TEXTO PARA DISCUSSÃO

No. 674

Online Action Learning in High Dimensions: A New Exploration Rule for Contextual εt-Greedy Heuristics

Claudio C. Flores
Marcelo C. Medeiros

PUC RIO

DEPARTAMENTO DE ECONOMIA
www.econ.puc-rio.br

Online Action Learning in High Dimensions: A New Exploration Rule for Contextual ϵ_t-Greedy Heuristics

Claudio C. Flores
Pontifical Catholic University of Rio de Janeiro
claudioflores@hotmail.com

Marcelo C. Medeiros
Pontifical Catholic University of Rio de Janeiro
mcm@econ.puc-rio.br

September 28, 2020

Abstract

Bandit problems are pervasive in various fields of research and are also present in several practical applications. Examples, including dynamic pricing and assortment and the design of auctions and incentives, permeate a large number of sequential treatment experiments. Different applications impose distinct levels of restrictions on viable actions. Some favor diversity of outcomes, while others require harmful actions to be closely monitored or mainly avoided. In this paper, we extend one of the most popular bandit solutions, the original ϵ_t-greedy heuristics, to high-dimensional contexts. Moreover, we introduce a competing exploration mechanism that counts with searching sets based on order statistics. We view our proposals as alternatives for cases where pluralism is valued or, in the opposite direction, cases where the end-user should carefully tune the range of exploration of new actions. We find reasonable bounds for the cumulative regret of a decaying ϵ_t-greedy heuristic in both cases, and we provide an upper bound for the initialization phase that implies the regret bounds when order statistics are considered to be at most equal but mostly better than the case when random searching is the sole exploration mechanism. Additionally, we show that end-users have sufficient flexibility to avoid harmful actions since any cardinality for the higher-order statistics can be used to achieve stricter upper bound. In a simulation exercise, we show that the algorithms proposed in this paper outperform simple and adapted counterparts.

Keywords: Bandit, sequential treatment, high dimensions, LASSO, regret

1 Introduction

In this paper, we combine contextual decaying ϵ_t-greedy heuristics, among the most popular bandit solutions, with high-dimensional setups. We propose two algorithms to address different levels of restrictions regarding the exploration of new actions. Our framework is especially useful for situations where an agent must learn the best course of action to maximize some reward through experience and the observation of a large pool of covariates.

A multiarmed bandit problem can be interpreted as a sequential treatment, where a limited set of resources must be allocated between alternative choices to maximize utility. The properties of the choices are not fully known at the time of allocation and may become better understood as time passes, provided a learning rule with theoretical guarantees is available. A particularly useful extension of the bandit problem is called the contextual multiarmed bandit problem, where observed covariates yield important information to the learning process in the sense that the supposed best policy may be predicted; see, for instance, Auer (2003); Li et al. (2010); Langford and Zhang (2008); Agrawal and Goyal (2012).

Contextual multiarmed bandit problems have applications in various areas. For instance, large online retailers must decide on real-time prices for products and differentiate among distinct areas without complete demand information; see, for example, Dani et al. (2008) and den Boer (2013). Arriving customers may take purchase decisions among offered products based on maximizing their utility. If information on consumers' utility is available, the seller could learn which subset of products to offer (Sauré and Zeevi (2013)). Further, the reserve price of auctions could be better designed to maximize revenue (Cesa-Bianchi et al. (2013)). Mechanisms design in the case where agents may not know their true value functions but the mechanism is repeated for multiple rounds can take advantage of accumulated experience (Kandasamy et al. (2020)). Also, sequential experiments or programs, including public policies (Tran-Thanh et al. (2010) devises an algorithm that consider costly policies), may be assigned under the scope of learning problems. In this regard, excellent works can be found in Kock and Thyrsgaard (2017), Kock et al. (2018) and Kock et al. (2020).

1.1 Motivation and Comparison with the Literature

Designing a sequence of policies to minimize error is a difficult task and, for a considerable period in the past, was also a computationally intractable goal. In this respect, several heuristics with well-behaved properties have emerged in the literature, such as Thompson sampling (Agrawal and Goyal, 2012; Russo and Van Roy, 2016), upper confidence bounds (Dani et al., 2008; Abbasi-Yadkori et al., 2011) and greedy algorithms (Auer, 2003; Bastani et al., 2017; Goldenshluger and Zeevi, 2013). In a very recent work, Chen et al. (2020) established asymptotic normality in a ϵ-greedy contextual bandit.

Provided that a large pool of characteristics has been collected from the target population, a sparse setup has the potential to catalyze the benefits of large information sets into strong predictive power of what the rewards would be for a chosen action in a contextual bandit framework. This possible superior performance translates directly into better exploitation steps and, in our case, improved exploration.

Few papers consider high-dimensional bandit setups. These studies include Carpentier and Munos (2012), Abbasi-Yadkori and D. Pál (2012), Deshpande and Montanari (2012), Bouneffouf et al. (2017), Bastani and Bayati (2019), and Krishnamurthy and Athey (2020). However, most of the previous papers either extend a version of the upper confidence bound (UCB) algorithm to a sparse high-dimensional setup or put emphasis on exploitation-only solutions.

Our contribution enriches the current set of high-dimensional algorithms by providing solutions that consider distinct levels of restrictions in exploration. Consider, for example, a recommendation system in which some accidental discoveries have a positive impact on user experience. In this case, one of our algorithms can provide a better outcome than its counterparts in the high-dimensional literature since it explores fully at random and not based on some previously imputed restrictions.

The abovementioned solution favors diversity, which is clearly suitable for some applications. However, in many practical situations, exploration at random can be unethical, unfeasible, or harmful, for example, in medical trials, where an optimal dosage of a new drug is to be determined. Therefore, we also provide a solution to these cases that leverages the fact that nonharmful actions should belong to a very intuitive set comprising the most promising actions based, for example, on the predicted output's highest-order statistics.

Our work is related to Bastani and Bayati (2019). The studied algorithm, however, gathers parameter information by forcing arms to be played at specific instants of time and chooses the best action based on full sample estimation. This procedure belongs to the class of exploitation-only algorithms in the sense that it always chooses the best predicted policy. Since distinct practical applications dictate the boundaries and weights one should place on exploration and exploitation, we view our algorithms as alternatives that extend the work in Bastani and Bayati (2019) in the direction of greater exploration.

Krishnamurthy and Athey (2020) propose a variation of the traditional LinUCB algorithm depicted in Li et al. (2010) to select personalized questions for surveys based on observed responses. Regret properties are proved when ridge regression and elastic net are used for parameter estimation. In general, UCB rules adhere to the principle that when it comes to select nongreedy actions, it would be better to select the action according to its potential of being optimal. Typically, this task is achieved through the construction of confidence sets, which may favor some actions over others. Given the environment's limitations, the proposed methods provide greater diversity or, when required, greater flexibility to select nonharmful actions.

1.2 Main Takeaways

We show that distinct levels of restrictions in the exploration of new actions can be settled by using variations of the original multiarmed ϵ_t-greedy heuristic. Our contributions are twofold: the extension of the ϵ_t-greedy heuristics to the high-dimensional context and its subsequent refinement via the introduction of a competing exploration mechanism that counts with a high-order statistics searching set.

Specifically, the first contribution is suitable for applications with no restrictions on exploration. In these cases, at each time, the algorithm selects a random action (exploration) with probability ϵ_t, which is just the original ϵ_t-greedy rule. Despite of its simplicity, this rule efficiently introduces diversity in outcomes. Our second contribution is designed for cases with mild restrictions in exploration. That is, we construct, at each period, a set consisting of the most promising actions based on the predicted output's highest-order statistics. The implicit assumption is that a promising action should not be harmful to the activity.

Although not mandatory, both algorithms are equipped with an initialization phase where information about the parameters is gathered by attempting distinct actions. We show that the cumulative regrets of both contextual ϵ_t-greedy algorithms are reasonably bounded, implying that even in the high-dimensional setup, effective learning is achieved by employing the aforementioned rules. We also provide an upper bound for the initialization phase that implies regret bounds when order statistics are considered to be at most equal to or better than the case when random search is the unique exploration mechanism. This is especially important for cases when random exploration at its full extent is not advisable. Practitioners can use this result to setup an acceptable initialization phase that guarantees that exploring at a selective set of actions would yield better results.

In addition, we show that it is viable to pick any cardinality for the high-order statistics set and still respect the limits established in this paper. This approach introduces flexibility, as it is possible to choose the range of alternatives to explore. On the other hand, we show that competition between searching mechanisms may not be optimal, as the dominance of one or another should be considered as a function of the initialization phase. In a simulation exercise, we show that the algorithms proposed in this paper outperform simple and adapted (to the high-dimensional context) counterparts.

To the best of our knowledge, no previous works have specifically addressed a decaying ϵ_t-greedy algorithm in this manner.

1.3 Organization of the Paper

The rest of this paper is structured as follows. Section 2 establishes the framework and the main assumptions for the regret analysis, while Section 3 depicts the proposed algorithms. Section 4 exhibits the main theoretical results of the paper. Section 5 provides a sensitivity analysis of the algorithms with relation to parameters set by the user, a comparison among simple and adapted algorithms and suggestions for the application of our rules in two practical problems. Section 6 concludes this work. All proofs and auxiliary results are relegated to the Appendix. Supplementary Material provides additional results.

1.4 Notation

Bold capital letters $\boldsymbol{X}$ represent matrices and small bold letters $\boldsymbol{x}$ represent vectors. $\|\cdot\|$ is the vector norm, while # is the cardinality of a set. Matrices or vectors followed by subscript or superscript parentheses denote specific elements. For example, $\boldsymbol{X}_{kt}^{(j)}$ is the j-th column of matrix $\boldsymbol{X}_{kt}$. Likewise, $\boldsymbol{x}_{t,(j)}$ is the j-th scalar element of vector $\boldsymbol{x}_t$. Finally, $\mathbb{1}_{\{x \in X\}}$ is the indicator function that takes a value of 1 when $x \in X$. Additional notation is presented throughout the paper whenever necessary.

2 Setup and Assumptions

Contextual bandit problems are intrinsic to various fields of research. In this paper, we use the nomenclature derived from the treatment effects literature such that terms as "arms" and "rewards" may be replaced by "policies" and "treatment effects", respectively.

Consider an institution, for example, a central planner or a firm, that offers a finite sequential program. The planner has to choose, at each instant of time, the best possible policy (arm) to implement, for example, a deterministic function of covariates to assign units to the treated and nontreated groups. The goal is to set a sequence of policies to maximize some measure of treatment effects (rewards).

In many cases, sequential programs have inherent costs that may be enhanced by frequent changes in the treatment assignment rules. Furthermore, even without costs, one may use fairness arguments to rule out the possibility that changes in the program are so substantial that they alter the program's nature. In this case, participants would be in a position of a total lack of knowledge as the next policy could be completely divergent from the previous ones. Assumption 1 formalizes a more realistic case that will be our subject of study.

Assumption 1 (Policy Assumptions). *i. Let $(\mathcal{W}, \|\cdot\|)$ be any normed vector space. The planner has at its disposal a finite set of policies to be tested $\mathcal{W}_p \equiv \{\boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_{w-1}\}$, $\mathcal{W}_p \subset \mathcal{W}$, where $\boldsymbol{\omega}_k \in \mathbb{R}^c$ and c is arbitrary.*

ii. The planner starts from a pretty good idea of a reasonable initial policy $\boldsymbol{\omega}_0$ and selects alternatives inside the ball $\mathcal{B}(\boldsymbol{\omega}_0, \tau)$, provided that $\mathbf{0} \in \mathcal{B}(\boldsymbol{\omega}_0, \tau)$.

Remark 1. *We could relax the normed vector space requirement to any metric space in general. For our purposes, it is imperative to compare policies that, provided a metric is in place, are achievable in more complex spaces. Without loss of generality, we say that $\boldsymbol{\omega}_k \in \mathbb{R}^c$ because euclidean spaces are simple to work with. In this case, $\mathcal{B}(\boldsymbol{\omega}_0, \tau) = \{\boldsymbol{\omega}_k \in \mathbb{R}^c | \, \|\boldsymbol{\omega}_k - \boldsymbol{\omega}_0\| \leq \tau\}$. As a simple example, w_k could be a scalar chosen by the planner as a cutoff for a regression discontinuity design. Units with covariates above the cutoff would be assigned to the treated group, and those below the cutoff would not be treated. The finite set of policies, its cardinality, $\#\mathcal{W} = w$, and the level of dissimilarity among policies, measured by τ, directly depend on the degree of slackness in the environment. In most practical situations, a large number of very unlikely policies is not a plausible option. However, we do not restrain them in any manner. Furthermore, it is possible to connect Assumption 1 to the case in which the planner may initially have a pretty good idea of a reasonable policy to employ but, for example, may be interested in fine-tuning activities to improve the results. The requirement that $\mathbf{0} \in \mathcal{B}(\boldsymbol{\omega}_0, \tau)$ is easily satisfied for a standardized set of policies.*

Let $(\boldsymbol{\Omega}, \mathcal{F}, \mathbb{P})$ be a probability space. At an arbitrary time instant $t \in \mathcal{T} \equiv \{1, 2, \ldots, T\}$, the planner observes covariates, e.g., individual characteristics of its target sample, as well as the sequence of its past realizations $\{\boldsymbol{x}_t\}_{t=1}^{t}$ identically and independently distributed (iid) from $\mathbb{P}$. She also knows the past rewards[1] $\{y_{kt}\}_{t=1}^{t-1}$, for all values of k as long as $\boldsymbol{\omega}_k$ have been implemented before t. Then, at each time t, the planner must choose a policy $\boldsymbol{\omega}_k$ from $\mathcal{W}$ to maximize some key variable (reward or some measure of treatment effect)[2]. The range of y_{kt} is a subset of $\mathcal{Y} \subset \mathbb{R}$, while that of $\boldsymbol{x}_t$ is a subset of $\mathcal{X} \subset \mathbb{R}^p$, where p may grow with the sample size. However, to simplify the notation, in the rest of this paper, we do not exhibit this dependence (between p and T) explicitly.

