Bandit Problems under Censored Feedback

by

Gauthier Marc Benoit Guinet

Diplôme d'ingénieur, Ecole Polytechnique (2020)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Sloan School of Management
June 30, 2022

Certified by . . . .
Saurabh Amin
Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by . . . .
Patrick Jaillet
Dugald C. Jackson Professor, Department of Electrical Engineering and
Computer Science
Thesis Supervisor

Accepted by . . . .
Georgia Perakis
William F. Pounds Professor of Management Science
Co-director, Operations Research Center

Bandit Problems under Censored Feedback

by

Gauthier Marc Benoit Guinet

Submitted to the Sloan School of Management
on June 30, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

In this thesis, we study sequential decision-making models where the feedback received by the principal depends on strategic uncertainty (e.g., agents' willingness to follow a recommendation) and/or random uncertainty (e.g., loss or delay in arrival of information). Such challenges often arise in AI-driven platforms, with applications in recommender systems, revenue management or transportation. We model and study this class of problems through the lens of multi-armed and contextual bandits evolving in censored environments. Our goal is to estimate the performance loss due to censorship in the context of classical algorithms designed for uncensored environments. Our main contributions include the introduction of a broad class of censorship models and their analysis in terms of the *effective dimension* of the problem – a natural measure of its underlying statistical complexity and main driver of the regret bound. In particular, the effective dimension allows us to maintain the structure of the original problem at first order, while embedding it in a bigger space, and thus naturally leads to results analogous to uncensored settings. Our analysis involves a continuous generalization of the Elliptical Potential Inequality, which we believe is of independent interest. We also discover an interesting property of decision-making under censorship: a transient phase during which initial misspecification of censorship is self-corrected at an extra cost; followed by a stationary phase that reflects the inherent slowdown of learning governed by the effective dimension.

Thesis Supervisor: Saurabh Amin
Title: Associate Professor of Civil and Environmental Engineering

Thesis Supervisor: Patrick Jaillet
Title: Dugald C. Jackson Professor, Department of Electrical Engineering and Computer Science

Acknowledgments

Throughout my time in the Operations Research Center at MIT, I am extremely fortunate to have been co-advised by Saurabh Amin and Patrick Jaillet. I would like to express my sincere thanks to them for their unwavering guidance and support over the past two years. Patrick, thank you for pushing me to work on impactful problems, for always raising insightful and challenging questions, and for believing in my success even when I did not. Saurabh, thank you for the countless hours we spent together exploring new research directions or discussing, for always being so understanding and engaging, and for advising me to work on the problems that matter most. Thank you both for not only being incredibly inspiring academic advisors, but also caring mentors.

I would also like to thank my other outstanding collaborator, Prem Talwai. Thank you for all the hours spent discussing research, Big math, and life, as well as keeping my hopes and motivation alive.

My time at MIT would not have been the same without all the friends I met during my stay. I would like to thank the team of French companions with whom I undertook this trip: Jean, Victor, Raphaelle, Arie, Greg, Raphael, Antoine, Rémi, Hedi, Marjo, Paul, Martin, Pauline, Julien, Léa as well as the "adopted" French Ani, Shaksham, Emily, Sam, Lindsey and Paula. A very big thank you also to all my friends at the ORC, LIDS and RESIL, and in particular: Moise, Amine, Romain, Manuel, Angelos, Andy, Sohil, Aaron, Rahman, Jean-Baptiste and Samarth. A special thanks to Andala, who was always there when it was needed. Acknowledging my friends from France also goes without saying.

Finally, I would like to thanks all that my family has done and continues to do for me. To my parents, thank you for giving me both the opportunity and the will to study and work in fields I am passionate about, for always being a supportive and loving presence, and for pushing me to become a better person. To my sisters Celeste and Clemence, thank you as well for all the lively discussions, laughter, and adventures we've had together.

Finally, this thesis is based upon work supported by the AFOSR FA9550-19-1-0263 “Building attack resilience into complex networks” Grant.

.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Focus

Bandit problems are prototypical models of sequential decision-making under uncertainty. They are widely studied due to their applications in recommender systems, online advertising, medical treatment assignment, revenue management, network routing and control [29, 42]. Our work is motivated by settings in which the feedback received by the decision-maker in each round of decision is censored by a stochastic process that depends on the current action as well as past history of feedbacks and actions. For instance, in typical missing data problems, the decision-maker needs to deal with frequent losses of information (or delays in arrival of information) due to exogeneous failures such as faulty and/or unreliable communication. Missing observations in dynamical interactions with the environment are a common concern in diverse fields ranging from operations management to health sciences to physical sciences [48, 23, 32]. In other settings, such as AI-driven platforms for health alerts, route guidance, and product recommendations [6, 49], the reception of feedback depends on whether or not the decision (or recommendation) is adopted by strategic agents (e.g. patients, customers or drivers) with private valuations. Thus, from the platform's viewpoint, the adoption behavior of heterogeneous agents can be regarded as a stochastic censorship process.

