
&\leq \tilde{C} \int_u \left\|\nabla f_1(x_1) - \nabla g_1(x_1)\right\|^2 d\mathbf{P}_u\left\{u\right\} \;=\; \tilde{C}\mathbb{E}_f^\pi \left\|\nabla f_1(x_1) - \nabla g_1(x_1)\right\|^2.
\end{aligned}$$

Hence, we have established that for any $\tau \geq 1$:

$$\mathcal{K}\left(\mathbb{P}_f^{\pi,\tau} \| \mathbb{P}_g^{\pi,\tau}\right) \;\leq\; \tilde{C} \sum_{t=1}^{\tau} \mathbb{E}_f^{\pi} \left\| \nabla f_t(x_t) - g_t(x_t) \right\|^2 .$$

Finally, following the steps above, the proof of Lemma A.3 (for the feedback structure $\phi = \phi^{(0)}$) is immediate, using the notation introduced in the proof of Theorem 5 for cost feedback structure, along with Assumption 2.2. This concludes the proof. □

Proof of Theorem 2.6. Fix $T \geq 1$, and $1 \leq V_T \leq T$. Let π be the OGD algorithm with $\eta_t = \eta = \sqrt{V_T/T}$ for any $t = 2, \ldots, T$. Fix $\Delta_T \in \{1, \ldots, T\}$ (to be specified below), and define a partition of $\mathcal{T}$ into batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size Δ_T each (except perhaps $\mathcal{T}_m$) according to (3.2). We note that the partition and the selection of Δ_T are only for analysis purposes, and do not affect the policy.

Fix $f \in \mathcal{V}_s$. Following the proof of Proposition 2.2, we have for $C = \max\left\{\frac{G}{2}, 2\right\}$,

$$\begin{aligned}
\mathbb{E}^{\pi}\left[\sum_{t=1}^{T} f_t(X_t)\right] - \sum_{t=1}^{T} f_t(x_t^*) \;&\leq\; \sum_{j=1}^{m}\left(\mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j} f_t(X_t)\right] - \inf_{x\in\mathcal{X}}\left\{\sum_{t\in\mathcal{T}_j} f_t(x)\right\}\right) + C\cdot\Delta_T V_T \\
&\overset{(a)}{\leq}\; \frac{1}{2}\sum_{j=1}^{m}\sum_{t\in\mathcal{T}_j}\left(\mathbb{E}^{\pi}\left\|X_t - x^*\right\|^2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H\right) + \left(G^2+\sigma^2\right)\eta_{t+1}\right) + C\cdot\Delta_T V_T \\
&\overset{(b)}{\leq}\; \frac{1}{2}\sum_{j=1}^{m}\left(\left(G^2+\sigma^2\right)\sum_{t\in\mathcal{T}_j}\sqrt{\frac{V_T}{T}}\right) + C\cdot\Delta_T V_T \\
&\leq\; \frac{\left(G^2+\sigma^2\right)}{2}\cdot\left(\frac{T}{\Delta_T}+1\right)\left(\Delta_T\cdot\sqrt{\frac{V_T}{T}}\right) + C\cdot\Delta_T V_T \\
&=\; \frac{\left(G^2+\sigma^2\right)}{2}\cdot\sqrt{V_T T} + \frac{\left(G^2+\sigma^2\right)\cdot\Delta_T}{2}\cdot\sqrt{\frac{V_T}{T}} + C\cdot\Delta_T V_T
\end{aligned}$$

for any $1 \leq \Delta_T \leq T$, where (a) follows the proof of Lemma A.5 (Appendix C, see equation (A.29)), and (b) follows since $\eta_t = \sqrt{V_T/T}$ for any $t = 2, \ldots, T$, and $H > 0$. Taking $\Delta_T = \left\lfloor \sqrt{T/V_T} \right\rfloor$ we have:

$$\begin{aligned}
\mathbb{E}^{\pi}\left[\sum_{t=1}^{T} f_t(X_t)\right] - \sum_{t=1}^{T} f_t(x_t^*) \;&\leq\; \left(\frac{\left(G^2+\sigma^2\right)}{2} + C\right)\cdot\sqrt{V_T T} + \frac{\left(G^2+\sigma^2\right)}{2} \\
&\leq\; \left(G^2+\sigma^2+C\right)\cdot\sqrt{V_T T},
\end{aligned}$$

where the last inequality follows form $1 \leq V_T \leq T$. Since the above holds for any $f \in \mathcal{V}_s$, we

established:

$$\mathcal{R}^{\pi}_{\phi^{(1)}}(\mathcal{V}_s, T) \;\leq\; \left(G^2 + \sigma^2 + \max\left\{\frac{G}{2}, 2\right\}\right) \cdot \sqrt{V_T T}.$$

This concludes the proof. □

Proof of claims made in Example 2.3. Fix $T \geq 1$. Let $\mathcal{X} = [-1, 3]$ (we assume that ν, appearing in (2.2), is smaller than 1) and consider the following two functions: $g^1(x) = (x - \alpha)^2$, and $g^2(x) = x^2$. We assume that in each epoch t, after selecting an action x_t, there is a noiseless access to the gradient of the cost function, evaluated at point x_t. The the deterministic actions are generated by an OGD algorithm:

$$x_{t+1} = P_{\mathcal{X}}\left(x_t - \eta_{t+1} \cdot f_t'(x_t)\right), \qquad \text{for all } t \geq 1,$$

with the initial selection $x_1 = 1$. In the first part we consider the case of $\eta_t = \eta = C/\sqrt{T}$, and in a second part we consider the case of $\eta_t = C/t$. The structure of both parts is similar: first we analyze the variation of the instance, showing it is sublinear. Then, by analyzing the sequence of decisions $\{x_t\}_{t=1}^T$ that is generated by the Online Gradient Descent policy, we show that in a linear portion of the horizon there is a constant C_2 such that $|x_t - x_t^*| > C_2$, and therefore a linear regret is incurred.

Part 1. Assume that $\eta_t = \eta = C/\sqrt{T} \leq 1/2$. Select $\Delta_T = \left\lfloor 1 + \frac{1}{2\eta} \right\rfloor$, and set $\alpha = 1 + (1 - 2\eta)^{\Delta_T}$ (note that $1 \leq \alpha \leq 2$). We assume that nature selects the cost function to be $g_1(\cdot)$ in the even batches and $g_2(\cdot)$ in the odd batches. We start by analyzing the variation along the horizon:

$$\begin{aligned}
\sum_{t=2}^{T} \sup_{x \in \mathcal{X}} |f_t(x) - f_{t-1}(x)| &\leq \left(\left\lceil \frac{T}{\Delta_T} \right\rceil - 1\right) \cdot \sup_{x \in \mathcal{X}} |g_2(x) - g_1(x)| \\
&\leq \frac{T}{\Delta_T} \cdot \sup_{x \in \mathcal{X}} \left|\alpha^2 - 2\alpha x\right| \\
&\overset{(a)}{\leq} \frac{8T}{\Delta_T} = \frac{8T}{\left\lfloor 2 + \frac{1}{2\eta} \right\rfloor} \\
&\leq 16T\eta = 16C \cdot \sqrt{T},
\end{aligned}$$

where (a) follows from $1 \leq \alpha \leq 2$ and $-1 \leq x \leq 3$. Next, we analyze the incurred regret. We start by analyzing decisions generated by the OGD algorithm throughout the first two batches.

Recalling that $x_1 = 1$ and that $g_2(\cdot)$ is the cost function throughout the first batch, one has for any $2 \leq t \leq \Delta_T + 1$:

$$\begin{aligned} x_t &= x_{t-1} - \eta \cdot f'(x_{t-1}) \\ &= x_{t-1} - \eta \cdot 2x_{t-1} \;=\; x_{t-1}\,(1-2\eta) \\ &= x_1\,(1-2\eta)^{t-1} \;=\; (1-2\eta)^{t-1} \\ &= \exp\left\{(t-1)\ln(1-2\eta)\right\} \\ &\overset{(a)}{\geq} \exp\left\{(t-1)(-2\eta-2\eta^2)\right\} \\ &\overset{(b)}{\geq} \exp\left\{-1-\eta\right\} \\ &\overset{(c)}{>} \frac{1}{e^2}, \end{aligned}$$

where: (a) follows since for any $-1 < x \leq 1$ one has $\ln(1+x) \geq x - \frac{x^2}{2}$; (b) follows from $t \leq \Delta_T \leq 1 + \frac{1}{2\eta}$; and (c) follows from $\eta \leq \frac{1}{2} < 1$. Since $x_t^* = 0$ for any $1 \leq t \leq \Delta_T$, one has:

$$x_t - x_t^* \;>\; \frac{1}{e^2},$$

for any $1 \leq t \leq \Delta_T$. At the end of the first batch the cost function changes from $f(\cdot)$ to $g(\cdot)$. Note that the first action of the second batch is $x_{\Delta_T+1} = (1-2\eta)^{\Delta_T}$. Since $g_1(\cdot)$ is the cost function throughout the second batch, for any $\Delta_T + 2 \leq t \leq 2\Delta_T + 1$ one has:

$$\begin{aligned} x_t &= x_{t-1} - \eta \cdot g'(x_{t-1}) \\ &= x_{t-1} - \eta \cdot 2\,(x_{t-1} - \alpha)\,. \end{aligned}$$

Using the transformation $y_t = x_t - \alpha$ for all t, one has:

$$\begin{aligned}
y_t &= y_{t-1} - \eta \cdot 2y_{t-1} \;=\; y_{t-1}\,(1-2\eta) \\
&= y_{\Delta_T+1}\,(1-2\eta)^{t-\Delta_T-1} \\
&= x_{\Delta_T+1}\,(1-2\eta)^{t-\Delta_T-1} - \alpha\,(1-2\eta)^{t-\Delta_T-1} \\
&= (1-2\eta)^{t-1} - (1-2\eta)^{t-\Delta_T-1} - (1-2\eta)^{t-1} \\
&= -(1-2\eta)^{t-\Delta_T-1} \\
&= -\exp\left\{(t-\Delta_T-1)\ln(1-2\eta)\right\} \\
&\overset{(a)}{\leq} -\exp\left\{(t-\Delta_T-1)(-2\eta-2\eta^2)\right\} \\
&\overset{(b)}{\leq} -\exp\left\{-1-\eta\right\} \\
&\overset{(c)}{<} -\frac{1}{e^2},
\end{aligned}$$

where: (a) holds since for any $-1 < x \leq 1$ one has $\ln(1+x) \geq x - \frac{x^2}{2}$; (b) follows from $t \leq 2\Delta_T \leq 1 + \frac{1}{2\eta} + \Delta_T$; and (c) follows from $\eta \leq \frac{1}{2} < 1$. Finally, recalling that $x_t^* = \alpha$ and using the transformation $y_t = x_t - \alpha$, one has for any $\Delta_T + 1 \leq t \leq 2\Delta_T$:

$$x_t^* - x_t \;=\; y_t \;<\; -\frac{1}{e^2}.$$

In the beginning of the third batch $g_2(\cdot)$ becomes the cost function once again. We note that the first action of the third batch is the same as the first action of the first batch:

$$x_{2\Delta_T+1} \;=\; \alpha + y_{2\Delta_T+1} \;=\; \alpha - (1-2\eta)^{2\Delta_T+1-\Delta_T-1} \;=\; \alpha - (1-2\eta)^{\Delta_T} \;=\; 1 \;=\; x_1,$$

and therefore the actions taken in the first two batches are repeated throughout the horizon. We conclude that for any $1 \leq t \leq T$,

$$|x_t - x_t^*| \;>\; \frac{1}{e^2}.$$

Finally, we calculate the regret incurred throughout the horizon. Using Taylor expansion, one has

$$\sum_{t=1}^{T} \left(f_t(x_t) - f_t(x_t^*)\right) \;=\; \sum_{t=1}^{T} (x_t - x_t^*)^2 \;>\; \sum_{t=1}^{T} \frac{1}{e^4} \;=\; \frac{T}{e^4}.$$