The connection between covariates and rewards is stated as follows:

Assumption 2 (Contextual Linear Bandit)**.** *There is a linear relationship between rewards and covariates of the form:*

$$y_{kt} = \boldsymbol{\beta}_k' \boldsymbol{x}_t + \epsilon_{kt}, \tag{1}$$

where y_{kt} is some measure of treatment effects at time t, as a result of the implementation of policy k, conditional on the covariates $\boldsymbol{x}_t$ (and all their past realizations) and an idiosyncratic shock ϵ_{kt}. $\forall k$, $\boldsymbol{\beta}_k$ belongs to the parametric space $\mathcal{B} \subset \mathbb{R}^p$. Furthermore:

i. $\forall t \in \mathcal{T}$, $|\boldsymbol{x}_{t,(j)}| \leq \theta_x$, $j \in \{1, ..., p\}$.

ii. $\forall k \in \{0, ..., w-1\}, t \in \mathcal{T}$, *the sequence* $\{\epsilon_{kt}\}$ *is composed of independent centered random variables with variance* $\mathbb{E}(\epsilon_i^2) < \sigma^2$.

Remark 2. *Assumption 2 restrains our setup to linear bandit problems. Rewards are policy-/time-dependent in the sense that the dynamics of $\boldsymbol{x}_t$ interfere with the level of reward. However, depending on the chosen policy $\boldsymbol{\omega}_k$, the mechanism ($\boldsymbol{\beta}_k$) that "links" covariates to rewards is different. Moreover, in contrast to several papers that make specific distributional assumptions concerning the covariates and the error term, we require only that covariates be*

[1] The planner observes $\boldsymbol{x}_t$ at each time t but does not yet know y_t.

[2] For ease of notation, in our setup, y_t is a scalar random variable, but the reader will recognize throughout this paper that this choice is not restrictive. Multivariate versions are allowed.

bounded in absolute terms. Regarding the sequence of errors, we only bound their variances. Both assumptions are necessary to guarantee that instantaneous regrets (defined below) do not have explosive behavior.

Clearly, two pieces of nomenclature have been used: policies chosen by the planner and "mechanisms" through which these policies operate. Assumption 3 connects them:

Assumption 3 (Metric Spaces)**.** *There is an h-Lipschitz function* $h_f : \mathbb{R}^c \to \mathcal{B}$

Remark 3. *Assumption 3 is a restriction on the joint behavior of the two relevant metric spaces we are working with,* $\mathcal{B}$ *and* $\mathbb{R}^c$*. It is advisable to impose some healthy patterns to avoid the possibility that small changes in mechanisms could result in substantial changes in policies, which would not be expected in most practical situations. In the case considered by Assumption 3 we have that* $d_{\mathcal{B}}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_j) \leq h d_c(\boldsymbol{\omega}_k - \boldsymbol{\omega}_j)$*, for* $h \in \mathbb{R}^+$ *the Lipschitz constant and* d_c *and* $d_{\mathcal{B}}$*, relevant metrics for the two spaces.*

One of the most useful instruments to assess the effectiveness of bandit algorithms is the regret function, which, in general, may be studied in its instantaneous or cumulative version. Regret represents the difference (in a naive sense) between the expected reward obtained by choosing an arbitrary policy and that obtained by picking the best policy. Clearly, the term best policy does not refer to the absolute sense but the best conditional on the fact that it belongs to the available set of policies. Definition 1 formalizes these concepts.

Definition 1. *(Regret Functions) The instantaneous* (r_t) *regret function of implementing any policy* $\boldsymbol{\omega}_k \in \mathcal{W}$ *at time* $t \in \mathcal{T}$*, leading to the reward* y_{kt}*, and the respective cumulative* (R_T) *regret until time* T *are defined as:*

$$r_t = \mathbb{E}\left[\max_{j \in \{0,\ldots,w-1\}} (y_{jt} - y_{kt})\right] \quad \text{and} \quad R_T = \sum_{t=1}^{T-1} r_t$$

Motivated by the high-dimensional context, we perform Lasso estimation in the following sections. These estimators operate on a familiar and imperative assumption of sparsity, i.e., that in the true model, not all covariates are important to explain a given dependent variable. Regarding this aspect, we define the sparsity index in Definition 2 and impose the well-known compatibility condition for random matrices in the Assumption 4, which is standard in the high-dimensional literature.

Definition 2. *(Sparsity Index) For any $p > 0$ and $k \in \{1, \ldots, p\}$, define $S_0 \equiv \{k | \beta_k^0 \neq 0\}$ and the sparsity index as $s_0 = \#S_0$.*

It is important to establish an additional piece of notation. Define $\beta_k[S_0] \equiv \beta_k \mathbb{1}\{k \in S_0\}$ and $\beta_k[S_0^C] \equiv \beta_k \mathbb{1}\{k \notin S_0\}$.

Assumption 4. *(Compatibility Condition) For an arbitrary $(n \times p)$-matrix $\boldsymbol{X}$ and $\forall \boldsymbol{\beta} \in \mathbb{R}^p$, such that $\left\|\boldsymbol{\beta}[S_0^C]\right\|_1 \leq 3 \left\|\boldsymbol{\beta}[S_0]\right\|_1$, for some S_0, $\exists\, \phi_0 > \sqrt{32 b s_0} > 0$, with $b \geq \max_{j,k} |(\widehat{\boldsymbol{\Sigma}})_{j,k} - (\boldsymbol{\Sigma})_{j,k}|$ such that:*

$$\left\|\boldsymbol{\beta}[S_0]\right\|_1^2 \leq \frac{s_0 \boldsymbol{\beta}' \boldsymbol{\Sigma} \boldsymbol{\beta}}{\phi_0^2},$$

where $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ are the empirical and population covariance matrices of $\boldsymbol{X}$, respectively.

Finally, we impose a bounding condition for the density of covariates near a decision boundary, as in Tsybakov (2004), Goldenshluger and Zeevi (2013) and Bastani and Bayati (2019), among others.

Assumption 5. *(Margin Condition) For $k \in \mathbb{R}^+$, $\exists\, C_m \in \mathbb{R}^+$, $C_m \leq \frac{\phi_0^2}{8\theta_x s_0 \lambda}$, such that for $i, j \in \{0, \ldots, w-1\}$, $\mathbb{P}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i) \leq k\right] \leq C_m k$.*

Remark 4. *Assumption 5 is related to the behavior of the distribution of $\boldsymbol{x}_t$ "near" a decision boundary. In these cases, there is a possibility for rewards to be so similar that small deviations in estimation procedures could lead to suboptimal policies being selected by the algorithms. With this assumption, we impose that even in small balls of similar policies, there is a strictly positive probability that rewards ($\boldsymbol{x}_t' \boldsymbol{\beta}_i$) for a given policy $\boldsymbol{\omega}_i$ are strictly greater than those of any other policy $\boldsymbol{\omega}_k$. That is, there is no doubt about what policy is the best. In contrast to other papers, we establish an upper bound for the constant C_m as a function of the intrinsic parameters of the problem.*

3 Algorithms and Estimation Procedures

Choosing any policy at each instant of time generates the well-known problem of bandit feedback, which in general terms, relates to the fact that a planner following an arbitrary algorithm obtains feedback for only the chosen policy. Other possible rewards are simply not observable and the best possible one, at each time t, remains unknown to the planner. This intrinsic characteristic can lead to incorrect premature conclusions, for example, in cases when a policy had not been frequently tested in the past. In this case, it may be labeled as a suboptimal policy, while in fact, it simply did not have sufficient opportunity to prove itself. Additionally, bandit feedback poses serious problems for the evaluation of different policies and comparison of algorithms using real data sets. If a target policy, different than the implemented one, is to be evaluated, difficulties arise, leading to alternatives such as counterfactual estimation (Agarwal et al., 2017).

Another somewhat different feature, but equally vital for the efficiency of bandit algorithms, relates to the way that policies are selected. After some time, the planner has already formed her opinion about the implemented policies. Then, a crucial decision must be made: exploit and use the most profitable policy, in a predicted sense in the case of contextual multiarmed bandits, or explore and implement a new one, taking advantage of the fact that in a changing world, the past may not reflect the future, thereby preventing the algorithm from becoming stuck on suboptimal policies (best only in the past). This exploitation-exploration trade-off is well-known in the bandit literature and dictates the properties of the regret function; see, for example, Auer (2003) and Langford and Zhang (2008).

In general terms, all bandit algorithms take the abovementioned problems into consideration while pursuing their main goal of a bounded well-behaved regret function. The ϵ_t-greedy heuristic is no different. It is important, however, to properly review since the proposed algorithms in this paper reflect and extend its ideas.

Define the action function $I : \mathcal{T} \to \mathcal{W}$, such that for each $t \in \mathcal{T}$, $I(t) = \boldsymbol{\omega}_k$ represents the policy selected by the planner. Then, Definition 3 presents the ϵ_t-greedy algorithm, which is the same one established in Auer et al. (2002).

Definition 3 (ϵ_t-Greedy Heuristic)**.** *Let $c > 0$ and $0 < d < 1$. Let $w \in \mathbb{N}+$, $w > 1$ and define the sequence $\epsilon_t \in (0, 1]$, $t \in \mathcal{T}$, by $\epsilon_t \equiv \min\left\{1, \frac{cw}{d^2 t}\right\}$. Then, the ϵ_t-greedy algorithm is*

Algorithm 1: ϵ_t-Greedy Heuristic

```
input parameters: c, d, w
for t ∈ T do
    ε_t ⟵ min{1, cw/(d²t)};
    q_t ⟵ U(0, 1);
    if q_t ≤ ε_t then
        a_t ⟵ U(0, w);
        I(t) ⟵ ω_{a_t};
    else
        b_t ⟵ argmax_{j∈{0,...,w−1}} (1/(t−1)) Σ_{i=1}^{t−1} y_{ji};
        I(t) ⟵ ω_{b_t};
    end
end
```

The ϵ_t-greedy heuristic gives random weights to the exploration-exploitation trade-off. That is, with probability ϵ_t, it explores selecting a random policy in the set $\mathcal{W}$ and, with probability $1 - \epsilon_t$ it exploits selecting the best empirical policy (in the average sense).

Note the total absence of covariates, which makes Definition 3 appropriate for multi-armed bandits (without context). To extend this framework to cases where covariates play an important role, we expand the ideas in Bastani et al. (2017) to consider not only the probability exploration-exploitation trade-off but also to allow for high dimensions.

Define the partition of the set $\mathcal{T}$, $\{\mathcal{T}^{(l)}, \mathcal{T}^{(l^c)}\}$, where $\mathcal{T}^{(l)} \equiv \{t \in \mathcal{T} | t \leq l\}$ and $\mathcal{T}^{(l^c)}$ is the respective complement. $l \in \mathbb{N}^+$, $l > 1$ is the length of the initialization phase, and we require that $l = vw$, which implies that every policy in $\mathcal{W}$ is implemented v times in this phase. Definition 4 formalizes the contextual lasso greedy (CLG) algorithm.

Definition 4 (CLG Algorithm)**.** *Let $c > 0$, $0 < d < 1$ and ϵ_t be defined in the same way as in Definition 3. Let $w \in \mathbb{N}+$, $w > 1$ and $v \in \mathbb{N}^+$. Then, the CLG algorithm is:*

Algorithm 2: CLG Algorithm

---

input parameters: c, d, w,v

Inicialization;

for $i \in \{1, 2, ..., v\}$ **do**

  for $j \in \{1, 2, ..., w\}$ **do**

    $I(t) \longleftarrow \boldsymbol{\omega}_j$;

    Update $\widehat{\boldsymbol{\beta}}_j$;

  end

end

Exploration-Exploitation;

for $t \in \mathcal{T}^{(l^c)}$ **do**

  $\epsilon_t \longleftarrow \min\left\{1, \frac{cw}{d^2 t}\right\}$; $q_t \longleftarrow \mathsf{U}(0, 1)$;

  if $_t \leq \epsilon_t$ **then**

    $a_t \longleftarrow \mathsf{U}(0, w)$; $I(t) \longleftarrow \boldsymbol{\omega}_{a_t}$;

    Update $\widehat{\boldsymbol{\beta}}_{a_t}$;

  else

    $b_t \longleftarrow \underset{j \in \{0,...,w-1\}}{\arg\max}\ \widehat{y}_{jt}$; $I(t) \longleftarrow \boldsymbol{\omega}_{b_t}$;

    Update $\widehat{\boldsymbol{\beta}}_{b_t}$;

  end

end

The CLG algorithm is a natural expansion of the ϵ_t-greedy solution to contextual settings. It is equipped with an initialization phase, and the exploited selected policy is given by the best estimated/predicted reward.

