Contextual Bandits with Knapsacks for a Conversion Model

Zhen Li, Gilles Stoltz

▶ **To cite this version:**

Zhen Li, Gilles Stoltz. Contextual Bandits with Knapsacks for a Conversion Model. Thirty-sixth Conference on Neural Information Processing Systems, 2022, New Orleans, United States. ⟨hal-03683289v2⟩

HAL Id: hal-03683289

https://hal.science/hal-03683289v2

Submitted on 29 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Contextual Bandits with Knapsacks for a Conversion Model

Zhen Li
BNP Paribas, 16 boulevard des Italiens, 75009 Paris, France
zhen.li@bnpparibas.com

Gilles Stoltz
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France
gilles.stoltz@universite-paris-saclay.fr
HEC Paris, 1 rue de la Libération, 78350 Jouy-en-Josas, France
stoltz@hec.fr

Abstract

We consider contextual bandits with knapsacks, with an underlying structure between rewards generated and cost vectors suffered. We do so motivated by sales with commercial discounts. At each round, given the stochastic i.i.d. context $\boldsymbol{x}_t$ and the arm picked a_t (corresponding, e.g., to a discount level), a customer conversion may be obtained, in which case a reward $r(a, \boldsymbol{x}_t)$ is gained and vector costs $\boldsymbol{c}(a_t, \boldsymbol{x}_t)$ are suffered (corresponding, e.g., to losses of earnings). Otherwise, in the absence of a conversion, the reward and costs are null. The reward and costs achieved are thus coupled through the binary variable measuring conversion or the absence thereof. This underlying structure between rewards and costs is different from the linear structures considered by Agrawal and Devanur [2016] (but we show that the techniques introduced in the present article may also be applied to the case of these linear structures). The adaptive policies exhibited solve at each round a linear program based on upper-confidence estimates of the probabilities of conversion given a and $\boldsymbol{x}$. This kind of policy is most natural and achieves a regret bound of the typical order $(\mathrm{OPT}/B)\sqrt{T}$, where B is the total budget allowed, OPT is the optimal expected reward achievable by a static policy, and T is the number of rounds.

1 Introduction and Literature Review

We consider the framework of stochastic multi-armed bandits, which has been extensively studied since the early works by Thompson [1933] and Robbins [1952]. Two recent (and complementary) surveys summarizing the latest research in the field were written by Lattimore and Szepesvári [2020] and Slivkins [2019]. On the one hand, we are particularly interested in the setting of *contextual* stochastic multi-armed bandits, preferably with some structural assumptions on the dependency between rewards and contexts: linear models (again, a rich literature, see, among many others, Chu et al. [2011] and Abbasi-Yadkori et al. [2011], whose work marked a turning point), and, for $[0, 1]$-valued rewards, logistic models (Filippi et al. [2010] and Faury et al. [2020]). On the other hand, we are also particularly interested in stochastic multi-armed bandits *with knapsacks*, i.e., with cumulative vector-cost constraints to be abided by on top of maximizing the accumulated rewards. The setting was introduced by Badanidiyuru et al. [2013, 2018] and a comprehensive summary of the results achieved since then may be found in Slivkins [2019, Chapter 10]. The intersection of these two frameworks of interest is called *contextual bandits with knapsacks* [CBwK] and is the focus of the present article.

Preprint. Under review.

Literature review on CBwK. The first approach to CBwK, by Badanidiyuru et al. [2014] and Agrawal et al. [2016], assumes a joint stochastic generation of triplets of contexts-rewards-costs, with no specific underlying structure, and makes the problem tractable by using as a benchmark a finite set of static policies. As noted by Agrawal and Devanur [2016], picking this finite set may be uneasy, which is why they introduce instead a structural assumption of linear modeling: the (unknown) expected rewards and cost vectors depend linearly on the contexts.

We consider a different modeling assumption, motivated by sales with commercial discounts (see Appendix A): general (known) reward and cost functions are considered but they are coupled via a 0/1–valued factor, called a (customer) conversion, obtained as the realization of a Bernoulli variable with parameter $P(a, \boldsymbol{x})$ depending on the context $\boldsymbol{x}$ observed (customer's characteristics) and the action a taken (discount level offered). The probabilities $P(a, \boldsymbol{x})$ are themselves modeled by a logistic regression, whose parameters may be learned through an adaptation of the techniques by Filippi et al. [2010] and Faury et al. [2020]. We do so in the first phase of the adaptive policy introduced in this article. More details on the comparison of the new setting considered to known settings of CBwK may be found in Section 2.2.

Primal-dual approach. The second phase of the adaptive policy exhibited uses the primal-dual approach to a convex optimization problem—actually, a simple optimization problem given by a linear program. This approach was already used in various ways for bandits with knapsacks, including CBwK, to define policies based on the dual problem: this is explicit in the LagrangeBwK policy of Immorlica et al. [2019] and is implicit in the reward-minus-weighted-cost approach of Agrawal and Devanur [2016] and Agrawal et al. [2016], as we underline in the proof sketch of Section 4 as well as in the discussion of Section 6. However, we only use the primal-dual approach in the analysis and state our adaptive policy directly in terms of the primal problem, where we substituted upper-confidence estimates of the probabilities $P(a, \boldsymbol{x})$. We therefore end up with a most natural adaptive policy, which mimics the optimal static policy used as a benchmark. This direct primal statement of the policy actually also works for the setting of linear CBwK studied by Agrawal and Devanur [2016], as we show in Section 6. Policies based on such direct primal statements were already considered for bandits with knapsacks (see Li et al. [2021] and references therein) but do not seem easily extendable to CBwK.

Outline and main contributions. The first contribution of this article is a new structured setting of CBwK, based on a coupling between general rewards and cost vectors through conversions modeled based on a logistic regression; we present and discuss it in Section 2.1 (and explain its origins in Appendix A of the supplementary material). The adaptive policy introduced is described in Section 3. Its first phase consists of learning the parameter of logistic regression and is adapted from Faury et al. [2020]. Its second phase—and this is the second contribution of this article—directly solves a primal problem with optimistic conversion probabilities. The analysis, which we believe is concise, elegant, and natural, is provided in Sections 4 (when the context distribution ν is known) and 5 (when ν is unknown). As mentioned above, Section 6 draws the consequences of our second contribution for linear CBwK.

Notation. Throughout the article, vectors are denoted with bold symbols. In particular, $\mathbf{0}$ and $\mathbf{1}$ denote the vectors with all components equal to 0 and 1, respectively. With no additional subscript, $\|\mathbf{v}\|$ denotes the Euclidean norm of a vector $\mathbf{v}$, while a subscript given by a non-negative symmetric matrix M refers to $\|\mathbf{v}\|_M = \sqrt{\mathbf{v}^{\mathrm{T}} M \mathbf{v}}$.

2 Learning Protocol and Motivation

We describe the learning protocol and objectives considered (Section 2.1) and explain why it is not covered by earlier works (Section 2.2). We also detail (Appendix A in the supplementary material) how this learning protocol was defined based on an industrial motivation in the banking sector: market share expansion for loans by granting discounts, under commercial budget constraints.

2.1 Learning Protocol and Modeling Assumptions

We consider a finite action set $\mathcal{A}$, including a special action a_{null} called no-op, and a finite context set $\mathcal{X} \subseteq \mathbb{R}^n$. (We discuss and mitigate finiteness of $\mathcal{X}$ in Section 2.2.) A scalar reward function $r : \mathcal{A} \times \mathcal{X} \to [0, 1]$ and a vector-valued cost function $\boldsymbol{c} : \mathcal{A} \times \mathcal{X} \to [0, 1]^d$ evaluate the performance

of actions given the contexts. There are several sources of costs to control: each corresponds to a component of $\boldsymbol{c}$. We assume that these function are known, and (with no loss of generality) that their ranges are $[0,1]$ and $[0,1]^d$. The no-op action induces null reward and costs: $r(a_{\text{null}}, \boldsymbol{x}) = 0$ and $\boldsymbol{c}(a_{\text{null}}, \boldsymbol{x}) = \mathbf{0}$ for all $\boldsymbol{x} \in \mathcal{X}$.

Contexts—which correspond, for instance, to customers' characteristics, see Appendix A—are drawn sequentially according to some distribution ν, which may be known or unknown (we will deal with both cases). At each round $t \geqslant 1$, upon observing the context $\boldsymbol{x}_t \in \mathcal{X}$ drawn, the learner picks an action $a_t \in \mathcal{A}$, which corresponds, for instance, to an offer made to the customer t. If the latter accepts the offer, an event which we denote $y_t = 1$, then the learner obtains a reward $r(a_t, \boldsymbol{x}_t)$ and suffers some costs $\boldsymbol{c}(a_t, \boldsymbol{x}_t)$. When the customer declines the offer, we set $y_t = 0$, and null reward and costs are obtained. Thus, in both cases, the reward and costs may be written as $r(a_t, \boldsymbol{x}_t)\, y_t$ and $\boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t$. We call y_t the conversion and now explain how it is modeled.

Modeling conversions. We model each conversion y_t as an independent Bernoulli random drawn, with parameter $P(a_t, \boldsymbol{x}_t)$ depending on the context $\boldsymbol{x}_t$ and action $a_t \neq a_{\text{null}}$. We further assume that these probabilities may be written as a logistic regression model, i.e., there exists a known transfer function $\boldsymbol{\varphi} : \mathcal{A} \setminus \{a_{\text{null}}\} \times \mathcal{X} \to \mathbb{R}^m$ and some unknown parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^m$ such that

$$\forall \boldsymbol{x} \in \mathcal{X},\ \forall a \in \mathcal{A} \setminus \{a_{\text{null}}\}, \qquad P(a, \boldsymbol{x}) = \eta\big(\boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)\,, \qquad \text{where} \quad \eta(x) = 1/(1+\mathrm{e}^{-x})\,. \tag{1}$$

We assume that $\boldsymbol{\varphi}$ is normalized in a way that its Euclidean norm satisfies $\|\boldsymbol{\varphi}\| \leqslant 1$ and that a bounded convex set Θ containing $\boldsymbol{\theta}_\star$ is known. Such a modeling is natural and opens the toolbox of logistic bandits; see Faury et al. [2020] and references cited therein. We however note (and discuss this fact in Appendix C) that the logistic regression model above is slightly different from the one by Faury et al. [2020].

The concept of a conversion y for a round when the no-op action a_{null} is played is void, and thus, we leave the probabilities $P(a_{\text{null}}, \boldsymbol{x})$ undefined, though by an abuse of notation, these quantities might appear but always multiplied by a 0, given, e.g., by indicator functions like $\mathbb{1}_{\{a \neq a_{\text{null}}\}}$, null rewards $r(a_{\text{null}}, \boldsymbol{x})$, or null costs $\boldsymbol{c}(a_{\text{null}}, \boldsymbol{x})$.

Policies: static vs. adaptive. The learner is given a number of rounds T and a maximal budget B (the same for all cost components, with no loss of generality: up to some normalization). A static policy is a function $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of probability distributions over $\mathcal{A}$. As is traditional in the literature of CBwK (we recall below why this is the case), we take as benchmark the static policy $\pi^\star$ with largest expected cumulative rewards under the condition that its cumulative costs abide by the budget constraints in expectation. More formally, $\pi^\star$ achieves the maximum defining

$$\begin{aligned} \textsc{opt}(\nu, P, B) = \max_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} & T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, P(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right] \\ \text{under} \qquad & T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, P(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right] \leqslant B\mathbf{1}\,, \end{aligned} \tag{2}$$

where $\mathbb{E}_{\boldsymbol{X} \sim \nu}$ denotes an expectation solely over random contexts $\boldsymbol{X}$ following distribution ν, where $\pi_a(\boldsymbol{X})$ denotes the probability mass put by $\pi(\boldsymbol{X})$ on $a \in \mathcal{A}$, and where $\leqslant$ is understood componentwise. Of course, the sums in the two expectations above are taken indifferently over $\mathcal{A}$ or $\mathcal{A} \setminus \{a_{\text{null}}\}$.

The learner uses an adaptive policy, i.e., a sequence of measurable functions $\boldsymbol{p}_t : \mathcal{H}^{t-1} \times \mathcal{X} \to \mathcal{P}(\mathcal{A})$ indexed by $t \geqslant 1$, where $\mathcal{H} = \mathcal{X} \times \mathcal{A} \times \{0,1\}$. Indeed, the history available to the learner at the beginning of the round $t \geqslant 2$ is summarized by $h_{t-1} = (\boldsymbol{x}_s, a_s,\, y_s)_{s \leqslant t-1}$, and we define h_0 as the empty vector. Such a policy draws the action a_t for round $t \geqslant 1$ independently at random according to $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$. We impose hard budget constraints on adaptive policies: they must satisfy

$$\sum_{t \leqslant T} \boldsymbol{c}(a_t, \boldsymbol{x}_t) y_t \leqslant B\mathbf{1} \quad \text{a.s.}$$

Such adaptive policies are called feasible in the literature. To abide by these hard constraints, we may restrict our attention to adaptive policies that pick Dirac masses on a_{null} whenever one component of the cumulative costs is larger than $B - 1$. At the same time, an adaptive policy should maximize the cumulative rewards obtained or, equivalently, minimize its regret:

$$R_T = \textsc{opt}(\nu, P, B) - \sum_{t \leqslant T} r(a_t, \boldsymbol{x}_t) y_t\,.$$

BOX A: CONTEXTUAL BANDITS WITH KNAPSACKS [CBWK] FOR A CONVERSION MODEL

Known parameters: finite action set $\mathcal{A}$ including a no-op action a_{null}; finite context set $\mathcal{X} \subseteq \mathbb{R}^n$; scalar reward function $r : \mathcal{A} \times \mathcal{X} \to [0,1]$; vector-valued cost function $\boldsymbol{c} : \mathcal{A} \times \mathcal{X} \to [0,1]^d$; number T of rounds; total budget constraint $B > 0$.

Possibly unknown parameters: context distribution ν on $\mathcal{X}$; probability of conversion given action and context $P : \mathcal{A} \setminus \{a_{\text{null}}\} \times \mathcal{X} \to [0,1]$, modeled as $P(a, \boldsymbol{x}) = \eta\big(\boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)$ for some known transfer function $\boldsymbol{\varphi} : \mathcal{A} \setminus \{a_{\text{null}}\} \times \mathcal{X} \to \mathbb{R}^m$, with $\|\boldsymbol{\varphi}\| \leqslant 1$, and some unknown parameter $\boldsymbol{\theta}_\star \in \mathbb{R}^m$, lying in a known bounded convex set Θ.

For rounds $t = 1, 2, 3, \ldots, T$:

1. Context $\boldsymbol{x}_t \sim \nu$ is drawn independently of the past;
2. Learner observes $\boldsymbol{x}_t$ and picks an action $a_t \in \mathcal{A}$;
3. Conversion $y_t \in \{0,1\}$ is drawn according to $\mathrm{Ber}\big(P(a_t, \boldsymbol{x}_t)\big)$;
4. Learner observes y_t, gets reward $r(a_t, \boldsymbol{x}_t)\, y_t$, and suffers costs $\boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t$.

Goals: Maximize $\displaystyle\sum_{t \leqslant T} r(a_t, \boldsymbol{x}_t)\, y_t$ while controlling $\displaystyle\sum_{t \leqslant T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t \leqslant B\mathbf{1}$

It may be proved (along the same lines as Agrawal and Devanur [2016, Appendix B] do for a different model) that the optimal static policy $\pi^\star$ obtains, on average and in expectation, a cumulative reward at least as good as the best feasible adaptive policy.

Summary. A summary of the learning protocol and of the goals is provided in Box A. We note here that rewards gained and vector costs suffered at round t in the case $y_t = 1$ of a conversion could be stochastic with expectations $r(a_t, \boldsymbol{x}_t)$ and $\boldsymbol{c}(a_t, \boldsymbol{x}_t)$: our analysis and the regret bounds would be unchanged, as long as the expectation functions r and $\boldsymbol{c}$ are known.

2.2 Discussion and Comparison to Existing Learning Protocols

The setting described above may be reduced to the general setting of CBwK, as introduced by Badanidiyuru et al. [2014] and Agrawal et al. [2016]. Indeed, introduce independent Bernoulli variables $y_{t,a}$ with parameters $P(a, \boldsymbol{x}_t)$, for all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$, and set $y_{t,a_{\text{null}}} = 0$. The vectors

$$\Big(\boldsymbol{x}_t,\, \big(r_t(a)\big)_{a \in \mathcal{A}},\, \big(\boldsymbol{c}_t(a)\big)_{a \in \mathcal{A}}\Big), \qquad \text{where} \quad r_t(a) = r(a, \boldsymbol{x}_t)\, y_{t,a} \quad \text{and} \quad \boldsymbol{c}_t(a) = \boldsymbol{c}(a, \boldsymbol{x}_t)\, y_{t,a}$$

are i.i.d., and upon picking action $a_t \in \mathcal{A}$, the obtained and observed rewards and cost vectors equal $r_t(a_t)$ and $\boldsymbol{c}_t(a_t)$. When $\mathcal{X}$ is discrete, we may consider the set Π of base policies that map $\mathcal{X}$ to $\{\delta_a : a \in \mathcal{A}\}$, the set of Dirac masses at some $a \in \mathcal{A}$. The convex hull of Π is the set of all static policies $\mathcal{X} \to \mathcal{P}(\mathcal{A})$, against which we would like our policy to compete; but the adaptive policies by Badanidiyuru et al. [2014] and Agrawal et al. [2016] only compete with respect to the best single element in Π, not the best convex combination of elements of Π.

The setting of linear CBwK (Agrawal and Devanur [2016]) provides a structural link between contexts and expected rewards and cost vectors, but in a linear way that is incomparable to the setting of CBwK for a conversion model introduced above. More details are given in Section 6. We also mention that linear and logistic structural links between contexts (prices) and rewards or costs were studied in a non-contextual setting (i.e., not in CBwK) by Miao et al. [2021]. Their strategy bears some resemblance to the one by Agrawal and Devanur [2016], in particular, both consider an online convex optimization strategy as a subroutine.

All mentioned references consider a no-op action a_{null}. (It could be replaced by the existence of a standard action $a_{\text{no-cost}}$ always achieving null costs and possibly some positive rewards.)

On the contrary, none of the mentioned references assumes that the context $\mathcal{X}$ set is finite. This is a technical necessity for a part of the adaptive policy introduced; see the discussion of computational complexity at the end of Section 3. But somehow, considering a finite set Π of policies, as in Badanidiyuru et al. [2014] and Agrawal et al. [2016], is a counterpart to assuming finiteness of $\mathcal{X}$. Also, Appendix F actually mitigates this restriction that $\mathcal{X}$ is finite: learning the logistic parameter $\boldsymbol{\theta}_\star$ may be achieved with continuous contexts (see Phase 1 in Section 3); only the subsequent

optimization part (Phase 2 in Section 3) requires finiteness of $\mathcal{X}$. We may thus well discretize only $\mathcal{X}$ for this Phase 2, which is exactly what Appendix F performs. This mitigation comes with possible theoretical guarantees as Sections 4 and 5 reveal that the errors $\varepsilon_t(a, \boldsymbol{x})$ for learning $\theta_\star$ and P, obtained as outcomes of the first step of the analyses, are carried over in the subsequent steps, where the optimization part is evaluated.

3 Description of the Adaptive Policy Considered

At each stage $t \geqslant 1$, the policy first updates an estimator $\widehat{\boldsymbol{\theta}}_{t-1}$ of $\boldsymbol{\theta}_\star$ based on the history h_{t-1} available so far, based on an adaptation of the Logistic-UCB1 algorithm by Faury et al. [2020], and deduces estimators $\widehat{P}_{t-1}(a, \boldsymbol{x})$ and upper confidence bounds $U_{t-1}(a, \boldsymbol{x})$ of the probabilities $P(a, \boldsymbol{x})$. The policy then solves the corresponding estimated version of the optimization problem (2). We now describe the corresponding two steps. In the description below, quantities that depend on information available at round $t-1$ (respectively, t) are indexed by $t-1$ (respectively, t).

Phase 0: In case the cost constraints are about to be violated. To make sure cost constraints are never violated, whenever at least one of the components of the current cumulative costs is larger than $B-1$ and could possibly be larger than B at the end of round t, we play a_{null} (and we actually do so for the rest of the rounds). This corresponds to defining $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) = \delta_{a_{\text{null}}}$ for all $\boldsymbol{x} \in \mathcal{X}$, where $\delta_{a_{\text{null}}}$ denotes the Dirac mass on a_{null}. Otherwise, we proceed as described below in Phase 1 and Phase 2.

Phase 1: Learning $\boldsymbol{\theta}_\star$ via an adapted Logistic-UCB1. This first phase depends on a regularization parameter $\lambda > 0$ and on upper-confidence bonuses $\varepsilon_t(a, \boldsymbol{x}) > 0$, both to be specified by the analysis.

At rounds $t \geqslant 2$, we first maximize a regularized log-likelihood of the history h_{t-1}:

$$\tilde{\boldsymbol{\theta}}_{t-1} \in \operatorname*{argmax}_{\boldsymbol{\theta} \in \mathbb{R}^m} \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \left(y_s \ln \eta\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} \boldsymbol{\theta}\big) + (1-y_s) \ln\Big(1 - \eta\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} \boldsymbol{\theta}\big)\Big) \right) - \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 . \tag{3}$$

In the expression above, we read that we only gather information about $\boldsymbol{\theta}_\star$ at those rounds s when $a_s \neq a_{\text{null}}$. When $\tilde{\boldsymbol{\theta}}_{t-1}$ does not belong to Θ, an ad hoc projection step corrects for this, if needed:

$$\widehat{\boldsymbol{\theta}}_{t-1} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \Big\| \Psi_{t-1}(\boldsymbol{\theta}) - \Psi_{t-1}\big(\tilde{\boldsymbol{\theta}}_{t-1}\big) \Big\|_{W_{t-1}(\boldsymbol{\theta})^{-1}} , \tag{4}$$

$$\text{where} \qquad \Psi_{t-1}(\boldsymbol{\theta}) = \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \, \eta\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} \boldsymbol{\theta}\big) \, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s) + \lambda \boldsymbol{\theta}$$

$$\text{and} \qquad W_{t-1}(\boldsymbol{\theta}) = \lambda \, \mathrm{I}_m + \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \, \dot{\eta}\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} \boldsymbol{\theta}\big) \, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} . \tag{5}$$

We recall that the function $\dot{\eta}$ denotes the derivative of η, i.e., $\dot{\eta}(x) = \mathrm{e}^{-x} / \big(1 + \mathrm{e}^{-x}\big)^2$. We have $\dot{\eta} = \eta(1-\eta)$.

By plug-in, we finally define estimators and upper-confidence bounds of the probabilities $P(a, \boldsymbol{x})$ for $a \neq a_{\text{null}}$ and all $\boldsymbol{x} \in \mathcal{X}$:

$$\widehat{P}_{t-1}(a, \boldsymbol{x}) = \eta\big(\boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}} \widehat{\boldsymbol{\theta}}_{t-1}\big) \qquad \text{and} \qquad U_{t-1}(a, \boldsymbol{x}) = \min\big\{ \widehat{P}_{t-1}(a, \boldsymbol{x}) + \varepsilon_{t-1}(a, \boldsymbol{x}),\, 1 \big\} .$$

For a_{null}, no estimators or upper-confidence bounds need to be defined, as the quantities $P(a_{\text{null}}, \boldsymbol{x})$ are actually undefined.

Phase 2: Sampling, via solving an optimization problem with expected constraints. This phase relies on a conservative-budget parameter denoted by B_T, which is only slightly smaller than B and whose exact value is to be specified by the analysis.

We start with the case of a known context distribution ν. At round $t = 1$, we play an arbitrary action in $\mathcal{A} \setminus \{a_{\text{null}}\}$. At rounds $t \geqslant 2$, if at least one component of the cumulative vector costs suffered so far is larger than $B-1$, we pick $a_t = a_{\text{null}}$. Otherwise, we pick for $\boldsymbol{p}_t(h_{t-1}, \cdot)$ the solution of the optimization problem $\textsc{opt}(\nu, U_{t-1}, B_T)$ defined[1] in (2), and draw a_t according to $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$.

[1] In the definition (2) of $\textsc{opt}(\nu, U_{t-1}, B_T)$, expectations are only over $\boldsymbol{X} \sim \nu$ and not over the random variable U_{t-1}; more comments and explanations on this fact may be found in Appendix B.3.

BOX B: LOGISTIC-UCB1 FOR DIRECT SOLUTIONS TO OPT PROBLEMS

Parameters: regularization parameter $\lambda > 0$; conservative-budget parameter B_T; upper-confidence bonuses $\varepsilon_s(a, \boldsymbol{x}) > 0$, for $s \geqslant 1$ and $(a, \boldsymbol{x}) \in \big(\mathcal{A} \setminus \{a_{\text{null}}\}\big) \times \mathcal{X}$.