Part 2. For concreteness we assume in this part that T is even and larger than 2. We show that linear regret can be incurred when $\eta_t = \frac{C}{t}$. Set $\alpha = 1$ and $\Delta_T = T/2$ (therefore we have two batches). Assume that nature selects $g_1(\cdot)$ to be the cost function in the first batch, $g_2(\cdot)$ to

be the cost function in the second batch. We start by analyzing the variation along the horizon. Recalling that there is only one change in the cost function, one has:

$$\begin{aligned}\sum_{t=2}^{T}\sup_{x\in\mathcal{X}}|f_t(x)-f_{t-1}(x)| &= \sup_{x\in\mathcal{X}}|g_2(x)-g_1(x)| \\ &= \sup_{x\in\mathcal{X}}\left|\alpha^2-2\alpha x\right| = \sup_{x\in\mathcal{X}}|1-2x| \overset{(a)}{=} 5,\end{aligned}$$

where (a) holds because $-1\leq x\leq 3$. Since $x_1=1$, and $g_1'(1)=0$, one obtains $x_t=1$ for all $1\leq t\leq \left\lceil\frac{T}{2}\right\rceil+1$. After $\lceil T/2\rceil$ epochs, the cost function changes from $g_1(\cdot)$ to $g_2(\cdot)$, and for all $\left\lceil\frac{T}{2}\right\rceil+2\leq t\leq T$ one has:

$$\begin{aligned}x_t &= x_{t-1}-\eta_t\cdot g_2'(x_{t-1}) \\ &= x_{t-1}-\eta_t\cdot 2x_{t-1} = x_{t-1}\left(1-2\eta_t\right) \\ &= x_{\frac{T}{2}+1}\prod_{t'=\frac{T}{2}+1}^{t}\left(1-2\eta_{t'}\right) = \prod_{t'=\frac{T}{2}+1}^{t}\left(1-2\eta_{t'}\right) \\ &\overset{(a)}{\geq} \left(1-2\eta_{\frac{T}{2}+2}\right)^{t-\frac{T}{2}-1} = \left(1-\frac{4C}{T+4}\right)^{t-\frac{T}{2}-1} \\ &= \exp\left\{\left(t-\frac{T}{2}-1\right)\ln\left(1-\frac{4C}{T+4}\right)\right\} \\ &\overset{(b)}{\geq} \exp\left\{\left(t-\frac{T}{2}-1\right)\left(-\frac{4C}{T+4}-\frac{8C}{(T+4)^2}\right)\right\} \\ &\overset{(c)}{\geq} \exp\left\{-4C-\frac{8C^2}{T+4}\right\} > \exp\left\{-4C-2C^2\right\},\end{aligned}$$

where: (a) holds since $\{\eta_t\}$ is a decreasing sequence; (b) holds since $\ln(1+x)\geq x-\frac{x^2}{2}$ for any $-1<x\leq 1$; and (c) is obtained using $t<T+\frac{T}{2}+5$. Since $x_t^*=0$ for any $\frac{T}{2}+1\leq t\leq T$, one has:

$$x_t-x_t^* > \frac{1}{e^{2C(2+C)}},$$

for all $\frac{T}{2}+1\leq t\leq T$. Finally, we calculate the regret incurred throughout the horizon. Recalling that throughout the first batch no regret is incurred, and using Taylor expansion, one has:

$$\sum_{t=1}^{T}\left(f_t(x_t)-f_t(x_t^*)\right) = \sum_{t=\frac{T}{2}+1}^{T}\left(f(x_t)-f(x_t^*)\right) = \sum_{t=\frac{T}{2}+1}^{T}\left(x_t-x_t^*\right)^2 \geq \sum_{t=\frac{T}{2}+1}^{T}\frac{1}{e^{4C(2+C)}} = \frac{T}{2e^{4C(2+C)}}.$$

This concludes the proof. □

A.2 Auxiliary results for the OCO setting

A.2.1 Preliminaries

In this section we develop auxiliary results that provide bounds on the regret with respect to the single best action in the adversarial setting. As discussed in §1, the OCO literature most often considers few different feedback structures; typical examples include full access to the cost/gradient after the action X_t is selected, as well as a noiseless access to the cost/gradient evaluated at X_t. However, in this section we consider the feedback structures $\phi^{(0)}$ and $\phi^{(1)}$, where noisy access to the cost/gradient is granted.

We define admissible online algorithms exactly as admissible policies are defined in §2.[1] More precisely, letting U be a random variable defined over a probability space $(\mathbb{U}, \mathcal{U}, \mathbf{P}_u)$, we let $\mathcal{A}_1 : \mathbb{U} \to \mathbb{R}^d$ and $\mathcal{A}_t : \mathbb{R}^{(t-1)k} \times \mathbb{U} \to \mathbb{R}^d$ for $t = 2, 3, \ldots$ be measurable functions, such that X_t, the action at time t, is given by

$$X_t = \begin{cases} \mathcal{A}_1(U) & t = 1, \\ \mathcal{A}_t\left(\phi_{t-1}\left(X_{t-1}, f_{t-1}\right), \ldots, \phi_1\left(X_1, f_1\right), U\right) & t = 2, 3, \ldots, \end{cases}$$

where $k = 1$ if $\phi = \phi^{(0)}$, and $k = d$ if $\phi = \phi^{(1)}$. The mappings $\{\mathcal{A}_t : t = 1, \ldots, T\}$ together with the distribution $\mathbf{P}_u$ define the class of admissible online algorithms with respect to feedback ϕ, which is exactly the class $\mathcal{P}_\phi$. The filtration $\{\mathcal{H}_t,\ t = 1, \ldots, T\}$ is defined exactly as in §2. Given a feedback structure $\phi \in \left\{\phi^{(0)}, \phi^{(1)}\right\}$, the objective is to minimize the regret compared to the single best action:

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

We note that while most results in the OCO literature allow sequences that can adjust the cost function adversarially at each epoch, we consider the above setting where nature commits to a sequence of functions in advance. This, along with the setting of noisy cost/gradient observations, is done for the sake of consistency with the non-stationary stochastic framework we propose in chapter 2.

[1] We use the different terminology and notation only to highlight the different objectives: a policy π is designed to minimize regret with respect to the dynamic oracle, while an online algorithm $\mathcal{A}$ is designed to minimize regret compared to the static single best action benchmark.

A.2.2 Upper bounds

The first two results of this section, Lemma A.4 and Lemma A.5, analyze the performance of the EGS algorithm (given in §5) under structure $(\mathcal{F}_s, \phi^{(0)})$ and the OGD algorithm (given in §4) under structure $(\mathcal{F}_s, \phi^{(1)})$, respectively. To the best of our knowledge, the upper bound in Lemma A.4 is not documented in the Online Convex Optimization literature[2]. Lemma A.5 adapts Theorem 1 in Hazan et al. (2007) (that considered noiseless access to the gradient) to the feedback structure $\phi^{(1)}$.

Lemma A.4. (Performance of EGS in the adversarial setting) *Consider the feedback structure $\phi = \phi^{(0)}$. Let $\mathcal{A}$ be the EGS algorithm given in §5.2, with $a_t = 2d/Ht$ and $\delta_t = h_t = a_t^{1/4}$ for all $t \in \{1, \ldots, T-1\}$. Then, there exists a constant $\bar{C}$, independent of T such that for any $T \geq 1$,*

$$\mathcal{G}_{\phi}^{\mathcal{A}}(\mathcal{F}_s, T) \;\leq\; \bar{C}\sqrt{T}.$$

Proof. Let $\phi = \phi^{(0)}$. Fix $T \geq 1$ and $f \in \mathcal{F}_s$. Let $\mathcal{A}$ be the EGS algorithm, with the selection $a_t = 2d/Ht$ and $\delta_t = h_t = a_t^{1/4}$ for all $t \in \{1, \ldots, T-1\}$. We assume that $\delta_t \leq \nu$ for all $t \in \mathcal{T}$; in the end of the proof we discuss the case in which the former does not hold. For the sequence $\{\delta_t\}_{t=1}^T$, we denote by $\mathcal{X}_{\delta_t}$ the δ_t-interior of the action set $\mathcal{X}$: $\mathcal{X}_{\delta_t} = \{x \in \mathcal{X} \mid \mathbf{B}_{\delta_t}(x) \subseteq \mathcal{X}\}$. We have for all $f_t \in \mathcal{F}_s$:

$$\mathbb{E}\left[\phi_t^{(0)}(X_t, f_t) \;|X_t = x\right] \;=\; f_t(x) \quad \text{and} \quad \sup_{x \in \mathcal{X}} \left\{\mathbb{E}\left[\left(\phi_t^{(0)}(x, f_t)\right)^2\right]\right\} \;\leq\; G^2 + \sigma^2, \tag{A.20}$$

for some $\sigma \geq 0$. At any $t \in \mathcal{T}$ the gradient estimator is:

$$\hat{\nabla}_{h_t} f_t(X_t) \;=\; \frac{\phi_t^{(0)}(X_t + h_t\psi_t, f_t)\,\psi_t}{h_t},$$

for a fixed $h_t > 0$, and where $\{\psi_t\}$ is a sequence of iid random variables, drawn uniformly over the set $\left\{\pm e^{(1)}, \ldots, \pm e^{(d)}\right\}$, where $e^{(k)}$ denotes the unit vector with 1 at the k^{th} coordinate. In particular, we denote $\psi_t = Y_t W_t$, where Y_t and W_t are independent random variables, $\mathbb{P}\{y_t = 1\} = \mathbb{P}\{y_t = -1\} = 1/2$, and $W_t = e^{(k)}$ with probability $1/d$ for all $k \in \{1, \ldots, d\}$. The estimated gradient step is

$$Z_{t+1} \;=\; P_{\mathcal{X}_{\delta_t}}\left(Z_t - a_t \hat{\nabla}_{h_t} f_t(Z_t)\right), \qquad X_{t+1} \;=\; Z_{t+1} + h_{t+1}\psi_t,$$

[2]The feasibility of an upper bound of order $\sqrt{T}$ on the regret in an adversarial setting with noisy access to the cost and with strictly convex cost functions was suggested by Agarwal et al. (2010) without further details or proof.

where $P_{\mathcal{X}_{\delta_t}}$ denotes the Euclidean projection operator over the set $\mathcal{X}_{\delta_t}$. Note that $Z_t \in \mathcal{X}$, $X_t \in \mathcal{X}$, and $X_t + h_t\psi_t \in \mathcal{X}$ for all $t \in \mathcal{T}$. Since $\|\psi_t\| = 1$ for all $t \in \mathcal{T}$, one has:

$$\mathbb{E}\left[\left\|\hat{\nabla}_{h_t} f_t(Z_t)\right\|^2 \,|Z_t = z\right] \;=\; \frac{\mathbb{E}\left[\left(\phi_t^{(0)}\left(z + h_t\psi_t, f_t\right)\right)^2\right]}{h_t^2} \;\leq\; \frac{G^2 + \sigma^2}{h_t^2} \qquad \text{for all } z \in \mathcal{X}, \quad \text{(A.21)}$$

using (A.20). Then,

$$\mathbb{E}\left[\hat{\nabla}_{h_t} f_t(Z_t) \,|Z_t = z, \psi_t = \psi\right] \;=\; \frac{\mathbb{E}\left[\phi_t^{(0)}\left(Z_t + h_t\psi_t, f_t\right)\psi_t \,|Z_t = z, \psi_t = \psi\right]}{h_t} \;=\; \frac{f_t\left(z + h_t\psi\right)\psi}{h_t}.$$

Therefore, taking expectation with respect to ψ, one has

$$\begin{aligned}
\mathbb{E}\left[\hat{\nabla}_{h_t} f_t(Z_t) \,|Z_t = z\right] \;&=\; \mathbb{E}_{Y,W}\left[\frac{f_t\left(z + h_t\psi\right)\psi}{h_t}\right] \;=\; \frac{1}{d}\sum_{k=1}^{d} \frac{\left(f_t(z + h_t e^{(k)}) - f_t(z - h_t e^{(k)})\right) e^{(k)}}{2h_t} \\
&\overset{(a)}{\geq}\; \frac{1}{d}\sum_{k=1}^{d}\left(\nabla f_t(z - h_t e^{(k)}) \cdot e^{(k)}\right) e^{(k)} \\
&\overset{(b)}{\geq}\; \frac{1}{d}\sum_{k=1}^{d}\left(\nabla f_t(z) \cdot e^{(k)} - Hh_t\right) e^{(k)} \;=\; \frac{1}{d}\nabla f_t(z) - \frac{Hh_t}{d} \cdot \bar{e},
\end{aligned}$$

where $\bar{e}$ denotes a vector of ones. The equalities and inequalities above hold componentwise, where (a) follows from a Taylor expansion and the convexity of f_t: $f_t(z + h_t e^{(k)}) - f_t(z - h_t e^{(k)}) \geq \nabla f_t(z - h_t e^{(k)}) \cdot \left(2h_t e^{(k)}\right)$, for any $1 \leq k \leq d$, and (b) follows from a Taylor expansion, the convexity of f_t, and (2.11):

$$\nabla f_t(z - h_t e^{(k)}) \cdot e^{(k)} \;\geq\; \nabla f_t(z) \cdot e^{(k)} - \left(he^{(k)}\right) \cdot \left(\nabla^2 f_t\right) e^{(k)} \;\geq\; \nabla f_t(z) \cdot e^{(k)} - Gh_t,$$

for any $1 \leq k \leq d$. Therefore, for all $z \in \mathcal{X}$ and for all $t \in \mathcal{T}$:

$$\left\|\frac{1}{d}\nabla f_t(z) - \mathbb{E}\left[\hat{\nabla}_{h_t} f_t(Z_t) \,|Z_t = z\right]\right\| \;\leq\; \frac{Gh_t}{\sqrt{d}}. \qquad \text{(A.22)}$$