An important part of CLG is that the planner is required to update $\widehat{\boldsymbol{\beta}}_k$ only when $I(t) = \boldsymbol{\omega}_k$. Define $\mathcal{A}_{kt} \equiv \{t \in \mathcal{T} | I(t) = \boldsymbol{\omega}_k\}$, and let $n_{kt} \equiv \#\mathcal{A}_{kt}$ be the number of times an arbitrary policy $\boldsymbol{\omega}_k$ has been tested until time t. Let $\boldsymbol{X}_{kt}$ be an $n_{kt} \times p$ matrix containing all unit characteristics until time t, provided that $t \in \mathcal{A}_{kt}$. $\boldsymbol{y}_{kt}$ and $\boldsymbol{\epsilon}_{kt}$ are the $n_{kt} \times 1$ rewards and error terms, respectively. Then, we update $\widehat{\boldsymbol{\beta}}_k$ as:

$$\widehat{\boldsymbol{\beta}}_k = \arg\min_{\beta \in \mathcal{B}} \frac{1}{n_{kt}} \left\| \boldsymbol{y}_{kt} - \boldsymbol{X}_{kt}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 , \tag{2}$$

where $\lambda > 0$ is a penalty parameter.

Finally, without changing the exploration-exploitation probabilistic nature in the CLG algorithm, we augment it with an alternative option. Recall the usual definition of order statistics, which for the case of estimated rewards, take the form:

$$\widehat{y}_{(1:n_{kt})t} \equiv \min_{j \in \{1,\ldots,n_{kt}\}} \widehat{y}_{jt} \leq \widehat{y}_{(2:n_{kt})t} \leq \cdots \leq \widehat{y}_{(n_{kt}:n_{kt})t} \equiv \max_{j \in \{1,\ldots,n_{kt}\}} \widehat{y}_{jt}$$

Let $\mathcal{H}_{kt}^{(\kappa_t)} \equiv \{\widehat{y}_{kt} | \widehat{y}_{kt} \geq \widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t}\}$ be the set of the κ_t higher-order statistics considered as new options for exploration, such that $\forall t > vw$, $\kappa_t = \#\mathcal{H}_{kt}^{(\kappa_t)}$. $\kappa_t \in (1, v/2]$ to avoid extremes. In fact, if for some t, $\kappa_t = 1$, the overall effect would be to increase the CLG weight to exploit the policy with the best estimated reward, and we would again be under the scope of the CLG algorithm. On the other hand, since the higher-order statistics set is important only after the initialization phase, requiring that $\kappa_t \leq v/2$ implies that $\kappa_t \leq n_{kt}/2$ as, for $t > vw$, $n_{kt} \geq v$. In this sense, the upper bound on κ_t serves the purpose of adherence to the term "highest"-order statistics for the algorithm.

Definition 5 presents the CLG algorithm coupled with a κ-higher-order statistics search set (CLG-κHOS algorithm). The only difference between the CLG and the CLG-κHOS algorithms is the degree of randomness in the exploration phase. Note that we do not impose any choice of s_t and κ_t.

Definition 5 (CLG-κHOS Algorithm)**.** *Let $c > 0$, $0 < d < 1$ and ϵ_t be defined as in Definition 3. Let $w \in \mathbb{N}+$, $w > 1$, $v \in \mathbb{N}^+$, $\kappa_t \in \mathbb{N}^+$, $1 < \kappa_t \leq v/2$ and $s_t \in (0, 1)$. Then, the*

CLG-κHOS algorithm is:

Algorithm 3: CLG-κHOS algorithm

```
input parameters: c, d, w, v, κ_t, s_t
Inicialization;
for i ∈ {1, 2, ..., v} do
    for j ∈ {1, 2, ..., w} do
        I(t) ⟵ ω_j;
        Update β̂_j;
    end
end
Exploration-Exploitation;
for t ∈ T^(l^c) do
    ε_t ⟵ min{1, cw/(d²t)}; q_t ⟵ U(0,1); r_t ⟵ U(0,1);
    if q_t ≤ ε_t then
        if r_t ≤ s_t then
            Build H_kt^(κ_t);
            u_t ⟵ U(0, κ_t); I(t) ⟵ ω_{u_t} in H_kt^(κ_t);
            Update β̂_{u_t};
        else
            a_t ⟵ U(0, w); I(t) ⟵ ω_{a_t};
            Update β̂_{a_t};
        end
    else
        b_t ⟵ argmax_{j∈{0,...,w−1}} ŷ_jt; I(t) ⟵ ω_{b_t};
        Update β̂_{b_t};
    end
end
```

4 Finite Sample Properties of Regret Functions

Theorem 1 is the main result of the paper, as it provides the bounds on the cumulative regret functions for CLG and CLG-κHOS. In the case of this last algorithm, we use the fact that, by the results of Theorem 2, optimal choices for $s_t \equiv s$ may not be time dependent.

Theorem 1 (Cumulative Regret of both CLG and CLG-κHOS algorithms)**.** *Provided that the conditions required by Lemmas 3, 4, 5 and Theorem 2 are satisfied, the cumulative regret until time T of both the CLG (R_t^{clg}) and CLG-κHOS (R_t^{hos}) algorithms is bounded as:*

$$\begin{aligned} R_{T-1}^{clg} &\leq 2w\theta_x h\tau \left\{ v + v\log\left(\frac{T-1}{vw}\right) + \frac{16C_m\theta_x s_0\lambda\left[(T-1)^{3/2} - (vw+1)^{3/2}\right]}{3\phi_0^2\sqrt{vw}} \right\} \\ &= \mathcal{O}\left\{\max\left[\log(T), s_0\lambda T^{3/2}\right]\right\} \\ R_{T-1}^{hos} &\leq R_{T-1}^{clg} + 2w\theta_x h\tau \left[vs\log\left(\frac{T-1}{vw}\right)\left(w\exp\left\{-\frac{2}{v}\left[v\left(1-P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2\right\} - 1\right)\right] \\ &= \mathcal{O}\left\{\max\left[\log(T), s_0\lambda T^{3/2}\right]\right\}, \end{aligned}$$

where $P_{\boldsymbol{\beta}}$, $\mathcal{X}$, s and C_m are provided in Lemmas 2, 5, Theorem 2 and Assumption 5, respectively.

Note that both cumulative regrets provided in Theorem 1 respect the same growth order. This fact is a recognition that the second term of R_{T-1}^{hos} does not grow at a faster rate than the first one.

Remark 5. *The suggestions provided in the literature for the growth rate of λ and s_0 could provide more intuitive bounds. For example, Carvalho et al. (2018) comment that in the Gaussian case, it is common to assume $\lambda = O\left(\sqrt{\frac{\log p}{T}}\right)$ and $\frac{s_0 \log p}{\sqrt{T}} = o(1)$. Consequently, in such cases, $s_0\lambda\sqrt{\log p}$ is $o(1)$, implying that, with high probability, $s_0\lambda$ overrides the growth in p, which grows with T. In summary, assumptions like these proactively counteract the $T^{3/2}$ growth rate.*

Theorem 2 (Flexibility and Dominance of CLG-κHOS algorithm)**.** *Provided that the conditions required by Lemmas 4 and 5 are satisfied, the least upper bound for the specification of* *the algorithm described in Definition 5 does not depend on* κ_t*, and for any sufficiently small* $\delta > 0$*, it is optimal to set* $s_t \equiv s = (1-\delta)\mathbb{1}_{\{vw<8\mathcal{X}^2\}} + \delta\mathbb{1}_{\{vw\geq 8\mathcal{X}^2\}}$*. Moreover,* $\forall vw < 8\mathcal{X}^2$*:*

$$\max_{t\in\mathcal{T}\cap\{t|t>vw\}}(r_t^{hos}) < \max_{t\in\mathcal{T}\cap\{t|t>vw\}}(r_t^{clg}),$$

where r_t^{clg} *is provided in Lemma 4 and* r_t^{hos} *and* $\mathcal{X}$ *in Lemma 5.*

Theorem 2 shows that a not too large initialization phase, $vw < 8\mathcal{X}^2$, guarantees that the bound on the CLGκ-HOS is at most equal to but mostly better than that of the CLG algorithm. Also, concerning the best choice of s_t and κ_t in terms of achieving the best/least possible upper bounds, one can see that $\forall t > vw$ the bounds on the regrets do not depend on κ_t, while s_t should be set close to 0 or close to 1 and not necessarily be time-varying, if the initialization period is higher or lower than $8\mathcal{X}^2$, respectively. That is, the least upper bound is achievable when there are exploration at random competes and into a higher-order statistics searching set of any cardinality less than or equal to $v/2$ (by our imposition in Section 2). Moreover, competition between searching mechanisms apparently does not converge to optimum bounds since as a function of the length of the initialization period, one mechanism dominates the other.

Remark 6. *Theorem 2 represents our additional contribution to the high-dimensional bandit literature by providing supplemental guarantees for practitioners with mild restrictions in exploration of new actions. In these cases, limitations imposed by practical applications naturally bound exploration to be confined in a restrictive, possibly time-varying, set of actions. In these cases, it would be preferable to have some flexibility in the action screening process without impacting the algorithm's properties. Theorem 2 can be helpful in this regard since it provides the flexibility to explore groups of different sizes according to the users' needs and, additionally, it suggests that this approach would be the best course of action for a reasonable duration of the initialization phase. That is, it would not only be advisable (by operational limitations in real applications) but better to look into a set of promising actions than otherwise.*

5 Simulations and Sensitivity Analysis

There are two types of variables in the CLG-κHOS algorithm: those that the planner can only observe and those the she can control and serve as inputs to the algorithm. In this section, we provide a sensitivity analysis for the set of variables in the last group. The results corroborate our theoretical findings.

We evaluate the sensitivity of the algorithm with respect to changes in the following inputs: (1) the number of available policies, w; (2) the weight attributed for the confinement of exploration to a higher-order statistics set, s_t; and (3) the cardinality of the higher-order statistics set, κ_t. We focus on the CLG-κHOS algorithm since the CLG algorithm should have similar behavior, at least for changes in w.[3] Moreover, we also compare the CLG and the CLG-κHOS algorithms with a few related alternatives.

General Setup: We set $T = 2000$; covariates $\boldsymbol{x}_t$ are generated from a truncated Gaussian distribution such that Assumption 2.i translates to $|\boldsymbol{x}_{t,(j)}| \leq 1$, $\forall t$. The dimension of $\boldsymbol{x}_t$ is $p = 200$, and the sparsity index is $s_0 = 5$. $\epsilon_{kt} \sim \mathsf{N}(0, 0.05)$, $\forall k \in \{0, \dots, w-1\}$ and $\forall t$. We consider $v = 30$ as the number of times that each policy is implemented in the initialization phase.[4] Each policy $\boldsymbol{\omega}_k$ has its own mechanism $\boldsymbol{\beta}_k$ drawn independently from a $\mathsf{U}(0, 1)$ probability distribution. The simulation is repeated $n_{sim} = 50$ times, and the results are presented as the average regret. That is, the instantaneous regret at a specific time t is the average of 50 simulated instantaneous regrets at this time.

Sensitivity to $\boldsymbol{w}$: We set $w \in \{5, 10, 15\}$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 1/w$.

Cumulative regret is increasing in w, as the greater the number of policies tested, the more difficult it is for the algorithm to select the best policy. Another implication specific to our formulation is that the higher the value of w is, the longer the initialization phase, which means that the logarithmic growth of the exploration vs. exploitation phase bound would

[3] Recall that the CLG algorithm does not count with s_t or κ_t.

[4] We do not explicitly test the sensitivity of the algorithm to v since, given our specification, the variable affects only the initialization phase and the precision of the mechanism's estimates. However, we tested the sensitivity to w.

take longer to operate and the levels of cumulative regret would be higher. Furthermore, recall that in relatively longer initialization phases, we do not have guarantees that the bound on CLGκ-HOS is lower than that on CLG, highlighting the importance of w. These arguments are illustrated in Figure 1.

Sensitivity to s_t: Figure 2a illustrates the proof of Theorem 2. That is, given values for w and κ_t, with all else being constant, the upper bound of the CLGκ-HOS algorithm is increasing with s_t, provided that $vw \geq 8\mathcal{X}^2$. In this setting, it is optimal to reduce the weight of searching the higher-order statistics set, which is why we chose small values for s_t. Theorem 2 suggests a small $\delta > 0$ in this case. Figure 2b is just an amplification of Figure 2a for $t > vw$. Simulations are conducted for $w = 10$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s \in \{0.5/w, 1/w, 1.5/w\}$.