Round $t = 1$: play an arbitrary action $a_1 \in \mathcal{A} \setminus \{a_{\text{null}}\}$

At rounds $t \geqslant 2$:

Phase 0 If $\displaystyle\sum_{s \leqslant t-1} \boldsymbol{c}(a_s, \boldsymbol{x}_s)\, y_s \leqslant (B-1)\mathbf{1}$ is violated, then $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) = \delta_{a_{\text{null}}}$ for all $\boldsymbol{x}$

Phase 1 Otherwise, compute a maximum-likelihood estimator $\tilde{\boldsymbol{\theta}}_{t-1}$ of $\boldsymbol{\theta}_\star$ according to (3), compute its projection $\widehat{\boldsymbol{\theta}}_{t-1}$ onto Θ according to (4), and define, for $a \neq a_{\text{null}}$:

$$\widehat{P}_{t-1}(a, \boldsymbol{x}) = \eta\big(\boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathsf{T}}\, \widehat{\boldsymbol{\theta}}_{t-1}\big) \quad \text{and} \quad U_{t-1}(a, \boldsymbol{x}) = \min\Big\{\widehat{P}_{t-1}(a, \boldsymbol{x}) + \varepsilon_{t-1}(a, \boldsymbol{x}),\ 1\Big\}$$

Phase 2 Compute the solution $\boldsymbol{p}_t(h_{t-1}, \cdot\,)$ of

$$\textsc{opt}\big(\tilde{\nu}, U_{t-1}, B_T\big) = \max_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} T\, \mathbb{E}_{\boldsymbol{X} \sim \tilde{\nu}}\left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X})\right]$$

$$\text{under} \qquad T\, \mathbb{E}_{\boldsymbol{X} \sim \tilde{\nu}}\left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X})\right] \leqslant B_T \mathbf{1}\,,$$

where $\tilde{\nu}$ denotes either ν (when it is known) or its empirical estimate $\widehat{\nu}_t$ in (6)

Draw an arm $a_t \sim \boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$.

When the context distribution is unknown, we rather pick for $\boldsymbol{p}_t(h_{t-1}, \cdot\,)$ the solution of the optimization problem $\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, B_T\big)$, where

$$\widehat{\nu}_t = \frac{1}{t} \sum_{s=1}^{t} \delta_{\boldsymbol{x}_s}\,, \tag{6}$$

with $\delta_{\boldsymbol{x}}$ denoting the Dirac mass at $\boldsymbol{x} \in \mathcal{X}$. Since $\boldsymbol{x}_t$ is revealed at the beginning of round t, before we pick an action, we may indeed use $\widehat{\nu}_t$ at round t.

Summary and discussion of the computational complexity. We summarize the considered adaptive policy in Box B and now discuss its computational complexity.

As $\ln \varphi$ and $\ln(1-\varphi)$ are strictly concave and smooth, the maximum-likelihood step (3) of Phase 1 consists of maximizing a strictly concave and smooth function over $\mathbb{R}^m$, which may be performed efficiently. The projection step (4) of Phase 1 is however an issue, both with the version of Logistic-UCB1 discussed here and with the earlier approach by Filippi et al. [2010, Section 3]. The latter and Faury et al. [2020, Section 4.1] both underline that the projection step (4) is a complex optimization problem that however does not often need to be solved in practice, as they usually observe $\tilde{\boldsymbol{\theta}}_{t-1} \in \Theta$. Our numerical experiments concur with this statement (but admittedly, they rely on choosing a rather large value of Θ).

On the contrary, Phase 2 of the adaptive policy consists of solving a linear program with $|\mathcal{X}| \times |\mathcal{A}|$ constraints, where where $|\mathcal{X}|$ and $\mathcal{A}$ denote the cardinality of $\mathcal{X}$ and $\mathcal{A}$, respectively—see the detailed rewriting (13) in the supplementary material. Therefore, the computational complexity of Phase 2 is polynomial (of weak order) in $|\mathcal{X}| \times |\mathcal{A}|$. To achieve this acceptable complexity we had however to restrict our attention to finite sets of contexts $\mathcal{X}$, which requires in practice segmenting countable or continuous context sets into finitely many clusters, for instance. We do so in our experiments.

Simulation study. A simulation study on partially simulated but realistic data may be found in Appendix F. The underlying dataset is the standard "default of credit card clients" dataset of UCI [2016], initially provided by Yeh and Lien [2009]. (It may be used under a Creative Commons Attribution 4.0 International [CC BY 4.0] license.)

4 Analysis for a Known Context Distribution ν

Since Θ is bounded, the following quantity, standardly introduced in the context of logistic bandits (see Faury et al. [2020] and references therein), is finite, though possibly large:

$$\kappa = \sup\left\{ \frac{1}{\dot{\eta}\big(\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\,\boldsymbol{\theta}\big)} : \boldsymbol{x} \in \mathcal{X},\ a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \boldsymbol{\theta} \in \Theta \right\} < +\infty\,.$$

We denote by $\|\Theta\| = \max\big\{\|\boldsymbol{\theta}\| :\ \boldsymbol{\theta} \in \Theta\big\}$ the maximal Euclidean norm of an element in Θ.

By construction, given that individual cost vectors lie in $[0,1]^d$ and due to its "Phase 0", the adaptive policy considered always satisfies the budget constraints. The bound on rewards reads as follows.

Theorem 1. *In the setting of Box A of Section 2.1, we consider the adaptive policy of Box B of Section 3 assuming that the distribution of the contexts is known, i.e., with $\tilde{\nu} = \nu$. We set a confidence level $1-\delta \in (0,1)$ and use parameters $\lambda = m\ln(1+T/m)$,*

$$B_T = B - 2 - \sqrt{2T\ln(4d/\delta)}\,,$$

and $\varepsilon_t(a,\boldsymbol{x})$ stated in (9) *of the supplementary material. Then, provided that $T \geqslant 2m$ and $B > 4 + 2\sqrt{2T\ln(4d/\delta)}$, we have, with probability at least $1-2\delta$,*

$$\textsc{opt}(\nu,P,B) - \sum_{t\leqslant T} r(a_t,\boldsymbol{x}_t)\,y_t \leqslant \left(4 + 2\sqrt{2T\ln\frac{4d}{\delta}}\right)\frac{\textsc{opt}(\nu,P,B)}{B} + E_T + \sqrt{2T\ln\frac{4}{\delta}} + 1\,,$$

where the closed-form expression of $E_T = \mathcal{O}\big(m\sqrt{T}\ln T\big)$ is in (35) *of the supplementary material.*

We will rather discuss the bound of the more general Theorem 2 (to be stated and proved in Section 5) than the one of Theorem 1. We provide a proof sketch in Section 4.1 and discuss the main technical novelty in Section 4.2.

4.1 Proof Sketch for Theorem 1

The detailed proof of Theorem 1 may be found in Appendix B. We provide here an overview thereof, highlighting the four main ingredients. The third and fourth steps benefited from some inspiration drawn from the proof techniques of Agrawal and Devanur [2016]. The first step is an adaptation of Lemmas 1 and 2 by Faury et al. [2020].

First, the mentioned adaptation provides values of the parameters $\varepsilon_t(a,\boldsymbol{x})$ such that, with probability at least $1-\delta$,

$$\forall t \geqslant 1,\ \forall a \in \mathcal{A}\setminus\{a_{\text{null}}\},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \big|\widehat{P}_t(a,\boldsymbol{x}) - P(a,\boldsymbol{x})\big| \leqslant \varepsilon_t(a,\boldsymbol{x})\,,$$

$$\text{hence} \qquad U_t(a,\boldsymbol{x}) - 2\varepsilon_t(a,\boldsymbol{x}) \leqslant P(a,\boldsymbol{x}) \leqslant U_t(a,\boldsymbol{x})\,,$$

while $\displaystyle\sum_{t\leqslant T} \varepsilon_{t-1}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t \neq a_{\text{null}}\}}$ is of order $\sqrt{T}$ up to poly-logarithmic terms.

Second, the Phase 2 formulation of the strategy, in a primal form, is equivalently restated in a dual form. For each round $t \geqslant 2$, strong duality holds and entails the existence of a vector $\boldsymbol{\beta}_t^{\text{budg},\star} \in \mathbb{R}^d$ such that $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$ may be identified as the argmax over $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ of

$$\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[T\sum_{a\in\mathcal{A}}\Big(r(a,\boldsymbol{X}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{X})\Big)\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X}) + \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{a\in\mathcal{A}}\beta_{\boldsymbol{x},a}^{\text{p-pos},\star}\,\pi_a(\boldsymbol{x})\right].$$

By exploiting the KKT conditions, we are able to get rid of the double sum above and finally get a $\mathcal{X}$–pointwise characterization of $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$: for all $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}\boldsymbol{p}_t(h_{t-1},\boldsymbol{x}) &\in \operatorname*{argmax}_{\boldsymbol{q}\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}}\big(r(a,\boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{x})\big)\,U_{t-1}(a,\boldsymbol{x})\,q_a \\ &= \operatorname*{argmax}_{\boldsymbol{q}\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}}\big(r(a,\boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{x})\big)_+\,U_{t-1}(a,\boldsymbol{x})\,q_a\,.\end{aligned}$$

Non-negative parts $(\,\cdot\,)_+$ may be introduced thanks to the existence of the no-op action a_{null}. The distributions $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x})$ may therefore be interpreted as maximizing some upper-confidence bound on penalized gains (rewards minus some scalarized costs); the dual variables $\boldsymbol{\beta}_t^{\text{budg},\star}$ play a role similar to the Z parameter of Agrawal and Devanur [2016, Section 3.3] in terms of weighing gains versus costs. In passing, we also prove

$$\textsc{opt}(\nu, U_{t-1}, B_T) \geqslant B_T (\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}} \mathbf{1}$$

based on the KKT conditions. The latter inequality is comparable in spirit to the bound of Agrawal and Devanur [2016, Corollary 3], relating Z to $\textsc{opt}(\nu, P, B)/B$.

Third, for $t \geqslant 2$, whenever the policy $\boldsymbol{p}_t(h_{t-1}, \,\cdot\,)$ is obtained by solving the optimization problem $\textsc{opt}(\nu, U_{t-1}, B_T)$ of Phase 2 and by independence of $\boldsymbol{x}_t$ and h_{t-1}, we have

$$\begin{aligned}\frac{\textsc{opt}(\nu, U_{t-1}, B_T)}{T} &= \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, p_{t,a}(h_{t-1}, \boldsymbol{X})\right] \\ &= \mathbb{E}\big[r(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t) \,\big|\, h_{t-1}\big]\,.\end{aligned}$$

Therefore, repeated applications of the Hoeffding-Azuma inequality and the inequalities of the first step entail that, up to quantities of the order of $\sqrt{T}$,

$$\begin{aligned}\sum_{t=2}^{T} \frac{\textsc{opt}(\nu, U_{t-1}, B_T)}{T} &\approx \sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t) U_{t-1}(a_t, \boldsymbol{x}_t) \\ &\lesssim \sum_{t=2}^{T} \varepsilon_{t-1}(a_t, \boldsymbol{x}_t) \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} + \sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t) P(a_t, \boldsymbol{x}_t) \lesssim \sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t)\, y_t\,.\end{aligned}$$

We thus only need to control $\textsc{opt}(\nu, P, B) - \displaystyle\sum_{t=2}^{T} \frac{\textsc{opt}(\nu, U_{t-1}, B_T)}{T}$, which may be assumed $\geqslant 0$.

The value $B_T = B - 2 - \sqrt{2T\ln(4d/\delta)}$ and similar Hoeffding-Azuma-based arguments show that with high probability, the budget limit $B - 1$ is indeed never reached and that we always compute $\boldsymbol{p}_t(h_{t-1}, \,\cdot\,)$ in the way indicated by Phase 2.

Fourth, we collect all bounds together. We start with

$$\sum_{t=2}^{T} \frac{B_T}{T} (\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}} \mathbf{1} \leqslant \sum_{t=2}^{T} \frac{\textsc{opt}(\nu, U_{t-1}, B_T)}{T} \leqslant \textsc{opt}(\nu, P, B)\,.$$

We the exploit the dual characterization of $\boldsymbol{p}_t(h_{t-1}, \,\cdot\,)$ and the control $P \leqslant U_{t-1}$ to get that with high probability, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}&\sum_{a\in\mathcal{A}} \big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x})\big)\, U_{t-1}(a, \boldsymbol{x})\, p_{t,a}(h_{t-1}, \boldsymbol{x}) \\ &\qquad\qquad \geqslant \sum_{a\in\mathcal{A}} \big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x})\big)\, P(a, \boldsymbol{x})\, \pi_a^\star(\boldsymbol{x})\,.\end{aligned}$$

After integration over $\boldsymbol{X} \sim \nu$ and substituting of the definitions of $\pi^\star$ and $\boldsymbol{p}_{t,a}(h_{t-1}, \,\cdot\,)$, as well as the equality stemming from the KKT conditions, we have

$$\begin{aligned}&\overbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \boldsymbol{p}_{t,a}(h_{t-1}, \boldsymbol{X})\right]}^{=\textsc{opt}(\nu, U_{t-1}, B_T)/T} \\ &\quad \underbrace{- \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \boldsymbol{p}_{t,a}(h_{t-1}, \boldsymbol{X})\right]}_{(B_T/T)(\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}}\mathbf{1}} \\ \geqslant\ &\underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a, \boldsymbol{X})\, P(a, \boldsymbol{X})\, \pi_a^\star(\boldsymbol{X})\right]}_{=\textsc{opt}(\nu, P, B)/T} - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, P(a, \boldsymbol{X})\, \pi_a^\star(\boldsymbol{X})\right]}_{\leqslant (B/T)\mathbf{1}}\,.\end{aligned}$$

Rearranging and summing over $2 \leqslant t \leqslant T$, we obtain

$$\sum_{t=2}^{T} \frac{\textsc{opt}(\nu, P, B) - \textsc{opt}(\nu, U_{t-1}, B_T)}{T} \leqslant \sum_{t=2}^{T} \frac{B - B_T}{T} (\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}} \mathbf{1} \leqslant \left(\frac{B}{B_T} - 1 \right) \textsc{opt}(\nu, P, B) \,,$$

where we substituted the first inequality stated in this fourth step. This concludes the proof.

4.2 Discussion on the Main Technical Novelties

As should be clear from the comments at the beginning of Section 4.1, the technical novelties lies in the second step of the proof of Theorem 1. On the one hand, we are able to directly analyze a strategy stated in a primal form, which is a more natural formulation. On the other hand, doing so, we are also able to avoid the issues that come with dual formulations, relying, e.g., on some critical parameter Z, as in Agrawal and Devanur [2016, Theorem 3], to trade off rewards and costs. This parameter Z should be of order $\textsc{opt}/B$ and has to be learned, e.g., through $\sqrt{T}$ initial exploration rounds. (More details are to be found in Section E.3.) In our analysis, this parameter Z is superseded by dual optimal variables $\boldsymbol{\beta}_t^{\text{budg},\star} \geqslant \mathbf{0}$, that are only used in the analysis and not to state the policy, unlike in Agrawal and Devanur [2016]. Put differently, the clever use in this context of KKT conditions is the main technical novelty. On a side note, we are also able to take care in an explicit and detailed fashion of the no-op action a_{null}, whose specific treatment is often unaddressed in the literature.

5 Analysis for an Unknown Context Distribution ν

When the context distribution ν is unknown, we simply estimate it through its empirical frequencies (6). The regret bound is almost unchanged: an additional mild factor of, e.g., $2|\mathcal{X}|\sqrt{2T\ln(2T|\mathcal{X}|/\delta)}$ appears in the $\sqrt{T}$ term multiplying $\textsc{opt}(\nu, P, B)/B$. This term comes from some uniform deviation argument stated in (7) and 8.

Theorem 2. *In the setting of Box A of Section 2.1, we consider the adaptive policy of Box B of Section 3 with $\tilde{\nu} = \widehat{\nu}_t$ at rounds $t \geqslant 2$. We set a confidence level $1 - \delta \in (0, 1)$ and use parameters $\lambda = m \ln(1 + T/m)$, a working budget of*

$$B - b_T \,, \qquad \text{where} \qquad b_T = 2 + \sqrt{2T \ln(4d/\delta)} + |\mathcal{X}| \sqrt{2T \ln\bigl(2T|\mathcal{X}|/\delta\bigr)} \,,$$

and $\varepsilon_t(a, \boldsymbol{x})$ stated in (9) *of the supplementary material. Then, provided that $T \geqslant 2m$ and $B > 2b_T$, we have, with probability at least $1 - 3\delta$,*

$$\textsc{opt}(\nu, P, B) - \sum_{t \leqslant T} r(a_t, \boldsymbol{x}_t)\, y_t \leqslant 2 b_T \left(1 + \frac{\textsc{opt}(\nu, P, B)}{B} \right) + E_T \,,$$

where the expression of $E_T = \mathcal{O}\bigl(m\sqrt{T}\ln T\bigr)$ may be found in (35) *of the supplementary material.*

The order of magnitude of the regret bound is $\bigl(m + |\mathcal{X}|\textsc{opt}(\nu, P, B)/B\bigr)\sqrt{T}\ln T$, which is reminiscent of all known regret upper bounds for CBwK (e.g., the ones by Badanidiyuru et al. [2014] and Agrawal et al. [2016], for general CBwK, and Agrawal and Devanur [2016] for linear CBwK, see Section 6). The factor $|\mathcal{X}|$ may be improved, see below, but this is a detail. A discussion on exhibiting corresponding lower bounds is to be found at the end of Section 6.

A detailed proof of Theorem 2 is provided in Appendix D of the supplementary material. It follows closely the proof of Theorem 1, with modifications mostly consisting of relating quantities of the form

$$\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t}\bigl[f(\boldsymbol{X})\bigr] \;\text{ vs. }\; \mathbb{E}_{\boldsymbol{X} \sim \nu}\bigl[f(\boldsymbol{X})\bigr] \,, \;\text{ where, e.g., }\; f(\boldsymbol{X}) = \sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \boldsymbol{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \,.$$

To do so, we use that for all functions $f : \mathcal{X} \to [0, 1]$,

$$\forall\, t \leqslant T, \qquad \Bigl| \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t}\bigl[f(\boldsymbol{X})\bigr] - \mathbb{E}_{\boldsymbol{X} \sim \nu}\bigl[f(\boldsymbol{X})\bigr] \Bigr| \leqslant \sum_{x \in \mathcal{X}} \bigl| \widehat{\nu}_t(\boldsymbol{x}) - \nu(\boldsymbol{x}) \bigr| \stackrel{\text{def}}{=} \bigl\| \widehat{\nu}_t - \nu \bigr\|_1 \,, \tag{7}$$

where $\bigl\| \widehat{\nu}_t - \nu \bigr\|_1$ is the total variation distance between $\widehat{\nu}_t$ and ν. In Appendix D, we upper bound the latter, for the sake of simplicity, in a crude way by applying $T|\mathcal{X}|$ times the Hoeffding-Azuma inequality (once for each $1 \leqslant t \leqslant T$ and $\boldsymbol{x} \in \mathcal{X}$) and obtain that with probability at least $1 - \delta$,

$$\forall\, t \leqslant T, \qquad \bigl\| \widehat{\nu}_t - \nu \bigr\|_1 \leqslant |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \,. \tag{8}$$

The $|\mathcal{X}|\sqrt{2T\ln(2T|\mathcal{X}|/\delta)}$ term in the regret bound of Theorem 2 appears as the sum over $t \leqslant T$ of the deviation bounds (8). The bounds (8) may actually be improved into bounds of the order of $\sqrt{|\mathcal{X}|/t}$, via some Cauchy-Schwarz bound and a deviation argument in Banach spaces by Pinelis [1994], or by more direct techniques described by Devroye [1983, Lemma 3] and Berend and Kontorovich [2012]. In any case, the regret bound of Theorem 2 automatically benefits from such improvements, by replacing the $2|\mathcal{X}|\sqrt{2T\ln(2T|\mathcal{X}|/\delta)}$ term in the bound therein by the (sum over $t \leqslant T$ of the) better uniform deviation bounds.

6 Extension to Linear Contextual Bandits with Knapsacks

This section is a brief summary of Appendix E. We explain therein how the adaptive policy of Box B may be adapted to the setting of linear CBwK, introduced by Agrawal and Devanur [2016], where the bounded rewards r_t and vector costs $\boldsymbol{c}_t$ are independently generated at each round according to bounded distributions with respective expectations $\overline{r}(a_t, \boldsymbol{x}_t)$ and $\overline{\boldsymbol{c}}(a_t, \boldsymbol{x}_t)$, depending linearly on (a transfer function φ of) the contexts: for all $a \neq a_{\text{null}}$ and $\boldsymbol{x} \in \mathcal{X}$, for all components i of $\overline{\boldsymbol{c}}$,

$$\overline{r}(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\mu}_\star \qquad \text{and} \qquad \overline{c}_i(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\theta}_{\star,i} \, .$$

We consider the same benchmark $\textsc{opt}(\nu, \overline{r}, \overline{\boldsymbol{c}}, B)$ as Agrawal and Devanur [2016] and are able to exhibit a similar $\big(\textsc{opt}(\nu, \overline{r}, \overline{\boldsymbol{c}}, B)/B\big) m\sqrt{T}\ln T$ regret bound, with however a slight relaxation on the order of magnitude required for B. We do so with a strategy that we deem more direct and natural, inspired from the one of Box B, where in Phase 1 a LinUCB-type (Abbasi-Yadkori et al. [2011]) estimation of the parameters is performed, and where in Phase 2, a direct solution to an OPT problem with estimated parameters is performed. The parameters are upper-confidence functions U_{t-1} on $\overline{r}$ and lower-confidence vector functions $\boldsymbol{L}_{t-1}$ on $\overline{\boldsymbol{c}}$.

The main advantage of our approach in the case of linear contextual bandits is exactly as described in Section 4.2: avoiding the critical parameter Z of Agrawal and Devanur [2016, Theorem 3], which is used to trade off rewards and costs. The main limitation of our approach is the assumption of a finite context set $\mathcal{X}$, which is required to make the Phase-2 linear program tractable.

7 Future Work

We conclude this article with a list of issues to be further investigated.

First, as discussed in Section 2.2, the restrictions of finiteness should be alleviated: finiteness of the context set in the setting of this article, or finiteness of the set of benchmark policies in other settings (see Badanidiyuru et al., 2014 and Agrawal et al., 2016).

Second, we only dealt with $\leqslant$ budget constraints (and do does the literature so far). Direct approaches to constraints of the form $\geqslant$ remain to be further investigated.

A third series of questions to be clarifies concerns regret lower bounds, and more generally, the tightness of the results—in particular, the required conditions on budget sizes. Earlier references for contextual bandits with knapsacks did also not provide lower bounds statements that were simultaneously optimal (i.e., matching the obtained upper bounds) and general (i.e., valid for all problems with a given number of rounds T, a given budget B, and a given value OPT for the optimal expected reward achievable by a static policy). Badanidiyuru et al. [2014, comments after Theorem 1] merely indicates that the obtained regret upper bound is optimal in some regimes, e.g., when the budget B grows linearly with the number of rounds T. Agrawal and Devanur [2016, comments after Theorem 1] only compares the obtained regret upper bound to the case of no budget constraints. In particular, as far as the orders of magnitude in T are concerned, it is unclear whether the $(\textsc{opt}/B)\sqrt{T}$ rates achieved (up to poly-logarithmic factors) in the present article and in the two mentioned references are optimal. These rates do not match the optimal rates in the case of no contexts, which were stated and proved by Badanidiyuru et al. [2013].

Acknowledgments and Disclosure of Funding

Zhen Li and Gilles Stoltz have no direct funding to acknowledge other than the salaries paid by their employers, BNP Paribas, CNRS, and HEC Paris. They have no competing interests to declare.

References

Default of credit card clients. UCI Machine Learning Repository, 2016. URL `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`.

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS'11)*, volume 24, 2011.

S. Agrawal and N. Devanur. Linear contextual bandits with knapsacks. In *Advances in Neural Information Processing Systems (NeurIPS'16)*, volume 29, 2016.

S. Agrawal, N.R. Devanur, and L. Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT'16)*, volume PMLR:49, pages 4–18, 2016.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with knapsacks. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS'13)*, pages 207–216, 2013. Latest extended version available at arXiv:1305.2545, dated September 2017.

A. Badanidiyuru, J. Langford, and A. Slivkins. Resourceful contextual bandits. In *Proceedings of the 27th Conference on Learning Theory (COLT'14)*, volume PMLR:35, pages 1109–1134, 2014.

A. Badanidiyuru, R. Kleinberg, and A. Slivkins. Bandits with global convex constraints and objective. *Journal of the ACM*, 65(3):1–55, 2018.

D. Berend and A. Kontorovich. On the convergence of the empirical distribution, 2012. Preprint, arXiv:1205.6711.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AIStats'11)*, volume 15, pages 208–214, 2011.

L. Devroye. The equivalence of weak, strong and complete convergence in l_1 for kernel density estimates. *Annals of Statistics*, 11(3):896–904, 1983.

L. Faury, M. Abeille, C. Calauzenes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, volume PMLR:119, pages 3052–3060, 2020.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems (NeurIPS'10)*, volume 23, 2010.

N. Immorlica, K.A. Sankararaman, R. Schapire, and A. Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS'19)*, pages 202–219, 2019.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

X. Li, C. Sun, and Y. Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, pages 6483–6492, 2021.

S. Miao, Y. Wang, and J. Zhang. A general framework for resource constrained revenue management with demand learning and large action space. *Available at SSRN 3841273*, 2021.

I. Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Annals of Probability*, 22(4):1679–1706, 1994.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12 (1-2):1–286, 2019.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Z. Xu and V.-A. Truong. Reoptimization algorithms for contextual bandits with knapsack constraints, 2019. URL `http://www.columbia.edu/~vt2196/OnlineLearningAllocation8.pdf`.