Define x^* as the single best action: $x^* = \arg\min_{x \in \mathcal{X}}\left\{\sum_{t=1}^{T} f_t(x)\right\}$. Then, for any $t \in \mathcal{T}$, one has

$$f_t(x^*) \;\geq\; f_t(Z_t) + \nabla f_t(Z_t) \cdot (x^* - Z_t) + \frac{1}{2}H\left\|x^* - Z_t\right\|^2,$$

and hence:

$$f_t(Z_t) - f_t(x^*) \;\leq\; \nabla f_t(Z_t) \cdot (Z_t - x^*) - \frac{1}{2}H\left\|Z_t - x^*\right\|^2. \qquad \text{(A.23)}$$

Next, using the estimated gradient step, one has

$$\begin{aligned}
\|Z_{t+1}-x^*\|^2 &= \left\|P_{\mathcal{X}_{\delta_t}}\left(Z_t - a_t\hat{\nabla}_{h_t}f_t(Z_t)\right)-x^*\right\|^2 \\
&\overset{(a)}{\leq} \left\|Z_t - a_t\hat{\nabla}_{h_t}f_t(Z_t)-x^*\right\|^2 \\
&= \|Z_t-x^*\|^2 - 2a_t\left(Z_t-x^*\right)\cdot\hat{\nabla}_{h_t}f_t(Z_t) + a_t^2\left\|\hat{\nabla}_{h_t}f_t(Z_t)\right\|^2 \\
&= \|Z_t-x^*\|^2 - \frac{2a_t}{d}\cdot\left(Z_t-x^*\right)\cdot\nabla f_t(Z_t) + a_t^2\left\|\hat{\nabla}_{h_t}f_t(Z_t)\right\|^2 \\
&\quad + 2a_t\left(Z_t-x^*\right)\cdot\left(\frac{1}{d}\nabla f_t(Z_t)-\hat{\nabla}_{h_t}f_t(Z_t)\right) \\
&\leq \|Z_t-x^*\|^2 - \frac{2a_t}{d}\cdot\left(Z_t-x^*\right)\cdot\nabla f_t(Z_t) + a_t^2\left\|\hat{\nabla}_{h_t}f_t(Z_t)\right\|^2 \\
&\quad + 2a_t\|Z_t-x^*\|\cdot\left\|\frac{1}{d}\nabla f_t(Z_t)-\hat{\nabla}_{h_t}f_t(Z_t)\right\|,
\end{aligned}$$

where (a) follows from a standard contraction property of the Euclidean projection operator. Taking expectation with respect to ψ_t and conditioning on Z_t, we follow (A.21) and (A.22) to obtain

$$\mathbb{E}\left[\|Z_{t+1}-x^*\|^2\ |Z_t\right] \leq \|Z_t-x^*\|^2 - \frac{2a_t}{d}\cdot\left(Z_t-x^*\right)\cdot\nabla f_t(Z_t) + \frac{a_t^2\left(G^2+\sigma^2\right)}{h_t^2} + \frac{2Ga_th_t}{\sqrt{d}}\cdot\|Z_t-x^*\|.$$

Taking another expectation, with respect to Z_t, we get

$$\mathbb{E}\left[\|Z_{t+1}-x^*\|^2\right] \leq \mathbb{E}\left[\|Z_t-x^*\|^2\right] - \frac{2a_t}{d}\cdot\mathbb{E}\left[\left(Z_t-x^*\right)\cdot\nabla f_t(Z_t)\right] + \frac{a_t^2\left(G^2+\sigma^2\right)}{h_t^2} + \frac{2Ga_th_t}{\sqrt{d}}\cdot\mathbb{E}\|Z_t-x^*\|,$$

and therefore, fixing some $\gamma>0$, we have for all $t\in\{1,\ldots,T-1\}$:

$$\begin{aligned}
\mathbb{E}\left[\left(Z_t-x^*\right)\cdot\nabla f_t(X_t)\right] &\leq \frac{d}{2a_t}\left(\mathbb{E}\left[\|Z_t-x^*\|^2\right]-\mathbb{E}\left[\|Z_{t+1}-x^*\|^2\right]\right) + \frac{\left(G^2+\sigma^2\right)a_td}{2h_t^2} \\
&\quad +\gamma\cdot\frac{1}{\gamma}\cdot Gh_t\sqrt{d}\cdot\mathbb{E}\|Z_t-x^*\| \\
&\overset{(a)}{\leq} \frac{d}{2a_t}\left(\mathbb{E}\left[\|Z_t-x^*\|^2\right]-\mathbb{E}\left[\|Z_{t+1}-x^*\|^2\right]\right) + \frac{\left(G^2+\sigma^2\right)a_td}{2h_t^2} \\
&\quad + \frac{\gamma^2}{2}\cdot\mathbb{E}\left[\|Z_t-x^*\|^2\right] + \frac{G^2h_t^2d}{2\gamma^2}, \qquad \text{(A.24)}
\end{aligned}$$

where (a) holds by $ab\leq\left(a^2+b^2\right)/2$, and by Jensen's inequality. In addition, one has for any

$t \in \mathcal{T}$:

$$\begin{aligned}
\mathbb{E}\left[f_t(X_t)\right] &= \mathbb{E}\left[\mathbb{E}\left[f_t(X_t)|Z_t\right]\right] = \mathbb{E}\left[\frac{1}{2}\left(f_t(Z_t+h_t)+f_t(Z_t-h_t)\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[2f_t(Z_t)+h_t\left(\nabla f_t(Z_t+h_t)-\nabla f_t(Z_t-h_t)\right)-Hh_t^2\right] \\
&\leq \mathbb{E}\left[f_t(Z_t)+\frac{1}{2}Hh_t^2\right]. \qquad \text{(A.25)}
\end{aligned}$$

The regret with respect to the single best action is:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}^{\mathcal{A}}\left[f_t(X_t)-f_t(x^*)\right] &\leq 2G+\sum_{t=1}^{T-1}\mathbb{E}^{\pi}\left[f_t(Z_t)-f_t(x^*)+\frac{1}{2}Hh_t^2\right] \\
&\overset{(a)}{\leq} 2G+\sum_{t=1}^{T-1}\mathbb{E}\left[\nabla f_t(Z_t)\cdot(Z_t-x^*)-\frac{1}{2}H\left\|Z_t-x^*\right\|^2+\frac{1}{2}Hh_t^2\right] \\
&\overset{(b)}{\leq} \mathbb{E}\left[\sum_{t=1}^{T-1}\left(\frac{d}{2a_t}\left(\left\|Z_t-x^*\right\|^2-\left\|Z_{t+1}-x^*\right\|^2\right)+\frac{\left(\gamma^2-H\right)}{2}\cdot\left\|Z_t-x^*\right\|^2\right)\right] \\
&\quad +2G+\frac{\left(G^2+\sigma^2\right)}{2}\sum_{t=1}^{T-1}\left(\frac{a_t d}{h_t^2}+\frac{h_t^2 d}{\gamma^2}\right)+\frac{H}{2}\sum_{t=1}^{T-1}h_t^2 \\
&\overset{(c)}{=} \frac{1}{2}\cdot\mathbb{E}\left[\sum_{t=2}^{T}\left\|Z_t-x^*\right\|^2\underbrace{\left(\frac{d}{a_t}-\frac{d}{a_{t-1}}+\left(\gamma^2-H\right)\right)}_{I_t}\right]+\left\|Z_1-x^*\right\|^2\underbrace{\left(\frac{d}{2a_1}+\frac{\gamma^2-H}{2}\right)}_{I_1} \\
&\quad -\left\|Z_T-x^*\right\|^2\frac{d}{2a_{T-1}}+2G+\frac{\left(G^2+\sigma^2\right)}{2}\sum_{t=1}^{T-1}\left(\frac{a_t d}{h_t^2}+\frac{h_t^2 d}{\gamma^2}\right)+\frac{H}{2}\sum_{t=1}^{T-1}h_t^2,
\end{aligned}$$

where (a) holds by (A.23), (b) holds by (A.24), and (c) holds by rearranging the summation. By selecting $\gamma^2=\frac{H}{2}$, $a_t=\frac{d}{(H-\gamma^2)t}$, and $h_t=\delta_t=a_t^{1/4}$, we have $I_t=0$ for all $t\in\mathcal{T}$, and:

$$\mathbb{E}^{\mathcal{A}}\left[\sum_{t=1}^{T}f_t(X_t)\right]-\inf_{x\in\mathcal{X}}\left\{\sum_{t=1}^{T}f_t(x)\right\} \leq 2G+\frac{\left(G^2+\sigma^2+H\right)d^{3/2}}{\sqrt{2H}}\cdot\sqrt{T}.$$

Since the above holds for any $f\in\mathcal{F}_s$, we conclude that

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}\left(\mathcal{F}_s,T\right) \leq 2G+\frac{\left(G^2+\sigma^2+H\right)d^{3/2}}{\sqrt{2H}}\cdot\sqrt{T}.$$

Finally, we consider the case in which there exists at least one time epoch t such that $\delta_t>\nu$. Then, for any such time epoch we select $h'_t=\delta'_t=\min\{\nu,\delta_t\}$. We note that the sequence $\{\delta_t\}$ is converging to 0, and therefore for any number ν there is some epoch t_ν, independent of T, such

that $\delta_t \leq \nu$ for any $t \geq t_\nu$. Therefore there can be no more than t_ν such epochs. In particular, it follows that such a case could add to the regret above no more than a constant (independent of T), that depends solely on ν, the dimension d, and the second derivative bound H. This concludes the proof. □

Lemma A.5. (Performance of OGD in the adversarial setting) *Consider the feedback structure $\phi = \phi^{(1)}$. Let $\mathcal{A}$ be the OGD algorithm given in §4, with the selection $\eta_t = 1/Ht$ for $t = 2, \dots T$. Then, there exists a constant $\bar{C}$, independent of T such that for any $T \geq 1$,*

$$\mathcal{G}_\phi^{\mathcal{A}}(\mathcal{F}_s, T) \;\leq\; \bar{C} \log T.$$

Proof. We adapt the proof of Theorem 1 in Hazan et al. (2007) to the feedback $\phi^{(1)}$. Fix $\phi = \phi^{(1)}$, $T \geq 1$, and $f \in \mathcal{F}_s$. Selecting $\eta_t = 1/Ht$ for any $t = 2, \dots T$, one has that for any $x \in \mathcal{X}$ and f_t,

$$\mathbb{E}\left[\phi_t^{(1)}(X_t, f_t) \;|X_t = x\right] \;=\; \nabla f_t(x), \quad \text{and} \quad \mathbb{E}\left[\left\|\phi_t^{(1)}(x, f_t)\right\|^2\right] \;\leq\; G^2 + \sigma^2, \tag{A.26}$$

for some $\sigma \geq 0$. Define x^* as the single best action in hindsight: $x^* = \arg\min_{x \in \mathcal{X}} \left\{\sum_{t=1}^T f_t(x)\right\}$. Then, by a Taylor expansion, for any $x \in \mathcal{X}$ there is a point $\tilde{x}$ on the segment between x and x^* such that:

$$\begin{aligned} f_t(x^*) &= f_t(x) + \nabla f_t(x) \cdot (x^* - x) + \frac{1}{2}(x^* - x) \cdot \nabla^2 f_t(\tilde{x})(x^* - x) \\ &\overset{(a)}{\geq} f_t(x) + \nabla f_t(x) \cdot (x^* - x) + \frac{H}{2}\|x^* - x\|^2, \end{aligned}$$

for any $t \in \mathcal{T}$, where (a) holds by (2.11). Substituting X_t in the above and taking expectation with respect to X_t, one has:

$$\mathbb{E}[f_t(X_t)] - f_t(x^*) \;\leq\; \mathbb{E}[\nabla f_t(x) \cdot (x^* - X_t)] + \frac{H}{2}\mathbb{E}\|x^* - X_t\|^2, \tag{A.27}$$

for any $t \in \mathcal{T}$. By the OGD step,

$$\|X_{t+1} - x^*\|^2 \;=\; \left\|P_{\mathcal{X}}\left(X_t - \eta_{t+1}\phi_t^{(1)}(X_t, f_t)\right) - x^*\right\|^2 \;\overset{(a)}{\leq}\; \left\|X_t - \eta_{t+1}\phi_t^{(1)}(X_t, f_t) - x^*\right\|^2,$$

where (a) follows from a standard contraction property of the Euclidean projection operator. Taking expectation with respect to X_t, one has:

$$\begin{aligned} \mathbb{E}\|X_{t+1} - x^*\|^2 &\leq \mathbb{E}\|X_t - x^*\|^2 + \eta_{t+1}^2 \mathbb{E}\left\|\phi_t^{(1)}(X_t, f_t)\right\|^2 - 2\eta_{t+1}\mathbb{E}\left[\left(\phi_t^{(1)}(X_t, f_t)\right) \cdot (X_t - x^*)\right] \\ &\overset{(a)}{\leq} \mathbb{E}\|X_t - x^*\|^2 + \eta_{t+1}^2\left(G^2 + \sigma^2\right) - 2\eta_{t+1}\mathbb{E}\left[(\nabla f_t(X_t)) \cdot (X_t - x^*)\right], \end{aligned}$$

where (a) follows from (A.26). Therefore, for any $t \in \mathcal{T}$, we get:

$$\mathbb{E}\left[\nabla f_t(X_t) \cdot (X_t - x^*)\right] \;\leq\; \frac{\mathbb{E}\left\|X_t - x^*\right\|^2 - \mathbb{E}\left\|X_{t+1} - x^*\right\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1}}{2}\left(G^2 + \sigma^2\right). \tag{A.28}$$