Sensitivity to κ_t: Figures 3a and 3b present the sensitivity of the algorithm to values of $\kappa_t \equiv \kappa \in \{0.2w, 0.3w, 0.5w\}$, where $w = 10$ and $s_t \equiv s = 0.8\kappa_t/w$. The results are also a reflection of Theorem 2 with respect to κ_t since the regrets are not κ_t-dependent. The first panel comprises all time steps, and the second is for $t > vw$.

The cumulative regret is by far the most important measure in problems like that studied in this paper. However, it is instructive to investigate other measures. Figure 4a presents the difference in the policies selected by CLGκ-HOS and the best policies at each t, exemplified by a simulation for $w = 5$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 1/w$. The main feature to observe in this figure is that, compared to the initialization period, the exploration-exploitation phase makes fewer mistakes.

Figure 4b exhibits the average (across simulations and across the time horizon) frequency of hits for the CLG-κHOS algorithm for varying parameters. For 50 simulations and 2000 time steps, the worst specification achieves 90% (because w is large) correct policies selected on average, while the best one reaches 95%.

5.1 Comparison to Related and Adapted Algorithms

Comparison of bandit algorithms using real data sets is challenging considering the bandit feedback problems already mentioned in this paper. Therefore, we employ a simulation exercise in which we compare the cumulative regret of the CLG-κHOS algorithm to the regret of the CLG and three additional algorithms: OLS-CG, OLS-CG-κHOS and Expfirst. As these algorithms are, in general, adapted for the high-dimensional case, we briefly discuss the last three. The general setup assumed in the beginning of this section is expanded to consider $w = 10$, $s_t \equiv s = 0.5$ and $\kappa_t \equiv \kappa = w/2$. The same initialization phase is implemented for all algorithms.

OLS-CG and OLG-CG-κHOS: These algorithms are the OLS-contextual greedy and the OLS-contextual greedy with κ-higher-order statistics algorithms. Both are counterparts of CLG and CLG-κHOS that use OLS as the estimation methodology to update the estimated mechanisms $\hat{\boldsymbol{\beta}}_k$ when $\boldsymbol{\omega}_k$ is selected. In a high-dimensional sparse context, we would expect lasso to outperform a poorly defined OLS estimator. Inclusion of these algorithms in the comparison set serves to compare the estimation performance and its implications on the regret function.

ExpFirst: This is a kind of exploitation-only algorithm. The initialization phase is the same as that in the CLG-κHOS, that is, estimation of $\boldsymbol{\beta}_k$ for selected policies is the same as in the high-dimensional case and lasso is employed. However, this algorithm does not explore. After initialization, it always selects the policy that presented the minimum regret in the initialization. In a different setting, provided that some new assumptions are in place, Bastani et al. (2017) have shown that exploitation-only algorithms can achieve logarithmic growth in the OLS-estimation context.

Figure 5a shows that the CLG-κHOS algorithm largely outperforms its peers, except for the *CLG*, in which case the improvement in cumulative regret is more modest. Figure 5b amplifies Figure 5a and presents the comparison only for these two algorithms. In these simulations, we use $\kappa_t \equiv \kappa = 2$, $w = 10$ and $s_t \equiv s = 1/w$.

5.2 Two Potential Practical Applications

5.2.1 Recommendation Systems

Recommendation systems are used to match consumers to firms in an environment where preferences are fully or partially unknown. Preferences and priorities must be learned from available features related to both the users and the products being purchased. In general, businesses can benefit from a recommendation system, provided that a large set of costumers' characteristics is available. In these cases, algorithms can be more cost-efficient than humans, given the complexity and the size of the problem. Sectors such as those involved in e-commerce, retail, banking and media can potentially leverage their revenue if a reasonable recommendation system is in place. A report from Mckinsey in 2013 states that the recommendation feature contributed to 35% and 23.7% growth in revenue for Amazon and BestBuy, respectively. The report also stated that 75% of video consumption and 60% of views on web services Netflix and YouTube, respectively, are due to recommendations. [5]

A simple example would comprise several vendors (e.g., restaurants) and consumers. Available data related to rating, geographical distance from vendors to consumers, preparation time, delivery time, gender, and promotions could be used by the algorithms to learn how consumers build their preferences. In other words, the problem can be viewed as making acceptable online predictions of with which vendors consumers are willing to establish a commercial relationship.

5.2.2 Job Application Screening

In a very recent work, Li et al. (2020) use, perhaps for the first time, a bandit rule to select individual job applicants for an interview, which is an important part of the studied firm hiring process. The data they use suggest that the vast majority of applicants are not even considered for interviews. The whole process is costly because interview slots are scarce

[5] For more details, refer to https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers.

and, historically, supervised learning algorithms may introduce some human bias in the recruitment process by selecting groups with proven track records rather than taking risks on nontraditional applicants. That is, there is not a proper exploitation-exploration trade-off, and good actions may never be taken by the firm. To remedy this fact, the authors introduce a variant of the upper confidence bound algorithm as a way to introduce exploration.

Ultimately, firms would like to select quality and ability from the several resumes they receive based on a history of hired professionals. Considering that this information is intrinsically unobservable, a high-dimensional context could be fruitful in the exploitation phase to alleviate any bias that might arise. However, although there is a clear applicability of our algorithms to this problem, more work should be done to consider a broader set of actions (arms) than the authors considered in their work. The two actions considered, to invite or not to invite for an interview, are not directly applicable to our setup.

6 Concluding Remarks

In this paper, we contribute to reducing the gap related to the lack of research related to contextual bandits in high-dimensional scenarios. To this end, we extend a historically popular multiarmed bandit heuristic, the ϵ_t-greedy heuristic, to consider not only high-dimensional contexts but also a competing exploration mechanism. To the best of our knowledge, no previous work has specifically addressed the ϵ_t-greedy algorithm in this manner.

For a decreasing ϵ_t-greedy multiarmed bandit, we find that adding a high-dimensional context to the original setting does not substantially jeopardize the original results, except that in our case, regret not only grows reasonably with time but also depends on the covariate dimensions, as the latter grows with the former in high-dimensional problems.

Specifically, the consideration of a higher-order statistics searching set as an alternative to random exploration also leads to reasonable upper bounds on the time horizon. As by-products, we show that the regret bounds when order statistics are considered are at most equal to but mostly better than the case when random searching is the sole exploration mechanism, provided that the initialization phase is not excessively long. Furthermore, we show that in order to achieve the least upper bound for the cumulative regret function of the CLGκ-HOS algorithm, one should not exert effort designing the cardinality of the higher-order statistics searching set. In a simulation exercise, we show that the algorithms proposed in this paper outperform simple and adapted counterparts.

A Auxiliary Lemmas

Lemmas 1 and 2 establish the properties for the Lasso estimation.

Lemma 1 (Finite-Sample Properties of $\widehat{\boldsymbol{\beta}}_k$)**.** *Define:*

$$\mathcal{G}_{kt} := \left\{ \frac{2}{n_{kt}} \max_{1 \leq j \leq p} \left| \boldsymbol{\epsilon}'_{kt} \boldsymbol{X}^{(j)}_{kt} \right| \leq a \right\}$$

If $\widehat{\boldsymbol{\beta}}_k$ is the solution of (2), Assumption 4 holds. Furthermore, provided that $\lambda \geq 2a$, on $\mathcal{G}_{kt}$, it is true that:

$$\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|_1 \leq \frac{\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}}}{\lambda} + \frac{4\lambda s_0}{\phi_0^2},$$

where $\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}} \equiv (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)' \widehat{\boldsymbol{\Sigma}}_{kt} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ and $\widehat{\boldsymbol{\Sigma}}_{kt} \equiv \frac{1}{n_{kt}} \boldsymbol{X}'_{kt} \boldsymbol{X}_{kt}$.

Proof. See the Supplementary Material. □

Lemma 2 (Finite-Sample Properties of $\widehat{\boldsymbol{\beta}}_k$ - Continuation)**.** *Given that Assumptions 2 and 4 and the conditions of Lemma 1 are satisfied, then:*

$$\mathbb{P}\left(\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|_1 > \frac{4 s_0 \lambda}{\phi_0^2} \right) \leq \frac{\log(2p)}{n_{kt}} \left\{ \frac{C_1}{n_{kt}} + C_2 + C_3 \left[\frac{\log(2p)}{n_{kt}} \right]^{-1/2} \right\} =: P_{\boldsymbol{\beta}}, \text{ where}$$

$C_1 := C_1(\sigma, \theta_x, \lambda) = \dfrac{128 \sigma^2 \theta_x^2}{\lambda^2}$, $C_2 := C_2(s_0, \theta_x, \phi_0) = \dfrac{32 s_0 \theta_x^2}{\phi_0^2}$, and $C_3 := C_3(s_0, \theta_x, \phi_0) = \sqrt{2} C_2$.

Proof. Provided that $\lambda \geq 2a$, on $\mathcal{G}_{kt}$, that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{i,j}|(\widehat{\boldsymbol{\Sigma}}_{kt})_{i,j} - (\boldsymbol{\Sigma}_{kt})_{i,j}|$, Lemma 1 indicates that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{4s_0\lambda}{\phi_0^2}$. Then,

$$\begin{aligned}
\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 > \frac{4s_0\lambda}{\phi_0^2}\right) &= \mathbb{P}\left[\left(\mathcal{G}_{kt} \cap \max_{i,j}|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}| \leq b\right)^c\right] \\
&= \mathbb{P}(\mathcal{G}_{kt}^c \cup \max_{i,j}|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}| > b) \\
&\leq \mathbb{P}(\mathcal{G}_{kt}^c) + \mathbb{P}(\max_{i,j}|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}| > b) \qquad (3) \\
&= \mathbb{P}\left(\frac{2}{n_{kt}}\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}^{(j)}| > \frac{\lambda}{2}\right) \\
&\quad + \mathbb{P}(\max_{i,j}|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}| > b)
\end{aligned}$$

where the second equality is De Morgan's law and the first inequality is an application of the union bound.

For the first term of (3), given that $\{\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}^{(j)}|\}$, $j = 1, \ldots, p$ is a sequence of positive random variables, for $r > 0$, we employ the Markov inequality to obtain:

$$\begin{aligned}
\mathbb{P}\left(\frac{2}{n_{kt}}\max_{1\leq j\leq p}\left|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}^{(j)}\right| > \frac{\lambda}{2}\right) &\leq 4^r\frac{\mathbb{E}\left(\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}^{(j)}|^r\right)}{(n_{kt}\lambda)^r} \\
&= 4^r\frac{\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n_{kt}}\epsilon_{kt,(i)}x_{kt,(i,j)}/n_{kt}\right|^r\right)}{n_{kt}^{r-1}\lambda^r}. \qquad (4)
\end{aligned}$$

Since (4) holds for any value of $r > 0$, take $r = 2$. Therefore, by Lemma 7:

$$\begin{aligned}
16\frac{\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n_{kt}}\epsilon_{kt,(i)}x_{kt,(i,j)}/n_{kt}\right|^2\right)}{n_{kt}\lambda^2} &\leq \frac{128}{n_{kt}^3\lambda^2}\sigma^2\log(2p)\sum_{i=1}^{n_{kt}}\left(\max_{1\leq j\leq p}|x_{kt,(i,j)}|\right)^2 \\
&\leq \frac{128}{(n_{kt}\lambda)^2}\sigma^2\log(2p)\theta_x^2 \qquad (5)
\end{aligned}$$

For the second term in (3), we also have that $\{\max_{i,j}|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}|\}$ is a sequence of positive random variables. Then, by the Markov inequality, for $r > 0$, provided that

$\frac{32bs_0}{\phi_0^2} \leq 1$:

$$\begin{aligned}\mathbb{P}\left(\max_{i,j}\left|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}\right| > b\right) &\leq \mathbb{P}\left(\max_{i,j}\left|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}\right| > \frac{\phi_0^2}{32s_0}\right) \\ &\leq \frac{32s_0}{\phi_0^2}\mathbb{E}\left(\max_{i,j}\left|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}\right|\right).\end{aligned} \tag{6}$$

Recall that $\widehat{\boldsymbol{\Sigma}}_{kt} := \frac{1}{n_{kt}}\boldsymbol{X'}_{kt}\boldsymbol{X}_{kt}$. Then, its elements are given by:

$$\widehat{\Sigma}_{kt,(i,j)} = \frac{1}{n_{kt}}\sum_{b=1}^{n_{kt}} X_{kt,(i,b)}^2.$$

Define the function $\gamma : \mathbb{R} \to \mathbb{R}$, such that for a bounded random variable $x \in \mathbb{R}$, $\gamma(x) = \frac{x^2 - \mathbb{E}(x^2)}{[\max(x)]^2}$. Then, equation (6) can be rewritten as:

$$\frac{32s_0}{\phi_0^2}\mathbb{E}\left(\max_{i,j}\left|\widehat{\Sigma}_{kt,(i,j)} - \Sigma_{kt,(i,j)}\right|\right) = \frac{32s_0}{\phi_0^2}\mathbb{E}\left[\max_{i,j}\left|\frac{1}{n_{kt}}\sum_{b=1}^{n_{kt}}\theta_x^2\gamma(\boldsymbol{X}_{kt,(i,b)})\right|\right]. \tag{7}$$