I.C. Yeh and C.H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, volume PMLR:119, pages 3052–3060, 2003.

Supplementary Material for

"Contextual Bandits with Knapsacks for a Conversion Model"

The appendices of this article contain the following elements.

- Appendix A provides the industrial motivation behind the setting described in Section 2.1, namely, within the banking industry, a market share expansion for loans.
- Appendix B contains the proof of Theorem 1, i.e., the regret bound in case of a known distribution ν, except for a lemma on learning the parameter of logistic bandits, provided in Appendix C.
- Appendix C states and proves the indicated lemma; both the statement and the proof are mere adaptations of Faury et al. [2020, Lemmas 1 and 2].
- Appendix D contains the proof of Theorem 2, i.e., explains how to adapt the proof of Appendix B to the case of an unknown distribution ν.
- Appendix E details the claims of Section 6, i.e., the extension of the techniques introduced to the setting of linear CBwK.
- Appendix F reports a simulation study based on realistic data.

A Industrial Motivation: Market Share Expansion for Loans

We describe the industrial problem we faced and which led to the setting described in Section 2.1.

Incentives and discounts are common practices in many industries to achieve some business objectives; however, there is usually a limit on the number or/and total volume of discounts that can be granted, so companies need to select carefully who should receive them. We consider, for instance, the banking industry, with a business objective of market share expansion: achieving the highest possible volume of loan subscription (total subscribed amount). Note that in practice, all loan applications need to go first through a risk-assessment process, and offers are only made if the bank considers that the loan will not put the customer's solvability at risk. We assume that all the clients concerned here have already gone through this process and are eligible for getting a loan from a given bank. We formulate the problem in a sequential manner as follows.

At each round $t \geqslant 1$, a client asks for a credit product. Her/his characteristics are denoted by $\boldsymbol{x}_t \in \mathbb{R}^n$, and encompass the socio-demographic profile, the loan request (amount $x_{\text{am},t}$, duration $x_{\text{dur},t}$), etc. It is reasonable to model these characteristics as independent draws from a common (possibly unknown) distribution ν. Based on $\boldsymbol{x}_t$, the bank will suggest some standard interest rate $\mathrm{ir}(\boldsymbol{x}_t)$ based on its pricing rules; the detail of the rules is not relevant and we assume that the underlying function ir is given. If client t accepts the offered rate $\mathrm{ir}(\boldsymbol{x}_t)$ and subscribes to the loan, an event which we denote by $y_t = 1$ and call a conversion, the bank gets a sales performance (gain on volume) $x_{\text{am},t}$. Otherwise, the client declines the offer, which we denote by $y_t = 0$ and the bank gets a null reward.

Actually, to improve the chances of a conversion, the bank may also offer a discount $a_t \in (0, 1]$, or lack of discount $a_t = 0$, on the interest rate. If it offers a discount $a_t > 0$ and $y_t = 1$, the bank will suffer a loss of earnings, equal to $a_t \,\mathrm{ir}(\boldsymbol{x}_t)\,\mathrm{out}(\boldsymbol{x}_t)$, where $\mathrm{out}(\boldsymbol{x}_t)$ denotes the total outstanding amounts. This loss of earnings is considered a promotion cost. These promotion costs are summed up to previous such costs and should usually not exceed a fixed-in-advance budget $B_2 > 0$. Also, there is usually a fixed-in-advance limit $B_1 > 0$ on the total number of clients who can subscribe with a discount.

Given that the customers' characteristics are i.i.d., it is indeed reasonable to assume that y_t follows some Bernoulli distribution with unknown probability $P(a_t, \boldsymbol{x}_t)$. Of course, the higher the discount, the higher the probability of a conversion.

We summarize the setting with the notation of Section 2.1. We assume that discounts are picked in a finite grid $\mathcal{D} = \big\{j/D : j \in \{0, \dots, D\}\big\}$, so that the action set equals $\mathcal{A} = \mathcal{D} \cup \{a_{\text{null}}\}$. At each round $t \geqslant 1$, given the customer's characteristics $\boldsymbol{x}_t$ and the discount $a_t \in [0, 1]$ picked by the bank, the latter receives the following reward and suffers the following costs:

$$r(a_t, \boldsymbol{x}_t)\, y_t\,, \qquad \text{where} \qquad r : (a, \boldsymbol{x}) \mapsto x_{\text{am}}\,,$$
$$\text{and} \qquad \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t\,, \qquad \text{where} \qquad \boldsymbol{c} : (a, \boldsymbol{x}) \mapsto \big(\mathbb{1}_{\{a \neq 0\}},\, a\,\mathrm{ir}(\boldsymbol{x}_t)\,\mathrm{out}(\boldsymbol{x}_t)\big)\,.$$

The first component of the cumulative cost vector measures the total number subscriptions with discounts, and the second component reports the total promotion costs. The bank wants to enforce

$$\sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t = \sum_{t=1}^{T} \big(\mathbb{1}_{\{a_t \neq 0\}},\, a_t\,\mathrm{ir}(\boldsymbol{x}_t)\,\mathrm{out}(\boldsymbol{x}_t)\big)\, y_t \leqslant (B_1,\, B_2)$$

while maximizing the sum of the achieved rewards.

Normalizations in $[0, 1]$ both for rewards and cost components may be achieved by considering the maximal amount M_{am} and outstanding M_{out} that the bank would allow, and by considering

$$r : (a, \boldsymbol{x}) \mapsto x_{\text{am}}/M_{\text{am}} \qquad \text{and} \qquad \boldsymbol{c} : (a, \boldsymbol{x}) \mapsto \big(\mathbb{1}_{\{a \neq 0\}},\, a\,\mathrm{ir}(\boldsymbol{x}_t)\,\mathrm{out}(\boldsymbol{x}_t)/M_{\text{out}}\big)$$

with the alternative budget $(B_1,\, B_2/M_{\text{out}})$. A single budget parameter $B = \min\big\{B_1,\, B_2/M_{\text{out}}\big\}$ may be considered by a final normalization: by dividing the first cost component by $B_1/B > 1$ if $B_1 > B$, or the second cost component by $(B_2/M_{\text{out}})/B > 1$ if $B_2/M_{\text{out}} > B$, respectively.

B Detailed Proof of the Regret Bound in Case of a Known Distribution ν: Proof of Theorem 1

The proof is divided into four steps.

B.1 First Step: Defining Confidence Intervals on the Probabilities $P(a, \boldsymbol{x})$

The keystone of this step is the following lemma, adapted from Faury et al. [2020]: it provides guarantees for the adapted version of Logistic-UCB1 defined in Phase 1 of the adaptive policy studied in this article. The lemma actually holds for any sampling strategy of the arms, not just the one used in Phase 2 of the adaptive policy.

The reasons of the adaptations, lying in different settings being considered, as well as a detailed proof, are provided in Appendix C. We recall that we denote

$$\kappa = \sup\left\{ \frac{1}{\dot{\eta}\big(\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\theta}\big)} : \boldsymbol{x} \in \mathcal{X},\ a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \boldsymbol{\theta} \in \Theta \right\},$$

and that since $\dot{\eta} = \eta(1-\eta) \in [0, 1/4]$, we always have $\kappa \geqslant 4$. We also recall the notation for the maximal Euclidean norm of an element in Θ:

$$\|\Theta\| = \max_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\theta}\|\,.$$

Lemma 1 (combination of Lemmas 1 and 2 of Faury et al. [2020] with minor adjustments, detailed in Appendix C)**.** *Assume that $\kappa < +\infty$. Fix any sampling strategy and consider the version of Logistic-UCB1 given by* (3)–(5)*. For all $\delta \in (0,1)$, there exists an event $\mathcal{E}_{\text{prob},\delta}$ with probability at least $1-\delta$ and such that over $\mathcal{E}_{\text{prob},\delta}$:*

$$\forall t \geqslant 1,\ \forall a \in \mathcal{A} \setminus \{a_{null}\},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \big|\widehat{P}_t(a,\boldsymbol{x}) - P(a,\boldsymbol{x})\big| \leqslant \gamma_{t,\lambda,\delta}\sqrt{\kappa\big(\|\Theta\|+1/2\big)}\,\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{V_t^{-1}}\,,$$

$$\textit{where} \qquad \gamma_{t,\lambda,\delta} = \sqrt{\lambda}\big(\|\Theta\|+1/2\big) + \frac{2}{\sqrt{\lambda}}\ln\left(\frac{2^m}{\delta}\left(1+\frac{t}{4m\lambda}\right)^{m/2}\right)$$

$$\textit{and} \qquad V_t = \sum_{s=1}^{t} \boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\mathbb{1}_{\{a_s \neq a_{null}\}} + \kappa\lambda\,\mathrm{I}_m\ .$$

Associated confidence intervals for the $P(a, \boldsymbol{x})$. Thanks to this lemma, we consider the upper-confidence bonuses

$$\varepsilon_t(a,\boldsymbol{x}) = \gamma_{t,\lambda,\delta}\sqrt{\kappa\big(\|\Theta\|+1/2\big)}\,\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{V_t^{-1}}\,. \tag{9}$$

On the event $\mathcal{E}_{\text{prob},\delta}$ of Lemma 1, we have, for all $t \geqslant 1$, all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$, and all $\boldsymbol{x} \in \mathcal{X}$: on the one hand,

$$U_t(a,\boldsymbol{x}) = \min\Big\{\widehat{P}_t(a,\boldsymbol{x}) + \varepsilon_t(a,\boldsymbol{x}),\ 1\Big\} \geqslant \min\big\{P(a,\boldsymbol{x}),\ 1\big\} = P(a,\boldsymbol{x}) \tag{10}$$

and on the other hand,

$$U_t(a,\boldsymbol{x}) = \min\Big\{\widehat{P}_t(a,\boldsymbol{x}) + \varepsilon_t(a,\boldsymbol{x}),\ 1\Big\} \leqslant \widehat{P}_t(a,\boldsymbol{x}) + \varepsilon_t(a,\boldsymbol{x}) \leqslant P(a,\boldsymbol{x}) + 2\varepsilon_t(a,\boldsymbol{x})\,. \tag{11}$$

Control of the sum of upper-confidence bonuses. According to the proof sketch of Section 4, it only remains to control the sum of the upper-confidence bonuses at observed contexts $\boldsymbol{x}_t$ and played actions a_t. Note that at rounds $t \geqslant 2$, we use the bonuses $\varepsilon_{t-1}(a,\boldsymbol{x})$. We prove that

$$2\sum_{t=2}^{T} \varepsilon_{t-1}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t \neq a_{\text{null}}\}} \leqslant \underbrace{\gamma_{T,\lambda,\delta}\sqrt{\kappa\big(4\|\Theta\|+2\big)}\sqrt{2mT\max\left\{1,\ \frac{1}{\kappa\lambda}\right\}\ln\left(1+\frac{T}{\kappa\lambda m}\right)}}_{\stackrel{\text{def}}{=} E_T}\,. \tag{12}$$

To that end, we first note that $\gamma_{t-1,\lambda,\delta} \leqslant \gamma_{T,\lambda,\delta}$:

$$2\sum_{t=2}^{T} \varepsilon_{t-1}(a_t, \boldsymbol{x}_t) \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} \leqslant \gamma_{T,\lambda,\delta} \sqrt{\kappa\big(4\|\Theta\| + 2\big)} \sum_{t=2}^{T} \big\| \boldsymbol{\varphi}(a_t, \boldsymbol{x}_t) \big\|_{V_{t-1}^{-1}} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} \,.$$

Given that $\|\boldsymbol{\varphi}\| \leqslant 1$ by assumption, a direct application of Lemma 2 below (a classical result of basic algebra for linear bandits) ensures that

$$\sum_{t=1}^{T} \big\| \boldsymbol{\varphi}(a_t, \boldsymbol{x}_t) \big\|^2_{V_{t-1}^{-1}} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} \leqslant 2m \max\left\{1,\ \frac{1}{\kappa\lambda}\right\} \ln\left(1 + \frac{\sum_{t=1}^{T} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}}}{\kappa\lambda m}\right).$$

A Cauchy-Schwarz inequality thus entails

$$\begin{aligned}
\sum_{t=1}^{T} \big\| \boldsymbol{\varphi}(a_t, \boldsymbol{x}_t) \big\|_{V_{t-1}^{-1}} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} &\leqslant \sqrt{\sum_{t=1}^{T} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}}} \sqrt{2m \max\left\{1,\ \frac{1}{\kappa\lambda}\right\} \ln\left(1 + \frac{\sum_{t=1}^{T} \mathbb{1}_{\{a_t \neq a_{\text{null}}\}}}{\kappa\lambda m}\right)} \\
&\leqslant \sqrt{T} \sqrt{2m \max\left\{1,\ \frac{1}{\kappa\lambda}\right\} \ln\left(1 + \frac{T}{\kappa\lambda m}\right)},
\end{aligned}$$

which concludes this step.

Lemma 2 (Elliptic potential and determinant-trace inequality, cf. Lemmas 10 and 11 of Abbasi-Yadkori et al. [2011], Lemmas 15 and 16 of Faury et al. [2020])**.** *For all $\lambda > 0$ and all sequences $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ of vectors in $\mathbb{R}^m$ with $\|\boldsymbol{u}_s\| \leqslant 1$, defining $U_0 = \lambda \,\mathrm{I}_m$ and for $t \geqslant 1$,*

$$U_t = \lambda \,\mathrm{I}_m + \sum_{s=1}^{t} \boldsymbol{u}_t (\boldsymbol{u}_t)^{\mathrm{T}} \,,$$

we have, for all $\tau \geqslant 1$:

$$\sum_{t=1}^{\tau} \|\boldsymbol{u}_t\|^2_{U_{t-1}^{-1}} \leqslant 2m \max\left\{1,\ \frac{1}{\lambda}\right\} \ln\left(1 + \frac{\tau}{\lambda m}\right).$$

B.2 Second Step: Dual Formulation of the Sampling Phase (Phase 2) and Consequences

In this step, we consider a round $t \geqslant 2$ for which the cost constraints of Phase 0 of the adaptive policy are not violated and the optimization problem $\textsc{opt}(\nu,\, U_{t-1}, B_{t,T})$ is to be solved; its solution is the policy $\boldsymbol{p}_t(h_{t-1},\, \cdot\,)$ used to sample a_t according to $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$.

We first rewrite in its dual form the optimization problem $\textsc{opt}(\nu,\, U_{t-1}, B_{t,T})$ and show that strong duality holds. As a consequence, there exists a vector $\boldsymbol{\beta}_t^{\text{budg},\star} \in \mathbb{R}^d$ such that $\boldsymbol{p}_t(h_{t-1},\, \cdot\,)$ may be identified as

$$\operatorname*{argmax}_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} \ \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[T \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big) \, U_{t-1}(a, \boldsymbol{X}) \, \pi_a(\boldsymbol{X}) + \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}} \beta_{\boldsymbol{x},a}^{\text{p-pos},\star} \, \pi_a(\boldsymbol{x}) \right].$$

By exploiting the KKT conditions, we are able to get rid of the double sum above and finally get a $\mathcal{X}$–pointwise characterization of $\boldsymbol{p}_t(h_{t-1},\, \cdot\,)$: for all $\boldsymbol{x} \in \mathcal{X}$,

$$\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big) \, U_{t-1}(a, \boldsymbol{x}) \, q_a \,,$$

where, with no impact, we may replace the $r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x})$ by their non-negative parts. The distributions $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x})$ may therefore be interpreted as maximizing some upper-confidence bound on penalized gains (rewards minus some scalarized costs).

We also prove $\textsc{opt}(\nu, U_{t-1}, B_T) \geqslant B_T (\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}} \mathbf{1}$ based on the KKT conditions.

Primal form of the optimization problem $\textsc{opt}(\nu,\, U_{t-1}, B_T)$. Since $\mathcal{X}$ is a finite set (this is actually the key place where we need this assumption), the optimization problem $\textsc{opt}(\nu,\, U_{t-1}, B_T)$ may be stated as the *opposite* of

$$\begin{aligned}
\min_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \; & -T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] \\
\text{under} \quad & T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] \leqslant B_T\mathbf{1}\,, \\
& \forall a\in\mathcal{A},\ \forall \boldsymbol{x}\in\mathcal{X}, \quad \pi_a(\boldsymbol{x}) \geqslant 0\,, \\
& \forall \boldsymbol{x}\in\mathcal{X}, \quad \sum_{a\in\mathcal{A}} \pi_a(\boldsymbol{x}) = 1\,.
\end{aligned}$$

Thanks to the no-op action $a_{\text{null}}\in\mathcal{A}$, which is used to model abstention and results in null rewards and costs, i.e., $r(a_{\text{null}},\boldsymbol{x})=0$ and $\boldsymbol{c}(a_{\text{null}},\boldsymbol{x})=\mathbf{0}$ for all $\boldsymbol{x}\in\mathcal{X}$, we may relax the third constraint into

$$\forall \boldsymbol{x}\in\mathcal{X}, \quad \sum_{a\in\mathcal{A}} \pi_a(\boldsymbol{x}) \leqslant 1\,.$$

Indeed, any vector $\big(\pi_a(\boldsymbol{x})\big)\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}$ satisfying the constraint with $\leqslant 1$ can be transformed into a vector $\big(\pi'_a(\boldsymbol{x})\big)\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}$ for which the expected reward and the first and second constraints remain identical while the third constraint is satisfied with $=1$: by adding the necessary probability mass to the components $\pi_{a_{\text{null}}}(\boldsymbol{x})$.

In the sequel, we consider this primal problem with the $\leqslant 1$ constraint:

$$\begin{aligned}
-\textsc{opt}(\nu,\, U_{t-1}, B_T) = \min_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \; & -T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] \\
\text{under} \quad & T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] \leqslant B_T\mathbf{1}\,, \\
& \forall a\in\mathcal{A},\ \forall \boldsymbol{x}\in\mathcal{X}, \quad \pi_a(\boldsymbol{x}) \geqslant 0\,, \\
& \forall \boldsymbol{x}\in\mathcal{X}, \quad \sum_{a\in\mathcal{A}} \pi_a(\boldsymbol{x}) \leqslant 1\,.
\end{aligned} \tag{13}$$

It forms a convex optimization problem, as its objective and its constraints are all affine (Boyd and Vandenberghe [2004, Section 4.2.1]).

Lagrangian (dual) form of the optimization problem. Denote by $\boldsymbol{\beta}^{\text{budg}}$, $\boldsymbol{\beta}^{\text{p-pos}}$, $\boldsymbol{\beta}^{\text{p-sum}}$ the vector dual variables associated with the constraints on budget [budg], non-negative probability [p-pos], and sum of probabilities [p-sum], respectively. The vectors $\boldsymbol{\beta}^{\text{p-pos}}$ and $\boldsymbol{\beta}^{\text{p-sum}}$ have components $\beta^{\text{p-sum}}_{\boldsymbol{x}}$ and $\beta^{\text{p-pos}}_{\boldsymbol{x},a}$ indexed by $\boldsymbol{x}\in\mathcal{X}$ and $a\in\mathcal{A}$.

We define the Lagrangian associated with our primal problem:

$$\begin{aligned}
&\mathcal{L}_t\Big(\big(\pi_a(\boldsymbol{x})\big)_{a,\boldsymbol{x}},\, \boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}\Big) \\
&\qquad = -\,T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] \\
&\qquad\quad + \big(\boldsymbol{\beta}^{\text{budg}}\big)^{\mathsf{T}}\left(T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X})\right] - B_T\mathbf{1}\right) \\
&\qquad\quad + \sum_{\boldsymbol{x}\in\mathcal{X}} \beta^{\text{p-sum}}_{\boldsymbol{x}}\left(\sum_{a\in\mathcal{A}} \pi_a(\boldsymbol{x}) - 1\right) - \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{a\in\mathcal{A}} \beta^{\text{p-pos}}_{\boldsymbol{x},a}\,\pi_a(\boldsymbol{x})\,.
\end{aligned} \tag{14}$$

The dual problem consists of maximizing

$$\inf_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \mathcal{L}_t\Big(\big(\pi_a(\boldsymbol{x})\big)_{a,\boldsymbol{x}}, \boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}\Big)$$

under the constraints that all components of the $\boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}$ are non-negative, which we denote in a vector-wise manner by

$$\boldsymbol{\beta}^{\text{budg}} \geqslant \mathbf{0}\,, \qquad \boldsymbol{\beta}^{\text{p-sum}} \geqslant \mathbf{0}\,, \qquad \boldsymbol{\beta}^{\text{p-pos}} \geqslant \mathbf{0}\,.$$

Strong duality and consequences. We explain why the so-called Slater's condition (Boyd and Vandenberghe, 2004, Section 5.2.3) holds; it entails that the value $-\text{OPT}\big(\nu,\, U_{t-1}, B_T\big)$ of the primal problem equals the value of the Lagrangian dual problem. The primal problem is convex, all its constraints are affine, with domain $\mathbb{R}^{\mathcal{A}\times\mathcal{X}}$: Slater's condition therefore reduces to feasibility. And feasibility of the constraints is clear by taking Dirac masses on a_{null}, i.e., $\pi_a(\boldsymbol{x}) = \delta_{a_{\text{null}}}$ for all $\boldsymbol{x} \in \mathcal{X}$. Since the values of the primal and dual problems are clearly larger than $-T > -\infty$, Slater's condition also implies that the dual optimal value is achieved at a dual feasible set of parameters, i.e., that the the constrained supremum and the infimum defining the dual problem are a minimum and a maximum, respectively. We may therefore summarize the consequences of Slater's condition as follows:

$$\begin{aligned} -\text{OPT}(\nu, U_{t-1}, B_T) = \max_{\boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}} \; \min_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \; & \mathcal{L}_t\Big(\big(\pi_a(\boldsymbol{x})\big)_{a,\boldsymbol{x}}, \boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}\Big) \\ \text{under} \qquad & \boldsymbol{\beta}^{\text{budg}} \geqslant \mathbf{0}\,, \qquad \boldsymbol{\beta}^{\text{p-sum}} \geqslant \mathbf{0}\,, \qquad \boldsymbol{\beta}^{\text{p-pos}} \geqslant \mathbf{0}\,. \end{aligned}$$

Because of strong duality and the existence of a dual feasible set of parameters, the max-min above equals its min-max counterpart (Boyd and Vandenberghe [2004, Sections 5.4.1 and 5.4.2]):

$$-\text{OPT}(\nu, U_{t-1}, B_T) = \min_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \; \max_{\boldsymbol{\beta}^{\text{budg}}\geqslant\mathbf{0},\, \boldsymbol{\beta}^{\text{p-sum}}\geqslant\mathbf{0},\, \boldsymbol{\beta}^{\text{p-pos}}\geqslant\mathbf{0}} \mathcal{L}_t\Big(\big(\pi_a(\boldsymbol{x})\big)_{a,\boldsymbol{x}}, \boldsymbol{\beta}^{\text{budg}}, \boldsymbol{\beta}^{\text{p-sum}}, \boldsymbol{\beta}^{\text{p-pos}}\Big)\,.$$

We let $\boldsymbol{\beta}_t^{\text{budg},\star} \geqslant \mathbf{0}$, $\boldsymbol{\beta}_t^{\text{p-sum},\star} \geqslant \mathbf{0}$, and $\boldsymbol{\beta}_t^{\text{p-pos},\star} \geqslant \mathbf{0}$ be an optimal dual solution and recall that $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$ denote an optimal primal solution, which, with no loss of generality, may be assumed satisfying the $\leqslant 1$ constraints with equality. From Boyd and Vandenberghe [2004, Section 5.4.2], this pair of solutions forms a saddle-point for the Lagrangian; in particular,

$$\begin{aligned} -\text{OPT}(\nu, U_{t-1}, B_T) &= \mathcal{L}_t\big(\boldsymbol{p}_t(h_{t-1},\,\cdot\,), \boldsymbol{\beta}_t^{\text{budg},\star}, \boldsymbol{\beta}_t^{\text{p-sum},\star}, \boldsymbol{\beta}_t^{\text{p-pos},\star}\big) \\ &= \min_{(\pi_a(\boldsymbol{x}))\in\mathbb{R}^{\mathcal{A}\times\mathcal{X}}} \mathcal{L}_t\Big(\big(\pi_a(\boldsymbol{x})\big), \boldsymbol{\beta}_t^{\text{budg},\star}, \boldsymbol{\beta}_t^{\text{p-sum},\star}, \boldsymbol{\beta}_t^{\text{p-pos},\star}\Big) \\ &= \min_{\pi:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \mathcal{L}_t(\pi, \boldsymbol{\beta}_t^{\text{budg},\star}, \boldsymbol{\beta}_t^{\text{p-sum},\star}, \boldsymbol{\beta}_t^{\text{p-pos},\star})\,. \end{aligned} \tag{15}$$

The distribution $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$ played thus appears as the argument of the minimum above.