Summing (A.27) over the horizon and using (A.28), one has:

$$\begin{aligned}\sum_{t=1}^{T}\left(\mathbb{E}\left[f_t(X_t)\right] - f_t(x^*)\right) \;&\leq\; \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left\|X_t - x^*\right\|^2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H\right) + \frac{\left(G^2+\sigma^2\right)}{2}\sum_{t=1}^{T}\eta_t \qquad \text{(A.29)}\\ &\overset{(a)}{=}\; \frac{\left(G^2+\sigma^2\right)}{2}\sum_{t=1}^{T}\frac{1}{Ht} \;\leq\; \frac{\left(G^2+\sigma^2\right)}{2H}\left(1+\log T\right),\end{aligned}$$

where (a) holds using $\eta_t = 1/Ht$. Since the above holds for any sequence of functions in $\mathcal{F}_s$ we have that

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}\left(\mathcal{F}_s, T\right) \;\leq\; \frac{\left(G^2+\sigma^2\right)}{2H}\left(1+\log T\right),$$

which concludes the proof. □

A.2.3 Lower bounds

The last two results of this section, Lemma A.6 and Lemma A.7, establish lower bounds on the best achievable performance in the adversarial setting, under the structures $(\mathcal{F}_s, \phi^{(0)})$, and $(\mathcal{F}, \phi^{(1)})$, respectively. Lemma A.6 provides a lower bound that (together with the upper bound in Lemma A.4) establishes that the EGS algorithm is rate optimal in a setting with strongly convex cost functions and noisy cost observations. Lemma A.7 provides a lower bound that matches the upper bound in Lemma 3.1 in Flaxman et al. (2005), establishing that the OGD algorithm (with a careful selection of step-sizes), is rate optimal in a setting with general convex cost functions and noisy gradient observations.

Lemma A.6. *Let Assumption 2.2 hold. Then, there exists a constant C, independent of T such that for any online algorithm $\mathcal{A} \in \mathcal{P}_{\phi^{(0)}}$ and for all $T \geq 1$:*

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}\left(\mathcal{F}_s, T\right) \;\geq\; C\sqrt{T}.$$

Proof. Let $\mathcal{X} = [0,1]$. Consider the quadratic functions f^1 and f^2 in (A.12), used in the proof of Theorems 2.4 and 2.5. (note that δ will be selected differently). Fix some algorithm $\mathcal{A} \in \mathcal{P}_{\phi^{(0)}}$. Let $\tilde{f}$ be a random sequence where in the beginning of the horizon nature draws (according to

a uniform discrete distribution) a cost function from $\left\{f^1, f^2\right\}$, and applies it throughout the horizon. Taking expectation over the random sequence $\tilde{f}$ one has

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T) \;\geq\; \frac{1}{2}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(f^1(X_t) - f^1(x_1^*)\right)\right] + \frac{1}{2}\mathbb{E}^{\mathcal{A}}_{f^2}\left[\sum_{t=1}^{T}\left(f^2(X_t) - f^2(x_2^*)\right)\right],$$

where the inequality follows as in step 3 of the proof of theorem 2.3. In the following we use notation described at the proof of Theorem 2.5, for the online algorithm $\mathcal{A}$. We start by bounding the Kullback-Leibler divergence between $\mathbb{P}^{\mathcal{A},\tau}_{f^1}$ and $\mathbb{P}^{\mathcal{A},\tau}_{f^2}$ for all $\tau \in \mathcal{T}$:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}^{\mathcal{A},T}_{f^1} \| \mathbb{P}^{\mathcal{A},T}_{f^2}\right) &\overset{(a)}{\leq} \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(f^1(X_t) - f^2(X_t)\right)^2\right] \;=\; \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(\delta X_t - \frac{\delta}{2}\right)^2\right] \\
&= \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\delta^2\sum_{t=1}^{T}\left(X_t - x_1^*\right)^2\right] \\
&\overset{(b)}{=} \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[2\delta^2\sum_{t=1}^{T}\left(f^1(X_t) - f^1(x_1^*)\right)\right] \overset{(c)}{\leq} 4\tilde{C}\delta^2\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T), \qquad \text{(A.30)}
\end{aligned}$$

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

$$f^1(x) - f^1(x_1^*) \;=\; \nabla f^1(x_1^*)\cdot(x - x_1^*) + \frac{1}{2}\cdot\nabla f^1(x_1^*)\cdot(x - x_1^*)^2 \;=\; \frac{1}{2}(x - x_1^*)^2$$

for any $x \in \mathcal{X}$; and (c) holds by

$$\begin{aligned}
\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T) &\geq \frac{1}{2}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(f^1(X_t) - f^1(x_1^*)\right)\right] + \frac{1}{2}\mathbb{E}^{\mathcal{A}}_{f^2}\left[\sum_{t=1}^{T}\left(f^2(X_t) - f^2(x_2^*)\right)\right] \\
&\geq \frac{1}{2}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(f^1(X_t) - f^1(x_1^*)\right)\right]. \qquad \text{(A.31)}
\end{aligned}$$

Therefore, for any $x_0 \in \mathcal{X}$, by Lemma A.2 with $\varphi_t = \mathbb{1}\{X_t > x_0\}$, we have:

$$\max\left\{\mathbb{P}^{\mathcal{A}}_{f^1}\left\{X_\tau > x_0\right\}, \mathbb{P}^{\mathcal{A}}_{f^2}\left\{X_\tau \leq x_0\right\}\right\} \;\geq\; \frac{1}{4}\exp\left\{-4\tilde{C}\delta^2\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T)\right\} \qquad \text{for all } \tau \in \mathcal{T}. \qquad \text{(A.32)}$$

Set $x_0 = \frac{1}{2}\left(x_1^* + x_2^*\right) = 1/2 + \delta/4$. Then, following step 3 in the proof of Theorem 2.5, one has:

$$\begin{aligned}
\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T) &\geq \frac{1}{2}\sum_{t=1}^{T}\left(f^1(x_0) - f^1(x_1^*)\right)\mathbb{P}^{\mathcal{A}}_{f^1}\left\{X_t > x_0\right\} + \frac{1}{2}\sum_{t=1}^{T}\left(f^2(x_0) - f^2(x_2^*)\right)\mathbb{P}^{\mathcal{A}}_{f^2}\left\{X_t \leq x_0\right\} \\
&\geq \frac{\delta^2}{16}\sum_{t=1}^{T}\left(\mathbb{P}^{\mathcal{A}}_{f^1}\left\{X_t > x_0\right\} + \mathbb{P}^{\mathcal{A}}_{f^2}\left\{X_t \leq x_0\right\}\right) \\
&\geq \frac{\delta^2}{16}\sum_{t=1}^{T}\max\left\{\mathbb{P}^{\mathcal{A}}_{f^1}\left\{X_t > x_0\right\}, \mathbb{P}^{\mathcal{A}}_{f^2}\left\{X_t \leq x_0\right\}\right\} \\
&\overset{(a)}{\geq} \frac{\delta^2}{16}\sum_{t=1}^{T}\frac{1}{4}\exp\left\{-4\tilde{C}\delta^2\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T)\right\} \;=\; \frac{\delta^2 T}{16}\exp\left\{-4\tilde{C}\delta^2\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T)\right\}
\end{aligned}$$

where (a) holds by (A.32). Set $\delta = \left(\frac{4}{\tilde{C}T}\right)^{1/4}$. Then, one has for $\beta = 8\sqrt{\tilde{C}/T}$:

$$\beta\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T) \;\geq\; \exp\left\{-\beta\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T)\right\}. \tag{A.33}$$

Let y_0 be the unique solution to the equation $y = \exp\{-y\}$. Then, (A.33) implies $\beta\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{S}, T) \geq y_0$. In particular, since $y_0 > 1/2$ this implies

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(0)}}(\mathcal{F}_s, T) \;\geq\; 1/\left(2\beta\right) \;=\; \frac{1}{16\sqrt{\tilde{C}}}\cdot\sqrt{T}.$$

This concludes the proof. □

Lemma A.7. *Let Assumption 2.1 hold. Then, there exists a constant C, independent of T, such that for any online algorithm $\mathcal{A} \in \mathcal{P}_{\phi^{(1)}}$ and for all $T \geq 1$:*

$$\mathcal{G}^{\mathcal{A}}_{\phi^{(1)}}\left(\mathcal{F}, T\right) \;\geq\; C\sqrt{T}.$$

Proof. Fix $T \geq 1$. Let $\mathcal{X} = [0, 1]$, and consider functions f^1 and f^2 that are given in (A.8), and used in the proof of Theorem 2.3 (note that δ will be selected differently). Let $\tilde{f}$ be a random sequence of cost functions, where in the beginning of the time horizon nature draws (from a uniform discrete distribution) a function from $\left\{f^1, f^2\right\}$, and applies it throughout the horizon.

Fix $\mathcal{A} \in \mathcal{P}_{\phi^{(1)}}$. In the following we use notation described in the proof of Theorem 2.3, as well as in Lemma A.6. Set $\delta = 1/\sqrt{16\tilde{C}T}$, where $\tilde{C}$ is the constant that appears in Assumption 2.1. Then:

$$\begin{aligned}
\mathcal{K}\left(\mathbb{P}^{\mathcal{A},T}_{f^1} \| \mathbb{P}^{\mathcal{A},T}_{f^2}\right) &\overset{(a)}{\leq} \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}\left(\nabla f^1(X_t) - \nabla f^2(X_t)\right)^2\right] \\
&= \tilde{C}\mathbb{E}^{\mathcal{A}}_{f^1}\left[\sum_{t=1}^{T}16\delta^2 X_t^2\right] \;\leq\; 16\tilde{C}T\delta^2 \;\overset{(b)}{\leq}\; 1,
\end{aligned} \tag{A.34}$$

where (a) follows from Lemma A.1, and (b) holds by $\delta = 1/\sqrt{16\tilde{C}T}$. Since $\mathcal{K}(\mathbb{P}_1^{\mathcal{A},\tau} \| \mathbb{P}_2^{\mathcal{A},\tau})$ is non-decreasing in τ throughout the horizon, we deduce that the Kullback-Leibler divergence is bounded by 1 throughout the horizon. Therefore, for any $x_0 \in \mathcal{X}$, by Lemma A.2 with $\varphi_\tau = \mathbb{1}\{X_\tau \leq x_0\}$ and $\beta = 1$, one has:

$$\max\left\{\mathbb{P}_{f^1}^{\mathcal{A}}\left\{X_\tau \leq x_0\right\}, \mathbb{P}_{f^2}^{\mathcal{A}}\left\{X_t > x_0\right\}\right\} \geq \frac{1}{4e} \qquad \text{for all } \tau \in \mathcal{T}. \tag{A.35}$$