A typical application is dynamic decision-making in logistics systems where an operator (principal) aims to maximize a cumulative reward metric (e.g. timeliness or fuel usage efficiency) by recommending routes to drivers (agents). At a given time (stage), the principal can only revise estimates on specific routes based on the data from agents who follow its recommendation to take those routes. The choice model of the agents endogenizes the censorship process. Additionally, censorship can also arise due to unreliable or insecure communication between principal and agents. The "optimality" in this decision-making problem depends on how fast the underlying latent condition of the network that governs the stage-wise rewards can be learned. The challenge arises from the fact that the data generating process is mediated by agents' behavior and the data available is incomplete due to censorship. The question then is to develop efficient algorithms that account for censorship and estimate the performance loss (relative to no censorship benchmark).

Similar questions were also raised in the recent *value alignment* literature in order to study the extent to which an artificial agent can infer a agent's preferences and beliefs from her choices or decisions i.e. the learning from human feedback question. A variety of settings, ranging from pure learning [21, 11] to sequential and/or cooperative games under partial information[20, 9], as well as the training of Large Language Models were raised. This thesis contributes to this field by establishing some first theoretical foundations for evaluating the statistical value of received feedback and by providing model based insights to algorithmic designers in experimental settings.

In static environments, the bias induced by the presence of randomly missing information has been thoroughly studied [32, 36]. However, in online settings, the dynamics of learning and acting are inherently coupled: since censorship mediates current information of the environment, it impacts the outcome of data-driven decision process; this in turn conditions the future decisions and future censored feedback, creating a complex and endogenous joint temporal dependency. Our work contributes to the analysis of such phenomena for a broad classes of decision and censorship models. Importantly, it is the first normative inquiry of how censorship impacts the statistical complexity of bandit problems. We develop an analysis approach that is useful for both estimating the performance loss due to censorship and refining the classical algorithms designed for uncensored environments. We also arrive at worst-case guarantees on the performance of resulting strategies. This effort contributes to a systematic study of sequential decision making problems in strategic and adversarial environments.

1.2 Related Work

Within the extensive bandits literature, well-surveyed in [29, 42], our work is most closely related to stochastic delayed bandits. Initially, this line of work focused on the joint evolution of actions and information in settings where the reception of the latter is delayed [17]. Of particular interest is the packet loss model recently introduced in [27], which provides the regret bound $\mathcal{O}(\frac{1}{p}R_T)$ where R_T is the uncensored regret and p the censorship probability. Analogous results have been shown in the context of Combinatorial Multi-Armed Bandits with probabilistically triggered arms; see for example, [10] and [46]. Our work provides a systematic approach to study more general censorship models, and sheds light on how the impact of coupled feedback and censorship realizations on the expected regret can be evaluated in terms of the *effective dimension* of the problem.

Importantly, we also tackle the contextual bandit problems, where relatively few results are available on the regret under missing or censored feedback. A notable exception is the work of [45], who focus on a different information structure and obtain a scaling of $1/p$ (see Remark 4). A related contribution by [2] provides both a potential-based analysis of UCB for multi-armed bandits and an algorithmic variant leveraging Kaplan-Meier estimator, although their censorship setting is different than ours. In particular, our results are applicable to settings when delay is significantly large (possibly infinite). This is in contrast to prior results on bandits with delayed information structure which assume either that the delay is *constant*, *upper bounded*, has a *finite mean*, or simply provide regret guarantees that are *linear in the cumulative delays* up to time T [17, 25, 33, 50, 37]. Under such assumptions on delay, one usually gets an additive dependency of the regret in terms of delay parameters.

The abovementioned works primarily focus on modifying well-known bandit algorithms to account for delays, or propose new delay-robust algorithms which may be difficult to implement in practice; a notable exception includes [47]. In our work, we instead focus on estimating the performance loss due to censorship and derive insights on the behavior of well-known UCB class of algorithms [30, 12, 1]. These algorithms are widely used in practice; moreover, their theoretical study has been shown to be useful for analysis of broader class of algorithms (notably Thompson Sampling [3, 40] and Information-Directed Sampling [41, 26]).