Substituting the definition (14) of $\mathcal{L}_t$ into the characterization (15), rearranging the first two terms of $\mathcal{L}_t$, noting that the third term of $\mathcal{L}_t$ is null, and discarding the constant term $B_T\big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\mathbf{1}$, we get:

$$\begin{aligned} &\boldsymbol{p}_t(h_{t-1},\,\cdot\,) \in \qquad\qquad (16) \\ &\operatorname*{argmin}_{\pi:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[-T\sum_{a\in\mathcal{A}}\Big(r(a,\boldsymbol{X}) + \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{X})\Big)\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X}) - \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{a\in\mathcal{A}}\beta_{\boldsymbol{x},a}^{\text{p-pos},\star}\,\pi_a(\boldsymbol{x})\right] \\ = &\operatorname*{argmax}_{\pi:\mathcal{X}\to\mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[T\sum_{a\in\mathcal{A}}\Big(r(a,\boldsymbol{X}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{X})\Big)\,U_{t-1}(a,\boldsymbol{X})\,\pi_a(\boldsymbol{X}) + \sum_{\boldsymbol{x}\in\mathcal{X}}\sum_{a\in\mathcal{A}}\beta_{\boldsymbol{x},a}^{\text{p-pos},\star}\,\pi_a(\boldsymbol{x})\right]. \end{aligned}$$

We further simplify this alternative definition by showing that the sums of the $\beta_{\boldsymbol{x},a}^{\text{p-pos}}\,\pi_a(\boldsymbol{x})$ may be omitted. We do so by exploiting the KKT conditions.

KKT conditions: statement. The Karush–Kuhn–Tucker (KKT) conditions (Boyd and Vandenberghe [2004, Section 5.5.3]) for the primal optimal $\boldsymbol{p}_t(h_{t-1},\,\cdot\,) : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ and the dual optimal $\boldsymbol{\beta}_t^{\text{budg},\star} \geqslant \mathbf{0}$, $\boldsymbol{\beta}_t^{\text{p-sum},\star} \geqslant \mathbf{0}$, and $\boldsymbol{\beta}_t^{\text{p-pos},\star} \geqslant \mathbf{0}$ imply the following conditions: first, complementary slackness, which reads

$$\forall a \in \mathcal{A},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \beta_{t,\boldsymbol{x},a}^{\text{p-pos},\star}\; p_{t,a}(h_{t-1},\boldsymbol{x}) = 0 \tag{17}$$

and

$$\big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}}\boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] = \frac{B_T}{T}\,\big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}}\mathbf{1}\,; \tag{18}$$

second, a stationary condition, based on the fact that the gradient of the Lagrangian function (14) with respect to the $\pi_a(\boldsymbol{x})$ vanishes. Denoting by $\nu(\boldsymbol{x})$ the probability mass put by ν on $\boldsymbol{x}$, we have:

$$\begin{aligned}\forall a \in \mathcal{A},\ \ \forall \boldsymbol{x} \in \mathcal{X}, \qquad & T\, r(a, \boldsymbol{x})\, U_{t-1}(a, \boldsymbol{x})\, \nu(\boldsymbol{x}) \\ & = \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \Big(T\, \boldsymbol{c}(a, \boldsymbol{x})\, U_{t-1}(a, \boldsymbol{x})\, \nu(\boldsymbol{x}) \Big) + \beta_{t,\boldsymbol{x}}^{\text{p-sum},\star} - \beta_{t,\boldsymbol{x},a}^{\text{p-pos},\star}\,. \end{aligned} \tag{19}$$

KKT conditions: first consequence—final characterization of $\boldsymbol{p}_t(h_{t-1}, \cdot)$. For now, we only exploit (17): this equality and the fact that $\beta_{t,\boldsymbol{x},a}^{\text{p-pos},\star}\, \pi_a(\boldsymbol{x})$ is always non-negative show that, as announced, the characterization (16) may be further simplified into

$$\begin{aligned}\boldsymbol{p}_t(h_{t-1}, \cdot) &\in \operatorname*{argmax}_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[T \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big)\, U_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right] \\ &= \operatorname*{argmax}_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big)\, U_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right]. \end{aligned} \tag{20}$$

In the characterization (20), as we got rid of the cross terms, the maximization may be carried out in a $\mathcal{X}$–pointwise manner, i.e., by separately computing each probability distribution $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) \in \mathcal{P}(\mathcal{A})$. More formally, for each $\boldsymbol{x} \in \mathcal{X}$,

$$\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)\, U_{t-1}(a, \boldsymbol{x})\, q_a\,. \tag{21}$$

In this characterization, $r(a, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x})$ appears as a penalized gain in case a client with characteristics $\boldsymbol{x}$ converts, and $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x})$ is obtained by combining an upper-confidence-bound estimation $U_{t-1}(a, \boldsymbol{x})$ of the conversion rate $P(a, \boldsymbol{x})$ with this penalized gain.

Denote by $(z)_+ = \max\{z,\, 0\}$ the non-negative part of $z \in \mathbb{R}$ and fix $\boldsymbol{x} \in \mathcal{X}$. Given that the no-op action a_{null} is such that $r(a_{\text{null}}, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a_{\text{null}}, \boldsymbol{x}) = 0$ and $U_{t-1}(a, \boldsymbol{x}) \geqslant 0$ for all $a \in \mathcal{A}$, in view of the objective, any distribution $\boldsymbol{q}$ should move the probability mass q_a on an action $a \in \mathcal{A}$ with $r(a, \boldsymbol{x}) - (\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) < 0$ to a_{null}. As a consequence,

$$\begin{aligned}\max_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} & \Big(r(a, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)\, U_{t-1}(a, \boldsymbol{x})\, q_a(\boldsymbol{x}) \\ & = \max_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ U_{t-1}(a, \boldsymbol{x})\, q_a(\boldsymbol{x}) \end{aligned} \tag{22}$$

and

$$\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ U_{t-1}(a, \boldsymbol{x})\, q_a\,. \tag{23}$$

KKT conditions: a second consequence. We first exploit the stationary condition (19). Multiplying both sides of this equality by $p_{t,a}(h_{t-1}, \boldsymbol{x})$, summing over $\boldsymbol{x} \in \mathcal{X}$ and $a \in A$, we obtain an equality between expectations with respect to $\boldsymbol{X} \sim \nu$:

$$\begin{aligned} & T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, p_{t,a}(h_{t-1}, \boldsymbol{X}) \right] \\ & = \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \left(T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, p_{t,a}(h_{t-1}, \boldsymbol{X}) \right] \right) \\ & \quad + \sum_{\boldsymbol{x} \in \mathcal{X}} \beta_{t,\boldsymbol{x}}^{\text{p-sum},\star} \underbrace{\sum_{a \in \mathcal{A}} p_{t,a}(h_{t-1}, \boldsymbol{x})}_{=1} - \sum_{\boldsymbol{x} \in \mathcal{X}} \sum_{a \in \mathcal{A}} \underbrace{\beta_{t,\boldsymbol{x},a}^{\text{p-pos},\star}\, p_{t,a}(h_{t-1}, \boldsymbol{x})}_{=0}, \end{aligned} \tag{24}$$

where the equality to 0 indicated in the right-hand side correspond to (17). We now substitute (18) into (24) and obtain

$$T\,\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] = B_T(\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathsf{T}}\,\mathbf{1} + \sum_{\boldsymbol{x}\in\mathcal{X}} \beta_{t,\boldsymbol{x}}^{\text{p-sum},\star}\,.$$

The left-hand side equals $\textsc{opt}(\nu, U_{t-1}, B_T)$ since $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$ is the solution of the primal problem. Thus, the equality above entails, by $\boldsymbol{\beta}_t^{\text{p-sum},\star} \geqslant \mathbf{0}$, that

$$\textsc{opt}(\nu, U_{t-1}, B_T) \geqslant B_T(\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathsf{T}}\,\mathbf{1}\,. \tag{25}$$

B.3 Third Step: Various High-Probability Controls

We prove below that on the intersection between the event $\mathcal{E}_{\text{prob},\delta}$ of Lemma 1 and another event $\mathcal{E}_{\text{HAz},\delta}$, also of probability at least $1-\delta$, we have simultaneously that for all rounds $t \geqslant 2$, the policy $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$ is obtained by Phase 2, i.e., by solving $\textsc{opt}(\nu, U_{t-1}, B_T)$, and that

$$\begin{aligned} &\sum_{t=1}^{T} r(a_t,\boldsymbol{x}_t)\,y_t \geqslant \sum_{t=2}^{T} \frac{\textsc{opt}(\nu, U_{t-1}, B_T)}{T} - \sqrt{2T\ln\frac{4}{\delta}} - E_T \qquad (26)\\ \text{and}\qquad &\sum_{t=1}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\,y_t \leqslant \left(B_T + 1 + \sqrt{2T\ln\frac{4d}{\delta}}\right)\mathbf{1} = (B-1)\mathbf{1}\,, \end{aligned}$$

where the bound E_T was defined in (12) and where we used the definition of B_T, namely,

$$B_T = B - 2 - \sqrt{2T\ln\frac{4d}{\delta}}\,. \tag{27}$$

This definition requires that

$$B > c\left(2 + \sqrt{2T\ln\frac{4d}{\delta}}\right) \tag{28}$$

for $c = 1$, but we will rather assume that the inequality holds with $c = 2$.

Applications of the Hoeffding-Azuma inequality to handle the conversions y_t. We recall that we defined h_0 as the empty vector and $h_t = (\boldsymbol{x}_s, a_s,\, y_s)_{s\leqslant t}$ for $t \geqslant 1$. We introduce the filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geqslant 0}$, with $\mathcal{F}_0 = \sigma(a_1, \boldsymbol{x}_1)$ and for $t \geqslant 1$,

$$\mathcal{F}_t = \sigma\big(h_t,\, a_{t+1},\, \boldsymbol{x}_{t+1}\big)\,.$$

Then, for all $t \geqslant 1$, the variables $r(a_t,\boldsymbol{x}_t)\,y_t$ and $\boldsymbol{c}(a_t,\boldsymbol{x}_t)\,y_t$ are $\mathcal{F}_t$–measurable, with conditional expectations with respect to $\mathcal{F}_{t-1}$ equal to

$$\begin{aligned} &\mathbb{E}\big[r(a_t,\boldsymbol{x}_t)\,y_t \,\big|\, \mathcal{F}_{t-1}\big] = r(a_t,\boldsymbol{x}_t)\,P(a_t,\boldsymbol{x}_t)\\ \text{and}\qquad &\mathbb{E}\big[\boldsymbol{c}(a_t,\boldsymbol{x}_t)\,y_t \,\big|\, \mathcal{F}_{t-1}\big] = \boldsymbol{c}(a_t,\boldsymbol{x}_t)\,P(a_t,\boldsymbol{x}_t)\,. \end{aligned}$$

Indeed, the conditioning by $\mathcal{F}_{t-1}$ fixes a_t and $\boldsymbol{x}_t$, but not y_t, and exactly means, when $a_t \neq a_{\text{null}}$, integrating over $y_t \sim P(a_t,\boldsymbol{x}_t)$. When $a_t = a_{\text{null}}$, all equalities above remain valid thanks to the abuse of notation discussed in Section 2.1. Given that r takes values in $[0,1]$ and $\boldsymbol{c}$ in $[0,1]^d$, we may apply $d+1$ times the Hoeffding-Azuma inequality (once for r and each component of c) together with a union bound: there exist two events $\mathcal{E}_{r,P,\delta}$ and $\mathcal{E}_{\boldsymbol{c},P,\delta}$, each of probability at least $1-\delta/4$, such that

$$\begin{aligned} &\text{on } \mathcal{E}_{r,P,\delta}, \qquad \sum_{t=1}^{T} r(a_t,\boldsymbol{x}_t)\,y_t \geqslant \sum_{t=1}^{T} r(a_t,\boldsymbol{x}_t)\,P(a_t,\boldsymbol{x}_t) - \sqrt{\frac{T}{2}\ln\frac{4}{\delta}}\\ \text{and}\qquad &\text{on } \mathcal{E}_{\boldsymbol{c},P,\delta}, \qquad \sum_{t=1}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\,y_t \leqslant \sqrt{\frac{T}{2}\ln\frac{4d}{\delta}}\;\mathbf{1} + \sum_{t=1}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\,P(a_t,\boldsymbol{x}_t)\,. \end{aligned}$$

Applications of the Hoeffding-Azuma inequality to use the properties of the $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$. For this sub-step, we rather condition directly by h_{t-1} (instead of $\mathcal{F}_{t-1}$ as in the previous step), where $t \geqslant 2$. Conditioning by h_{t-1} amounts to integrating first over $a_t \sim \boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$ and then over $\boldsymbol{x}_t \sim \nu$: this is because of the definition of the random draw of a_t according to $\boldsymbol{p}_t(h_{t-1}, \boldsymbol{x}_t)$ independently from everything else, and because $\boldsymbol{x}_t$ is drawn independent from the past according to ν. More precisely, for each $t \geqslant 2$, given that U_{t-1} is h_{t-1}–measurable, we thus have the following equalities:

$$\begin{aligned}
\mathbb{E}\big[r(a_t,\boldsymbol{x}_t)\,U_{t-1}(a_t,\boldsymbol{x}_t)\,\big|\,h_{t-1}\big] &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{x}_t)\,U_{t-1}(a,\boldsymbol{x}_t)\,p_{t,a}(h_{t-1},\boldsymbol{x}_t)\,\middle|\,h_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] \\
\text{and}\qquad \mathbb{E}\big[\boldsymbol{c}(a_t,\boldsymbol{x}_t)\,U_{t-1}(a_t,\boldsymbol{x}_t)\,\big|\,h_{t-1}\big] &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{x}_t)\,U_{t-1}(a,\boldsymbol{x}_t)\,p_{t,a}(h_{t-1},\boldsymbol{x}_t)\,\middle|\,h_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right],
\end{aligned}$$

where we recall that $\mathbb{E}_{\boldsymbol{X}\sim\nu}$ denotes an integration solely over $\boldsymbol{X}\sim\nu$.

By definition of $\boldsymbol{p}_t(h_{t-1},\,\cdot\,)$, whenever the adaptive policy reaches Phase 2 at a given round $t \geqslant 2$:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] = \frac{\textsc{opt}(\nu,U_{t-1},B_T)}{T} \\
\text{and}\qquad &\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] \leqslant \frac{B_T}{T}\mathbf{1}\,.
\end{aligned}$$

Otherwise, the adaptive policy is stuck in Phase 0, because at least one cost constraint is larger than $B-1$; in this case, $\boldsymbol{p}_t(h_{t-1},\boldsymbol{x}) = \delta_{a_{\text{null}}}$ for all $\boldsymbol{x}\in\mathcal{X}$ and both expectations above are null. We may summarize all cases by introducing the indicator function that all cost constraints are smaller than $B-1$,

$$\mathbb{1}_{\{\boldsymbol{C}_{t-1}\leqslant(B-1)\mathbf{1}\}}\,, \qquad \text{where} \qquad \boldsymbol{C}_{t-1} = \sum_{s=1}^{t-1}\boldsymbol{c}(a_s,\boldsymbol{x}_s)\,y_s\,,$$

and stating that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] \geqslant \mathbb{1}_{\{\boldsymbol{C}_{t-1}\leqslant(B-1)\mathbf{1}\}}\frac{\textsc{opt}(\nu,U_{t-1},B_T)}{T} \\
\text{and}\qquad &\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,p_{t,a}(h_{t-1},\boldsymbol{X})\right] \leqslant \frac{B_T}{T}\mathbf{1}\,.
\end{aligned}$$

Therefore, a second series of applications of the Hoeffding-Azuma inequality (at times $2 \leqslant t \leqslant T$, i.e., excluding the first round) entails the existence of two events $\mathcal{E}_{r,U,\delta}$ and $\mathcal{E}_{\boldsymbol{c},U,\delta}$, each of probability at least $1-\delta/4$, such that

$$\begin{aligned}
&\text{on } \mathcal{E}_{r,U,\delta}, \qquad \sum_{t=2}^{T} r(a_t,\boldsymbol{x}_t)\,U_{t-1}(a_t,\boldsymbol{x}_t) \geqslant \sum_{t=2}^{T}\mathbb{1}_{\{\boldsymbol{C}_{t-1}\leqslant(B-1)\mathbf{1}\}}\frac{\textsc{opt}(\nu,U_{t-1},B_T)}{T} - \sqrt{\frac{T}{2}\ln\frac{4}{\delta}} \\
&\text{and on } \mathcal{E}_{\boldsymbol{c},U,\delta}, \qquad \sum_{t=2}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\,U_{t-1}(a_t,\boldsymbol{x}_t) \leqslant \left(B_T + \sqrt{\frac{T}{2}\ln\frac{4d}{\delta}}\right)\mathbf{1}\,.
\end{aligned}$$

Appeal to results of Appendix B.1. The inequalities (10) and (11) state that on the event $\mathcal{E}_{\text{prob},\delta}$ of Lemma 1, we have

$$\forall t \geqslant 1,\ \ \forall a \in \mathcal{A}\setminus\{a_{\text{null}}\},\ \ \forall \boldsymbol{x}\in\mathcal{X}, \qquad P(a,\boldsymbol{x}) \leqslant U_t(a,\boldsymbol{x}) \leqslant P(a,\boldsymbol{x}) + 2\varepsilon_t(a,\boldsymbol{x})\,.$$

Thus, on $\mathcal{E}_{\text{prob},\delta}$,

$$\sum_{t=1}^{T} r(a_t, \boldsymbol{x}_t)\, P(a_t, \boldsymbol{x}_t) \geqslant \sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t) - 2\sum_{t=2}^{T} \varepsilon_{t-1}(a_t, \boldsymbol{x}_t) \mathbb{1}_{\{a_t \neq a_{\text{null}}\}}$$
$$\geqslant \sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t) - E_T$$

and $$\sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, P(a_t, \boldsymbol{x}_t) \leqslant \mathbf{1} + \sum_{t=2}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t)\,,$$

where we recall that the bound E_T was defined in (12).

Conclusion of this step. We define

$$\mathcal{E}_{\text{HAz},\delta} = \mathcal{E}_{r,P,\delta} \cap \mathcal{E}_{\boldsymbol{c},P,\delta} \cap \mathcal{E}_{r,U,\delta} \cap \mathcal{E}_{\boldsymbol{c},U,\delta}\,,$$

which is an event of probability at least $1-\delta$. On the intersection of $\mathcal{E}_{\text{HAz},\delta}$ and $\mathcal{E}_{\text{prob},\delta}$, by collecting all bounds together, we have

$$\sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t \leqslant \left(B_T + 1 + \sqrt{2T \ln \frac{4d}{\delta}} \right) \mathbf{1} = (B-1)\mathbf{1}\,.$$

This shows that on the intersection of $\mathcal{E}_{\text{HAz},\delta}$ and $\mathcal{E}_{\text{prob},\delta}$, the indicator functions $\mathbb{1}_{\{\boldsymbol{C}_{t-1} \leqslant (B-1)\mathbf{1}\}}$ all equal 1, and that the policies $\boldsymbol{p}_t(h_{t-1}, \cdot\,)$ are always obtained by solving the optimization problems of Phase 2. We conclude this step by collecting the obtained bounds for rewards and by legitimately replacing the indicator functions therein by 1.

B.4 Fourth Step: Conclusion

In this step, we merely combine the bounds exhibited in the first three steps to obtain the closed-form expression of the regret bound. We then propose suitable orders of magnitude for the parameters.

Collecting all bounds to get a closed-form regret bound. By considering $\boldsymbol{q} = \pi^\star(\boldsymbol{x})$ for each $\boldsymbol{x} \in \mathcal{X}$ in (23), where we recall that $\pi^\star$ is the optimal static policy, and by the equality (22), we note that, for all $t \geqslant 2$,

$$\forall \boldsymbol{x} \in \mathcal{X}, \qquad \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)\, U_{t-1}(a, \boldsymbol{x})\, p_{t,a}(h_{t-1}, \boldsymbol{x})$$
$$\geqslant \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ U_{t-1}(a, \boldsymbol{x})\, \pi_a^\star(\boldsymbol{x})\,.$$

On the event $\mathcal{E}_{\text{prob},\delta}$ of Lemma 1, the inequality (10) states that $U_{t-1}(a, \boldsymbol{x}) \geqslant P(a, \boldsymbol{x})$, which we may substitute in the inequality above (thanks to the non-negative part in the right-hand side) to get

$$\forall \boldsymbol{x} \in \mathcal{X}, \qquad \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)\, U_{t-1}(a, \boldsymbol{x})\, p_{t,a}(h_{t-1}, \boldsymbol{x})$$
$$\geqslant \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ P(a, \boldsymbol{x})\, \pi_a^\star(\boldsymbol{x})$$
$$\geqslant \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\boldsymbol{\beta}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)\, P(a, \boldsymbol{x})\, \pi_a^\star(\boldsymbol{x})\,.$$

We replace the individual $\boldsymbol{x}$ by a random variable $\boldsymbol{X} \sim \nu$ and integrate over $\boldsymbol{X}$: on $\mathcal{E}_{\text{prob},\delta}$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, p_{t,a}(h_{t-1},\boldsymbol{X})\right] \qquad (29)\\
&\qquad - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, p_{t,a}(h_{t-1},\boldsymbol{X})\right]\\
\geqslant\ & \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\, P(a,\boldsymbol{X})\, \pi_a^\star(\boldsymbol{X})\right]}_{=\mathrm{OPT}(\nu,P,B)/T} - \left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \underbrace{\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\, P(a,\boldsymbol{X})\, \pi_a^\star(\boldsymbol{X})\right]}_{\leqslant (B/T)\mathbf{1}}.
\end{aligned}$$

The equality to $\mathrm{OPT}(\nu,P,B)/T$ and the inequality $\leqslant (B/T)\mathbf{1}$ above come from the very definition of $\pi^\star$ as the static policy solving $\mathrm{OPT}(\nu,P,B)$. Similarly, Appendix B.3 shows that on the event $\mathcal{E}_{\text{HAz},\delta}$, for all $2 \leqslant t \leqslant T$, the policies $p_{t,a}(h_{t-1},\,\cdot\,)$ are obtained by solving $\mathrm{OPT}(\nu, U_{t-1}, B_T)$, so that, by definition,

$$\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, p_{t,a}(h_{t-1},\boldsymbol{X})\right] = \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T}\,.$$

Substituting this equality and the KKT condition (18) into (29) and rearranging, we see that we proved so far that on the intersection of $\mathcal{E}_{\text{prob},\delta}$ and $\mathcal{E}_{\text{HAz},\delta}$,

$$\forall\, 2 \leqslant t \leqslant T, \qquad \mathrm{OPT}(\nu,P,B) - \mathrm{OPT}(\nu,U_{t-1},B_T) \leqslant (B - B_T)\left(\boldsymbol{\beta}_t^{\text{budg},\star}\right)^{\mathrm{T}} \mathbf{1}\,.$$

We may also substitute (25), i.e., $(\boldsymbol{\beta}_t^{\text{budg},\star})^{\mathrm{T}}\,\mathbf{1} \leqslant \mathrm{OPT}(\nu,U_{t-1},B_T)/B_T$ and finally get that on the intersection of $\mathcal{E}_{\text{prob},\delta}$ and $\mathcal{E}_{\text{HAz},\delta}$,

$$\forall\, 2 \leqslant t \leqslant T, \qquad \mathrm{OPT}(\nu,P,B) - \mathrm{OPT}(\nu,U_{t-1},B_T) \leqslant \left(\frac{B}{B_T} - 1\right)\mathrm{OPT}(\nu,U_{t-1},B_T)\,.$$

Summing over $2 \leqslant t \leqslant T$ and using that $\mathrm{OPT}(\nu,P,B)/T \leqslant 1$ by definition and by the fact that r takes values in $[0,1]$, we obtain

$$\begin{aligned}
\mathrm{OPT}(\nu,P,B) - \sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T} &\leqslant 1 + \sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,P,B) - \mathrm{OPT}(\nu,U_{t-1},B_T)}{T}\\
&\leqslant 1 + \left(\frac{B}{B_T} - 1\right)\sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T}\,. \qquad (30)
\end{aligned}$$

Distinguishing the cases

$$\mathrm{OPT}(\nu,P,B) - \sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T} \leqslant 0 \quad \text{ and } \quad \mathrm{OPT}(\nu,P,B) - \sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T} \geqslant 0\,,$$

we see, based on (30), that in both cases

$$\mathrm{OPT}(\nu,P,B) - \sum_{t=2}^{T} \frac{\mathrm{OPT}(\nu,U_{t-1},B_T)}{T} \leqslant 1 + \left(\frac{B}{B_T} - 1\right)\mathrm{OPT}(\nu,P,B)\,.$$

We finally substitute (26) and have proved, as claimed, that on the intersection of $\mathcal{E}_{\text{prob},\delta}$ and $\mathcal{E}_{\text{HAz},\delta}$,

$$\mathrm{OPT}(\nu,P,B) - \sum_{t=1}^{T} r(a_t,\boldsymbol{x}_t)\, y_t \leqslant 1 + \left(\frac{B}{B_T} - 1\right)\mathrm{OPT}(\nu,P,B) + E_T + \sqrt{2T\ln\frac{4}{\delta}}\,, \qquad (31)$$

where we recall that E_T was defined in (12).

Improving readability and setting λ. As indicated, we require (28) with $c = 2$. By the definition (27) of B_T and the inequality $1/(1-u) \leqslant 1 + 2u$ for $u \in (0, 1/2)$, we have

$$\frac{B}{B_T} - 1 \leqslant \frac{2\big(2 + \sqrt{2T\ln(4d/\delta)}\big)}{B}\,, \tag{32}$$

which we may substitute in (31). It only remains to deal with a bound on E_T to conclude the proof of Theorem 1.