Set $x_0 = \frac{1}{2}\left(x_1^* + x_2^*\right) = \frac{1}{2}$. Taking expectation over $\tilde{f}$ and following step 3 in the proof of Theorem 2.3, one has:

$$\begin{aligned}
\mathcal{G}_{\phi^{(1)}}^{\mathcal{A}}(\mathcal{F}, T) &\geq \frac{1}{2}\mathbb{E}_{f^1}^{\mathcal{A}}\left[\sum_{t=1}^{T}\left(f^1(X_t) - f^1(x_1^*)\right)\right] + \frac{1}{2}\mathbb{E}_{f^2}^{\mathcal{A}}\left[\sum_{t=1}^{T}\left(f^2(X_t) - f^2(x_2^*)\right)\right] \\
&\geq \frac{1}{2}\sum_{t=1}^{T}\left(f^1(x_0) - f^1(x_1^*)\right)\mathbb{P}_{f^1}^{\mathcal{A}}\left\{X_t > x_0\right\} + \frac{1}{2}\sum_{t=1}^{T}\left(f^2(x_0) - f^2(x_2^*)\right)\mathbb{P}_{f^2}^{\mathcal{A}}\left\{X_t \leq x_0\right\} \\
&\geq \left(\frac{\delta}{4} + \frac{\delta^2}{2}\right)\sum_{t=1}^{T}\left(\mathbb{P}_{f^1}^{\mathcal{A}}\left\{X_t > x_0\right\} + \mathbb{P}_{f^2}^{\mathcal{A}}\left\{X_t \leq x_0\right\}\right) \\
&\geq \left(\frac{\delta}{4} + \frac{\delta^2}{2}\right)\sum_{t=1}^{T}\max\left\{\mathbb{P}_{f^1}^{\pi}\left\{X_t > x_0\right\}, \mathbb{P}_{f^2}^{\mathcal{A}}\left\{X_t \leq x_0\right\}\right\} \\
&\overset{(a)}{\geq} \left(\frac{\delta}{4} + \frac{\delta^2}{2}\right)\sum_{t=1}^{T}\frac{1}{4}\exp\{-1\} \geq \frac{\delta T}{16e} \overset{(b)}{=} \frac{1}{64e\sqrt{\tilde{C}}} \cdot \sqrt{T},
\end{aligned}$$

where (a) holds by (A.35), and (b) holds by $\delta = 1/\sqrt{16\tilde{C}T}$. This concludes the proof. □

Appendix B

Appendix to Chapter 3

B.1 Proofs

Proof of Theorem 3.1. At a high level the proof adapts a general approach of identifying a worst-case nature "strategy" (see proof of Theorem 5.1 in Auer et al. (2002) which analyze the worst-case regret relative to a single best action benchmark in a fully adversarial environment), extending these ideas appropriately to our setting. Fix $T \geq 1$, $K \geq 2$, and $V_T \in \left[K^{-1}, K^{-1}T\right]$. In what follows we restrict nature to the class $\mathcal{V}' \subseteq \mathcal{V}$ that was described in §3, and show that when μ is drawn randomly from $\mathcal{V}'$, any policy in $\mathcal{P}$ must incur regret of order $(KV_T)^{1/3}\,T^{2/3}$.

Step 1 (Preliminaries). Define a partition of the decision horizon $\mathcal{T}$ to $m = \left\lceil \frac{T}{\tilde{\Delta}_T} \right\rceil$ batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ batches of size $\tilde{\Delta}_T$ each (except perhaps $\mathcal{T}_m$) according to (3.2). For some $\varepsilon > 0$ that will be specified shortly, define $\mathcal{V}'$ to be the set of reward vectors sequences μ such that:

- $\mu_t^k \in \{1/2, 1/2 + \varepsilon\}$ for all $k \in \mathcal{K}$, $t \in \mathcal{T}$
- $\sum_{k \in \mathcal{K}} \mu_t^k = K/2 + \varepsilon$ for all $t \in \mathcal{T}$
- $\mu_t^k = \mu_{t+1}^k$ for any $(j-1)\tilde{\Delta}_T + 1 \leq t \leq \min\left\{j\tilde{\Delta}_T, T\right\} - 1$, $j = 1, \ldots, m$, for all $k \in \mathcal{K}$

For each sequence in $\mathcal{V}'$ in any epoch there is exactly one arm with expected reward $1/2 + \varepsilon$ where the rest of the arms have expected reward $1/2$, and expected rewards cannot change within a batch. Let $\varepsilon = \min\left\{1/4, V_T\tilde{\Delta}_T/T\right\}$. Then, for any $\mu \in \mathcal{V}'$ one has:

$$\sum_{t=1}^{T-1} \sup_{k \in \mathcal{K}} \left|\mu_t^k - \mu_{t+1}^k\right| \;\leq\; \sum_{j=1}^{m-1} \varepsilon \;=\; \left(\left\lceil \frac{T}{\tilde{\Delta}_T} \right\rceil - 1\right) \cdot \varepsilon \;\leq \frac{T\varepsilon}{\tilde{\Delta}_T} \;\leq\; V_T,$$

where the first inequality follows from the structure of $\mathcal{V}'$. Therefore, $\mathcal{V}' \subset \mathcal{V}$.

Step 2 (Single batch analysis). Fix some policy $\pi \in \mathcal{P}$, and fix a batch $j \in \{1, \ldots, m\}$. We denote by $\mathbb{P}_k^j$ the probability distribution conditioned on arm k being the "good" arm in batch j, and by $\mathbb{P}_0$ the probability distribution with respect to random rewards (i.e. expected reward 1/2) for each arm. We further denote by $\mathbb{E}_k^j[\cdot]$ and $\mathbb{E}_0[\cdot]$ the respective expectations. Assuming binary rewards, we let X denote a vector of $|\mathcal{T}_j|$ rewards, i.e. $X \in \{0,1\}^{|\mathcal{T}_j|}$. We denote by N_k^j the number of times arm k was selected in batch j. In the proof we use Lemma A.1 from Auer et al. (2002) that characterizes the difference between the two different expectations of some function of the observed rewards vector:

Lemma B.1. *Let $f : \{0,1\}^{|\mathcal{T}_j|} \to [0, M]$ be a bounded real function. Then, for any $k \in \mathcal{K}$:*

$$\mathbb{E}_k^j [f(X)] - \mathbb{E}_0 [f(X)] \leq \frac{M}{2} \sqrt{-\mathbb{E}_0 \left[N_k^j\right] \log(1 - 4\varepsilon^2)}.$$

Let k_j denote the "good" arm of batch j. Then, one has

$$\mathbb{E}_{k_j}^j [\mu_t^\pi] \;=\; \left(\frac{1}{2} + \varepsilon\right) \mathbb{P}_{k_j}^j \{\pi_t = k_j\} + \frac{1}{2} \mathbb{P}_{k_j}^j \{\pi_t \neq k_j\} \;=\; \frac{1}{2} + \varepsilon \mathbb{P}_{k_j}^j \{\pi_t = k_j\},$$

and therefore,

$$\mathbb{E}_{k_j}^j \left[\sum_{t \in \mathcal{T}_j} \mu_t^\pi\right] \;=\; \frac{|\mathcal{T}_j|}{2} + \sum_{t \in \mathcal{T}_j} \varepsilon \mathbb{P}_{k_j}^j \{\pi_t = k_j\} \;=\; \frac{|\mathcal{T}_j|}{2} + \varepsilon \mathbb{E}_{k_j}^j \left[N_{k_j}^j\right]. \tag{B.1}$$

In addition, applying Lemma B.1 with $f(X) = N_{k_j}^j$ (clearly $N_{k_j}^j \in \{0, \ldots, |\mathcal{T}_j|\}$) we have:

$$\mathbb{E}_{k_j}^j \left[N_{k_j}^j\right] \;\leq\; \mathbb{E}_0 \left[N_{k_j}^j\right] + \frac{|\mathcal{T}_j|}{2} \sqrt{-\mathbb{E}_0 \left[N_{k_j}^j\right] \log(1 - 4\varepsilon^2)}.$$

Summing over arms, one has:

$$\begin{aligned}
\sum_{k_j=1}^K \mathbb{E}_{k_j}^j \left[N_{k_j}^j\right] \;&\leq\; \sum_{k_j=1}^K \mathbb{E}_0 \left[N_{k_j}^j\right] + \sum_{k_j=1}^K \frac{|\mathcal{T}_j|}{2} \sqrt{-\mathbb{E}_0 \left[N_{k_j}^j\right] \log(1 - 4\varepsilon^2)} \\
&\overset{(a)}{\leq}\; |\mathcal{T}_j| + \frac{|\mathcal{T}_j|}{2} \sqrt{-\log(1 - 4\varepsilon^2) |\mathcal{T}_j| K} \\
&\overset{(b)}{\leq}\; \tilde{\Delta}_T + \frac{\tilde{\Delta}_T}{2} \sqrt{-\log(1 - 4\varepsilon^2) \tilde{\Delta}_T K},
\end{aligned} \tag{B.2}$$

for any $j \in \{1, \ldots, m\}$, where: (a) holds since $\sum_{k_j=1}^K \mathbb{E}_0 \left[N_{k_j}^j\right] = |\mathcal{T}_j|$, and thus by Cauchy-Schwarz inequality $\sum_{k_j=1}^K \sqrt{\mathbb{E}_0 \left[N_{k_j}^j\right]} \leq \sqrt{|\mathcal{T}_j| K}$; and (b) holds since $|\mathcal{T}_j| \leq \tilde{\Delta}_T$ for all $j \in \{1, \ldots, m\}$.

Step 3 (Regret along the horizon). Let $\tilde{\mu}$ be a random sequence of expected rewards vectors, in which in every batch the "good" arm is drawn according to an independent uniform distribution over the set $\mathcal{K}$. Clearly, every realization of $\tilde{\mu}$ is in $\mathcal{V}'$. In particular, taking expectation over $\tilde{\mu}$, one has:

$$\begin{aligned}
\mathcal{R}^{\pi}(\mathcal{V}',T) &= \sup_{\mu\in\mathcal{V}'}\left\{\sum_{t=1}^{T}\mu_t^* - \mathbb{E}^{\pi}\left[\sum_{t=1}^{T}\mu_t^{\pi}\right]\right\} \;\geq\; \mathbb{E}^{\tilde{\mu}}\left[\sum_{t=1}^{T}\tilde{\mu}_t^* - \mathbb{E}^{\pi}\left[\sum_{t=1}^{T}\tilde{\mu}_t^{\pi}\right]\right] \\
&\geq \sum_{j=1}^{m}\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{1}{K}\sum_{k_j=1}^{K}\mathbb{E}^{\pi}\mathbb{E}^{j}_{k_j}\left[\sum_{t\in\mathcal{T}_j}\tilde{\mu}_t^{\pi}\right]\right) \\
&\overset{(a)}{\geq} \sum_{j=1}^{m}\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{1}{K}\sum_{k_j=1}^{K}\left(\frac{|\mathcal{T}_j|}{2}+\varepsilon\mathbb{E}^{\pi}\mathbb{E}^{j}_{k_j}\left[N^j_{k_j}\right]\right)\right) \\
&\geq \sum_{j=1}^{m}\left(\sum_{t\in\mathcal{T}_j}\left(\frac{1}{2}+\varepsilon\right) - \frac{|\mathcal{T}_j|}{2} - \frac{\varepsilon}{K}\mathbb{E}^{\pi}\sum_{k_j=1}^{K}\mathbb{E}^{j}_{k_j}\left[N^j_{k_j}\right]\right) \\
&\overset{(b)}{\geq} \sum_{j=1}^{m}\left(|\mathcal{T}_j|\,\varepsilon - \frac{\varepsilon}{K}\left(\tilde{\Delta}_T + \frac{\tilde{\Delta}_T}{2}\sqrt{-\log\left(1-4\varepsilon^2\right)\tilde{\Delta}_T K}\right)\right) \\
&\overset{(c)}{\geq} T\varepsilon - \frac{T\varepsilon}{K} - \frac{T\varepsilon}{2K}\sqrt{-\log\left(1-4\varepsilon^2\right)\tilde{\Delta}_T K} \\
&\overset{(d)}{\geq} \frac{T\varepsilon}{2} - \frac{T\varepsilon^2}{K}\sqrt{\log\left(4/3\right)\tilde{\Delta}_T K},
\end{aligned}$$

where: (a) holds by (B.1); (b) holds by (B.2); (c) holds since $\sum_{j=1}^{m}|\mathcal{T}_j| = T$ and since $m \geq T/\tilde{\Delta}_T$; and (d) holds since $4\varepsilon^2 \leq 1/4$, and since $-\log(1-x) \leq 4\log(4/3)x$ for all $x \in [0,1/4]$, and because $K \geq 2$. Set $\tilde{\Delta}_T = \left\lceil K^{1/3}\left(\frac{T}{V_T}\right)^{2/3}\right\rceil$. Recall that $\varepsilon = \min\left\{1/4, V_T\tilde{\Delta}_T/T\right\}$. Suppose first that $V_T\tilde{\Delta}_T/T \leq 1/4$. Then, $\varepsilon = V_T\tilde{\Delta}_T/T \geq (KV_T/T)^{1/3}$, and one has

$$\mathcal{R}^{\pi}(\mathcal{V}',T) \;\geq\; \frac{1}{2}\cdot(KV_T)^{1/3}T^{2/3} - \sqrt{\log(4/3)}\cdot(KV_T)^{1/3}T^{2/3} \;\geq\; \frac{1}{8}\cdot(KV_T)^{1/3}T^{2/3}.$$

On the other hand, if $V_T\tilde{\Delta}_T/T \geq 1/4$, one has $\varepsilon = 1/4$, and therefore

$$\mathcal{R}^{\pi}(\mathcal{V}',T) \;\geq\; \frac{\frac{T(KV_T)^{1/3}}{4} - \frac{T^{4/3}\sqrt{\log(4/3)}}{16}}{(KV_T)^{1/3}} \;\geq\; \frac{T^{4/3}}{8(KV_T)^{1/3}} \;\geq\; \frac{1}{8}\cdot(KV_T)^{1/3}T^{2/3},$$

where the last two inequalities hold by $T \geq KV_T$. Thus, since $\mathcal{V}' \subset \mathcal{V}$, we have established that:

$$\mathcal{R}^{\pi}(\mathcal{V},T) \;\geq\; \mathcal{R}^{\pi}(\mathcal{V}',T) \;\geq\; \frac{1}{8}\cdot(KV_T)^{1/3}T^{2/3}.$$

This concludes the proof. □

Proof of Theorem 3.2 The structure of the proof is as follows. First, breaking the decision horizon to a sequence of batches of size Δ_T each, we analyze the difference in performance between the the single best action and the performance of the dynamic oracle in a single batch. Then, we plug in a known performance guarantee for Exp3 relative to the single best action in the adversarial setting, and sum over batches to establish the regret of Rexp3 with respect to the dynamic oracle.