Now, note that for $m = 2, 3, 4, \ldots$, such that $m \leq 1 + \log(p)$:

$$\begin{aligned}&\mathbb{E}\left[\gamma\left(X_{kt,(i,b)}\right)\right] = \frac{1}{\theta_x^2}\left[\mathbb{E}\left(X_{kt,(i,b)}^2\right) - \mathbb{E}\left(X_{kt,(i,b)}^2\right)\right] = 0 \\ &\frac{1}{n_{kt}}\sum_{i=1}^{n_{kt}}\mathbb{E}\left[\left|\gamma(X_{kt,(i,b)})\right|^m\right] \leq \frac{1}{n_{kt}\theta_x^{2m}}\sum_{i=1}^{n_{kt}}\mathbb{E}\left[\left|X_{kt,(i,b)}^2 - \mathbb{E}\left(X_{kt,(i,b)}^2\right)\right|^m\right] \leq \frac{\theta_x^{2m}}{\theta_x^{2m}} = 1.\end{aligned}$$

Then, the conditions of Lemma 8 are satisfied, and we can apply it to (7) to find that:

$$\frac{32s_0}{\phi_0^2}\mathbb{E}\left(\max_{i,j}\left|\frac{1}{n_{kt}}\sum_{b=1}^{n_{kt}}\theta_x^2\gamma(X_{kt,(i,b)})\right|\right) \leq \frac{32s_0\theta_x^2}{\phi_0^2}\left[\frac{\log(2p)}{n_{kt}} + \sqrt{\frac{2\log(2p)}{n_{kt}}}\right]. \tag{8}$$

Merging (5) and (8), we have:

$$\begin{aligned}\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 > \frac{4s_0\lambda}{\phi_0^2}\right) &\leq \frac{128}{(n_{kt}\lambda)^2}\sigma^2\log(2p)\theta_x^2 + \frac{32s_0\theta_x^2}{\phi_0^2}\left[\frac{\log(2p)}{n_{kt}} + \sqrt{\frac{2\log(2p)}{n_{kt}}}\right] \\ &= \frac{\log(2p)}{n_{kt}}\left\{\frac{C_1}{n_{kt}} + C_2 + C_3\left[\frac{\log(2p)}{n_{kt}}\right]^{-1/2}\right\} =: P_{\boldsymbol{\beta}},\end{aligned}$$

where $C_1 = \frac{128\sigma^2\theta_x^2}{\lambda^2}$, $C_2 = \frac{32s_0\theta_x^2}{\phi_0^2}$ and $C_3 = \sqrt{2}C_2$. □

Regarding the regret behavior of the CLG and the CLGκ-HOS algorithms, Lemma 3 presents the cumulative regret immediately after the initialization phase ($t = vw$), which is common to both algorithms. On the other hand, Lemmas 4 and 5 exhibit the results for the instantaneous regret.

Lemma 3 (Initialization Regret)**.** *Given the duration $l = vw$ of the initialization phase, $v, w \in \mathbb{N}^+$, $w > 1$ and provided that Assumptions 1 to 3 are satisfied, the cumulative regret of the CLG and CLG-κHOS algorithms in the initialization phase (R_l) is bounded as $R_l \leq 2vw\theta_x h\tau$.*

Proof. The cumulative regret is established in Definition (1). The worst case is to select wrong policies for all $t \leq vw$. Define $j_t \in \{0, \dots, w-1\} \equiv \underset{j \in \{0,\dots,w-1\}}{\arg\max}\ y_{jt}$ to be the index that leads to the best rewards for each $t \leq vw$. Then, the regret for the initialization phase is $R_l = \sum_{t=1}^{vw} \mathbb{E}(y_{j_t t} - y_{kt})$, where in the worst case $k \neq j_t$, $\forall\, t \leq vw$. By Assumption 2

$$R_l = \sum_{t=1}^{vw} \mathbb{E}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_{j_t} - \boldsymbol{\beta}_k)\right]. \tag{9}$$

The right-hand side of equation (9) can be bounded in absolute terms as $|\boldsymbol{x}_t'(\boldsymbol{\beta}_{j_t} - \boldsymbol{\beta}_k)| \leq \max_{1 \leq j \leq p} |\boldsymbol{x}_{t,(j)}| \left\|\boldsymbol{\beta}_{j_t} - \boldsymbol{\beta}_k\right\|_1$.

Using Assumptions 1.ii, 2.i and 3, we find that $R_l \leq \sum_{t=1}^{vw} 2\theta_x h\tau \leq 2vw\theta_x h\tau$ since by the subadditivity of any metric, and for $\{\boldsymbol{\omega}_k, \boldsymbol{\omega}_j, \boldsymbol{\omega}_z\} \in \mathcal{W}$, $\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_1 \leq \|\boldsymbol{\omega}_j - \boldsymbol{\omega}_z\|_1 + \|\boldsymbol{\omega}_k - \boldsymbol{\omega}_z\|_1$. Taking $\boldsymbol{\omega}_z = \boldsymbol{\omega}_0$ then, $\|\boldsymbol{\omega}_k - \boldsymbol{\omega}_j\|_1 \leq 2\tau$. □

Lemma 4 (Instantaneous Regret of the CLG algorithm)**.** *Provided that $\lambda \geq 2a$, on $\mathcal{G}_{it}$, that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{j,k} |(\widehat{\boldsymbol{\Sigma}}_{it})_{j,k} - (\boldsymbol{\Sigma}_{it})_{j,k}|$, for $t > vw$, $v, w \in \mathbb{N}^+$, $w > 1$, and given that Assumptions 1 to 5 hold, the instantaneous regret of the CLG algorithm (r_t^{clg}) is bounded as $r_t^{clg} \leq 2w\theta_x h\tau P_{it}^{clg}$, where*

$$P_{it}^{clg} \leq \frac{v}{t} + \left(1 - \frac{vw}{t}\right) \frac{8C_m\theta_x s_0\lambda}{\phi_0^2}$$

and C_m is established in Assumption 5.

Proof. For $t > vw$, define j_t in the same way as in the proof of Lemma 3 and consider the definition of $I(t)$ in Section 3. Then, by the law of total expectation, the instantaneous regret r_t^{clg} of the CLG algorithm is:

$$r_t^{clg} = \sum_{i=0}^{w-1} \mathbb{E}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_{j_t} - \boldsymbol{\beta}_i)|I(t) = \boldsymbol{\omega}_i\right]\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right]. \tag{10}$$

In the CLG algorithm, we have that:

$$\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] = \frac{\epsilon_t}{w} + (1-\epsilon_t)\mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \geq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j\right), \quad \forall j \in \{0,\ldots,w-1\}. \tag{11}$$

From the properties of the maximum of a sequence of random variables, we have the following fact applied to the last term of (11):

$$\begin{aligned}\mathbb{P}\left(\max_{j\in\{0,\ldots,w-1\}} \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right) &= \mathbb{P}\left(\bigcap_{j=0}^{w-1} \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right)\\ &\leq \mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right) \quad \text{for some} \quad j \in \{0,\ldots,w-1\},\end{aligned}$$

since for any sequence of sets A_i, $i = 1,\ldots,n$, the event $\{\bigcap_{i=1}^n A_i\}$ is a subset of every A_i.

Note that

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right) &= \mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j - \boldsymbol{x}_t'\boldsymbol{\beta}_j + \boldsymbol{x}_t'\boldsymbol{\beta}_j - \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i + \boldsymbol{x}_t'\boldsymbol{\beta}_i - \boldsymbol{x}_t'\boldsymbol{\beta}_i \leq 0\right)\\ &= \mathbb{P}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) \leq \boldsymbol{x}_t'(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j) + \boldsymbol{x}_t'(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)\right]\end{aligned} \tag{12}$$

Bounding the term $\boldsymbol{x}_t'(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i) - \boldsymbol{x}_t'(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ in absolute value and using the triangle inequality, we find that:

$$\begin{aligned}|\boldsymbol{x}_t'(\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_j + \boldsymbol{\beta}_j)| &\leq \left(\max_{1\leq j\leq p}|\boldsymbol{x}_{t,(j)}|\right)\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i - \widehat{\boldsymbol{\beta}}_j + \boldsymbol{\beta}_j\right\|_1\\ &\leq \left(\max_{1\leq j\leq p}|\boldsymbol{x}_{t,(j)}|\right)\left(\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right)\\ &\leq \theta_x\left(\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right)\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right) \leq \mathbb{P}\left[x_t'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) \leq \theta_x\left(\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right)\right] \tag{13}$$

Provided that $\lambda \geq 2a$, on $\mathcal{G}_{it}$, that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{j,k}|(\widehat{\boldsymbol{\Sigma}}_{it})_{j,k} - (\boldsymbol{\Sigma}_{it})_{j,k}|$, Lemma 1 indicates that for an arbitrary $i \in \{0,\ldots,w-1\}$, $\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 \leq \frac{4s_0\lambda}{\phi_0^2}$. Using this fact in equation 13 and Assumption 5, we find that:

$$\mathbb{P}\left(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j \le \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i\right) \le \mathbb{P}\left[x_t'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_j) \le \frac{8\theta_x s_0 \lambda}{\phi_0^2}\right] \le \frac{8C_m\theta_x s_0\lambda}{\phi_0^2} \tag{14}$$

Inserting the result obtained in equation (14) into equation (11), we find that:

$$\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] \le \frac{\epsilon_t}{w} + (1-\epsilon_t)\frac{8C_m\theta_x s_0\lambda}{\phi_0^2} \tag{15}$$

Recall that the suggestion for ϵ_t contained in Auer et al. (2002) is for $\epsilon_t = \frac{cw}{d^2 t}$, for $c > 0$, $0 < d < 1$ and $t \ge \frac{cw}{d^2}$. Since equation (15) is valid for $t > vw$ it suffices to take c, d, such that $c/d^2 = v$. In this case:

$$\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] \le \frac{v}{t} + \left(1 - \frac{vw}{t}\right)\frac{8C_m\theta_x s_0\lambda}{\phi_0^2} =: P_{it}^{clg}$$

Finally, the instantaneous regret can be bounded as $r_t^{clg} \le 2\theta_x h\tau \sum_{i=0}^{w-1}\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] \le 2w\theta_x h\tau P_{it}^{clg}$. □

Lemma 5 (Instantaneous Regret of the CLG-κHOS algorithm)**.** *Provided that $\lambda \ge 2a$, on $\mathcal{G}_{it}$, that $\frac{32bs_0}{\phi_0^2} \le 1$, where $b \ge \max_{j,k}|(\widehat{\boldsymbol{\Sigma}}_{it})_{j,k} - (\boldsymbol{\Sigma}_{it})_{j,k}|$. Provided that $\mathcal{X} \le v(1 - P_{\boldsymbol{\beta}})$, where*

$$\mathcal{X} := \frac{4\theta_x s_0\lambda}{\phi_0^2} + 2\theta_x h\tau.$$

For $t > vw$, $v, w \in \mathbb{N}^+$, $w > 1$, and given that Assumptions 1–4 hold, the instantaneous regret of the CLG-κHOS algorithm (r_t^{hos}) is bounded as: $r_t^{hos} \le 2w\theta_x h\tau\left(P_{it}^{hos} - \frac{\epsilon_t s_t}{w} + P_{it}^{clg}\right)$, where

$$P_{it}^{hos} := \epsilon_t s_t \exp\left\{-\frac{2}{v}\left[\left(v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right)^2\right]\right\}. \tag{16}$$

$P_{\boldsymbol{\beta}}$ is the result of Lemma 2 and P_{it}^{clg} is provided in Lemma 4*.*

Proof. For $t > vw$, define j_t in the same way as in the proof of Lemma 3 and consider the definition of $I(t)$ in Section 3. Then, by the law of total expectation, the instantaneous regret r_t^{hos} of the CLG-κHOS algorithm is:

$$r_t^{hos} = \sum_{i=0}^{w-1}\mathbb{E}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_{j_t} - \boldsymbol{\beta}_i)|I(t) = \boldsymbol{\omega}_i\right]\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] \tag{17}$$

In the CLG-κHOS algorithm, we have that $\forall j \in \{0, \ldots, w-1\}$:

$$\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right] = \frac{\epsilon_t s_t}{\kappa_t}\mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \in \mathcal{H}_{it}^{(\kappa_t)}) + \frac{1}{w}\left[\epsilon_t(1-s_t)\right] + (1-\epsilon_t)\mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \geq \boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_j). \tag{18}$$

The last term of the right side of equation (18) is the same as last term of $\mathbb{P}\left[I(t) = \boldsymbol{\omega}_i\right]$ in the CLG algorithm. Regarding the first term of equation (18), by the definition of $\mathcal{H}_{kt}^{(\kappa_t)}$ (Section 3):

$$\mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \in \mathcal{H}_{it}^{(\kappa_t)}) = \mathbb{P}\left(\bigcup_{j=n_{kt}-\kappa_t}^{n_{kt}} \{\widehat{y}_{it} \geq \widehat{y}_{(j:n_{kt})t}\}\right) \tag{19}$$