We set below a value of λ larger than 1. Recalling that $\kappa \geqslant 4$, we may already bound E_T as

$$E_T \leqslant \gamma_{T,\lambda,\delta}\sqrt{\kappa\big(4\|\Theta\| + 2\big)}\sqrt{2mT\ln\left(1 + \frac{T}{4m}\right)} = 2\gamma_{T,\lambda,\delta}\sqrt{\kappa\big(2\|\Theta\| + 1\big)}\sqrt{mT\ln\left(1 + \frac{T}{4m}\right)}\,, \tag{33}$$

where $\gamma_{T,\lambda,\delta}$, defined in the statement of Lemma 1, may itself be bounded by

$$\gamma_{T,\lambda,\delta} \leqslant \sqrt{\lambda}\big(\|\Theta\| + 1/2\big) + \frac{2}{\sqrt{\lambda}}\ln\left(\frac{2^m}{\delta}\left(1 + \frac{T}{4m}\right)^{m/2}\right).$$

For the sake of simplicity, we set the value of λ by only optimizing the orders of magnitude in m and T of (this upper bound on) $\gamma_{T,\lambda,\delta}$, i.e., by considering

$$\sqrt{\lambda} + \frac{m}{\sqrt{\lambda}}\ln T\,.$$

We take $\lambda = m\ln(1+T/m)$, which is indeed larger than 1 given that $T \geqslant 2m$ and $\ln(1+T/m) \geqslant 1$. We have

$$\frac{2}{\sqrt{\lambda}}\ln\frac{1}{\delta} \leqslant \ln\frac{1}{\delta} \qquad \text{and} \qquad \frac{2}{\sqrt{\lambda}}\ln 2^m \leqslant (2\ln 2)\sqrt{m} \leqslant 2\sqrt{m}\,,$$

as well as

$$\frac{2}{\sqrt{\lambda}}\ln\left(\left(1 + \frac{T}{4m}\right)^{m/2}\right) \leqslant \frac{\sqrt{m}}{\sqrt{\ln(1+T/m)}}\ln\left(1 + \frac{T}{4m}\right) \leqslant \sqrt{m\ln\left(1 + \frac{T}{4m}\right)}\,.$$

All in all,

$$\gamma_{T,\lambda,\delta} \leqslant \big(\sqrt{m} + \|\Theta\| + 1/2\big)\sqrt{\ln\left(1 + \frac{T}{m}\right)} + 2\sqrt{m} + \ln\frac{1}{\delta}\,. \tag{34}$$

Combining (33) and (34), we showed:

$$E_T \leqslant 2\sqrt{\kappa\big(2\|\Theta\| + 1\big)}\sqrt{mT\ln\left(1 + \frac{T}{4m}\right)}\left(\big(\sqrt{m} + \|\Theta\| + 1/2\big)\sqrt{\ln\left(1 + \frac{T}{m}\right)} + 2\sqrt{m} + \ln\frac{1}{\delta}\right). \tag{35}$$

C Adaptation of the Logistic-UCB1 Strategy of Faury et al. [2020]: Proof of Lemma 1

The proof is copied from the proof of Faury et al. [2020, Lemmas 1 and 2], with minor adjustments. We mostly provide it for the sake of self-completeness.

The adjustments are required because the setting of Faury et al. [2020] is slightly different: their action set is a subset $\mathcal{X} \subseteq \mathbb{R}^n$, and when the learner picks an action $\boldsymbol{x}_t \in \mathcal{X}$ at round t, the obtained reward $r_t \in \{0,1\}$ is drawn independently at random according to a Bernoulli distribution with parameter $\eta\big(\boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{\theta}_\star\big)$, where $\boldsymbol{\theta}_\star \in \mathbb{R}^n$ is unknown. The learner then only observes r_t and not what would have been obtained with a different choice of action.

Therefore, the $\boldsymbol{x}_t$ and r_t of Faury et al. [2020] correspond to $\boldsymbol{\varphi}(a_t, \boldsymbol{x}_t)$ and y_t in our setting. The main difference is that while the learner has a full control over the choice of $\boldsymbol{x}_t \in \mathcal{X}$ in the setting of Faury et al. [2020], in our setting, $\boldsymbol{x}_t \in \mathcal{X}$ is drawn by the environment and the learner only picks $a_t \in \mathcal{A}$; the learner therefore does not have a full control over $\boldsymbol{\varphi}(a_t, \boldsymbol{x}_t)$. This is why we carefully check in the present appendix that the results by Faury et al. [2020] (namely, their Lemmas 1 and 2) extend to our setting.

Reminder — A tail inequality for self-normalized martingales. Theorem 1 of Faury et al. [2020] reads as follows. Let $\mathcal{F} = (\mathcal{F}_t)_{t \geqslant 0}$ be a filtration, $(\boldsymbol{U}_t)_{t \geqslant 1}$ an $\mathcal{F}$–adapted stochastic vector process in $\mathbb{R}^m$ such that $\|\boldsymbol{U}_t\| \leqslant 1$ a.s. for all $t \geqslant 1$, and $(\Delta_t)_{t \geqslant 1}$ an $\mathcal{F}$–martingale difference sequence with $|\Delta_t| \leqslant 1$ a.s. for all $t \geqslant 1$; i.e., for all $t \geqslant 1$, the random variable Δ_t is $\mathcal{F}_t$–measurable and

$$\mathbb{E}[\Delta_t \,|\, \mathcal{F}_{t-1}] = 0 \qquad \text{a.s.}$$

Denote $\sigma_t^2 = \mathbb{E}\big[\Delta_t^2 \,|\, \mathcal{F}_{t-1}\big]$, let $\lambda > 0$, and define, for $t \geqslant 1$:

$$S_t = \sum_{s=1}^{t} \Delta_s \boldsymbol{U}_s \qquad \text{and} \qquad M_t = \lambda \, \mathrm{I}_m + \sum_{s=1}^{t} \sigma_s^2 \boldsymbol{U}_s \boldsymbol{U}_s^{\mathrm{T}} \, .$$

Then, for all $\delta \in (0,1)$, with probability at least $1 - \delta$:

$$\forall t \geqslant 1, \qquad \|S_t\|_{M_t^{-1}} \leqslant \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}} \ln\!\left(\frac{2^m \det(M_t)^{1/2} \lambda^{-m/2}}{\delta} \right). \tag{36}$$

The result above is proved by Faury et al. [2020] based on Laplace's method on supermartingales, which is a standard argument to provide confidence bounds on self-normalized sums of conditionally centered random vectors and was previously introduced, in the context of linear contextual bandits, by Abbasi-Yadkori et al. [2011, Theorem 2]; see also the monograph by Lattimore and Szepesvári [2020, Theorem 20.2].

Step 1 — A martingale control. We apply (36) to the following elements. We take as filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geqslant 0}$, with $\mathcal{F}_0 = \sigma(a_1, \boldsymbol{x}_1)$ and for $t \geqslant 1$,

$$\mathcal{F}_t = \sigma\big(h_t,\, a_{t+1},\, \boldsymbol{x}_{t+1}\big) \, ,$$

where we recall $h_t = (\boldsymbol{x}_s, a_s,\, y_s)_{s \leqslant t}$. We set $\boldsymbol{U}_t = \boldsymbol{\varphi}(a_t, \boldsymbol{x}_t)$, which is indeed $\mathcal{F}_t$–measurable and with Euclidean norm smaller than 1 (thanks to the normalization assumed in Section 2.1). Finally, we set, for $t \geqslant 1$,

$$\Delta_t = \begin{cases} 0 & \text{if } a_t = a_{\text{null}}, \\ y_t - \eta\big(\boldsymbol{\varphi}(a_t, \boldsymbol{x}_t)^{\mathrm{T}} \boldsymbol{\theta}_\star\big) & \text{if } a_t \neq a_{\text{null}}, \end{cases}$$

which we rewrite, by the abuses of notation indicated in Section 2.1,

$$\Delta_t = \Big(y_t - \eta\big(\boldsymbol{\varphi}(a_t, \boldsymbol{x}_t)^{\mathrm{T}} \boldsymbol{\theta}_\star\big) \Big) \, \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} = \big(y_t - P(a_t, \boldsymbol{x}_t) \big) \, \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} \, .$$

The conditioning by $\mathcal{F}_{t-1}$ fixes a_t and $\boldsymbol{x}_t$, but not y_t, and exactly means, when $a_t \neq a_{\text{null}}$, integrating over $y_t \sim P(a_t, \boldsymbol{x}_t)$. We therefore have that Δ_t is $\mathcal{F}_t$–measurable, with

$$\mathbb{E}\big[\Delta_t \,|\, \mathcal{F}_{t-1}\big] = 0 \qquad \text{a.s.;}$$

that is, $(\Delta_t)_{t\geqslant 1}$ appears as an $\mathcal{F}$–martingale difference sequence, satisfying the boundedness-by-1 constraint. We may therefore apply the result (36). To do so, we first compute the conditional variances of the Δ_t: for $t \geqslant 1$,

$$\begin{aligned}\mathbb{E}\big[\Delta_t^2 \,|\, \mathcal{F}_{t-1}\big] &= \mathbb{E}\bigg[\Big(y_t - \eta\big(\boldsymbol{\varphi}(a_t,\boldsymbol{x}_t)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\Big)^2 \,\Big|\, \mathcal{F}_{t-1}\bigg]\,\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} \\ &= \eta\big(\boldsymbol{\varphi}(a_t,\boldsymbol{x}_t)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\,\Big(1-\eta\big(\boldsymbol{\varphi}(a_t,\boldsymbol{x}_t)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\Big)\,\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} = \dot\eta\big(\boldsymbol{\varphi}(a_t,\boldsymbol{x}_t)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\,\mathbb{1}_{\{a_t\neq a_{\text{null}}\}}\,,\end{aligned}$$

where we used the fact that $\eta(1-\eta)=\dot\eta$. We rewrite

$$S_t = \sum_{s=1}^t \Delta_t\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s) = \sum_{s=1}^t \Big(y_s - \eta\big(\boldsymbol{\varphi}(a_t,\boldsymbol{x}_t)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\Big)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\mathbb{1}_{\{a_s\neq a_{\text{null}}\}}$$

and note that $M_t = W_t(\boldsymbol{\theta})$ with the definition (5). The control (36) may be rewritten as follows: for all $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\forall t\geqslant 1,\qquad \|S_t\|_{W_t(\boldsymbol{\theta}_\star)^{-1}} \leqslant \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\,\ln\bigg(\frac{2^m \det\big(W_t(\boldsymbol{\theta}_\star)\big)^{1/2}\lambda^{-m/2}}{\delta}\bigg).$$

As $\dot\eta = \eta(1-\eta)\in[0,1/4]$ and as $\|\boldsymbol{\varphi}\|\leqslant 1$, we have, by a standard trace-determinant inequality (see, e.g., Abbasi-Yadkori et al. [2011, Lemma 10]),

$$\begin{aligned}\det\big(W_t(\boldsymbol{\theta}_\star)\big) &\leqslant \bigg(\lambda + \frac{1}{m}\sum_{s=1}^t \dot\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\,\big\|\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\big\|^2\,\mathbb{1}_{\{a_s\neq a_{\text{null}}\}}\bigg)^m \\ &\leqslant \bigg(\lambda + \frac{1}{4m}\sum_{s=1}^t \mathbb{1}_{\{a_s\neq a_{\text{null}}\}}\bigg)^m \leqslant \bigg(\lambda+\frac{t}{4m}\bigg)^m.\end{aligned}$$

Combining the two inequalities, we have proved that for all $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\forall t\geqslant 1,\qquad \|S_t\|_{W_t(\boldsymbol{\theta}_\star)^{-1}} \leqslant \gamma_{t,\lambda,\delta} - \sqrt{\lambda}\|\Theta\|\,, \tag{37}$$

where $\gamma_{t,\lambda,\delta}$ was defined in Lemma 1.

Step 2 — Application of the martingale control. The martingale control (37) is applied as follows. We show below that the definition (3) of $\tilde{\boldsymbol{\theta}}_t$ entails that

$$S_t - \lambda\,\boldsymbol{\theta}_\star = \Psi_t\big(\tilde{\boldsymbol{\theta}}_t\big) - \Psi_t(\boldsymbol{\theta}_\star)\,, \tag{38}$$

where Ψ_t was defined in (4). Taking the $\|\,\cdot\,\|_{W_t(\boldsymbol{\theta}_\star)^{-1}}$ norms of both sides, together with a triangle inequality (keeping in mind the boundedness of Θ) and noting that

$$\|\boldsymbol{\theta}\|_{W_t(\boldsymbol{\theta}_\star)^{-1}} \leqslant \|\boldsymbol{\theta}\|_{(\lambda\,\mathrm{I}_m)^{-1}} = \|\boldsymbol{\theta}\|/\sqrt{\lambda}\,,$$

finally yields that for all $\delta\in(0,1)$, with probability at least $1-\delta$,

$$\forall t\geqslant 1,\qquad \Big\|\Psi_t\big(\tilde{\boldsymbol{\theta}}_t\big) - \Psi_t(\boldsymbol{\theta}_\star)\Big\|_{W_t(\boldsymbol{\theta}_\star)^{-1}} \leqslant \|S_t\|_{W_t(\boldsymbol{\theta}_\star)^{-1}} + \sqrt{\lambda}\|\boldsymbol{\theta}\| \leqslant \gamma_{t,\lambda,\delta}\,. \tag{39}$$

We now show (38). The gradient of the continuously differentiable function

$$\boldsymbol{\theta}\in\mathbb{R}^m \longmapsto \sum_{s=1}^t \mathbb{1}_{\{a_s\neq a_{\text{null}}\}}\bigg(y_s\ln\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\boldsymbol{\theta}\big) + (1-y_s)\ln\Big(1-\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\boldsymbol{\theta}\big)\Big)\bigg) - \frac{\lambda}{2}\|\boldsymbol{\theta}\|$$

vanishes at the point $\tilde{\boldsymbol{\theta}}_t$ where it achieves its maximum, i.e., $\tilde{\boldsymbol{\theta}}_t$ defined in (3) satisfies

$$\begin{aligned}&\sum_{s=1}^t \Big(\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big) - y_s\Big)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\mathbb{1}_{\{a_s\neq a_{\text{null}}\}} \\ &= \sum_{s=1}^t \bigg(y_s\,\frac{\dot\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big)\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)}{\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big)} + (1-y_s)\,\frac{\dot\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big)\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)}{1-\eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big)}\bigg)\mathbb{1}_{\{a_s\neq a_{\text{null}}\}} = \lambda\,\tilde{\boldsymbol{\theta}}_t\,,\end{aligned}$$

where we used $\dot\eta=\eta(1-\eta)$ to get the first equality. In particular,

$$\Psi_t\big(\tilde{\boldsymbol{\theta}}_t\big) = \sum_{s=1}^t \eta\big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\tilde{\boldsymbol{\theta}}_t\big)\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\mathbb{1}_{\{a_s\neq a_{\text{null}}\}} + \lambda\,\tilde{\boldsymbol{\theta}}_t = \sum_{s=1}^t y_s\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\mathbb{1}_{\{a_s\neq a_{\text{null}}\}}\,,$$

hence the stated rewriting (38).

Step 3 — Bound on prediction error. We now proceed to bounding, for all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$ and $\boldsymbol{x}$:

$$\left|P(a,\boldsymbol{x}) - \widehat{P}_t(a,\boldsymbol{x})\right| = \left|\eta\big(\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big) - \eta\big(\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\,\widehat{\boldsymbol{\theta}}_t\big)\right|.$$

As η is $1/4$–Lipschitz (given $\dot{\eta} = \eta(1-\eta) \in [0,1/4]$),

$$\left|\widehat{P}_t(a,\boldsymbol{x}) - P(a,\boldsymbol{x})\right| \leqslant \frac{1}{4}\left|\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\,(\boldsymbol{\theta}_\star - \widehat{\boldsymbol{\theta}}_t)\right|. \tag{40}$$

For two $m \times m$ symmetric definite positive matrices M and M', we write $M \succeq M'$ whenever $\|\boldsymbol{v}\|_M \geqslant \|\boldsymbol{v}\|_{M'}$ for all $\boldsymbol{v} \in \mathbb{R}^m$. This inequality entails $(M')^{-1} \succeq M^{-1}$. We introduce below a symmetric definite positive matrix G_t such that

$$G_t \succeq \kappa^{-1} V_t\,, \qquad G_t \succeq \big(1+2\|\Theta\|\big)^{-1} W_t\big(\widehat{\boldsymbol{\theta}}_t\big)\,, \qquad G_t \succeq \big(1+2\|\Theta\|\big)^{-1} W_t(\boldsymbol{\theta}_\star) \tag{41}$$

and

$$\Psi_t\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi_t\big(\boldsymbol{\theta}_\star\big) = G_t\big(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\big)\,. \tag{42}$$

Based on all these properties, we get, by a Cauchy-Schwarz inequality in the norms $\|\,\cdot\,\|_{G_t}$ and $\|\,\cdot\,\|_{G_t^{-1}}$:

$$\begin{aligned}
&\left|\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}}\,(\boldsymbol{\theta}_\star - \widehat{\boldsymbol{\theta}}_t)\right| \\
&\qquad \leqslant \big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{G_t^{-1}}\,\big\|\boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_t\big\|_{G_t} \\
&\qquad = \big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{G_t^{-1}}\,\Big\|G_t\big(\boldsymbol{\theta}_\star - \tilde{\boldsymbol{\theta}}_t\big)\Big\|_{G_t^{-1}} \\
&\qquad = \big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{G_t^{-1}}\,\Big\|\Psi\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi\big(\boldsymbol{\theta}_\star\big)\Big\|_{G_t^{-1}} \leqslant \sqrt{\kappa}\,\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{V_t^{-1}}\,\Big\|\Psi\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi\big(\boldsymbol{\theta}_\star\big)\Big\|_{G_t^{-1}}\,.
\end{aligned}$$

A triangle inequality for the first inequality, followed by applying (41) for the second inequality, and applying the definition of (4) as a projection for the third inequality, shows that

$$\begin{aligned}
\Big\|\Psi\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi\big(\boldsymbol{\theta}_\star\big)\Big\|_{G_t^{-1}} &\leqslant \Big\|\Psi\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi\big(\tilde{\boldsymbol{\theta}}_t\big)\Big\|_{G_t^{-1}} + \Big\|\Psi\big(\boldsymbol{\theta}_\star\big) - \Psi\big(\tilde{\boldsymbol{\theta}}_t\big)\Big\|_{G_t^{-1}} \\
&\leqslant \sqrt{1+2\|\Theta\|}\left(\Big\|\Psi\big(\widehat{\boldsymbol{\theta}}_t\big) - \Psi\big(\tilde{\boldsymbol{\theta}}_t\big)\Big\|_{W_t(\widehat{\boldsymbol{\theta}}_t)^{-1}} + \Big\|\Psi(\boldsymbol{\theta}_\star) - \Psi\big(\tilde{\boldsymbol{\theta}}_t\big)\Big\|_{W_t(\boldsymbol{\theta}_\star)^{-1}}\right) \\
&\leqslant 2\sqrt{1+2\|\Theta\|}\,\Big\|\Psi(\boldsymbol{\theta}_\star) - \Psi\big(\tilde{\boldsymbol{\theta}}_t\big)\Big\|_{W_t(\boldsymbol{\theta}_\star)^{-1}}\,.
\end{aligned}$$

Substituting the martingale control (39) and collecting all bounds together finally yields

$$\left|\widehat{P}_t(a,\boldsymbol{x}) - P(a,\boldsymbol{x})\right| \leqslant \frac{\sqrt{\kappa}}{2}\,\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{V_t^{-1}}\,\sqrt{1+2\|\Theta\|}\;\gamma_{t,\lambda,\delta}\,,$$

as desired.

Note in particular that (39) holds for all $t \geqslant 1$ with probability $1-\delta$, and that we took care of the dependencies on a and $\boldsymbol{x}$ through the $\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{V_t^{-1}}$ term. This explains why the result of Lemma 1 holds with probability $1-\delta$ for all $t \geqslant 1$, all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$ and all $\boldsymbol{x} \in \mathcal{X}$.

Step 4 — Construction of the matrix G_t. It only remains to show that there exists a symmetric definite positive matrix G_t such that (41) and (42) hold. We define

$$\begin{aligned}
G_t &= \int_{[0,1]} W_t\big(v\,\widehat{\boldsymbol{\theta}}_t + (1-v)\boldsymbol{\theta}_\star\big)\,\mathrm{d}v \\
&= \lambda\,\mathrm{I}_m + \sum_{s=1}^{t}\left(\int_{[0,1]} \dot{\eta}\big(v\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\widehat{\boldsymbol{\theta}}_t + (1-v)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big)\,\mathrm{d}v\right)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\mathbb{1}_{\{a_s \neq a_{\text{null}}\}}\,.
\end{aligned}$$

The thus defined matrix G_t is indeed symmetric definite positive matrix. By definition of κ and the fact that Θ is convex, we have, for all $v \in [0,1]$,

$$\dot{\eta}\big(v\,\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\widehat{\boldsymbol{\theta}}_t + (1-v)\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\boldsymbol{\theta}_\star\big) = \dot{\eta}\Big(\boldsymbol{\varphi}(a_s,\boldsymbol{x}_s)^{\mathrm{T}}\,\big(v\widehat{\boldsymbol{\theta}}_t + (1-v)\boldsymbol{\theta}_\star\big)\Big) \geqslant \kappa^{-1}\,,$$

which also immediately entails the first inequality of (41). To prove (42), we introduce, for $s \geqslant 1$ such that $a_s \neq a_{\text{null}}$, the continuously differentiable function

$$f_{s,t} : v \in [0,1] \longmapsto f_{s,t}(v) = \eta\big(v\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \widehat{\boldsymbol{\theta}}_t + (1-v) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)\,,$$

with derivative

$$\dot{f}_{s,t}(v) = \dot{\eta}\big(v\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \widehat{\boldsymbol{\theta}}_t + (1-v) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)\ \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} \big(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\big)\,,$$

and we have

$$\eta\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \widehat{\boldsymbol{\theta}}_t\big) - \eta\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big) = f_{s,t}(1) - f_{s,t}(0) = \int_{[0,1]} \dot{f}_{s,t}(v)\, \mathrm{d}v\,,$$

These facts, combined with the definition of G_t, immediately entail (42).

It only remains to prove the third inequality of (41), namely,

$$G_t \succeq \big(1 + 2\|\Theta\|\big)^{-1}\, W_t(\boldsymbol{\theta}_\star)\,,$$

as the second one is obtained by symmetry from there, by exchanging the roles of $\boldsymbol{\theta}_\star$ and $\widehat{\boldsymbol{\theta}}_t$. To do so, we rely on the following inequality: for all $z_1, z_2 \in \mathbb{R}$,

$$\int_{[0,1]} \dot{\eta}\big(z_1 + v(z_2 - z_1)\big)\, \mathrm{d}v \geqslant \frac{\dot{\eta}(z_1)}{1 + |z_1 - z_2|}\,. \tag{43}$$

This inequality is proved, in the case $z_1 \neq z_2$, by noting that the second-order derivative of η equals

$$\ddot{\eta}(x) = \frac{\mathrm{e}^{-x} - 1}{1 + \mathrm{e}^{x}}\ \dot{\eta}(x)\,, \qquad \text{where} \qquad \frac{\mathrm{e}^{-x} - 1}{1 + \mathrm{e}^{x}} \in [-1, 1]\,,$$

so that for all $z, z' \in \mathbb{R}$,

$$\ln \dot{\eta}(z') - \ln \dot{\eta}(z) = \int_z^{z'} \frac{\ddot{\eta}(\tau)}{\dot{\eta}(\tau)}\, \mathrm{d}\tau \geqslant -|z' - z|\,, \qquad \text{i.e.,} \qquad \dot{\eta}(z') \geqslant \dot{\eta}(z)\, \mathrm{e}^{-|z'-z|}\,.$$

Therefore,

$$\int_{[0,1]} \dot{\eta}\big(z_1 + v(z_2 - z_1)\big)\, \mathrm{d}v \geqslant \int_{[0,1]} \dot{\eta}(z_1)\, \mathrm{e}^{-v|z_1 - z_2|}\, \mathrm{d}v = \dot{\eta}(z_1)\, \frac{1 - \mathrm{e}^{-|z_1 - z_2|}}{|z_1 - z_2|} \geqslant \frac{\dot{\eta}(z_1)}{1 + |z_1 - z_2|}\,,$$

where we applied the inequality $(1 - \mathrm{e}^{-x})/x \geqslant 1/(1+x)$, which holds for all $x \geqslant 0$.

We go back to proving the third inequality of (41). With (43) for the first inequality, followed by an application of the Cauchy-Schwarz inequality for the second inequality, and the fact that $\boldsymbol{\theta}_\star$, $\widehat{\boldsymbol{\theta}}_t \in \Theta$ have Euclidean norms smaller than $\|\Theta\|$, together with $\|\boldsymbol{\varphi}\| \leqslant 1$, we have, for all $s \geqslant 1$ with $a_s \neq a_{\text{null}}$,

$$\begin{aligned}
&\int_{[0,1]} \dot{\eta}\big(v\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \widehat{\boldsymbol{\theta}}_t + (1-v) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)\, \mathrm{d}v \\
&\qquad\qquad \geqslant \dot{\eta}\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big) \Big(1 + \big|\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, (\boldsymbol{\theta}_\star - \widehat{\boldsymbol{\theta}}_t)\, \big|\Big)^{-1} \\
&\qquad\qquad \geqslant \dot{\eta}\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big) \Big(1 + \big\|\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, \big\|\ \big\|\, \boldsymbol{\theta}_\star - \widehat{\boldsymbol{\theta}}_t\, \big\|\Big)^{-1} \\
&\qquad\qquad \geqslant \dot{\eta}\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big) \big(1 + 2\|\Theta\|\big)^{-1}\,.
\end{aligned}$$

As $\big(1 + 2\ \|\Theta\|\big)^{-1} \leqslant 1$, we can then conclude, from the definition of G_t, that

$$\begin{aligned}
G_t &\succeq (1 + 2\ \|\Theta\|)^{-1} \left(\lambda\, \mathrm{I}_m + \sum_{s=1}^{t} \dot{\eta}\big(\boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \boldsymbol{\theta}_\star\big)\, \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}}\, \mathbb{1}_{\{a_s \neq a_{\text{null}}\}}\right) \\
&= (1 + 2\ \|\Theta\|)^{-1}\, W_t(\boldsymbol{\theta}_\star)\,,
\end{aligned}$$

as announced. This concludes the proof.