Step 1 (Preliminaries). Fix $T \geq 1$, $K \geq 2$, and $V_T \in \left[K^{-1}, K^{-1}T\right]$. Let π be the Rexp3 policy described in §4, tuned by $\gamma = \min\left\{1\ ,\ \sqrt{\frac{K\log K}{(e-1)\Delta_T}}\right\}$ and a batch size $\Delta_T \in \{1,\ldots,T\}$ (to be specified later on). We break the horizon $\mathcal{T}$ into a sequence of batches $\mathcal{T}_1,\ldots,\mathcal{T}_m$ of size Δ_T each (except, possibly $\mathcal{T}_m$) according to (3.2). Let $\mu \in \mathcal{V}$, and fix $j \in \{1,\ldots,m\}$. We decompose the regret in batch j:

$$\mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\left(\mu_t^* - \mu_t^{\pi}\right)\right] \ = \ \underbrace{\sum_{t\in\mathcal{T}_j}\mu_t^* - \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right]}_{J_{1,j}} + \underbrace{\mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right] - \mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\mu_t^{\pi}\right]}_{J_{2,j}}. \tag{B.3}$$

The first component, $J_{1,j}$, corresponds to the expected loss associated with using a single action over the batch. The second component, $J_{2,j}$, corresponds to the expected regret with respect to the best static action in batch j.

Step 2 (Analysis of $J_{1,j}$ and $J_{2,j}$). Defining $\mu_{T+1}^k = \mu_T^k$ for all $k \in \mathcal{K}$, we denote by $V_j = \sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}\left|\mu_{t+1}^k - \mu_t^k\right|$ the variation in expected rewards along batch j. We note that

$$\sum_{j=1}^{m}V_j \ = \ \sum_{j=1}^{m}\sum_{t\in\mathcal{T}_j}\max_{k\in\mathcal{K}}\left|\mu_{t+1}^k - \mu_t^k\right| \ \leq \ V_T. \tag{B.4}$$

Let k_0 by an arm with the best expected performance (the best static strategy) over batch $\mathcal{T}_j$, i.e., $k_0 \in \ \arg\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\}$. Then,

$$\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} \ = \ \sum_{t\in\mathcal{T}_j}\mu_t^{k_0} \ = \ \mathbb{E}\left[\sum_{t\in\mathcal{T}_j}X_t^{k_0}\right] \ \leq \ \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right], \tag{B.5}$$

and therefore, one has:

$$\begin{aligned} J_{1,j} \ &= \ \sum_{t\in\mathcal{T}_j}\mu_t^* - \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}\right] \ \overset{(a)}{\leq} \ \sum_{t\in\mathcal{T}_j}\left(\mu_t^* - \mu_t^{k_0}\right) \\ &\leq \ \Delta_T\max_{t\in\mathcal{T}_j}\left\{\mu_t^* - \mu_t^{k_0}\right\} \ \overset{(b)}{\leq} \ 2V_j\Delta_T, \end{aligned} \tag{B.6}$$

for any $\mu \in \mathcal{V}$ and $j \in \{1, \ldots, m\}$, where (a) holds by (B.5) and (b) holds by the following argument: otherwise there is an epoch $t_0 \in \mathcal{T}_j$ for which $\mu^*_{t_0} - \mu^{k_0}_{t_0} > 2V_j$. Indeed, let $k_1 = \arg\max_{k\in\mathcal{K}} \mu^k_{t_0}$. In such case, for all $t \in \mathcal{T}_j$ one has $\mu^{k_1}_t \geq \mu^{k_1}_{t_0} - V_j > \mu^{k_0}_{t_0} + V_j \geq \mu^{k_0}_t$, since V_j is the maximal variation in batch $\mathcal{T}_j$. This however, implies that the expected reward of k_0 is dominated by an expected reward of another arm throughout the whole period, and contradicts the optimality of k_0 .

In addition, Corollary 3.2 in Auer et al. (2002) points out that the regret with respect to the single best action of the batch, that is incurred by Exp3 with the tuning parameter $\gamma = \min\left\{1\ ,\ \sqrt{\frac{K\log K}{(e-1)\Delta_T}}\right\}$, is bounded by $2\sqrt{e-1}\sqrt{\Delta_T K \log K}$. Therefore, for each $j \in \{1, \ldots, m\}$ one has

$$J_{2,j} \ = \ \mathbb{E}\left[\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j} X^k_t\right\} - \mathbb{E}^\pi\left[\sum_{t\in\mathcal{T}_j}\mu^\pi_t\right]\right] \ \overset{(a)}{\leq} \ 2\sqrt{e-1}\sqrt{\Delta_T K\log K}, \tag{B.7}$$

for any $\mu \in \mathcal{V}$, where (a) holds since within each batch arms are pulled according to Exp3(γ).

Step 3 (Regret throughout the horizon). Summing over $m = \lceil T/\Delta_T \rceil$ batches we have:

$$\begin{aligned}
\mathcal{R}^\pi(\mathcal{V}, T) \ &= \ \sup_{\mu\in\mathcal{V}}\left\{\sum_{t=1}^T \mu^*_t - \mathbb{E}^\pi\left[\sum_{t=1}^T \mu^\pi_t\right]\right\} \ \overset{(a)}{\leq} \ \sum_{j=1}^m \left(2\sqrt{e-1}\sqrt{\Delta_T K\log K} + 2V_j\Delta_T\right) \\
&\overset{(b)}{\leq} \ \left(\frac{T}{\Delta_T}+1\right)\cdot 2\sqrt{e-1}\sqrt{\Delta_T K\log K} + 2\Delta_T V_T. \\
&= \ \frac{2\sqrt{e-1}\sqrt{K\log K}\cdot T}{\sqrt{\Delta_T}} + 2\sqrt{e-1}\sqrt{\Delta_T K\log K} + 2\Delta_T V_T.
\end{aligned}$$

where: (a) holds by (B.3), (B.6), and (B.7); and (b) follows from (B.4). Selecting $\Delta_T = \left\lceil (K\log K)^{1/3}\,(T/V_T)^{2/3}\right\rceil$, we establish:

$$\begin{aligned}
\mathcal{R}^\pi(\mathcal{V}, T) \ &\leq \ 2\sqrt{e-1}\,(K\log K\cdot V_T)^{1/3}\,T^{2/3} + 2\sqrt{e-1}\sqrt{\left((K\log K)^{1/3}\,(T/V_T)^{2/3}+1\right)K\log K} \\
&\quad +2\left((K\log K)^{1/3}\,(T/V_T)^{2/3}+1\right)V_T \\
&\overset{(a)}{\leq} \ \left(6\sqrt{e-1}+4\right)(K\log K\cdot V_T)^{1/3}\,T^{2/3},
\end{aligned}$$

where (a) follows from $K \geq 2$ and $V_T \in \left[K^{-1}, K^{-1}T\right]$. This concludes the proof. □

Proof of Theorem 3.3 The structure of the proof is follows: First, we follow the proof of Theorem 3.2, breaking the decision horizon to a sequence of decision batches and analyzing the

difference in performance between the sequence of single best actions and the performance of the dynamic oracle. Then, we analyze the regret of the Exp3.S policy when compared to the sequence of single-best-actions which is composed of the single best action of each batch (this part of the proof roughly follows the proof lines of Theorem 8.1 of Auer, Cesa-Bianchi, Freund and Schapire (2002), while considering a possibly infinite number of changes in the identity of the best arm. Finally, we select tuning parameters that minimize the overall regret.

Step 1 (Preliminaries). Fix $T \geq 1$, $K \geq 2$, and $TK^{-1} \geq V_T \geq K^{-1}$. Let π be the Exp3.S policy (the tuning parameters with be set later). We define a partition of the decision horizon $\mathcal{T}$ to batches $\mathcal{T}_1, \ldots, \mathcal{T}_m$ of size Δ_T each (except perhaps $\mathcal{T}_m$), according to (3.2).

Step 2. Let $\mu \in \mathcal{V}$. We follow the proof of Theorem 3.2 (see the beginning of step 3) to obtain:

$$\begin{aligned} \mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}(\mu_t^* - \mu_t^{\pi})\right] &= \sum_{t\in\mathcal{T}_j}\mu_t^* - \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} + \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} - \mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\mu_t^{\pi}\right] \\ &\leq 2V_j\Delta_T + \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} - \mathbb{E}^{\pi}\left[\sum_{t\in\mathcal{T}_j}\mu_t^{\pi}\right], \end{aligned} \tag{B.8}$$

for each $j \in \{1, \ldots, m\}$ and for any $\mu \in \mathcal{V}$. Fix $j \in \{1, \ldots, m\}$. We next bound the difference between the performance of the single best action in $\mathcal{T}_j$ and that of the policy, throughout $\mathcal{T}_j$. Let t_j denote the first decision index of batch j, that is, $t_j = (j-1)\Delta_T + 1$. We W_t denote the sum of all weights at decision t: $W_t = \sum_{k=1}^{K} w_t^k$. Following the proof of Theorem 8.1 in Auer et al. (2002), one has:

$$\frac{W_{t+1}}{W_t} \leq 1 + \frac{\gamma/K}{1-\gamma}X_t^{\pi} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{k=1}^{K}\hat{X}_t^k + e\alpha. \tag{B.9}$$

Taking logarithms on both sides of (B.9) and summing over all $t \in \mathcal{T}_j$, we get

$$\log\left(\frac{W_{t_{j+1}}}{W_{t_j}}\right) \leq \frac{\gamma/K}{1-\gamma}\sum_{t\in\mathcal{T}_j}X_t^{\pi} + \frac{(e-2)(\gamma/K)^2}{1-\gamma}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\hat{X}_t^k + e\alpha\left|\mathcal{T}_j\right| \tag{B.10}$$

(for $\mathcal{T}_m$ set $W_{t_{m+1}} = W_T$). Let k_j be the best single action in $\mathcal{T}_j$: $k_j \in \arg\max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}X_t^k\right\}$.

Then,

$$\begin{aligned} w_{t_{j+1}}^{k_j} &\geq w_{t_j+1}^{k_j} \exp\left\{\frac{\gamma}{K}\sum_{t_j+1}^{t_{j+1}-1}\hat{X}_t^{k_j}\right\} \\ &\geq \frac{e\alpha}{K}W_{t_j}\exp\left\{\frac{\gamma}{K}\sum_{t_j+1}^{t_{j+1}-1}\hat{X}_t^{k_j}\right\} \overset{(a)}{\geq} \frac{\alpha}{K}W_{t_j}\exp\left\{\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\hat{X}_t^{k_j}\right\}, \end{aligned}$$

where (a) holds since $\gamma\hat{X}_t^{k_j}/K \leq 1$. Therefore,

$$\log\left(\frac{W_{t_{j+1}}}{W_{t_j}}\right) \geq \log\left(\frac{w_{t_{j+1}}^{k_j}}{W_{t_j}}\right) \geq \log\left(\frac{\alpha}{K}\right) + \frac{\gamma}{K}\sum_{t\in\mathcal{T}_j} X_t^{\pi}. \tag{B.11}$$

Taking (B.10) and (B.11) together, one has

$$\sum_{t\in\mathcal{T}_j} X_t^{\pi} \;\geq\; (1-\gamma)\sum_{t\in\mathcal{T}_j}\hat{X}_t^{k_j} - \frac{K\log(K/\alpha)}{\gamma} - (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\hat{X}_t^{k} - \frac{e\alpha K\left|\mathcal{T}_j\right|}{\gamma}.$$

Taking expectation with respect to the noisy rewards and the actions of Exp3.S we have:

$$\begin{aligned} \max_{k\in\mathcal{K}}\left\{\sum_{t\in\mathcal{T}_j}\mu_t^k\right\} - \mathbb{E}\left[\sum_{t\in\mathcal{T}_j}\mu_t^{\pi}\right] &\leq \sum_{t\in\mathcal{T}_j}\mu_t^{k_j} + \frac{K\log(K/\alpha)}{\gamma} + (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k + \frac{e\alpha K\left|\mathcal{T}_j\right|}{\gamma} - (1-\gamma)\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} \\ &= \gamma\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} + \frac{K\log(K/\alpha)}{\gamma} + (e-2)\frac{\gamma}{K}\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k + \frac{e\alpha K\left|\mathcal{T}_j\right|}{\gamma} \\ &\overset{(a)}{\leq} (e-1)\gamma\left|\mathcal{T}_j\right| + \frac{K\log(K/\alpha)}{\gamma} + \frac{e\alpha K\left|\mathcal{T}_j\right|}{\gamma}, \end{aligned} \tag{B.12}$$

for every batch $1 \leq j \leq m$, where (a) holds since $\sum_{t\in\mathcal{T}_j}\mu_t^{k_j} \leq \left|\mathcal{T}_j\right|$ and $\sum_{t\in\mathcal{T}_j}\sum_{k=1}^{K}\mu_t^k \leq K\left|\mathcal{T}_j\right|$.