Employing the union bound and noting that restricted to the set of κ_t higher-order statistics, the event $\{\widehat{y}_{it} \geq \widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t}\} \subset \{\widehat{y}_{it} \geq \widehat{y}_{(j:n_{kt})t}\}$, $j \in \{n_{kt}-\kappa_t, \ldots, n_{kt}\}$ is the most probable to occur since $\widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t}$ is the lowest possible order statistic; then:

$$\mathbb{P}\left(\bigcup_{j=n_{kt}-\kappa_t}^{n_{kt}} \{\widehat{y}_{it} \geq \widehat{y}_{(j:n_{kt})t}\}\right) \leq \kappa_t \mathbb{P}\left(\widehat{y}_{it} \geq \widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t}\right). \tag{20}$$

From Assumptions 1–3, it is clear that $|\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i| \leq \theta_x \left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 + \theta_x \|\boldsymbol{\beta}_i\|_1$. Moreover, on $\mathcal{G}_{it} \cap \max_{j,k}|\widehat{\boldsymbol{\Sigma}}_{it,(j,k)} - \boldsymbol{\Sigma}_{it,(j,k)}| \leq b$, Lemma 1 indicates that

$$|\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i| \leq \frac{4\theta_x s_0 \lambda}{\phi_0^2} + 2\theta_x h\tau =: \mathcal{X}$$

Then, equation (20) leads to:

$$\begin{aligned} \kappa_t \mathbb{P}(\widehat{y}_{it} \geq \widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t}) &\leq \kappa_t \mathbb{P}\Big(\widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t} \leq \mathcal{X}\Big) \\ &\leq \kappa_t \sum_{l=n_{kt}-\kappa_t}^{n_{kt}} \binom{n_{kt}}{l} \left[\mathbb{P}(\widehat{y}_{it} \leq \mathcal{X})\right]^l \left[\right]1 - \mathbb{P}(\widehat{y}_{it} \leq \mathcal{X})\right]^{n_{kt}-l}, \end{aligned} \tag{21}$$

since, as an intermediate-order statistic, $\widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t} \sim Bin[n_{kt}, p_{it}(y)]$, for $p_{it}(y) \equiv \mathbb{P}(\widehat{y}_{it} \leq y)$, which in this case, we can take $y = \mathcal{X}$.

For $\mathcal{X} \leq n_{kt}p_{it}(\mathcal{X})$, we can use Lemma 9 to bound equation (21) as:

$$\kappa_t \mathbb{P}\left(\widehat{y}_{(n_{kt}-\kappa_t:n_{kt})t} \leq \mathcal{X}\right) \leq \kappa_t \exp\left[-2\frac{(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})^2}{n_{kt}}\right].$$

Note that

$$p_{it}(\mathcal{X}) := \mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \leq \mathcal{X}) \geq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 \leq \frac{4s_0\lambda}{\phi_0^2} + 2h\tau\right)$$
$$\geq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\right\|_1 \leq \frac{4s_0\lambda}{\phi_0^2}\right) \geq 1 - P_{\boldsymbol{\beta}},$$

the solution of which has already been established in Lemma 2. Then,

$$\kappa_t \exp\left[-2\frac{(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})^2}{n_{kt}}\right] \leq \kappa_t \exp\left[-2\frac{(n_{kt}(1 - P_{\boldsymbol{\beta}}) - \mathcal{X})^2}{n_{kt}}\right] \tag{22}$$

Also, the first derivative of the right-hand side of equation (22) is negative with respect to n_{kt}. To see this, note that:

$$\begin{aligned} f'(n_{kt}) &= \kappa_t \exp\left[-2\frac{(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})^2}{n_{kt}}\right] \\ &\quad \times \left\{-2\left[\frac{2(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})p_{it}(\mathcal{X})n_{kt}}{n_{kt}^2} - \frac{(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})^2}{n_{kt}^2}\right]\right\} \\ &= \kappa_t \exp\left[-2\frac{(n_{kt}p_{it}(\mathcal{X}) - \mathcal{X})^2}{n_{kt}}\right]\left[-2\left(\frac{n_{kt}^2p_{it}^2(\mathcal{X}) - \mathcal{X}^2}{n_{kt}^2}\right)\right] \leq 0, \end{aligned} \tag{23}$$

where the last inequality uses the condition that $\mathcal{X} \leq n_{kt}p_{it}(\mathcal{X})$.

Therefore, for $t > vw$, $\mathcal{X} \leq v(1 - P_{\boldsymbol{\beta}})$ is sufficient to replace the above requisite of Lemma 9. Moreover, as the right-hand side of equation (22) is nonincreasing in n_{kt} and $p_{it}(\mathcal{X})$, we can restate Lemma 9 as:

$$\mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \in \mathcal{H}_{it}^{(\kappa_t)}) \leq \kappa_t \exp\left\{-\frac{2}{v}\left[v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2\right\}$$

Define $P_{it}^{hos} := \frac{\epsilon_t s_t}{k_t}\mathbb{P}(\boldsymbol{x}_t'\widehat{\boldsymbol{\beta}}_i \in \mathcal{H}_{it}^{(\kappa_t)})$ and the instantaneous regret of the CLG-κHOS algorithm, equation (17), can be bounded as: $r_t^{hos} \leq 2w\theta_x h\tau\left(P_{it}^{hos} - \frac{\epsilon_t s_t}{w} + P_{it}^{clg}\right)$. □

Note from Lemmas 3, 4 and 5 that all bounds are increasing with θ_x, τ and w. The intuition behind this fact is clear since the larger the level of dissimilarity among policies or the larger the number of policies to be tested is, the greater the difficulty for the algorithm to select the right policy.

B Proof of Theorem 1

Provided that the conditions required by Lemmas 3, 4, and 5 and Theorem 2 are satisfied, the cumulative regret until time T of both the CLG (R_t^{clg}) and CLG-κHOS (R_t^{hos}) algorithms is bounded as:

$$
\begin{aligned}
R_{T-1}^{clg} &\leq 2w\theta_x h\tau \left\{ v + v\log\left(\frac{T-1}{vw}\right) + \frac{16C_m\theta_x s_0\lambda\left[(T-1)^{3/2} - (vw+1)^{3/2}\right]}{3\phi_0^2\sqrt{vw}} \right\} \\
&= \mathcal{O}\left\{\max\left[\log(T), s_0\lambda T^{3/2}\right]\right\} \\
R_{T-1}^{hos} &\leq R_{T-1}^{clg} + 2w\theta_x h\tau \left[vs\log\left(\frac{T-1}{vw}\right)\left(w\exp\left\{-\frac{2}{v}\left[v\left(1-P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2\right\} - 1\right)\right] \\
&= \mathcal{O}\left\{\max\left[\log(T), s_0\lambda T^{3/2}\right]\right\},
\end{aligned}
$$

where $P_{\boldsymbol{\beta}}$, $\mathcal{X}$, s and C_m are provided in Lemmas 2 and 5, Theorem 2 and Assumption 5.

Proof. For $t \leq vw$, the cumulative regrets of both algorithms are given by Lemma 3. For $t > vw$, since $1/t$ and $1-1/t$ are obviously decreasing and increasing functions of t, respectively, we use Lemmas 10 and 11 in the Supplementary Material to obtain that

$$
\begin{aligned}
R_{T-1,t>vw}^{clg} \leq & 2w\theta_x h\tau \sum_{t=vw+1}^{T-1} P_{it}^{clg} \\
&\leq 2w\theta_x h\tau \sum_{t=vw+1}^{T-1} \frac{v}{t} + \left(1 - \frac{vw}{t}\right)\frac{8C_m\theta_x s_0\lambda}{\phi_0^2} \\
&\leq 2w\theta_x h\tau \left\{ v\log\left(\frac{T-1}{vw}\right) + \frac{16C_m\theta_x s_0\lambda[(T-1)^{3/2} - (vw+1)^{3/2}]}{3\phi_0^2\sqrt{vw}} \right\} \\
&= \mathcal{O}\{\max[\log(T), s_0\lambda T^{3/2}]\}.
\end{aligned}
$$

Therefore, the total cumulative regret for the CLG until time T is:

$$
R_{T-1}^{clg} \leq 2w\theta_x h\tau \left\{ v + v\log\left(\frac{T-1}{vw}\right) + \frac{16C_m\theta_x s_0\lambda[(T-1)^{3/2} - (vw+1)^{3/2}]}{3\phi_0^2\sqrt{vw}} \right\}
$$

For the CLG-κHOS algorithm, from 5:

$$R^{hos}_{T-1,t>vw} \leq 2w\theta_x h\tau \sum_{t=vw+1}^{T-1} \left(P^{hos}_{it} - \frac{\epsilon_t s_t}{w} + P^{clg}_{it}\right) \leq$$

$$R^{clg}_{T-1,t>vw} + 2w\theta_x h\tau \sum_{t=vw+1}^{T-1} \frac{vws_t}{t} \left(\exp\left\{-\frac{2}{v}\left[v\left(1-P_{\boldsymbol{\beta}}\right)-\mathcal{X}\right]^2\right\}\right) - \frac{vs_t}{t}$$

From the results in Theorem 2, we recognize that the optimal s_t is not time-dependent, and we use s from now on, referring to this optimal choice. Then, we find that:

$$\begin{aligned} R^{hos}_{T-1,t>vw} \leq & R^{clg}_{T-1,t>vw} + 2w\theta_x h\tau \sum_{t=vw+1}^{T-1} \frac{vws}{t} \exp\left\{-\frac{2}{v}\Big[v\left(1-P_{\boldsymbol{\beta}}\right)-\mathcal{X}\Big]^2\right\} - \frac{vs}{t} \leq \\ & R^{clg}_{T-1,t>vw} + 2w\theta_x h\tau \Big[vws \log\left(\frac{T-1}{vw}\right)\left(\exp\left\{-\frac{2}{v}\left[v\left(1-P_{\boldsymbol{\beta}}\right)-\mathcal{X}\right]^2\right\}\right) - \\ & vs \log\left(\frac{T-1}{vw}\right)\Big] \end{aligned}$$

Finally, the total cumulative regret for the CLG until time T is:

$$\begin{aligned} R^{hos}_{T-1} \leq & R^{clg}_{T-1} + 2w\theta_x h\tau \left[vs \log\left(\frac{T-1}{vw}\right)\left(w \exp\left\{-\frac{2}{v}\Big[\left[v\left(1-P_{\boldsymbol{\beta}}\right)-\mathcal{X}\right]^2\right\} - 1\right)\right] \\ & = \mathcal{O}\{\max[\log(T), s_0\lambda T^{3/2}]\}. \end{aligned}$$

Provided that the conditions of Theorem 2 hold:

$$w \exp\left\{-\frac{2}{v}\left[v\left(1-P_{\boldsymbol{\beta}}\right)-\mathcal{X}\right]^2\right\} < 0$$

and R^{hos}_{T-1} as a whole does not grow at a higher rate than R^{clg}_{T-1} itself. □

C Proof of Theorem 2

Provided that the conditions required by Lemmas 4 and 5 are satisfied, the least upper bound for the specification of the algorithm described in Definition 5 does not depend on κ_t and, for any sufficiently small $\delta > 0$, it is optimal to set $s_t \equiv s = (1-\delta)\mathbb{1}_{\{vw<8\mathcal{X}^2\}} + \delta\mathbb{1}_{\{vw\geq 8\mathcal{X}^2\}}$. Moreover, $\forall vw < 8\mathcal{X}^2$:

$$\max_{t\in\mathcal{T}\cap\{t|t>vw\}}(r^{hos}_t) < \max_{t\in\mathcal{T}\cap\{t|t>vw\}}(r^{clg}_t),$$

where r_t^{clg} is provided in Lemma 4 and r_t^{hos} and $\mathcal{X}$ in Lemma 5.

Proof. For the second part of the theorem, note from the results of Lemmas 4 and 5 that:

$$r_t^{hos} - r_t^{clg} \leq 2w\theta_x h\tau \left(P_{it}^{hos} - \frac{\epsilon_t s_t}{w}\right) = 2w\theta_x h\tau\epsilon_t s_t \left[\frac{1}{\kappa_t}\mathbb{P}(\boldsymbol{x}_t' \in \mathcal{H}_{it}^{(\kappa_t)}) - \frac{1}{w}\right],$$

since $2w\theta_x h\tau\epsilon_t s_t > 0$. It is sufficient for our result that:

$$w < \exp\left\{\frac{2}{v}\left[v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2\right\},$$

which is guaranteed when:

$$vw < 2\left[v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2, \tag{24}$$

since for any $a \geq 0$, $a < e^a$.

Note that the discriminant of (24) is given by:

$$\Delta = w^2 - 8w\mathcal{X}(1 - P_{\boldsymbol{\beta}}),$$

which is negative when $w < 8\mathcal{X}(1 - P_{\boldsymbol{\beta}})$, meaning that $2\left[v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2 - vw$ has no real roots and is always positive.