D Proof of the Regret Bound in Case of an Unknown Distribution ν: Proof of Theorem 2

We rather explain the differences to and modifications with respect to the proof of Appendix B.

To best do so, we use $\widehat{\ }$superscripts to index various quantities defined based on the estimations $\widehat{\nu}_t$, even though these quantities are not themselves estimators. In particular, the budget parameter is denoted by $\widehat{B}_T$; we refer to the policies computed at rounds $t \geqslant 2$ by $\widehat{\boldsymbol{p}}_t(h_{t-1}, \cdot)$, and by $\widehat{\mathcal{E}}_\delta$ the event of Lemma 1 for the sampling strategy pulling actions $a_t \sim \widehat{\boldsymbol{p}}_t(\boldsymbol{x}_t)$, which is exactly the strategy that we are analyzing here; and finally, the optimal dual variables linked to the budget are denoted by $\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}$.

D.1 Key New Building Block: Uniform Deviation Inequality

Throughout this appendix, we will need to relate quantities defined based on $\widehat{\nu}_t$ to the target quantities defined based on ν. All these quantities will be of the form: for $1 \leqslant t \leqslant T$ and for various functions $f : \mathcal{X} \to [0,1]$,

$$\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\big[f(\boldsymbol{X})\big] \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\big[f(\boldsymbol{X})\big]\,.$$

A simple (but probably slightly suboptimal) way to do so is to apply $T|\mathcal{X}|$ times the Hoeffding-Azuma inequality together with a union bound. We get that on an event $\widehat{\mathcal{E}}_{\text{unif},\delta}$ of probability at least $1-\delta$,

$$\forall\, 1 \leqslant t \leqslant T, \quad \forall \boldsymbol{x} \in \mathcal{X}, \qquad \big|\widehat{\nu}_t(\boldsymbol{x}) - \nu(\boldsymbol{x})\big| \leqslant \sqrt{\frac{1}{2t}\ln\frac{2T|\mathcal{X}|}{\delta}}\,.$$

In particular, with probability at least $1-\delta$, for all functions $f : \mathcal{X} \to [0,1]$ and all $1 \leqslant t \leqslant T$,

$$\Big|\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\big[f(\boldsymbol{X})\big] - \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\big[f(\boldsymbol{X})\big]\Big| \leqslant \sum_{x\in\mathcal{X}}\big|\widehat{\nu}_t(\boldsymbol{x}) - \nu(\boldsymbol{x})\big| \leqslant |\mathcal{X}|\sqrt{\frac{1}{2t}\ln\frac{2T|\mathcal{X}|}{\delta}}\,. \tag{44}$$

We now explain the adaptations required (or not required) for each of the four steps of the proof provided in Appendix B.

D.2 First and Second Steps: No Adaptation Required

These two steps do not require any adaptation; we merely re-state the useful results extracted therein, with the corresponding adapted notation.

The first step (Appendix B.1) held for any sampling strategy. Therefore, the same upper-confidence bonuses (9) and Lemma 1 entail that on an event $\widehat{\mathcal{E}}_\delta$ of probability at least $1-\delta$, for all $t \geqslant 1$, all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$, and all $\boldsymbol{x} \in \mathcal{X}$:

$$P(a,\boldsymbol{x}) \leqslant U_t(a,\boldsymbol{x}) \leqslant P(a,\boldsymbol{x}) + 2\varepsilon_t(a,\boldsymbol{x})\,. \tag{45}$$

The bound (12) also still holds, as it was obtained in a deterministic manner not using any specific feature of the sampling strategy; namely,

$$2\sum_{t=2}^{T}\varepsilon_{t-1}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} \leqslant E_T\,. \tag{46}$$

Similarly, the second step (Appendix B.2) actually yielded general results between the primal and the dual formulations of the OPT problems considered. The equality (18), the characterizations (22) and (23), as well as the inequality (25) may be instantiated with $\widehat{\nu}_t$ (in lieu of ν) and $\widehat{B}_T$ (in lieu of B_T) as follows. For each $t \geqslant 2$ such that the cost constraints of Phase 0 of the adaptive policy are not violated and the optimization problem $\text{OPT}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)$ is to be solved, there exists a vector $\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star} \geqslant \mathbf{0}$ such that first,

$$\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathsf{T}}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}}\boldsymbol{c}(a,\boldsymbol{X})\,U_{t-1}(a,\boldsymbol{X})\,\widehat{p}_{t,a}(\boldsymbol{X})\right] = \frac{\widehat{B}_T}{T}\,\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathsf{T}}\mathbf{1}\,. \tag{47}$$

Second, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\max_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big) U_{t-1}(a, \boldsymbol{x})\, q_a(\boldsymbol{x})$$

$$= \max_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ U_{t-1}(a, \boldsymbol{x})\, q_a(\boldsymbol{x}) \quad (48)$$

and

$$\widehat{\boldsymbol{p}}_t(h_{t-1}, \boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{q} \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ U_{t-1}(a, \boldsymbol{x})\, q_a \,. \quad (49)$$

Third,

$$\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big) \geqslant \widehat{B}_T \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbf{1} \,. \quad (50)$$

D.3 Third Step: Most of the Adaptations Lie Here

We show below that on the intersection between the events $\widehat{\mathcal{E}}_{\text{prob},\delta}$ and $\widehat{\mathcal{E}}_{\text{unif},\delta}$, introduced above, and another event $\widehat{\mathcal{E}}_{\text{HAz},\delta}$, also of probability at least $1 - \delta$, we have that

$$\sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t \leqslant (B - 1)\, \mathbf{1} \,.$$

Therefore, on this intersection of three events, for all rounds $t \geqslant 2$, the policy $\widehat{\boldsymbol{p}}_t(h_{t-1}, \cdot\,)$ is obtained by Phase 2, i.e., by solving $\textsc{opt}\big(\nu_t, U_{t-1}, B_T\big)$.

Similarly, we then prove below that on $\widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\sum_{t=1}^{T} r(a_t, \boldsymbol{x}_t)\, y_t \geqslant \sum_{t=2}^{T} \overbrace{\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right]}^{= \textsc{opt}(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T)/T}$$

$$- \left(E_T + \sqrt{2T \ln \frac{4d}{\delta}} + |\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}} \right) . \quad (51)$$

The inequalities that may be extracted from Appendix B.3. Several applications of the Hoeffding-Azuma inequality, together with a union bound, show that there exist an event $\widehat{\mathcal{E}}_{\text{HAz},\delta}$ of probability at least $1 - \delta$ such that, simultaneously:

$$\sum_{t=1}^{T} r(a_t, \boldsymbol{x}_t)\, y_t \geqslant \sum_{t=1}^{T} r(a_t, \boldsymbol{x}_t)\, P(a_t, \boldsymbol{x}_t) - \sqrt{\frac{T}{2} \ln \frac{4}{\delta}} \,, \quad (52)$$

$$\sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, y_t \leqslant \sum_{t=1}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, P(a_t, \boldsymbol{x}_t) + \sqrt{\frac{T}{2} \ln \frac{4d}{\delta}}\, \mathbf{1} \,, \quad (53)$$

$$\sum_{t=2}^{T} r(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t) \geqslant \sum_{t=2}^{T} \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] - \sqrt{\frac{T}{2} \ln \frac{4}{\delta}} \,, \quad (54)$$

$$\sum_{t=2}^{T} \boldsymbol{c}(a_t, \boldsymbol{x}_t)\, U_{t-1}(a_t, \boldsymbol{x}_t) \leqslant \sum_{t=2}^{T} \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] + \sqrt{\frac{T}{2} \ln \frac{4d}{\delta}}\, \mathbf{1} \,. \quad (55)$$

Dealing with the cost constraints. Now, for each $t \geqslant 2$, the definition of $\widehat{\boldsymbol{p}}_t(h_{t-1}, \cdot\,)$, no matter whether it is provided by Phase 0 or Phase 2 of the adaptive policy, ensures that

$$\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X})\, U_{t-1}(a, \boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] \leqslant \frac{\widehat{B}_T}{T}\, \mathbf{1} \,.$$

By (44) and the fact that the sums at stake below lie in $[0,1]$, we have, on $\widehat{\mathcal{E}}_{\text{unif},\delta}$, that for all $t \geqslant 2$,

$$\mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] \\ \leqslant \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} \boldsymbol{c}(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] + |\mathcal{X}|\sqrt{\frac{1}{2t}\ln\frac{2T|\mathcal{X}|}{\delta}}\,\mathbf{1}\,.$$

Combining the two inequalities above with (53) and (55), as well as with the bounds $P \leqslant U_{t-1}$ of (45), we proved so far that

$$\text{on } \widehat{\mathcal{E}}_{\text{prob},\delta}\cap\widehat{\mathcal{E}}_{\text{HAz},\delta}\cap\widehat{\mathcal{E}}_{\text{unif},\delta}\,, \quad \sum_{t=1}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\, y_t \leqslant \left(\widehat{B}_T + 1 + \sqrt{2T\ln\frac{4d}{\delta}} + \sum_{t=2}^{T}|\mathcal{X}|\sqrt{\frac{1}{2t}\ln\frac{2T|\mathcal{X}|}{\delta}}\right)\mathbf{1}\,.$$

Since $\sum_{t\leqslant T} 1/\sqrt{t} \leqslant 2\sqrt{T}$ and by the definition of $\widehat{B}_T$ (see the statement of Theorem 2), we proved that

$$\text{on } \widehat{\mathcal{E}}_{\text{prob},\delta}\cap\widehat{\mathcal{E}}_{\text{HAz},\delta}\cap\widehat{\mathcal{E}}_{\text{unif},\delta}\,, \qquad \sum_{t=1}^{T} \boldsymbol{c}(a_t,\boldsymbol{x}_t)\, y_t \leqslant (B-1)\mathbf{1}\,,$$

as claimed. Therefore, on $\widehat{\mathcal{E}}_{\text{prob},\delta}\cap\widehat{\mathcal{E}}_{\text{HAz},\delta}\cap\widehat{\mathcal{E}}_{\text{unif},\delta}$, the adaptive policy of Box B of Section 3 never stays in Phase 0 and instead solves, at each round $t \geqslant 2$, the Phase-2 problem $\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)$.

Dealing with the rewards. By (52) and (54), by the bound $U_{t-1} \leqslant P + 2\varepsilon_{t-1}$ of (45) together with the bound E_T of (46), and by the uniform control (44), we have similarly that on $\widehat{\mathcal{E}}_{\text{prob},\delta}\cap\widehat{\mathcal{E}}_{\text{HAz},\delta}\cap\widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\sum_{t=1}^{T} r(a_t,\boldsymbol{x}_t)\, y_t \geqslant \sum_{t=2}^{T}\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] \\ - \left(E_T + \sqrt{2T\ln\frac{4d}{\delta}} + |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}}\right).$$

Given that on the intersection of events considered, the adaptive policy solves $\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)$ for all $t \geqslant 2$, we have

$$\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} r(a,\boldsymbol{X})\, U_{t-1}(a,\boldsymbol{X})\, \widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] = \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T}\,. \tag{56}$$

This concludes the proof of (51), and hence, the adaptation of the third step.

D.4 Fourth Step: Some Adaptations are Also Required

In this final step, we collect the bounds from the previous three steps. Some (rather minor) adaptations are required, e.g., we would like to integrate (48) and (49) over $\widehat{\nu}_t$ in the left-hand sides and ν in the right-hand sides.

Main modification. We consider some $t \geqslant 2$. For each $\boldsymbol{x} \in \mathcal{X}$, we apply (48) and (49) with $\boldsymbol{q} = \pi^\star(\boldsymbol{x})$ and get

$$\sum_{a\in\mathcal{A}}\Big(r(a,\boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{x})\Big)\, U_{t-1}(a,\boldsymbol{x})\, \widehat{p}_{t,a}(h_{t-1},\boldsymbol{x}) \\ \geqslant \sum_{a\in\mathcal{A}}\Big(r(a,\boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{c}(a,\boldsymbol{x})\Big)_+ U_{t-1}(a,\boldsymbol{x})\, \pi_a^\star(\boldsymbol{x})\,.$$

The bound $U_{t-1} \geqslant P$ of (45) and the non-negative parts taken in the right-hand side then entail that

$$\text{on } \widehat{\mathcal{E}}_{\text{prob},\delta}, \qquad \forall \boldsymbol{x} \in \mathcal{X}, \qquad \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big) U_{t-1}(a, \boldsymbol{x}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{x}) \\ \geqslant \sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{x}) \Big)_+ P(a, \boldsymbol{x}) \, \pi_a^\star(\boldsymbol{x}) \, .$$

We replace the individual $\boldsymbol{x}$ by a random variable $\boldsymbol{X} \sim \widehat{\nu}_t$ and take expectations with respect to $\boldsymbol{X}$:

$$\text{on } \widehat{\mathcal{E}}_{\text{prob},\delta}, \qquad \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big) U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] \\ \geqslant \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big)_+ P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right] .$$

Thanks to the non-negative parts in the right-hand side, we identify some function $f(\boldsymbol{X})$ where f takes values in $[0, 1]$, so that we may apply the uniform control (44) and get

$$\begin{aligned} \text{on } \widehat{\mathcal{E}}_{\text{unif},\delta}, \qquad & \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big)_+ P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right] \\ & \geqslant \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big)_+ P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right] - |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \\ & \geqslant \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \Big(r(a, \boldsymbol{X}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{c}(a, \boldsymbol{X}) \Big) P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right] - |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \, , \end{aligned}$$

where for the second inequality, we simply dropped the non-negative parts.

The rest of the proof for this final step is basically unchanged. Combining all the bounds exhibited so far in this updated fourth step, we have, for each $t \geqslant 2$,

$$\begin{aligned} \text{on } \widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}, \qquad & \overbrace{\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X}) \, U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right]}^{= \mathrm{OPT}(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T)/T} \\ & \qquad - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X}) \, U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] \\ \geqslant \; & \overbrace{\mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} r(a, \boldsymbol{X}) \, P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right]}^{= \mathrm{OPT}(\nu, P, B)/T} \\ & \qquad - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \underbrace{\mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X}) \, P(a, \boldsymbol{X}) \, \pi_a^\star(\boldsymbol{X}) \right]}_{\leqslant (B/T)\mathbf{1}} - |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \, , \end{aligned}$$

where we substituted the inequalities stemming from the definition of $\pi^\star$ as well as the rewriting (56). Rearranging the inequality above and substituting (47), we get that on $\widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\begin{aligned} & \frac{\mathrm{OPT}(\nu, P, B)}{T} - \frac{\mathrm{OPT}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} \\ & \leqslant \frac{B\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbf{1}}{T} - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \boldsymbol{c}(a, \boldsymbol{X}) \, U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(\boldsymbol{X}) \right] + |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \\ & = \frac{B - \widehat{B}_T}{T} \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbf{1} + |\mathcal{X}| \sqrt{\frac{1}{2t} \ln \frac{2T|\mathcal{X}|}{\delta}} \, . \end{aligned} \tag{57}$$

Summing this bound over $2 \leqslant t \leqslant T$ and combining it with (50), we obtain that on $\widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\textsc{opt}(\nu, P, B) - \sum_{t=2}^{T} \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} \leqslant 1 + \frac{B - \widehat{B}_T}{\widehat{B}_T} \sum_{t=2}^{T} \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} + |\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}} \,.$$

By distinguishing the cases

$$\textsc{opt}(\nu, P, B) - \sum_{t=2}^{T} \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} \leqslant 0 \quad \text{and} \quad \textsc{opt}(\nu, P, B) - \sum_{t=2}^{T} \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} \geqslant 0 \,,$$

the inequality above entails that on $\widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\textsc{opt}(\nu, P, B) - \sum_{t=2}^{T} \frac{\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \widehat{B}_T\big)}{T} \leqslant \frac{B - \widehat{B}_T}{\widehat{B}_T} \textsc{opt}(\nu, P, B) + |\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}} + 1 \,.$$

Substituting this upper bound into (51), we finally obtain that on $\widehat{\mathcal{E}}_{\text{prob},\delta} \cap \widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\begin{aligned} \textsc{opt}(\nu, P, B) - \sum_{t=1}^{T} r(a_t, \boldsymbol{x}_t)\, y_t \leqslant \frac{B - \widehat{B}_T}{\widehat{B}_T} \textsc{opt}(\nu, P, B) \\ + E_T + \underbrace{\sqrt{2T \ln \frac{4d}{\delta}} + 2|\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}} + 1}_{\leqslant 2b_T} \,. \end{aligned}$$

We conclude the proof by the same modifications to improve readability as at the end of the proof of Theorem 1: namely, since the definition of E_T did not change, the bound (35) is still applicable, while

$$\frac{B - \widehat{B}_T}{\widehat{B}_T} \leqslant \frac{2b_T}{B} = \frac{1}{B} \left(4 + 2\sqrt{2T \ln \frac{4d}{\delta}} + 2|\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}} \right)$$

is obtained with the same techniques and similar conditions as for (32).

E Details for the Material of Section 6

In this section, we first recall (Appendix E.1) the setting of linear CBwK introduced by Agrawal and Devanur [2016]—and actually, slightly generalize it to match the setting of CBwK for a logistic-regression conversion model. We then state the adaptive policy considered (Appendix E.2), which relies on upper-confidence estimates of the rewards and lower-confidence estimates of the cost vectors. We also state the corresponding estimation guarantees in a key lemma (Lemma 3). The heart of this section is to state, discuss (Appendix E.3) and prove (Appendix E.4) a regret bound, matching the one of Agrawal and Devanur [2016, Theorem 3], with a slight improvement consisting of a relaxation of the budget constraints. For the sake of completeness, we finally recall (Appendix E.5) the statement of the adaptive policy of Agrawal and Devanur [2016].

E.1 Setting

The setting is the following. We consider a finite action set $\mathcal{A}$ including a no-op action a_{null}, a finite context set $\mathcal{X} \subseteq \mathbb{R}^n$, a number T of rounds and a total budget constraint $B > 0$. All these parameters are known. Contexts $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_T$ are drawn i.i.d. according to some distribution ν. At round $t \geqslant 1$, the learner observes the context $\boldsymbol{x}_t$, picks an action a_t and, conditionally to $\boldsymbol{x}_t$ and a_t, when $a_t \neq a_{\text{null}}$, obtains a reward $r_t \in [0,1]$ drawn independently at random according to a distribution with expectation $\overline{r}(a_t, \boldsymbol{x}_t)$, where

$$\forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \overline{r}(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\mu}_\star \,,$$

and suffers a vector cost $\boldsymbol{c}_t \in [0,1]^d$ drawn independently at random according to a distribution with vector of expectations $\overline{\boldsymbol{c}}(a_t, \boldsymbol{x}_t)$, where each component $\overline{c}_i$ of $\overline{\boldsymbol{c}}$, for $i \in \{1, \dots, d\}$, is given by

$$\forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \overline{c}_i(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}} \boldsymbol{\theta}_{\star,i} \,.$$

In the definitions above, $\boldsymbol{\varphi} : \mathcal{A} \setminus \{a_{\text{null}}\} \times \mathcal{X} \to \mathbb{R}^m$ is a known transfer function, with $\|\boldsymbol{\varphi}\| \leqslant 1$, and $\boldsymbol{\mu}_\star$ and the $\boldsymbol{\theta}_{\star,i}$ are unknown parameters in $\mathbb{R}^m$. We assume that these unknown parameters lie in some bounded set Θ, with maximal norm still denoted by $\|\Theta\|$. When $a_t = a_{\text{null}}$, the obtained reward and suffered costs are null: $r_t = 0$ and $\boldsymbol{c}_t = \boldsymbol{0}$.

Comparison to the canonical setting of linear CBwK. Note that in the original formulation of Agrawal and Devanur [2016], we have (where $\boldsymbol{x}_a$ also denote vectors):

$$\boldsymbol{x} = \big(\boldsymbol{x}_a\big)_{a \in \mathcal{A} \setminus \{a_{\text{null}}\}} \qquad \text{and} \qquad \boldsymbol{\varphi}(a, \boldsymbol{x}) = \boldsymbol{x}_a \,.$$

Benchmark and regret. The goal is still to maximize the accumulated rewards while controlling the costs:

$$\text{maximize} \quad \sum_{t \leqslant T} r_t \qquad \text{while controlling} \quad \sum_{t \leqslant T} \boldsymbol{c}_t \leqslant B\boldsymbol{1} \,.$$

The goal can be equivalently defined as the minimization of the regret while controlling the costs, where the regret equals

$$R_T = \textsc{opt}(\nu, \overline{r}, \overline{\boldsymbol{c}}, B) - \sum_{t \leqslant T} r_t$$

for the benchmark given by the static policy $\pi^\star$ achieving the largest expected cumulative rewards under the condition that its cumulative vector costs abide by the budget constraints in expectation, i.e.,

$$\begin{aligned} \textsc{opt}(\nu, \overline{r}, \overline{\boldsymbol{c}}, B) = \max_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} & \ T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \overline{r}(a, \boldsymbol{X}) \right] \\ \text{under} \quad & \ T\, \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \overline{\boldsymbol{c}}(a, \boldsymbol{x}) \right] \leqslant B\boldsymbol{1} \,. \end{aligned} \tag{58}$$

In the sequel, we will use the definition (58) of OPT with different quadruplets of parameters; see, for instance, the definition of the adaptive policy of Box C.

E.2 Statement of an Adaptive Policy

The considered adaptive policy is stated in Box C. It is adapted from the adaptive policy of Section 3. The (almost only) changes lie in Phase 1, which depends heavily of the model. Here, we resort (as Agrawal and Devanur [2016]) to a LinUCB-type estimation of the parameters of the stochastic linear bandits yielding rewards and costs. Based on these estimated parameters, we are able to issue, at each round $t \geqslant 2$, an upper-confidence expected reward function U_{t-1} and a lower-confidence expected vector-cost function $\boldsymbol{L}_{t-1}$. We also use the empirical estimate of the context distribution. In Phase 2, we solve the OPT problem on these estimates and with the conservative budget $\widehat{B}_T$.

The adaptive policy of Box C bears links, and actually generalizes, the one by Xu and Truong [2019]. The setting of the latter reference is more limited as more information is provided to the learner, such as the costs for taking each action and the distribution ν of contexts (clients in their case).

For the sake of completeness, we state in Appendix E.5 the adaptive policy introduced by Agrawal and Devanur [2016].

The main additional ingredient in the analysis of the policy of Box C, compared to the analyses of the adaptive policy of Section 3, is a guarantee on the outcomes of Phase 1. We recall that we assumed that the parameters $\boldsymbol{\mu}_\star$ and $\boldsymbol{\theta}_{\star,i}$ lie in a bounded set Θ with maximal norm denoted by $\|\Theta\|$.

Lemma 3 (direct adaptation from Abbasi-Yadkori et al. [2011, Theorem 2])**.** *Fix any sampling strategy and consider the version of LinUCB given by Phase 1 of Box C. For all $\delta \in (0,1)$, there exists an event $\mathcal{E}_{\text{lin},\delta}$ with probability at least $1-\delta$ and such that over $\mathcal{E}_{\text{lin},\delta}$:*

$$\forall t \geqslant 1,\ \forall a \in \mathcal{A} \setminus \{a_{null}\},\ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \left|\widehat{r}_t(a,\boldsymbol{x}) - \overline{r}(a,\boldsymbol{x})\right| \leqslant \gamma_{t,\lambda,\delta} \left\| \boldsymbol{\varphi}(a,\boldsymbol{x}) \right\|_{X_t^{-1}}$$
$$\textit{and} \qquad \left|\widehat{\boldsymbol{c}}_t(a,\boldsymbol{x}) - \overline{\boldsymbol{c}}(a,\boldsymbol{x})\right| \leqslant \gamma_{t,\lambda,\delta} \left\| \boldsymbol{\varphi}(a,\boldsymbol{x}) \right\|_{X_t^{-1}} \mathbf{1}\,,$$

where

$$\gamma_{t,\lambda,\delta} = \frac{1}{4}\sqrt{m \ln \frac{1+t/(\lambda m)}{\delta/(d+1)}} + \sqrt{\lambda}\|\Theta\|.$$

Proof sketch. We explain why the bound for $\overline{r}$ holds with probability at least $1-\delta/(d+1)$. The lemma follows by repeating the argument for the components of $\overline{\boldsymbol{c}}$ and resorting to a union bound.