Step 3. Taking (B.8) together with (B.12), and summing over $m = \lceil T/\Delta_T\rceil$ batches we have:

$$\begin{aligned} \mathcal{R}^{\pi}(\mathcal{V},T) &\leq \sum_{j=1}^{m}\left((e-1)\gamma\left|\mathcal{T}_j\right| + \frac{K\log(K/\alpha)}{\gamma} + \frac{e\alpha K\left|\mathcal{T}_j\right|}{\gamma} + 2V_j\Delta_T\right) \\ &\leq (e-1)\gamma T + \frac{e\alpha KT}{\gamma} + \left(\frac{T}{\Delta_T}+1\right)\frac{K\log(K/\alpha)}{\gamma} + 2V_T\Delta_T. \end{aligned} \tag{B.13}$$

Setting the tuning parameters to be $\alpha = \frac{1}{T}$ and $\gamma = \min\left\{1, \left(\frac{2V_T K\log(KT)}{(e-1)^2 T}\right)^{1/3}\right\}$, and selecting a batch size $\Delta_T = \left\lceil (\log(KT)\,K)^{1/3}\cdot\left(\frac{T}{2V_T}\right)^{2/3}\right\rceil$ one has:

$$\mathcal{R}^{\pi}(\mathcal{V},T) \;\leq\; 8(e-1)\left(KV_T\log(KT)\right)^{1/3}\cdot T^{2/3}.$$

Finally, whenever T is unknown, we can use Exp3.S as a subroutine over exponentially increasing pulls epochs $T_\ell = 2^\ell$, $\ell = 0, 1, 2, \ldots$, in a manner which is similar to the one described in Corollary 8.4 in Auer, Cesa-Bianchi, Freund and Schapire (2002) to show that since for any ℓ the regret incurred during T_ℓ is at most $C\left(KV_T \log\left(KT_\ell\right)\right)^{1/3} \cdot \ T_\ell^{2/3}$ (by tuning α and γ according to T_ℓ in each epoch ℓ), and for some absolute constant $\tilde{C}$, we get that $\mathcal{R}^\pi(\mathcal{V}, T) \ \leq \ \tilde{C}\left(\log\left(KT\right)\right)^{1/3}\left(KV_T\right)^{1/3} T^{2/3}$. This concludes the proof. $\square$

Appendix C

Appendix to Chapter 4

C.1 Theoretical results

This section provides three theoretical results that support ideas described in §4.1, §4.2, and §4.3.

Proposition C.1. *The CRP given by (4.1) is NP-hard.*

Proof. We establish that the CRP is NP-hard by showing that the Hamiltonian path problem (HPP), a known NP problem (Gary and Johnson 1979) can be reduced to a special case of the CRP. We denote by $\mathcal{G}(\mathcal{V}, \mathcal{E})$ a directed graph, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of arcs. An arc connecting one node v with another node v' is denoted by $e_{v,v'}$. When $v \in \mathcal{V}$ is connected to $v' \in \mathcal{V}$, one has $e_{v,v'} \in \mathcal{E}$. Given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, the HPP is to determine whether there exists a connected path of arcs in $\mathcal{E}$, that visits all the vertices in $\mathcal{V}$ exactly once.

We next show that the HPP can be reduced to a special case of the CRP. Fix a general, directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, and consider the following special case of the CRP, with a fixed x_0, in which:

- $T = |\mathcal{X}_0| - \ell$.
- $\mathcal{X}_1 = \mathcal{V}$, with $w(x) = 1$ for each article in $x \in \mathcal{X}_1$.
- $u_t = u_0 = u$ for all $t = 1, \ldots, T$ (the reader type is fixed, and in particular, independent of the length of her path and on the articles she visits along her path).

- $\ell = 1$, i.e., every recommendation consists of a single link. Whenever a recommendation A includes the link to article y that is placed at the bottom of an article x, we denote for simplicity $\mathbb{P}_{u,x,y}(A) = \mathbb{P}_{u,x}(y)$.

- $\mathbb{P}_{u,x}(y) \in \{0,1\}$ for all $y \in \mathcal{X}_t$, for all $t = 1, \ldots, T$ (the click probabilities for any potential recommended link are binary at any epoch). In particular, for any $x, y \in \mathcal{X}_1$ we set:

$$\mathbb{P}_{u,x}(y) \;=\; \begin{cases} 1 & \text{if } e_{x,y} \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

- $\mathbb{P}_{u,x_0}(y) = 1$ for all $y \in \mathcal{X}_1$ (the first link is clicked, regardless of the selected recommendation).

Then, given the landing article $x_0 \in \mathcal{X}_0$, the CRP takes the following form:

$$V_t^*(u, \mathcal{X}_t, x_{t-1}) \;=\; \max_{x_t \in \mathcal{X}_t} \left\{ \mathbb{P}_{u,x_{t-1}}(x_t) \left(1 + V_{t+1}^*(u, \mathcal{X}_{t+1}, x_t)\right) \right\},$$

for $t = 1, \ldots, T-1$, and

$$V_T^*(u, \mathcal{X}_T, x_{T-1}) \;=\; \max_{x_T \in \mathcal{X}_T} \left\{ \mathbb{P}_{u,x_{T-1}}(x_T) \right\}.$$

To complete the reduction argument, we observe that there exists a connected path of arcs in $\mathcal{E}$ that visits any vertex in $\mathcal{V}$ exactly once if and only if $V_t^*(u, \mathcal{X}_1, x_0) = T$, and therefore, by obtaining a solution to the CRP one solves the HPP. Since the HPP is NP-hard, the CRP must be NP-hard as well. This concludes the proof. □

We next analyze the performance gap between an optimal schedule of recommendation, and a sequence of myopic recommendations. To do so we focus on a special case of the CRP in which:

- $T = |\mathcal{X}_0| - \ell$.

- $u_t = u_0 = u$ for all $t = 1, \ldots, T$ (the reader type is fixed, and in particular, independent of the length of her path and on the articles she visits along her path).

- $\ell = 1$, i.e., every recommendation consists of a single link. Whenever a recommendation A placed at he bottom of an article x includes the link to article y, we denote for simplicity $\mathbb{P}_{u,x,y}(A) = \mathbb{P}_{u,x}(y)$.

- $w(x) = 1$ for any available article x.

Recall the definition of the CRP in (4.1), in this case it can be written as:

$$V_t^*(u, \mathcal{X}_t, x_{t-1}) \;=\; \max_{x_t \in \mathcal{X}_t)} \left\{ \mathbb{P}_{u,x_{t-1}}(x_t) \left(1 + V_{t+1}^*(u, \mathcal{X}_{t+1}, x_t)\right) \right\},$$

for $t = 1, \ldots, T-1$, and

$$V_T^*(u, \mathcal{X}_T, x_{T-1}) \;=\; \max_{x_T \in \mathcal{X}_T} \left\{ \mathbb{P}_{u,x_{T-1}}(x_T) \right\}.$$

Given a reader type u, initial set of available articles $\mathcal{X}_0$, and a landing article $x_0 \in \mathcal{X}_0$, we define the fraction of optimal performance recovered by the myopic policy as:

$$\Delta_T(u, \mathcal{X}_1, x_0) \;=\; \frac{V_1^m(u, \mathcal{X}_1, x_0)}{V_1^*(u, \mathcal{X}_1, x_0)}.$$

We note that $\Delta_T(u, \mathcal{X}_1, x_0) \in [0, 1]$ for any problem primitives u, $\mathcal{X}_0$, and $x_0 \in \mathcal{X}_0$. Let $\mathcal{G}_{T+1}$ denote the class of all sets of initial articles that includes $T+1$ articles. The following result shows that myopic recommendations may yield arbitrarily poor performance compared to optimal recommendations when the size of the network and the problem horizon grow.

Proposition C.2.

$$\inf_{\mathcal{X}_0 \in \mathcal{G}_{T+1}, x_0 \in \mathcal{X}_0, u \in \mathcal{U}} \Delta_T(u, \mathcal{X}_1, x_0) \to 0 \quad \textit{as} \quad T \to \infty$$

Proof. Fix $u \in \mathcal{U}$ and $\varepsilon \in (0, 1/2)$. Consider the next construction of a set of available articles $\mathcal{X}_0 \in \mathcal{G}_{T+1}$, with a selection $x_0 \in \mathcal{X}_0$, in which there exists a sequence of articles $x_0, \ldots, x_T$, such that:

$$\mathbb{P}_{u,x}(y) \;=\; \begin{cases} 1/2 - \varepsilon & \text{if } x = x_0 \text{ and } y = x_1 \\ 1 & \text{if } x = x_{t-1} \text{ and } y = x_t \text{ for some } t \in \{2, \ldots, T\} \\ 1/2 + \varepsilon & \text{if } x = x_0 \text{ and } y = x_T \\ 0 & \text{otherwise.} \end{cases}$$

Then, the optimal schedule of recommendation is to recommend article x_t at epoch t, generating $(1/2 - \varepsilon)\, T$ expected clicks. Moreover, any myopic schedule of recommendation will begin with recommending x_T at epoch $t = 1$, generating $(1/2 + \varepsilon)$ expected clicks. Therefore, one has:

$$\inf_{\mathcal{X}_0 \in \mathcal{G}_{T+1}, x_0 \in \mathcal{X}_0, u \in \mathcal{U}} \Delta_T(u, \mathcal{X}_1, x_0) \;\leq\; \Delta_T(u, \mathcal{X}_1, x_0) \;=\; \frac{\frac{1}{2} + \varepsilon}{\left(\frac{1}{2} - \varepsilon\right) T} \;\to 0 \quad \text{as} \quad T \to \infty,$$

which concludes the proof. □

The final result of this section analyzes the performance of the best one-step look-ahead recommendations (that solve (4.8)), compared to optimal recommendations (that solve (4.1)). For simplicity we focus on a special case of the CRP in which:

- $T = |\mathcal{X}_0| - \ell$.
- $u_t = u_0 = u$ for all $t = 1, \ldots, T$.
- $\ell = 1$, i.e., every recommendation consists of a single link. Whenever a recommendation A includes link to article y that is placed at the bottom of an article x, we denote for simplicity $\mathbb{P}_{u,x,y}(A) = \mathbb{P}_{u,x}(y)$.
- $w(x) = 1$ for any available article x.

We further assume the set of available articles $\mathcal{X}$ to be continuous and convex (and therefore it is not updated throughout the problem horizon). Specifically, we assume the set $\mathcal{X}$ is defined by

$$\mathcal{X} = \{(\gamma, \beta) \ : \ -1 \leq \gamma \leq 1, \ -1 \leq \beta \leq 1, \ \beta \leq 2 - \varepsilon - \gamma\},$$

for some $\varepsilon \in [0, 1]$. The set $\mathcal{X}$ is depicted in Figure C.1.

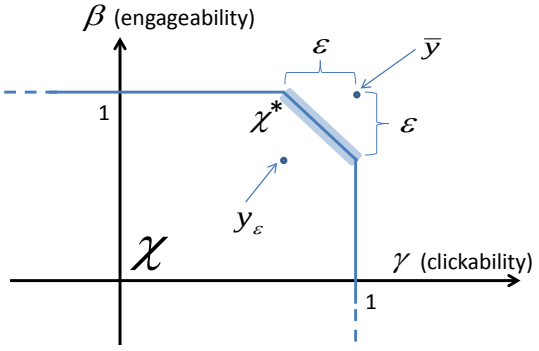


Figure C.1: **Convex set of available articles.** The optimal recommendation schedule is dominated by a policy that recommends $\bar{y}$ at any epoch (if $\bar{y}$ was an available article). The best one-step look-ahead schedule dominates a policy that recommends y_ε at each epoch.