For the proof of this part to be completed, we recall that the conditions required by Lemma 5 must be satisfied. In this case, $1 - P_{\boldsymbol{\beta}} \geq \frac{\mathcal{X}}{v}$

For the first part, we know that

$$\begin{aligned} r_t^{hos} \leq & 2w\theta_x h\tau \left(P_{it}^{hos} - \frac{\epsilon_t s_t}{w} + P_{it}^{clg}\right) \\ &= 2w\theta_x h\tau\epsilon_t s_t \left[\frac{1}{\kappa_t}\mathbb{P}(\boldsymbol{x}_t' \in \mathcal{H}_{it}^{(\kappa_t)}) - \frac{1}{w}\right] + 2w\theta_x h\tau P_{it}^{clg} \\ &\leq 2w\theta_x h\tau\epsilon_t s_t \left(\exp\left\{-\frac{2}{v}\left[v\left(1 - P_{\boldsymbol{\beta}}\right) - \mathcal{X}\right]^2\right\} - \frac{1}{w}\right) + 2w\theta_x h\tau P_{it}^{clg} \end{aligned} \tag{25}$$

Trivially, one can observe that none of the terms in inequality (25) depend on κ_t. Regarding s_t, it is optimal to set the highest s_t possible when $vw < 8\mathcal{X}^2$, provided that $s_t \in (0, 1)$, since in this case, according to the results of the second part of this theorem, we have that $\frac{1}{\kappa_t}\mathbb{P}\left(\boldsymbol{x}_t' \in \mathcal{H}_{it}^{(\kappa_t)}\right) - \frac{1}{w} < 0$, which completes the proof. □

References

Y. Abbasi-Yadkori and Dávid C. Szepesvári D. Pál. Online-to-confidence-set conversions and application to sparse stochastic bandits. *AISTATS*, 22:1–9, 2012.

Y. Abbasi-Yadkori, D. Pál, and C.Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2312–2320. 2011.

A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 687—696, 2017.

S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In S. Mannor, N. Srebro, and R.C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 2012.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2003.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

H. Bastani and M. Bayati. Online decision-making with high-dimensional covariates. *Operations Research*, 2019.

H. Bastani, M. Bayati, and K. Khosravi. Mostly exploration-free algorithms for contextual bandits. 2017.

D. Bouneffouf, I. Rish, G.A. Cecchi, and R. Féraud. Context attentive bandits: Contextual bandit with restricted context. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 1468–1475, 2017.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

A. Carpentier and R. Munos. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 190–198, 2012.

C.V. Carvalho, R.P. Masini, and M.C. Medeiros. ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics*, 207:352–380, 2018.

N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Regret minimization for reserve prices in second-price auctions. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1190–1204, 2013.

H. Chen, W. Lu, and R. Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 0(0):1–16, 2020.

V. Dani, S.M. Kakade, and T.P. Hayes. The price of bandit information for online optimization. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 345–352. Curran Associates, Inc., 2008.

Arnoud V. den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. 2013.

Yash Deshpande and Andrea Montanari. Linear bandits in high dimension and recommendation systems. In *Allerton Conference*, pages 1750–1754. IEEE, 2012.

A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stochastic Systems*, 3: 230–261, 2013.

Kirthevasan Kandasamy, Joseph E. Gonzalez, Michael I. Jordan, and Ion Stoica. Mechanism design with bandit feedback, 2020.

Anders Kock and Martin Thyrsgaard. Optimal dynamic treatment allocation. 05 2017.

Anders Bredahl Kock, David Preinerstorfer, and Bezirgen Veliyev. Functional sequential treatment allocation, 2018.

Anders Bredahl Kock, David Preinerstorfer, and Bezirgen Veliyev. Treatment recommendation with distributional targets, 2020.

Sanath Kumar Krishnamurthy and Susan Athey. Survey bandits with regret guarantees, 2020.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008.

D. Li, L. Raymond, and P. Bergman. Hiring as exploration. 2020. URL https://ssrn.com/abstract=3630630.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery.

Z. Lin and Z. Bai. *Probability Inequalities.* Springer-Verlag, 2011.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *J. Mach. Learn. Res.*, 17(1):2442–2471, 2016.

Denis Sauré and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

Long Tran-Thanh, Archie Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas Jennings. Epsilon–first policies for budget–limited multi-armed bandits. 01 2010.

Alexander Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32, 02 2004. doi: 10.1214/aos/1079120131.

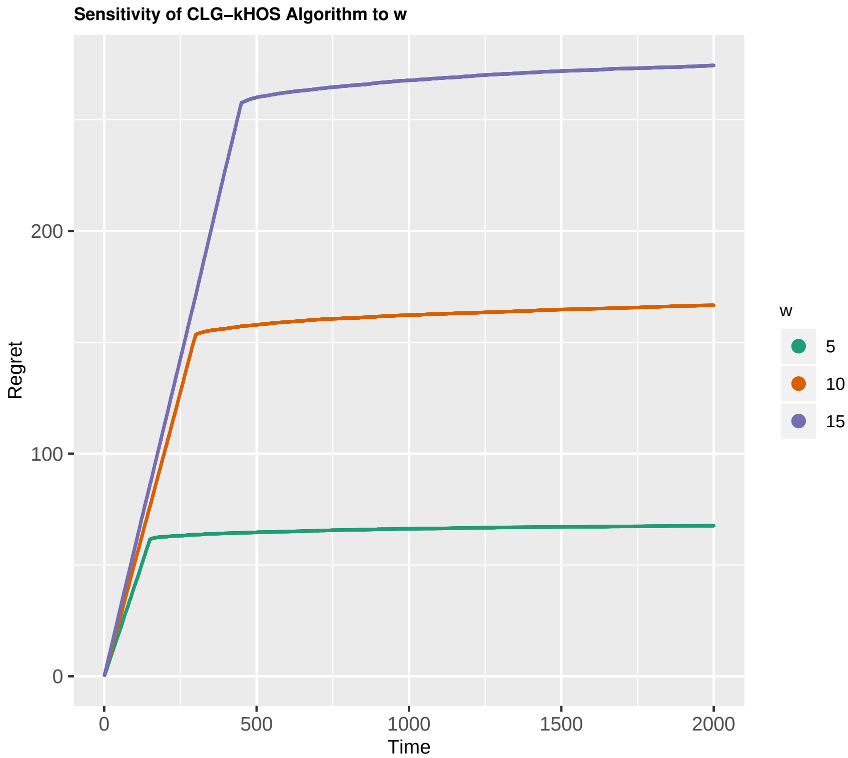
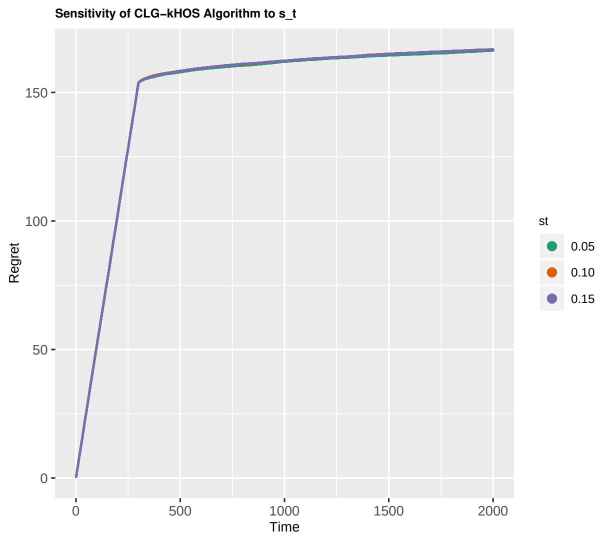
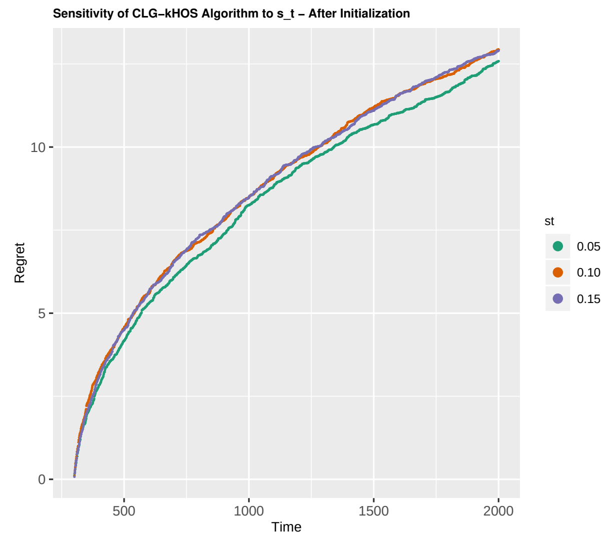


Figure 1: Comparison of Cumulative Regrets of the CLG-κHOS algorithm for values of $w \in \{5, 10, 15\}$, $s_t \equiv s = 0.5$ and $\kappa_t \equiv \kappa = w/2$.

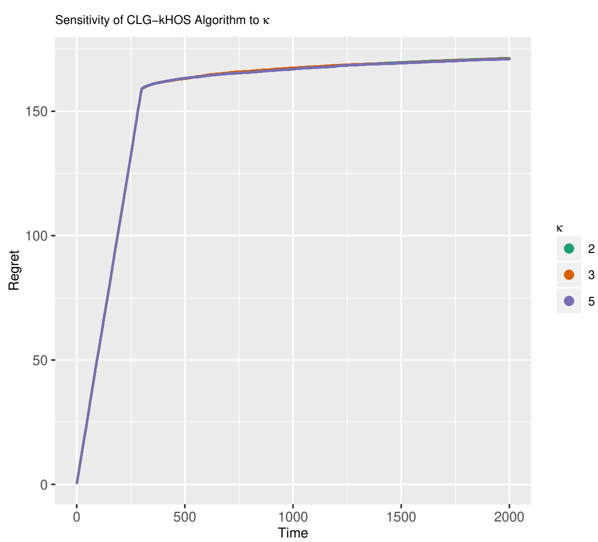


(a) Comparison of Cumulative Regrets of the CLG-κHOS algorithm for values of $s_t \equiv s \in \{0.25, 0.5, 0.75, 1\}$, $\kappa_t \equiv \kappa = w/2$ and $w = 10$.

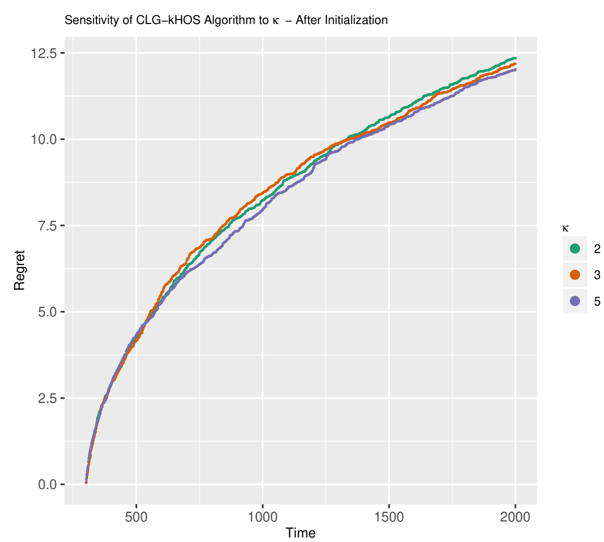


(b) Comparison of Cumulative Regrets of the CLG-κHOS algorithm, computed from $t = vw + 1$ to $t = T$, for values of $s_t \equiv s \in \{0.25, 0.5, 0.75, 1\}$, $\kappa_t \equiv \kappa = w/2$ and $w = 10$.

Figure 2: Sensitivity of CLG-κHOS algorithm to s_t.

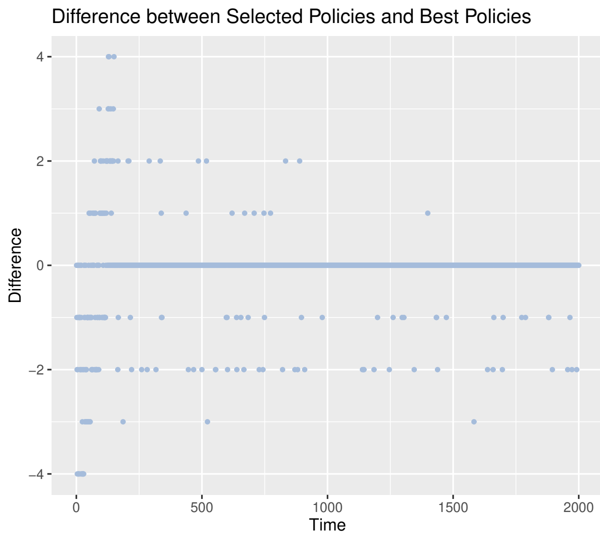


(a) Comparison of Cumulative Regrets of the CLG-κHOS algorithm for values of $\kappa_t \equiv \kappa \in \{0.2w, 0.3w, 0.5w\}$, where $w = 10$ and $s_t \equiv s = 0.8\kappa_t/w$.