Given that rewards lie in $[0,1]$ and are thus $1/4$–sub-Gaussian, the martingale analysis by Abbasi-Yadkori et al. [2011, Theorem 2], with the same adaptations as the ones carried out in Appendix C to take into account the rounds when $a_t = a_{\text{null}}$, shows that with probability at least $1-\delta/(d+1)$,

$$\forall t \geqslant 1, \qquad \left\| \boldsymbol{\mu}_t - \boldsymbol{\mu}_\star \right\|_{X_t} \leqslant \frac{1}{4}\sqrt{m \ln \frac{1+t/(\lambda m)}{\delta/(d+1)}} + \sqrt{\lambda}\|\Theta\| = \gamma_{t,\lambda,\delta}\,.$$

We then proceed similarly to the Cauchy-Schwarz inequalities following (42): for all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$ and $\boldsymbol{x} \in \mathcal{X}$,

$$\left|\widehat{r}_t(a,\boldsymbol{x}) - \overline{r}(a,\boldsymbol{x})\right| = \left|\boldsymbol{\varphi}(a,\boldsymbol{x})^{\mathrm{T}} \left(\boldsymbol{\mu}_{t-1} - \boldsymbol{\mu}_\star\right)\right| \leqslant \left\| \boldsymbol{\varphi}(a,\boldsymbol{x}) \right\|_{X_t^{-1}} \left\| \boldsymbol{\mu}_t - \boldsymbol{\mu}_\star \right\|_{X_t}.$$

This concludes the proof. □

As a consequence, we set, when defining the adaptive policy of Box C,

$$\varepsilon_s(a,\boldsymbol{x}) = \gamma_{t,\lambda,\delta} \left\| \boldsymbol{\varphi}(a,\boldsymbol{x}) \right\|_{X_s^{-1}}$$

for all $s \geqslant 1$, and denote by $\widehat{\mathcal{E}}_{\text{lin},\delta}$ the event of Lemma 3 for the sampling policy of Box C. This event is of probability at least $1-\delta$. We have:

$$\text{on } \widehat{\mathcal{E}}_{\text{lin},\delta}, \qquad \forall t \geqslant 1,\ \forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \forall \boldsymbol{x} \in \mathcal{X},$$
$$\overline{r}(a,\boldsymbol{x}) \leqslant U_t(a,\boldsymbol{x}) \leqslant \overline{r}(a,\boldsymbol{x}) + 2\varepsilon_t(a,\boldsymbol{x}) \tag{59}$$
$$\text{and} \qquad \boldsymbol{L}_t(a,\boldsymbol{x}) \leqslant \overline{\boldsymbol{c}}(a,\boldsymbol{x}) \leqslant \boldsymbol{L}_t(a,\boldsymbol{x}) + 2\varepsilon_t(a,\boldsymbol{x})\mathbf{1}\,. \tag{60}$$

BOX C: LINUCB FOR DIRECT SOLUTIONS TO OPT PROBLEMS

Parameters: regularization parameter $\lambda > 0$; conservative-budget parameter $\widehat{B}_T$; upper-confidence bonuses $\varepsilon_s(a, \boldsymbol{x}) > 0$, for $s \geqslant 1$ and $(a, \boldsymbol{x}) \in \big(\mathcal{A} \setminus \{a_{\text{null}}\}\big) \times \mathcal{X}$.

Round $t = 1$: play an arbitrary action $a_1 \in \mathcal{A} \setminus \{a_{\text{null}}\}$

At rounds $t \geqslant 2$:

Phase 0 If $\displaystyle\sum_{s \leqslant t-1} \boldsymbol{c}_s \leqslant (B-1)\mathbf{1}$ is violated, then $\widehat{\boldsymbol{p}}_t(h_{t-1}, \boldsymbol{x}) = \delta_{a_{\text{null}}}$ for all $\boldsymbol{x}$

Phase 1 Otherwise, estimate the parameters by

$$\boldsymbol{\mu}_{t-1} = X_{t-1}^{-1} \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, r_s$$

$$\text{and} \qquad \widehat{\boldsymbol{\theta}}_{t-1,i} = X_{t-1}^{-1} \sum_{s=1}^{t-1} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, c_{s,i}$$

$$\text{where} \qquad X_t = \sum_{s=1}^{t} \mathbb{1}_{\{a_s \neq a_{\text{null}}\}} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} + \lambda\, \mathrm{I}_m$$

Define the expected reward function $\widehat{r}$ and cost function $\widehat{\boldsymbol{c}}_{t-1} = \big(\widehat{c}_{t-1,i}\big)_{1 \leqslant i \leqslant d}$ as

$$\forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \ \forall \boldsymbol{x} \in \mathcal{X}, \qquad \widehat{r}_{t-1}(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\mu}_{t-1}$$
$$\text{and} \quad \forall 1 \leqslant i \leqslant d, \qquad \widehat{c}_{t-1,i} = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\theta}_{t-1,i}$$

Build the upper-confidence expected reward function U_{t-1} and the lower-confidence expected vector-cost function $\boldsymbol{L}_{t-1}$ as

$$\forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \ \forall \boldsymbol{x} \in \mathcal{X},$$
$$U_{t-1}(a, \boldsymbol{x}) = \max\Big\{\min\big\{\widehat{r}_{t-1}(a, \boldsymbol{x}) + \varepsilon_{t-1}(a, \boldsymbol{x}),\, 1\big\},\, 0\Big\}$$
$$\boldsymbol{L}_{t-1}(a, \boldsymbol{x}) = \max\Big\{\min\big\{\widehat{\boldsymbol{c}}_{t-1} - \varepsilon_{t-1}(a, \boldsymbol{x})\mathbf{1},\, \mathbf{1}\big\},\, \mathbf{0}\Big\}$$

where the maximum and minimum are taken pointwise in the definition of $\boldsymbol{L}_{t-1}$
Set $U_{t-1}(a_{\text{null}}, \boldsymbol{x}) = 0$ and $L_{t-1}(a_{\text{null}}, \boldsymbol{x}) = \mathbf{0}$ for all $\boldsymbol{x} \in \mathcal{X}$

Also estimate the context distribution by $\quad \widehat{\nu}_t = \dfrac{1}{t} \displaystyle\sum_{s=1}^{t} \delta_{\boldsymbol{x}_s}$

Phase 2 Compute the solution $\widehat{\boldsymbol{p}}_t(h_{t-1}, \cdot\,)$ of

$$\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, B_T\big) = \max_{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})} T\, \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} U_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right]$$
$$\text{under} \qquad T\, \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \boldsymbol{L}_{t-1}(a, \boldsymbol{X})\, \pi_a(\boldsymbol{X}) \right] \leqslant B_T \mathbf{1}$$

Draw an arm $a_t \sim \widehat{\boldsymbol{p}}_t(h_{t-1}, \boldsymbol{x}_t)$

E.3 Regret Bound

We sketch in Appendix E.4 below the proof of the following result. We use the $\varepsilon_s(a, \boldsymbol{x})$ indicated by Lemma 3.

Theorem 3. *In the setting of Appendix E.1, we consider the adaptive policy of Box C of Appendix E.2. We set a confidence level $1-\delta \in (0,1)$ and use parameters $\lambda = m\ln(1+T/m)$, a working budget of $\widehat{B}_T = B - b_T$, where*

$$b_T = 2 + m\big(2\sqrt{2}\|\Theta\| + 1\big)\sqrt{T}\ln\frac{1+T/m}{\delta/(d+1)} + \sqrt{2T\ln\frac{4d}{\delta}} + |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}}\,,$$

and $\varepsilon_t(a,\boldsymbol{x}) = \gamma_{t,\lambda,\delta}\,\big\|\boldsymbol{\varphi}(a,\boldsymbol{x})\big\|_{X_t^{-1}}$. Then, provided that $T \geqslant 2m$ and $B > 2b_T$, we have, with probability at least $1-3\delta$,

$$\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B) - \sum_{t=1}^{T} r_t \leqslant 2b_T\left(1 + \frac{\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)}{B}\right).$$

The order of magnitude of the regret bound, in terms of T, m, and B is

$$\frac{\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)}{B}\,m\sqrt{T}\ln T\,.$$

This matches the regret bound achieved by Agrawal and Devanur [2016, Theorem 3], except that the latter reference required a budget B of order $T^{3/4}$ up to logarithmic terms, while we relax the budget amount to $B \geqslant 2b_T$, i.e., B of order $\sqrt{T}$ up to logarithmic terms.

Also, and more importantly, we provide a natural strategy in Box C, whose parameters are easy to tune, while the fully adaptive algorithm underlying Agrawal and Devanur [2016, Theorem 3] has to estimate a critical parameter Z to trade off rewards and costs (the equivalent of our $\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}$ dual optimal variable). This Z should be of order $\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)/B$ and $\sqrt{T}$ initial rounds of the strategy underlying Agrawal and Devanur [2016, Theorem 3] are devoted to computing a suitable value of Z.

We also provide, in the analysis of Appendix E.4, a rigorous treatment of the use of the no-op action a_{null}.

However, the main advantage of Agrawal and Devanur [2016, Theorem 3] over Theorem 3 above lies in the absence of finiteness restriction on the context set $\mathcal{X}$, which we have to (somewhat artificially) introduce to ensure that the linear program of Phase 2 of the adaptive policy of Box C is tractable.

E.4 Proof Sketch of Theorem 3

We use the same $\widehat{\ }$conventions as in Appendix D. The main (but rather minor) changes with respect to the proofs of Appendices B and D are specifically underlined below. The reason why it is handy to consider instead a lower-confidence bound on the vector costs is to be found in Step 4 below.

Step 1. The first step corresponds to Lemma 3 above, together with the introduction of the bound E_T. Given that we pick $\lambda = m\ln(1+T/m) \geqslant 1 \geqslant 1$, we get the following counterpart of (12), by replacing κ by 1:

$$2\sum_{t=2}^{T}\varepsilon_{t-1}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} \leqslant 2\gamma_{T,\lambda,\delta}\sqrt{2mT\ln\left(1+\frac{T}{\lambda m}\right)} \stackrel{\text{def}}{=} E_T\,.$$

Substituting the value of λ and upper bounding $\gamma_{T,\lambda,\delta}$ by

$$\gamma_{T,\lambda,\delta} \leqslant \left(\|\Theta\| + \frac{1}{4}\right)\sqrt{m\ln\frac{1+T/m}{\delta/(d+1)}}\,,$$

we get

$$E_T \leqslant m\big(2\sqrt{2}\|\Theta\|+1\big)\sqrt{T}\ln\frac{1+T/m}{\delta/(d+1)}\,.$$

Step 2. There are three main outcomes of Step 2 (see the summary in Appendix D.2). Up to considering the new $U_t(a,\boldsymbol{x})$ and $\boldsymbol{L}_t(a,\boldsymbol{x})$ in lieu of the $r(a,\boldsymbol{x})\,U_t(a,\boldsymbol{x})$ and $\boldsymbol{c}(a,\boldsymbol{x})\,U_t(a,\boldsymbol{x})$, respectively, we have the following counterparts to (18), (20) and (25). For each $t \geqslant 2$ such that the cost constraints of Phase 0 of the adaptive policy of Box C are not violated and the optimization problem $\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T\big)$ is to be solved, there exists a vector $\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star} \geqslant \boldsymbol{0}$ such that the complementary slackness condition of KKT reads

$$\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} \boldsymbol{L}_{t-1}(a,\boldsymbol{X})\,\widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] = \frac{\widehat{B}_T}{T}\,\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{1}\,, \tag{61}$$

the policy $\widehat{\boldsymbol{p}}_t(h_{t-1},\,\cdot\,)$ satisfies

$$\widehat{\boldsymbol{p}}_t(h_{t-1},\boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{q}\in\mathcal{P}(\mathcal{A})} \sum_{a\in\mathcal{A}} \Big(U_{t-1}(a,\boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{L}_{t-1}(a,\boldsymbol{x})\Big)\,q_a\,, \tag{62}$$

and the value of the optimization problem is larger than

$$\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T\big) \geqslant \widehat{B}_T\big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}}\boldsymbol{1}\,. \tag{63}$$

Step 3. The uniform deviation argument (8), formulated equivalently as (44), still holds, on an event referred to as $\widehat{\mathcal{E}}_{\text{unif},\delta}$. Also, we assumed that rewards r_t and cost vectors $\boldsymbol{c}_t$ are bounded in $[0,1]$ and $[0,1]^d$, respectively. Several applications of the Hoeffding-Azuma inequality, together with a union bound, show that there exist an event $\widehat{\mathcal{E}}_{\text{HAz},\delta}$ of probability at least $1-\delta$ such that, simultaneously, various high-probability controls similar to (52)–(55) hold. We do not rewrite them explicitly.

On the intersection $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$, we have

$$\begin{aligned}
&\sum_{t=1}^T \boldsymbol{c}_t \\
&\leqslant \sum_{t=1}^T \overline{\boldsymbol{c}}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} + \sqrt{\frac{T}{2}\ln\frac{4d}{\delta}}\,\boldsymbol{1} \\
&\leqslant 1 + \sum_{t=2}^T \boldsymbol{L}_{t-1}(a_t,\boldsymbol{x}_t)\mathbb{1}_{\{a_t\neq a_{\text{null}}\}} + \left(E_T + \sqrt{\frac{T}{2}\ln\frac{4d}{\delta}}\right)\boldsymbol{1} \\
&\leqslant 1 + \sum_{t=2}^T \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}} \boldsymbol{L}_{t-1}(a,\boldsymbol{X})\,\widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] + \left(E_T + \sqrt{2T\ln\frac{4d}{\delta}}\right)\boldsymbol{1} \\
&\leqslant 1 + \sum_{t=2}^T \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} \boldsymbol{L}_{t-1}(a,\boldsymbol{X})\,\widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] + \left(E_T + \sqrt{2T\ln\frac{4d}{\delta}} + |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}}\right)\boldsymbol{1}\,,
\end{aligned}$$

where, among others, we used (60) and the definition of E_T for the second inequality. Note that the E_T term was not necessary in Appendices B and D as we were then using an upper-confidence bound on the vector costs, obtained thanks to an upper-confidence bound on the conversion rate. At each round $t \geqslant 2$, whether the strategy picks $\widehat{p}_{t,a}(h_{t-1},\,\cdot\,)$ in Phase 0 (in which case the left-hand side in the display below equals $\boldsymbol{0}$) or Phase 2 (in which case we have an equality in the display below), it holds that

$$\mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}} \boldsymbol{L}_{t-1}(a,\boldsymbol{X})\,\widehat{p}_{t,a}(h_{t-1},\boldsymbol{X})\right] \leqslant \frac{\widehat{B}_T}{T}\boldsymbol{1}\,.$$

Substituting the value of $\widehat{B}_T$, we proved that on $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$, which is an event of probability at least $1-3\delta$,

$$\sum_{t=1}^T \boldsymbol{c}_t \leqslant (B-1)\boldsymbol{1}\,.$$

This shows that on this intersection of events, the adaptive policy of Box C always resorts to Phase 2. We will consider this event for the rest of the proof, so that the results of Step 2 may be applied.

We may proceed similarly for rewards, based on (59): on $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$

$$
\begin{aligned}
&\sum_{t=1}^{T} r_t \\
&\geqslant \sum_{t=1}^{T} \overline{r}(a_t, \boldsymbol{x}_t) \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} - \sqrt{\frac{T}{2} \ln \frac{4}{\delta}} \\
&\geqslant \sum_{t=2}^{T} U_{t-1}(a_t, \boldsymbol{x}_t) \mathbb{1}_{\{a_t \neq a_{\text{null}}\}} - \left(E_T + \sqrt{\frac{T}{2} \ln \frac{4}{\delta}} \right) \\
&\geqslant \sum_{t=2}^{T} \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right] - \left(E_T + \sqrt{2T \ln \frac{4}{\delta}} \right) \\
&\geqslant \sum_{t=2}^{T} \underbrace{\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} U_{t-1}(a, \boldsymbol{X}) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{X}) \right]}_{= \textsc{opt}(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T)/T} - \Bigg(\underbrace{E_T + \sqrt{2T \ln \frac{4}{\delta}} + |\mathcal{X}| \sqrt{2T \ln \frac{2T|\mathcal{X}|}{\delta}}}_{\leqslant b_T} \Bigg),
\end{aligned}
\tag{64}
$$

where the indicated equality to $\textsc{opt}\big(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T\big)$ follows from the fact that on the considered intersection of events, the adaptive policy always resorts to Phase 2. We also used the piece of notation b_T introduced in the statement of Lemma 3.

Step 4. We build on (62) as follows. By the existence of a_{null}, the maximum in (62) can be taken with non-negative parts. We also substitute the upper confidence control (59) and the lower confidence control (60)—this piece of the proof is the very reason why such upper and lower confidence estimates were picked. We get: on $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$, for all $t \geqslant 2$, for all $\boldsymbol{x} \in \mathcal{X}$,

$$
\begin{aligned}
&\sum_{a \in \mathcal{A}} \Big(U_{t-1}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{L}_{t-1}(a, \boldsymbol{x}) \Big) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{x}) \\
&= \sum_{a \in \mathcal{A}} \Big(U_{t-1}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{L}_{t-1}(a, \boldsymbol{x}) \Big)_+ \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{x}) \\
&\geqslant \sum_{a \in \mathcal{A}} \Big(U_{t-1}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{L}_{t-1}(a, \boldsymbol{x}) \Big)_+ \pi_a^\star(\boldsymbol{x}) \\
&\geqslant \sum_{a \in \mathcal{A}} \Big(\overline{r}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \overline{\boldsymbol{c}}(a, \boldsymbol{x}) \Big)_+ \pi_a^\star(\boldsymbol{x}) \, .
\end{aligned}
$$

The rest of the proof follows the exact same logic as in Appendices B and D. By replacing the $\boldsymbol{x}$ by a random variable $\boldsymbol{X} \sim \widehat{\nu}_t$ and integrating, we have, on $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$, that for all $t \geqslant 2$,

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \Big(U_{t-1}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \boldsymbol{L}_{t-1}(a, \boldsymbol{x}) \Big) \, \widehat{p}_{t,a}(h_{t-1}, \boldsymbol{x}) \right] \\
&\geqslant \mathbb{E}_{\boldsymbol{X} \sim \widehat{\nu}_t} \left[\sum_{a \in \mathcal{A}} \Big(\overline{r}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \overline{\boldsymbol{c}}(a, \boldsymbol{x}) \Big)_+ \pi_a^\star(\boldsymbol{x}) \right] \\
&\geqslant \mathbb{E}_{\boldsymbol{X} \sim \nu} \left[\sum_{a \in \mathcal{A}} \Big(\overline{r}(a, \boldsymbol{x}) - \big(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\big)^{\mathrm{T}} \overline{\boldsymbol{c}}(a, \boldsymbol{x}) \Big)_+ \pi_a^\star(\boldsymbol{x}) \right] - |\mathcal{X}| \sqrt{\frac{t}{2} \ln \frac{2T|\mathcal{X}|}{\delta}} \, ,
\end{aligned}
$$

where the second inequality follows by (44), which is legitimately applied thanks the fact that the sum over a in the right-hand side takes values in $[0, 1]$, given the non-negative parts and the fact that $\overline{r}(a, \boldsymbol{x}) \leqslant 1$ by definition. Now, with (61) and the definition of $\pi^\star$: on $\widehat{\mathcal{E}}_{\text{HAz},\delta} \cap \mathcal{E}_{\text{lin},\delta} \cap \widehat{\mathcal{E}}_{\text{unif},\delta}$, for all

$t \geqslant 2$,

$$\begin{aligned}
&\frac{\textsc{opt}(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T)}{T} - \frac{\widehat{B}_T}{T}\left(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\right)^{\mathrm{T}}\mathbf{1} \\
&= \mathbb{E}_{\boldsymbol{X}\sim\widehat{\nu}_t}\left[\sum_{a\in\mathcal{A}}\Big(U_{t-1}(a,\boldsymbol{x}) - \left(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\right)^{\mathrm{T}}\boldsymbol{L}_{t-1}(a,\boldsymbol{x})\Big)\,\widehat{p}_{t,a}(h_{t-1},\boldsymbol{x})\right] \\
&\geqslant \mathbb{E}_{\boldsymbol{X}\sim\nu}\left[\sum_{a\in\mathcal{A}}\Big(\overline{r}(a,\boldsymbol{x}) - \left(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\right)^{\mathrm{T}}\overline{\boldsymbol{c}}(a,\boldsymbol{x})\Big)_+\pi_a^\star(\boldsymbol{x})\right] - |\mathcal{X}|\sqrt{\frac{t}{2}\ln\frac{2T|\mathcal{X}|}{\delta}} \\
&\geqslant \frac{\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)}{T} - \frac{B}{T}\left(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\right)^{\mathrm{T}}\mathbf{1} - |\mathcal{X}|\sqrt{\frac{t}{2}\ln\frac{2T|\mathcal{X}|}{\delta}}\,.
\end{aligned}$$

Based on these inequalities, we have

$$\begin{aligned}
&\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B) - \sum_{t=2}^{T}\frac{\textsc{opt}(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T)}{T} \\
&\leqslant |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}} + 1 + \sum_{t=2}^{T}\frac{B-\widehat{B}_T}{T}\left(\widehat{\boldsymbol{\beta}}_t^{\text{budg},\star}\right)^{\mathrm{T}}\mathbf{1} \\
&\leqslant |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}} + 1 + \frac{B-\widehat{B}_T}{\widehat{B}_T}\sum_{t=2}^{T}\frac{\textsc{opt}(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T)}{T}\,,
\end{aligned}$$

where we substituted (63) for the second inequality. By a case analysis, we finally proved that on $\widehat{\mathcal{E}}_{\text{HAz},\delta}\cap\mathcal{E}_{\text{lin},\delta}\cap\widehat{\mathcal{E}}_{\text{unif},\delta}$,

$$\begin{aligned}
&\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B) - \sum_{t=2}^{T}\frac{\textsc{opt}(\widehat{\nu}_t, U_{t-1}, \boldsymbol{L}_{t-1}, \widehat{B}_T)}{T} \\
&\qquad\qquad \leqslant \left(\frac{B}{\widehat{B}_T} - 1\right)\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B) + \underbrace{1 + |\mathcal{X}|\sqrt{2T\ln\frac{2T|\mathcal{X}|}{\delta}}}_{\leqslant b_T}\,.
\end{aligned}$$

The proof is concluded by combining the inequality above with the bound (64) on cumulative rewards:

$$\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B) - \sum_{t=1}^{T} r_t \leqslant 2b_T + \underbrace{\left(\frac{B}{\widehat{B}_T} - 1\right)}_{\leqslant 2b_T}\textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)\,,$$

where the bound on $B/\widehat{B}_T - 1$ is obtained with the same techniques and similar conditions as for (32).

E.5 Reminder: Algorithm 1 from Agrawal and Devanur [2016]

We recall (and actually slightly adapt to our setting) the adaptive policy of Agrawal and Devanur [2016] titled Algorithm 1 therein. We describe it in Box D. One of the adaptations is to state it with general upper-confidence bonuses $\varepsilon_s(a,\boldsymbol{x}) > 0$. As in Agrawal and Devanur [2016], who proceed as in the proof of Lemma 3, we will use the same values for $\varepsilon_s(a,\boldsymbol{x})$ as in Theorem 3. The same comment applies to λ. Another adaptation is that we specified the online convex optimization algorithm to be used and picked a simple strategy (instead of other possible choices discussed in Agrawal and Devanur [2016]), namely, the projected gradient descent introduced by Zinkevich [2003]. The latter relies on a learning rate $\eta > 0$. The drawback of the projected gradient descent is however that its dependency in the ambient dimension is suboptimal.

The final parameter of the adaptive policy of Box D is a parameter Z to trade off between rewards and costs. A recommended choice is, for instance, $Z = \textsc{opt}(\nu,\overline{r},\overline{\boldsymbol{c}},B)/B$, the issue being that, of course, the latter value is unknown. In the simulation study of Appendix F, we will provide a good value of Z to the adaptive policy, even though Agrawal and Devanur [2016] introduce a variant with a preliminary exploration phase meant to provide in an automatic way such a good value for Z.

BOX D: ADAPTATION OF ALGORITHM 1 FROM AGRAWAL AND DEVANUR [2016]

Parameters: regularization parameter $\lambda > 0$; trade-off parameter Z between reward and costs; upper-confidence bonuses $\varepsilon_s(a, \boldsymbol{x}) > 0$, for $s \geqslant 1$ and $(a, \boldsymbol{x}) \in \big(\mathcal{A} \setminus \{a_{\text{null}}\}\big) \times \mathcal{X}$; learning rate $\eta > 0$.