Proposition C.3. *Let $\mathcal{X}$ be the set of available articles, and assume that $\mathbb{P}_{u,x}(y) \leq \bar{p}$ for all $x \in \mathcal{X}$, $y \in \mathcal{X}$, and $u \in \mathcal{U}$. Then,*

$$\frac{V_1^{one}(u, \mathcal{X}, x_0)}{V_1^*(u, \mathcal{X}, x_0)} \ \geq \ e^{-2\varepsilon} \left(\frac{1 + e^{-2\varepsilon}\bar{p}}{1 + \bar{p}} \right)^{T-1},$$

for any $u \in \mathcal{U}$ and $x_0 \in \mathcal{X}$.

Proof. Let $\mathcal{X}^*$ be defined as:

$$\mathcal{X}^* = \{(\gamma, \beta) \ : \ -1 \leq \gamma \leq 1, \ -1 \leq \beta \leq 1, \ \beta = 2 - \varepsilon - \gamma\}.$$

The set $\mathcal{X}^*$ is depicted in Figure C.1. Consider the one-step look-ahead policy, defined by

$$V_t^{one}(u, \mathcal{X}, x_{t-1}) \ = \ \mathbb{P}_{u,x_{t-1}}(x_t) \left(1 + V_{t+1}^{one}(u, \mathcal{X}, x_t)\right),$$

for $t = 1, \ldots, T-1$, where

$$x_t \in \arg\max_{y \in \mathcal{X}} \left\{ \mathbb{P}_{u,x_{t-1}}(y) \left(1 + \max_{x_{t+1} \in \mathcal{X}} \{\mathbb{P}_{u,y}(x_{t+1})\}\right)\right\},$$

for $t = 1, \ldots, T-1$, and where the last recommendation is simply myopic. For each $t \in \{1, \ldots, T\}$ we denote by γ_t and β_t the clickability and the engageability values of x_t, the article which is recommended at step t. We denote the point $(1-\varepsilon, 1-\varepsilon)$ by y_ε (see Figure C.1). Since for any $u \in \mathcal{U}$, $V_t^{one}(u, \mathcal{X}, x_{t-1})$ is increasing in γ_t and β_t for all $t \in \{1, \ldots, T\}$, any point that is selected by the one-step lookahead policy belongs to the set $\mathcal{X}^*$. Moreover,

$$V_1^{one}(u, \mathcal{X}, x_0) \ \geq \ \mathbb{P}_{u,x_0}(y_\varepsilon) \prod_{t=2}^{T} \left(1 + \mathbb{P}_{u,y_\varepsilon}(y_\varepsilon)\right),$$

for any $u \in \mathcal{U}$ and $x_0 \in \mathcal{X}$. In words, the one-step look-ahead policy performs at least as well as a policy that selects y_ε at each epoch.

Next, consider the optimal recommendation schedule, defined by

$$V_t^*(u, \mathcal{X}, x_{t-1}) \ = \ \max_{y \in \mathcal{X}} \left\{\mathbb{P}_{u,x_{t-1}}(x_t) \left(1 + V_{t+1}^*(u, \mathcal{X}, x_t)\right)\right\},$$

for $t = 1, \ldots, T-1$, where the last recommendation is myopic. We denote the point $(1, 1)$ by $\bar{y}$ (see Figure C.1). Clearly, $\bar{y}$ does not belong to $\mathcal{X}$. Moreover, since $V_1^*(u, \mathcal{X}, x_0)$ is increasing in γ_t and β_t for all $t \in \{1, \ldots, T\}$, one has that

$$V_1^*(u, \mathcal{X}, x_0) \ \leq \ \mathbb{P}_{u,x_0}(\bar{y}) \prod_{t=2}^{T} \left(1 + \mathbb{P}_{u,\bar{y}}(\bar{y})\right),$$

for any $u \in \mathcal{U}$ and $x_0 \in \mathcal{X}$. In words, the optimal recommendation schedule performs at most as well as a policy that selects $\bar{y}$ at each epoch. Moreover, one has:

$$\begin{aligned} \frac{\mathbb{P}_{u,x_0}(y_\varepsilon)}{\mathbb{P}_{u,x_0}(\bar{y})} \ &= \ \frac{e^{\alpha+\theta_u+\beta_{x_0}+1-\varepsilon}}{1+e^{\alpha+\theta_u+\beta_{x_0}+1-\varepsilon}} \cdot \frac{1+e^{\alpha+\theta_u+\beta_{x_0}+1}}{e^{\alpha+\theta_u+\beta_{x_0}+1}} \\ &= \ e^{-\varepsilon} \cdot \frac{1+e^{\alpha+\theta_u+\beta_{x_0}+1}}{1+e^{\alpha+\theta_u+\beta_{x_0}+1-\varepsilon}} \ \geq \ e^{-\varepsilon}, \end{aligned} \tag{C.1}$$

for any $u \in \mathcal{U}$ and $x_0 \in \mathcal{X}$. In addition, we have:

$$\begin{aligned}\frac{\mathbb{P}_{u,y_\varepsilon}(y_\varepsilon)}{\mathbb{P}_{u,\bar{y}}(\bar{y})} &= \frac{e^{\alpha+\theta_u+2-2\varepsilon}}{1+e^{\alpha+\theta_u+2-2\varepsilon}} \cdot \frac{1+e^{\alpha+\theta_u+2}}{e^{\alpha+\theta_u+2}} \\ &= e^{-2\varepsilon} \cdot \frac{1+e^{\alpha+\theta_u+2}}{1+e^{\alpha+\theta_u+2-2\varepsilon}} \;\geq\; e^{-2\varepsilon}, \end{aligned} \tag{C.2}$$

for any $u \in \mathcal{U}$. Therefore, one has for any $u \in \mathcal{U}$, $x_0 \in \mathcal{X}$, and $\delta \geq \delta_\varepsilon$:

$$\begin{aligned}\frac{V_1^{one}(u,\mathcal{X},x_0)}{V_1^{*}(u,\mathcal{X},x_0)} &\geq \frac{\mathbb{P}_{u,x_0}(y_\varepsilon)\prod_{t=2}^{T}\left(1+\mathbb{P}_{u,y_\varepsilon}(y_\varepsilon)\right)}{\mathbb{P}_{u,x_0}(\bar{y})\prod_{t=2}^{T}\left(1+\mathbb{P}_{u,\bar{y}}(\bar{y})\right)} \\ &\overset{(a)}{\geq} e^{-2\varepsilon}\cdot\prod_{t=2}^{T}\left(\frac{1+\mathbb{P}_{u,y_\varepsilon}(y_\varepsilon)}{1+\mathbb{P}_{u,\bar{y}}(\bar{y})}\right) \\ &\overset{(b)}{\geq} e^{-2\varepsilon}\cdot\prod_{t=2}^{T}\left(\frac{1+e^{-2\varepsilon}\mathbb{P}_{u,\bar{y}}(\bar{y})}{1+\mathbb{P}_{u,\bar{y}}(\bar{y})}\right) \\ &\overset{(c)}{\geq} e^{-2\varepsilon}\cdot\prod_{t=2}^{T}\left(\frac{1+e^{-2\varepsilon}\bar{p}}{1+\bar{p}}\right) \;=\; e^{-2\varepsilon}\cdot\left(\frac{1+e^{-2\varepsilon}\bar{p}}{1+\bar{p}}\right)^{T-1}, \end{aligned}$$

where: (a) holds by (C.1); (b) holds by (C.2); and (c) holds since $\mathbb{P}_{u,\bar{y}}(\bar{y}) \leq \bar{p}$ for all $u \in \mathcal{U}$. This concludes the proof. □

C.2 Choice model and estimation

In this section we detail the estimation process described in §3. We start by a description of the control parameters that were used.

Given an assortment A, and an article y that appears in A, let $p : (y, A) \to \{0, \ldots, \ell-1\}$ denote the *position* of article y in the assortment A. If $p(y, A) = 0$, then y is recommended in the highest position in A, and if $p(y, A) = \ell - 1$, y is recommended in the lowest position in A. Then, given a user $u \in \mathcal{U}$ and an assortment A placed at a host article $x \in \mathcal{X}$ we define:

$$\mathbb{P}_{u,x,y}(A) \;=\; \begin{cases} \frac{\phi_{u,x,y}(A)}{1+\sum_{y' \in A}\phi_{u,x,y'}(A)} & \text{if } y \text{ appears in } A \\ 0 & \text{otherwise.} \end{cases}$$

Whenever y appears in A, we define:

$$\phi_{u,x,y}(A) \;=\; \exp\left\{\alpha + \theta_u + \beta_x + \gamma_y + \mu_{x,y} + \lambda_{p(y,A)}\right\}.$$

The host effect and the link effect are discussed in §4.2.1.

User effect. The parameter θ_u captures the effect of the user type, and in particular, of experienced users. We differentiate between two types of users: experienced users (that have clicked on an Outbrain recommendation before) and inexperienced users. Thus, we have $\theta_u \in \{\theta_{exp}, \theta_{inexp}\}$ for each $u \in \mathcal{U}$, where we normalize by setting $\theta_{inexp} = 0$ (treating unexperienced users as a benchmark) and estimate θ_{exp} from the data. Experienced readers were defined as ones that clicked on a recommendation during an initial period of of 10 days. The main motivation for distinguishing between experienced and inexperienced users is implied by an aggregated data analysis summarized in Table C.1, indicating that while most of the users are inexperienced, experienced users visit the publisher more than twice the times inexperienced ones visit it on average, and on average an experienced user clicks more than twice the times (per visit) an inexperienced one does .

User type	Population share	Visits share	Clicks per visit
Experienced	8.2%	16.9%	0.23
Inexperienced	91.8%	83.1%	0.10

Table C.1: **Experienced vs. Inexperienced users.** The table summarizes the differences between inexperienced users and experienced users, as was observed along the 30 days that followed the initial period.

Contextual relation effect. To formulate the effect of contextual connection between the host article and a recommended article we distinguished between cases in which the host and the recommended article are from the same topic category (using the classification to 84 topic categories), from cases in which the two articles are from different categories. Thus, we have $\mu_{x,y} \in \{\mu_{related}, \mu_{unrelated}\}$ for each $x, y \in \mathcal{X}$, where we normalize by letting $\mu_{unrelated} = 0$ (treating recommendations in which both articles are not in the same category as a benchmark) and estimate $\mu_{related}$ from the data. We note that the contextual connection effect may be formulated in many different ways. One alternative that we addressed is a more general approach of estimating the entries of the 84 by 84 (non-symmetric) matrix that describes the explicit connection between each pair of topic categories. This clearly increases the number of model parameters, and by doing so we observed no improvement over the predictive power of the model (through the approach detailed in §4.2.3). Another potential approach is to use an alternative

classification to topics; testing a coarser classifications to 9 categories showed no improvement in the prediction power of the model.

Position effect. This effect is captured by the variables $\lambda_p \in \{\lambda_0, \ldots, \lambda_5\}$, that correspond to recommendations that list 6 internal recommendations, as in the data set that was used to estimate the model. We set $\lambda_0 = 0$ (treating the highest position as a benchmark), and estimate the other 5 parameters from the data to measure the effect of lower positions.

The estimation process. The model was estimated in each two-hour batch by applying a Newton step method to maximize the log-likelihood of the model. The estimation results in all 360 estimation batches were consistent. The values of the control parameters received from the estimation over the first batch is presented in Table C.2. The estimate of θ_{exp} quantifies the

Effect	Parameter	Estimate	Standard error
Intercept	α	-4.45	$3.9 \cdot 10^{-6}$
User	θ_{exp}	1.13	$1.7 \cdot 10^{-3}$
Contextual relation	$\mu_{related}$	-0.10	0.02
Position	λ_1	-1.10	$4.9 \cdot 10^{-4}$
	λ_2	-1.71	$1.4 \cdot 10^{-5}$
	λ_3	-2.03	$1.9 \cdot 10^{-5}$
	λ_4	-2.28	$2.1 \cdot 10^{-5}$
	λ_5	-2.29	$2.1 \cdot 10^{-5}$

Table C.2: **Estimation of auxiliary parameters.** The estimated values and standard errors for the control parameters over the first estimation batch. All estimates are at significance level $p < 0.01$.

positive effect of previous user experience on the likelihood to click. It is in tune with the statistics presented in Table 1: users that are aware of and familiar with the recommendation service tend to use it more often than inexperienced users do. The estimate of $\mu_{related}$ quantifies the effect of contextual relation between the host and recommended articles. Interestingly, it suggests that on average, users tend to click less when the recommended article directly relates to the article they just finished reading, relative to cases in which such direct relation does not exist. One potential bias here, is that links that directly relate to the host article are typically generated by a class of contextual-focused algorithms, that may be less successful than other classes of algorithms, for

example, behavior-focused ones. The estimates of $\lambda_1, \ldots, \lambda_5$ quantify the "cost of lower position" in the recommendation box, relative to the highest position. Not surprisingly, the lower the link is, the smaller is the likelihood of a reader to click on that link.