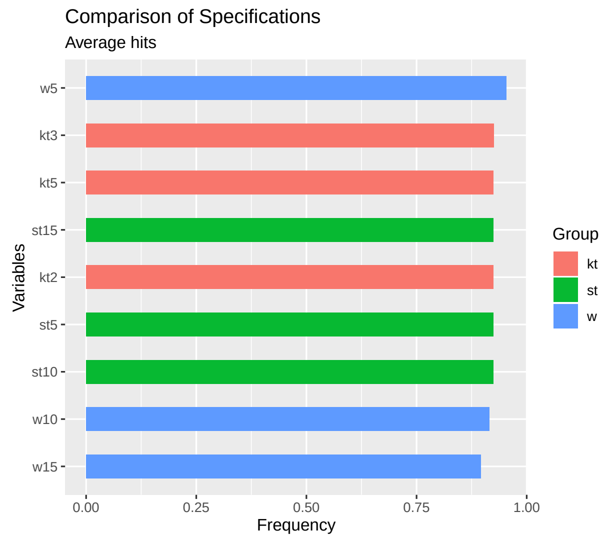


(b) Comparison of Cumulative Regrets of the CLG-κHOS algorithm, computed from $t = vw + 1$ to $t = T$, for values of $\kappa_t \equiv \kappa \in \{0.2w, 0.3w, 0.5w\}$, where $w = 10$ and $s_t \equiv s = 0.8\kappa_t/w$.

Figure 3: Sensitivity of CLG-κHOS algorithm to κ_t.

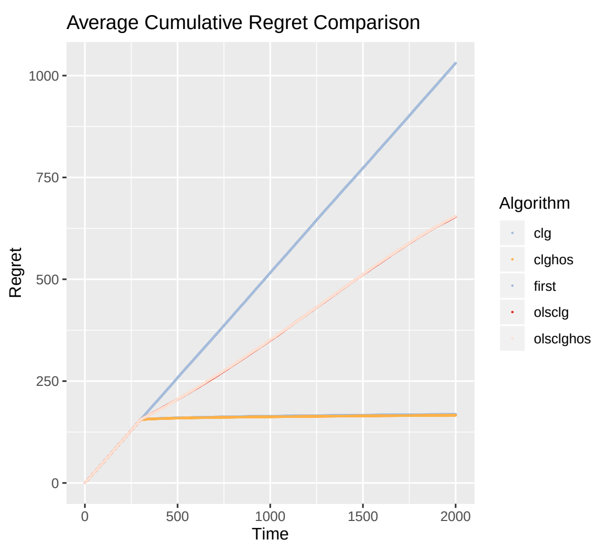


(a) Differences between the selected policy and the best policy for the CLG-κHOS algorithm for values of $w = 5$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 1/w$.

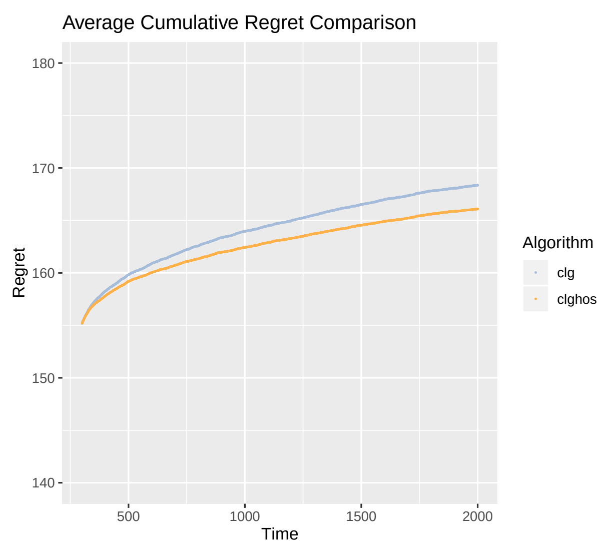


(b) Comparison of frequency of hits for the CLG-κHOS algorithm, computed from $t = vw + 1$ to $t = T$, for different specifications of s_t, κ_t and w.

Figure 4: Differences between selected and best policies and average hits for different specifications.



(a) Comparison of Cumulative Regrets of the CLG-κHOS algorithm with CLG, OLS-CG, OLS-CG-κHOS and ExpFirst for values of $\kappa_t \equiv \kappa = 2$, $w = 10$ and $s_t \equiv s = 1/w$.



(b) Comparison of Cumulative Regrets between the CLG-κHOS algorithm and CLG for values $\kappa_t \equiv \kappa = 2$, $w = 10$ and $s_t \equiv s = 1/w$.

Figure 5: Comparison of Algorithms.

Supplementary Material

Online Action Learning in High Dimensions: A New Exploration Rule for Contextual ϵ_t-Greedy Heuristics

Cláudio Flores and Marcelo C. Medeiros

S.1 Auxiliary Results

S.1.1 Proof of Lemma 1

Proof. This proof has been already provided in other papers, such as Carvalho et al. (2018). For the sake of completeness, we provide the main steps of the proof, even though it is a well-known result.

In equation (2), if $\widehat{\boldsymbol{\beta}}_k$ is the minimum of the optimization problem, then it is true that

$$\frac{1}{n_{kt}}\left\|\boldsymbol{y}_{kt}-\boldsymbol{X}_{kt}\widehat{\boldsymbol{\beta}}_k\right\|_2^2+\lambda\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1\leq\frac{1}{n_{kt}}\left\|\boldsymbol{y}_{kt}-\boldsymbol{X}_{kt}\boldsymbol{\beta}_k\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_k\right\|_1.$$

Using Assumption 2, we can replace $\boldsymbol{y}_{kt}$ in the above expression to obtain the basic inequality (see Bühlmann and van de Geer (2011) page 103):

$$\begin{aligned}&\frac{1}{n_{kt}}\left\|\boldsymbol{X}_{kt}(\boldsymbol{\beta}_k-\widehat{\boldsymbol{\beta}}_k)+\boldsymbol{\epsilon}_{kt}\right\|_2^2+\lambda\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1\leq\frac{1}{n_{kt}}\left\|\boldsymbol{\epsilon}_{kt}\right\|_2^2+\lambda\left\|\boldsymbol{\beta}_k\right\|_1\iff\\&\frac{1}{n_{kt}}\left\|\boldsymbol{X}_{kt}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)\right\|_2^2+\lambda\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1\leq\frac{2}{n_{kt}}\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)+\lambda\left\|\boldsymbol{\beta}_k\right\|_1\end{aligned}\tag{S.1}$$

Define $\left\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{kt}}^2\equiv(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)'\widehat{\boldsymbol{\Sigma}}_{kt}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)$, and the same for $\left\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\right\|_{\boldsymbol{\Sigma}_{kt}}^2$ replacing $\widehat{\boldsymbol{\Sigma}}_{kt}$ for $\boldsymbol{\Sigma}_{kt}$, where $\boldsymbol{\Sigma}_{kt}:=\mathbb{E}[\boldsymbol{X}_{kt}'\boldsymbol{X}_{kt}]$ and $\widehat{\boldsymbol{\Sigma}}_{kt}:=\frac{1}{n_{kt}}\boldsymbol{X}_{kt}'\boldsymbol{X}_{kt}$.

The first term on the right side of (S.1) can be bounded in absolute terms as:

$$\frac{2}{n_{kt}}|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)|\leq\left(\frac{2}{n_{kt}}\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{kt}'\boldsymbol{X}_{kt}^{(j)}|\right)\left\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\right\|_1.$$

On $\mathcal{G}_{kt}$, we have that

$$\left\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{kt}}^2+\lambda\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1\leq a\left\|\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k\right\|_1+\lambda\left\|\boldsymbol{\beta}_k\right\|_1\tag{S.2}$$

Using our previous definitions (see Section 2) for $\boldsymbol{\beta}_k[S_0]$ and $\boldsymbol{\beta}_k[S_0^c]$ and the respective counterparts for the estimators, by the triangle inequality of the left-hand side of equation (S.2), we have that:

$$\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1 \geq \|\boldsymbol{\beta}_k[S_0]\|_1 - \left\|(\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0])\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1$$

Using this result in (S.2) and the fact that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1$:

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}} + \lambda\Big(\|\boldsymbol{\beta}_k[S_0]\|_1 - \left\|(\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0])\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1 \Big) \leq \\
&a\Big(\left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1 \Big) + \lambda \|\boldsymbol{\beta}_k\|_1 \iff \\
&\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}} + (\lambda - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq 2\lambda \left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1.
\end{aligned}$$

By Assumption 4, we have that:

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}} + (\lambda - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{2\lambda\sqrt{s_0}}{\phi_0}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\boldsymbol{\Sigma}_{kt}} \tag{S.3}$$

Recall that Assumption 4 also requires that $\max_{i,j}|(\widehat{\boldsymbol{\Sigma}}_{kt})_{i,j} - (\boldsymbol{\Sigma}_{kt})_{i,j}| \leq b$. Then, using Lemma 6, provided that $\frac{32bs_0}{\phi_0^2} \leq 1$, we have that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\boldsymbol{\Sigma}_{kt}} \leq \sqrt{2}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{kt}}$. Substituting in (S.3):

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}} + (\lambda - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{2\sqrt{2}\lambda\sqrt{s_0}}{\phi_0}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{kt}}$$

Since $\lambda \geq 2a$, $a > 0$, multiplying the last expression by 2 and using this fact and that $4vu \leq u^2 + 4v^2$, we have:

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{kt}}}{\lambda} + \frac{4\lambda s_0}{\phi_0^2} \tag{S.4}$$

□

Lemma 6. *Suppose that the $\boldsymbol{\Sigma}_0$-compatibility condition holds for the set S with cardinality s with compatibility constant $\phi_{\boldsymbol{\Sigma}_0}(S)$ and that $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_\infty \leq \tilde{\lambda}$, where*

$$\frac{32\tilde{\lambda}s}{\phi^2_{\boldsymbol{\Sigma}_0}(S)} \leq 1.$$

Then, for the set S, the $\boldsymbol{\Sigma}_1$-compatibility condition holds as well, with $\phi^2_{\boldsymbol{\Sigma}_1}(S) \geq \phi^2_{\boldsymbol{\Sigma}_0}(S)/2$.

Proof. See Corollary 6.8 in Bühlmann and van de Geer (2011) □

Lemma 7. *For arbitrary n and p, consider independent centered random variables $\epsilon_1, \ldots, \epsilon_n$,* *such that* $\forall\, i$*, there is a* σ^2 *that bounds the variance as* $\mathbb{E}\left(\epsilon_i^2\right) \leq \sigma^2$*. Moreover, let* $\{x_{i,j} : i = 1, \ldots, n,\ j = 1, \ldots, p\}$ *be such that for* $i = 1, \ldots, n$*, there is a* $K_i := \underset{1 \leq j \leq p}{max} |x_{i,j}|$ *such that*

$$\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n}\frac{\epsilon_i x_{i,j}}{n}\right|^2\right) \leq \sigma^2\left[\frac{8\log(2p)}{n}\right]\left(\frac{\sum_{i=1}^{n} K_i^2}{n}\right)$$

Proof. See Lemma 14.24 in Bühlmann and van de Geer (2011) □

Lemma 8. *Let* $Z_1, \ldots, Z_n$ *be independent random variables and* $\gamma_1, \ldots, \gamma_p$ *be real-valued functions satisfied for* $j = 1, \ldots, p$*,*

$$\mathbb{E}[\gamma_j(Z_i)] = 0$$
$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|\gamma_j(Z_i)|^m] \leq \frac{m!}{2}K^{m-2}$$

for $K > 0$ *and* $m \leq 1 + log(p)$ *(easily satisfied for large* p*). Then,*

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^{n}\gamma_j(Z_i)\right|^m\right] \leq \left[\frac{K\log(2p)}{n} + \sqrt{\frac{2\log(2p)}{n}}\right]^m.$$

Proof. See Lemma 14.12 in Bühlmann and van de Geer (2011) □

Lemma 9. *Let* $X \sim Bin(n, p)$*. For* $k \leq np$*:*

$$\mathbb{P}(X \leq k) \leq \exp\left[-\frac{2(np-k)^2}{n}\right]$$

Proof. This is an application of Höffding's inequality to random variables that follow a binomial distribution. For more details, see Lemma 7.3 of Lin and Bai (2011) □

Lemma 10. *If* f *is a monotone decreasing function and* g *is a monotone increasing function, both integrable on the range* $[r-1, s]$*, then:*

$$\sum_{t=r}^{s} f(t) \leq \int_{r-1}^{s} f(t)dt \quad \text{and} \quad \sum_{t=r}^{s} g(t) \leq \int_{r}^{s} g(t)dt$$

Proof. This is a well-known fact for monotone functions linked to left and right Riemann sums. □

Lemma 11. *For* $a, t \in \mathbb{N}^+$ *and* $a < t$*:* $1 - \frac{a}{t} < \sqrt{\frac{t}{a}}$*.*

Proof. Since $t > a$: $a(t^2 + a^2) < t(t^2 + 2a^2)$.

Dividing both sides by $at^2 > 0$:

$$\frac{at^2 - 2a^2t + a^3 - t^3}{at^2} < 0.$$

In other terms:

$$1 - \frac{2a}{t} + \frac{a^2}{t^2} < \frac{t}{a},$$

which concludes the proof. □