Round $t = 1$: play an arbitrary action $a_1 \in \mathcal{A} \setminus \{a_{\text{null}}\}$; pick $\boldsymbol{\zeta}_1 = \mathbf{0}$

At rounds $t \geqslant 2$:

Phase 0 If $\displaystyle\sum_{s \leqslant t-1} \boldsymbol{c}_s \leqslant (B-1)\mathbf{1}$ is violated, then $a_t = a_{\text{null}}$

Phase 1 Otherwise, estimate the parameters by

$$\boldsymbol{\mu}_{t-1} = X_{t-1}^{-1} \sum_{s=1}^{t-1} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, r_s$$

$$\text{and} \qquad \widehat{\boldsymbol{\theta}}_{t-1,i} = X_{t-1}^{-1} \sum_{s=1}^{t-1} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)\, c_{s,i}$$

$$\text{where} \qquad X_t = \sum_{s=1}^{t} \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s) \boldsymbol{\varphi}(a_s, \boldsymbol{x}_s)^{\mathrm{T}} + \lambda\, \mathrm{I}_m$$

Define the expected reward function $\widehat{r}$ and cost function $\widehat{\boldsymbol{c}}_{t-1} = \big(\widehat{c}_{t-1,i}\big)_{1 \leqslant i \leqslant d}$ as

$$\begin{aligned} \forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \ \forall \boldsymbol{x} \in \mathcal{X}, \qquad & \widehat{r}_{t-1}(a, \boldsymbol{x}) = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\mu}_{t-1} \\ \text{and} \quad \forall 1 \leqslant i \leqslant d, \qquad & \widehat{c}_{t-1,i} = \boldsymbol{\varphi}(a, \boldsymbol{x})^{\mathrm{T}}\, \boldsymbol{\theta}_{t-1,i} \end{aligned}$$

Build the upper-confidence expected reward function U_{t-1} and the lower-confidence expected vector-cost function $\boldsymbol{L}_{t-1}$ as

$$\begin{aligned} &\forall a \in \mathcal{A} \setminus \{a_{\text{null}}\},\ \ \forall \boldsymbol{x} \in \mathcal{X}, \\ &\qquad U_{t-1}(a, \boldsymbol{x}) = \widehat{r}_{t-1}(a, \boldsymbol{x}) + \varepsilon_{t-1}(a, \boldsymbol{x}) \\ &\qquad \boldsymbol{L}_{t-1}(a, \boldsymbol{x}) = \widehat{\boldsymbol{c}}_{t-1} - \varepsilon_{t-1}(a, \boldsymbol{x})\mathbf{1} \end{aligned}$$

Phase 2 Play

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A} \setminus \{a_{\text{null}}\}} U_{t-1}(a, \boldsymbol{x}) - Z\left(\boldsymbol{\zeta}_{t-1}^{\mathrm{T}} \boldsymbol{L}_{t-1}(a, \boldsymbol{x})\right)$$

Compute

$$\boldsymbol{\zeta}_t = \Pi_{\text{unit}}\Big(\boldsymbol{\zeta}_{t-1} + \eta\big(\boldsymbol{c}_{t-1} - (B/T)\mathbf{1}\big)\Big)$$

where Π_{unit} denotes the Euclidean projection onto the unit ℓ_1–ball

$$\big\{\boldsymbol{\zeta} \in \mathbb{R}^d : \boldsymbol{\zeta} \geqslant \mathbf{0} \text{ and } \zeta_1 + \ldots + \zeta_d \leqslant 1\big\}$$

F Simulation Study

This appendix reports numerical simulations performed on partially simulated but realistic data (Appendix F.1), for the sake of illustration only. We describe (Appendix F.2) the specific experimental setting of CBwK for a conversion model considered—i.e., the features available, the parameters of the logistic regression, the reward and cost functions. A key point is that continuous variables are used to define rewards, costs, and even the conversion rate, while the adaptive policy of Box C must discretize these variables to solve its Phase 2 linear program. Though the experimental setting introduced is not a setting of linear CBwK, we may still apply the Box D adaptive policy (Appendix F.3), with the underlying idea that it provides linear approximations to non-linear reward and cost functions. We carefully explain how the hyper-parameters picked (Appendix F.4) before providing and discussing the outcomes of the simulations (Appendix F.5). The main outcome is that, as expected, the ad hoc adaptive policy of Box C outperforms the adaptive policy of Box D, which essentially linearly approximates non-linear rewards and costs. We end with a note (Appendix F.6) on the computation environment and time.

F.1 Data Preparation and Available Contexts

The underlying dataset for the simulations is the standard "default of credit card clients" dataset of UCI [2016], initially provided by Yeh and Lien [2009]. (It may be used under a Creative Commons Attribution 4.0 International [CC BY 4.0] license.) This dataset is originally for comparing algorithms predicting default probability of credit card clients. It includes some socio-demographic data, debt level, and payment/default history of the clients. It also includes a target measuring whether the client will default in the future (1-month ahead). We transform it to match our motivating application of market share expansion for loans, described in Appendix A. To do so, we consider each line of the dataset as a loan application. We then discard some variables (e.g., the target) and create new ones (requested amount, standard interest rate offered, risk score). Below, we begin with describing the variables that we keep as they are and explain next how we created the additional variables, based on existing ones.

Variables retrieved. The variable *Age* provides the age of a given client at the time of the loan application, in years. We discretize it into 5 levels with similar numbers of loan requests in each level. The cutoffs for each level are 27, 31, 37, and 43, respectively. This gives rise to a variable referred to as *Age–discrete*.

The variable *Education* reports the education level of a client; in the data there are 4 levels: "others" (level 1, representing 2% of clients), "high school" (level 2, with a share of 16%), "university degree" (level 3, with 47%), and "graduate school degree" (level 4, with 35%).

Finally, the variable *Marital status* provides the marital status of a client: "others" (level 1, accounting for 1.3% of clients), "single" (level 2, for 53.3% of clients), and "married" (level 3, for 45.4%).

Variables created based on existing ones. We create a variable *Requested amount*, in dollars ($), based on a variable provided in the data set that measures the current debt level, in dollars, of the clients: we do so by multiplying the debt level by 0.2. We cap the value of *Requested amount* to 100K$. We then discretize *Requested amount* into 5 levels with similar numbers of loan requests in each level; the obtained variable is referred to as *Requested amount–discrete*. The cutoffs for each level are 10K$, 20K$, 36K$, and 54K$, respectively.

For the final two variables, *Standard interest rate* and *Risk score*, we first build a probability-of-default model with the variables from the raw database as predictors and the occurrence of a default within the next month as a target. This probability-of-default model is based on XGBoost (Chen and Guestrin [2016]), run with no penalization, depth 3, learning rate 0.01, subsample parameter 0.8, min child weight 10, and number of trees $1{,}176$. We only set the number of trees by cross validation, while the rest of the hyper-parameters were set arbitrarily. As the default target is on a credit card, the predicted default rate seems high compared to what we deem as typical default rates on loans. We therefore divide the predicted probability of default by factor of 4 and cap this probability to 20%. This gives us a working variable called *PD*, for probability of default.

We build a *Risk score* rating the risk of a client's default, with 5 levels, coded from A (level 1) to E (level 5), where E represents the highest risk. It is created based on the 20% - 40% - 60% - 80% quantiles of *PD*.

We finally set the *Standard interest rate* variable as 0.9 times the *PD*, with a maximal value of 18% and a minimum one of 1%. This constitutes an oversimplification of risk-based pricing, since we do not take into account any loss given default; but we do not have enough information in the dataset to do so, which is why we basically assume that the loss given default is constant. Then, theory has it that the *Standard interest rate* can be considered proportional to *PD*, and we carefully picked the factor 0.9 to get realistic values. In the dataset, the thus created *Standard interest rate* variable exhibits an average of 4.9%, with median 3.3%. An important note is that this variable is continuous and does not take finitely many values, while the setting described in Section 2.1 imposes such a restriction (the linear setting of Section 6 does not require it). We discuss below how we take this fact into account: by only using this variable for the conversion model (i.e., in Phase 1 of the adaptive policy of Box C), but not to pick actions (i.e., not in Phase 2 of the adaptive policy of Box C).

Lastly, we filter the database to remove the outliers: the lines for which *Standard interest rate* times *Requested amount* is larger than 10K, which happens for 133 clients. Our final database contains $29{,}867$ loan applications, out of which we will bootstrap $T = 50{,}000$ applications.

Additional comments. Note that for *Requested amount* and *Age*, the discretizations performed aim to get five balanced classes; however, as some requests are with some specific boundary values, we do not get exact equal distributions over the classes.

All the parameters, constants, and cutoff/filter thresholds used in this data preparation step were decided arbitrarily and were not based on any real information. The context variables here were also selected somewhat arbitrarily (based on their availability), and solely for illustration purposes. In reality, the variables that can be used for commercial discounts need to comply with relevant laws, regulations and company's internal compliance rules.

Summary. The context $\boldsymbol{x}$ for a given client thus contains the following variables: *Age*, *Age-discrete*, *Education*, *Marital status*, *Requested amount*, *Requested amount–discrete*, *Risk score*, and *Standard interest rate*. Categorical variables are hot-one encoded via binary variables.

F.2 Specific Setting of CBwK for Logistic Conversions

We recall that our aim is to provide simulations matching the motivating example of market share expansion for loans described in Appendix A. We take as action set, i.e., as possible discount rates, $\mathcal{A} = \{a_{\text{null}}, 0.1, 0.2, 0.35, 0.55, 0.8\}$.

Features. The feature vectors $\varphi_{\text{conv}}(a, \boldsymbol{x})$ used are composed of only some of the variables defining the context $\boldsymbol{x}$, namely, *Age-discrete*, *Education*, *Marital status*, *Requested amount–discrete*, *Risk score*, and *Standard interest rate* (we recall that this variable is not discrete but will not use it in the linear program of Phase 2), with the addition of a new variable called *Final interest rate* equal to the discounted standard interest rate offered, i.e., *Final interest rate* = *Standard interest rate* $\times(1 - a)$.

Reward and vector cost functions. We use the following (normalized) reward and cost functions, inspired from Appendix A. We set a common duration for all loans, say, 2 years, so that the requested amount equals the outstanding. For all $a \in \mathcal{A} \setminus \{a_{\text{null}}\}$ and $\boldsymbol{x}$,

$$r(a, \boldsymbol{x}) = x_{\text{am}}/M_{\text{am}} \qquad \text{and} \qquad \boldsymbol{c}(a, \boldsymbol{x}) = \big(a/M_{\text{disc}},\, x_{\text{ir}} x_{\text{am}}/M_{\text{ir,am}}\big) \tag{65}$$

where x_{am} and x_{ir} denote the components of the context $\boldsymbol{x}$ containing the *Requested amount* and *Standard interest rate*, respectively, and the normalization factors equal $M_{\text{am}} = 10^5$, $M_{\text{disc}} = 7$, and $M_{\text{ir,am}} = 9{,}996$.

Note that the definition of the first cost here is different from that in Appendix A: we use a/M_{disc} here instead of $\mathbb{1}_{\{a \neq 0\}}$ there. We do so not to disfavor the Box D policy, see details in Appendix F.3 below.

Conversion rate function P. We model the conversion rate function P with the logistic-regression model stated in (1), with $\varphi = \varphi_{\text{conv}}$, and only need to provide the numerical value of $\boldsymbol{\theta}_\star$, which we do

Table 1: Coefficients picked for the logistic-regression model of P.

Intercept	0.8177				
Continuous variable	Single coefficient				
Final interest rate	-13.1101				
Discrete variables	Coefficients for each level				
	Level 1	Level 2	Level 3	Level 4	Level 5
Risk score	-0.3045	-0.0383	0.0515	0.1261	0.1636
Requested amount–discrete	0.7093	0.4703	0.1113	-0.2748	-1.0179
Age–discrete	-0.1837	-0.1392	-0.0476	0.1096	0.2592
Education	0.1836	0.0126	-0.0896	-0.1084	
Marital status	0.0799	0.0102	-0.0918		

in Table 1. This model, as well as the Phase 1 learning of $\boldsymbol{\theta}_\star$ described by (3)–(5) in Section 3, holds for possibly non-discrete contexts.

The numerical values picked for $\boldsymbol{\theta}_\star$ were so in some arbitrary way, to get somewhat realistic outcomes with a simple model structure. We imposed monotonicity constraints, as these are most natural: for instance, the conversion rate increases with the level of *Risk score* and *Age–discrete* increase, and decreases with the level of *Education*, *Requested amount–discrete*, and *Final interest rate*. The coefficients for *Marital status* indicates that conversions are more likely for clients that are single than for married clients.

The average conversion rate in the case $a = 0$ of no discount (i.e, by replacing *Final interest rate* by *Standard interest rate*) is around 50%.

Adaptive policy: based only on the discrete variables. As indicated above, the logistic-regression model and the learning of its parameters apply to continuous variables. The restriction that the context set $\mathcal{X}$ should be finite only came from Phase 2 of the Box C adaptive policy, i.e., the linear program—in particular, for it to be computationally tractable. Here, we thus restrict our attention to policies that map the discrete variables in $\boldsymbol{x}$ to distributions over $\mathcal{A}$: policies that ignore the variables *Age*, *Requested amount*, and *Standard interest rate*. For the first two variables, they may use their discretized versions *Age–discrete* and *Requested amount–discrete*. For *Standard interest rate*, given how it was constructed, *Risk Score* appears as its discretized version.

The aim of these simulations is to show, among others, that using discretizations only in Phase 2 is relevant and efficient.

Note that, on the theoretical side, the proof sketches provided in Sections 4 and 5 reveal that the errors $\varepsilon_t(a, \boldsymbol{x})$ for learning $\theta_\star$ and P, obtained as outcomes of the first step of the analyses, are carried over in the subsequent steps, where the optimization part is evaluated. Using discretizations only in Phase 2 does therefore not come at the price of loosing theoretical guarantees.

F.3 Consideration of the Box D Adaptive Policy for Linear CBwK

In these experiments, we also consider the Box D adaptive policy of Appendix E.5, which was introduced by Agrawal and Devanur [2016] in a different setting. To be as fair as possible to this adaptive policy, we do so with the extended features $\varphi_{\text{lin}}(a, \boldsymbol{x})$ consisting of the features $\varphi_{\text{conv}}(a, \boldsymbol{x})$ described above and three additional components: the discount a, the *Requested amount* x_{am}, and the product $x_{\text{ir}}x_{\text{am}}$ of the *Requested amount* by the *Standard interest rate*. Actually, to ensure that $\varphi_{\text{lin}}(a, \boldsymbol{x}) \in [0, 1]^m$, the last two components added are normalized: we rather use $x_{\text{am}}/M_{\text{am}}$ and $x_{\text{ir}}x_{\text{am}}/M_{\text{ir,am}}$. The reward and vector cost functions introduced in (65) are linear in $\varphi_{\text{lin}}(a, \boldsymbol{x})$. Even better, each component of $r(a, \boldsymbol{x})$ and $\boldsymbol{c}(a, \boldsymbol{x})$ is given directly by a component of $\varphi_{\text{lin}}(a, \boldsymbol{x})$—an extremely simple linear dependency on $\varphi_{\text{lin}}(a, \boldsymbol{x})$.

However, the expected reward and cost functions

$$\overline{r}(a, \boldsymbol{x}) = r(a, \boldsymbol{x})\, P(a, \boldsymbol{x}) \qquad \text{and} \qquad \overline{\boldsymbol{c}}(a, \boldsymbol{x}) = \boldsymbol{c}(a, x)\, P(a, \boldsymbol{x})\,,$$

which are the ones that should be linear in $\varphi_{\text{lin}}(a, \boldsymbol{x})$ according to the setting described in Appendix E.1, are not linear in these features. This is due to the $P(a, \boldsymbol{x})$ terms, which are given by logistic regressions.

The Box D adaptive policy is therefore disadvantaged. This is even more true as it models rewards and costs independently, while they are coupled through conversions. We nonetheless consider this linear-modeling policy because a typical justification for linear approximations is that they offer a typical and efficient first-stage approach to possibly complex problems. Another reason was the desire to have some competitor to our policy in the simulations, and Box D adaptive policy was an easy-to-implement strategy—unlike the policies by Badanidiyuru et al. [2014] and Agrawal et al. [2016], which rely on considering finitely many benchmark policies.

All in all, we report the performance of the Box D policy as well in our experiments, though, as expected, the ad-hoc Box C policy outperforms it.

F.4 Hyper-parameters Picked

We actually set the hyper-parameters based on the budget B, and therefore, first explain how we set its possible values. It turns out that setting $B \geqslant 3{,}650$ is equivalent to not imposing any constraint, while setting $B \geqslant 2{,}900$ is equivalent to no second budget constraint. To get meaningful results, we therefore picked $B = 1{,}600$ and $B = 2{,}200$ as the two possible values for B. We now describe how we tune each of the two adaptive policies considered.

Hyper-parameters common to the two adaptive policies. We take $T = 50{,}000$ clients in the experiments, by bootstrapping them from the enriched dataset prepared in Appendix F.1. We set initial 50 rounds as a warm start for the sequential logistic regression and sequential linear regression carried out in Phase 1 of the adaptive policies.

Both adaptive policies use upper-confidence bonuses $\varepsilon_s(a, \boldsymbol{x})$, which are roughly of the form (considering λ as a constant)

$$C\big(1 + \ln s\big) \left\| \boldsymbol{\varphi}(a, \boldsymbol{x}) \right\|_{X_s^{-1}},$$

where the matrix X_s was defined in Box C; for simplicity, we set $\kappa = 1$, so that the matrices V_s of Lemma 1 and X_s are equal, which explains the common form of the upper-confidence bonuses $\varepsilon_s(a, \boldsymbol{x})$. The hyper-parameter C controls the exploration: the higher C, the more exploration. We report in the simulations the results achieved for C in the range $\{0.025, 0.1, 0.3\}$. That range was set so that at round $s = 51$, which is the first round after the warm start, the $\varepsilon_{51}(a, \boldsymbol{x})$ take values around 0.05, 0.3, and 0.9, respectively.

For simplicity, we set $B_T = B$ as a working budget.

Hyper-parameters for the Box C adaptive policy. We feed this adaptive policy with a good value of λ, namely, $\lambda = 0.0129$. We obtained it by cross-validation on an independent T–sample of data, using the Phase 1 estimation. In the T–sample for estimating λ, at each round s, we take action from the optimal static policy and use the associated conversion y_s as target for estimation. We omit the projection step in Phase 1 by considering that a large enough set Θ was picked.

Hyper-parameters for the Box D adaptive policy. As discussed in Appendix E.5, we set $Z = \textsc{opt}(\nu, P, B)/B$, that is, $Z = 5.16$ for $B = 1{,}600$ and $Z = 3.87$ for $B = 2{,}200$. We also set $\lambda = 0.2452$ for $B = 1{,}600$ and $\lambda = 0.2765$ for $B = 2{,}200$. These values were obtained as weighted averages: the sum of 0.5 times the optimal λ for rewards and 0.25 times the optimal λ for each of the two cost components. These optimal λs for rewards and cost components were set by cross validation on an independent T–sample; with actions a_s taken at each round s from the optimal static policy and associated rewards r_s and costs $\boldsymbol{c}_s$ as targets.

Finally, the learning parameter η was selected in the range $\{0.005, 0.01, 0.05, 0.1, 0.2\}$. We did so given the other choices, by picking in hindsight the η with best performance; this of course, just like the clever choice of Z, should give an advantage to the Box D adaptive policy. Namely, when $B = 1{,}600$, for C equal 0.025, 0.1, and 0.3, we selected η equal to 0.05, 0.01, and 0.1, respectively; and when $B = 2{,}200$, we selected η equal to 0.01, 0.01, and 0.05, respectively. When performing these retrospective choices, we however noted that the performance was not significantly impacted by the choice of η.

F.5 Outcomes of the Simulations

We were limited by the computational power (see Appendix F.6) and could only perform 10 simulations for each pair of $B \in \{1{,}600,\ 2{,}200\}$ and $C \in \{0.025,\ 0.1,\ 0.3\}$. We report averages (strong lines) as well as ± 2 times the standard errors (shaded areas).

Figure 1 reports, in the first line of graphs, the regret achieved with respect to what achieves the optimal static policy, i.e.,

$$t \longmapsto \frac{t}{T}\textsc{opt}(\nu, P, B) - \sum_{s=1}^{t} r_s \,,$$

where $\textsc{opt}(\nu, P, B)$ is larger than $8{,}000$ for both values of B. This regret can take negative or positive values, but in expectation, it is non-negative. This is not immediately clear from the figure, which reports the empirical averages of the regret over 10 runs: these empirical averages are sometimes negative, but they always lie in confidence intervals containing the value 0.

The figures also reports, in the second and third lines, the difference between a constant linear increase of the costs (between a 0 initial cost and a final B cost) and the costs actually achieved by the adaptive policies. I.e., these graphs report the averages and standard errors of the following quantities: for each cost component $i \in \{1, 2\}$,

$$t \longmapsto \sum_{s=1}^{t} c_{s,i} - \frac{t}{T}B \,;$$

by design, the difference above must be non-positive. The second line of Figure 1 deals with the first cost component, and its third line reports the results for the second cost component.

The experiments reveal that while both adaptive policies seem to achieve sublinear regret, the Box D adaptive policy, which is suited for the CBwK setting for a conversion model, performs better than the Box C adaptive policy in terms of rewards: it achieves a smaller, sometimes negative, regret. In terms of costs, we globally see the same trend, with, for a given value of C, the Box D adaptive policy suffering smaller costs than the Box C adaptive policy while achieving higher rewards. This hints at a better use of the discounts.

The hyper-parameter C has an interesting impact: the lower C, the lower the regret (the higher the rewards) and the lower the costs. Rewards and costs go hand in hand: for a given adaptive policy, higher rewards are associated with higher costs.

F.6 Computation Time and Environment

As requested by the NeurIPS checklist, we provide details on the computation time and environment. Our experiments were ran on the following hardware environment: no GPU was required, CPU is 2.7 GHz Quad-Core with total of 8 threads and RAM is 16 GB 2133 MHz LPDDR3. We ran 5 simulations with different seeds on parallel each time. In the setting and for the data described above, it took us 8 hours for each such bunch of 5 runs of the adaptive policy of Box C, and 1.5 hours for Box D.

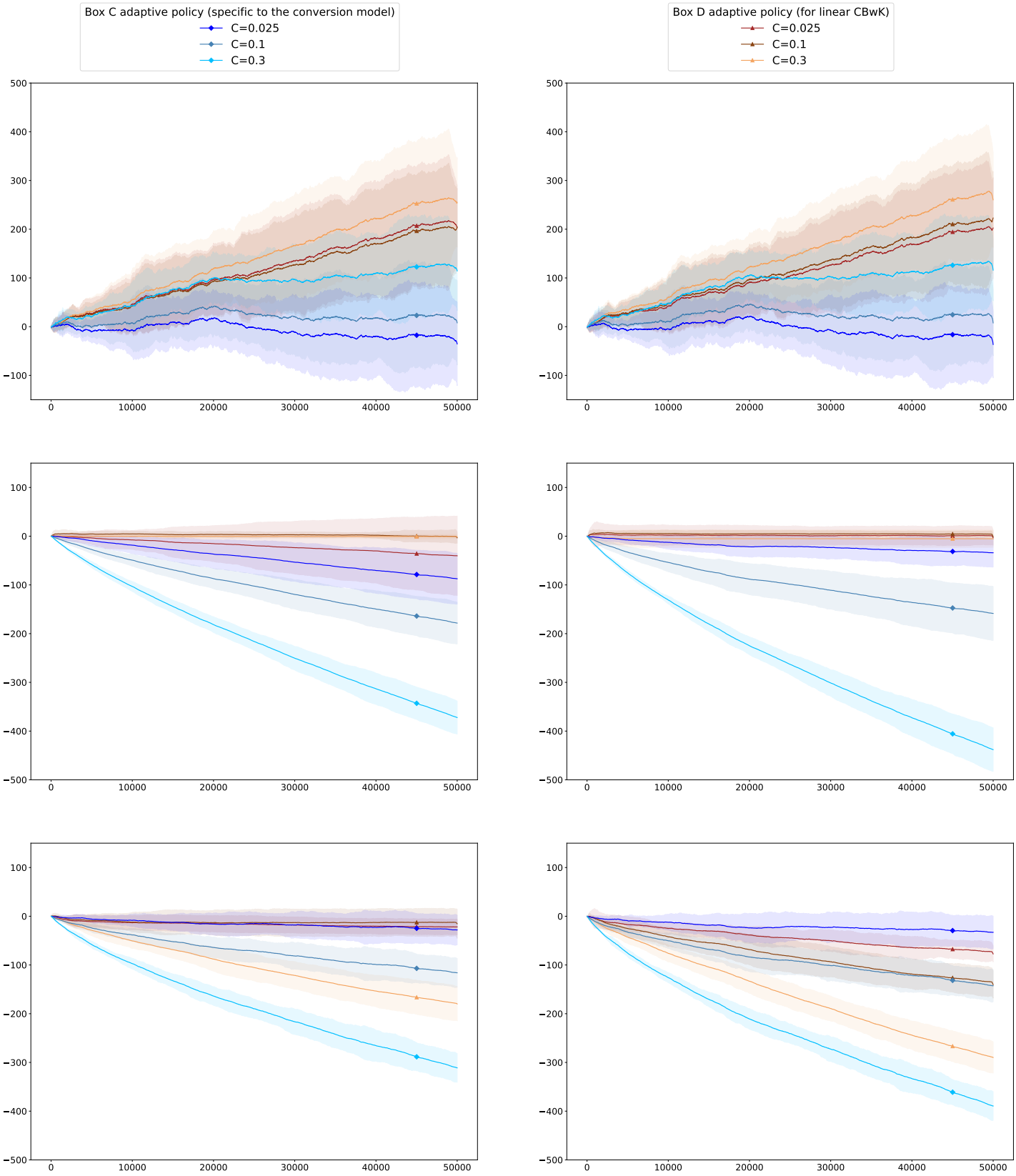

Figure 1: Averages (*solid lines*) and ± 2 times standard errors (*shaded areas*), achieved on 10 runs by the Box C (*blue*) and Box D (*orange*) adaptive policies: of the regret (*first line*), of the difference of the first cost component to tB/T (*second line*), and of the difference of the second cost component to tB/T (*third line*), by the values of the budget (B = 1,600 in the *first column*, B = 2,200 in the *second column*).