Our work contributes to the Generalized Linear Contextual Bandits literature [18, 31] in two ways: Firstly, through the use of these models in a sequential decision-making framework on which the impact of censorship is assessed in Sec. 4. Secondly, by showing that our multi-threshold censorship model $\mathcal{MT}$ induces, at first order, a non-linear structure that closely mirrors such models. Our results provide new tools to study this structure. It is useful to note that the notion of *effective dimension* has been well-studied in the statistical learning and kernels literature [43, 44] (where it is defined for a Gram matrix K_n and regularization λ as $d_{\text{eff}}^n(\lambda) = \text{tr}(K_n(K_n + \lambda \mathbb{I}_d)^{-1})$). Our work shows that an analogous quantity governs the regret bound of bandit problems in censored settings.

The literature on non-stochastic multi-armed bandit problems with delays problems [35, 8, 24] also tackles multiplicative dependency, although in a different setting than ours. Another related line of work is Partial Monitoring [5, 28] which deals with generic categorization of learnability, rather than a fine-grained analysis of dimensionality in relation to censorship. We also mention recent work on computational and statistical properties of estimators that work under truncated or censored samples [14, 15, 16]. Albeit the setting (mostly offline learning with Gaussian noise) and the tools (variant of stochastic gradient descent on modified log-likelihood) are quite different from our work, they showcase a growing interest in the study of censored learning.

Finally, there is a rich literature on classical missing and censored data problems, well surveyed in [32] or [36]. Our work contributes to this broad field since we deal with endogenous and sequentially generated censorship process.

1.3 Summary of Contributions

In Chap. 3, we consider Multi-Armed Bandit (MAB) models and prove that the regret scales as $\tilde{\mathcal{O}}(d_{\mathit{eff}}\sqrt{T})$ (Thm. 3.1.1), where d_{eff} is the effective dimension with value $\sum_{a\in[d]}\frac{1}{p_a}$, mirroring the $\tilde{\mathcal{O}}(d\sqrt{T})$ for uncensored case. In doing so, we recover and generalize related results from [27, 10] to more complex regularized settings and noise models. In particular, we prove that the effective dimension results from characterizing the so-called censored cumulative potential $\mathbb{V}_\alpha$. Our proof methodology easily allows to extend this result to the instance-dependence case, as demonstrated in Prop.3.1.2.

The second part of Chap. 3 focuses on a technical study of the adaptive nature of censorship for $\mathbb{V}_\alpha$. Interestingly, we show in Lemma 3.3.1 and Prop. 3.3.2 that the adaptivity only plays a second order role, that is, impact of censorship can be treated in an *offline* manner at first order. To the best of our knowledge, this set of results brings considerably new insights on the statistical nature of the censorship.

Importantly, our study of MAB under censorship instantiates an analysis framework which extends to Linear Contextual Bandits (LCB) (Chap. 4). Our main result provides that regret is still governed by the effective dimension, but now with a dependency of $\tilde{\mathcal{O}}(\sqrt{d\cdot d_{\mathit{eff}}}\sqrt{T})$ (Thm. 4.1.1), also mirroring the usual $\tilde{\mathcal{O}}(d\sqrt{T})$ for UCB in uncensored case. To the best of our knowledge, these regret bounds provide the first theoretical characterization in LCB with censorship, and contribute to the literature by evaluating the impact of censorship on the performance of UCB-type algorithms. Our second main contribution is identifying the effective dimension for a broad class of multi-threshold models $\mathcal{MT}$ as well as a precise understanding of the dynamic behavior induced by these models (Thm. 4.3.2). In particular, we find that censorship introduces a two-phase behavior: a transient phase during which the initial censoring misspecification is self-corrected at an additional cost; followed by a stationary phase that reflects the inherent slowdown of learning governed by the effective dimension. In extending our analysis from MAB to LCB, we also develop a continuous generalization of the widely used Elliptical Potential Inequality (Prop. 4.2.2), which we believe is also of independent interest.

Chapter 2

Preliminaries

In this chapter, we formally introduce the mathematical framework used throughout this work, as well as key notations. We first formally introduce the classes of Multi-Armed and Contextual bandit problems. We then give a generic definition of the censoring model and instantiate it specifically in chapters 3 and 4. We present the design principle of the upper confidence bound (UCB). We conclude by presenting the notion of regret, the performance criterion.

2.1 Bandit models

We successively consider stochastic multi-armed bandits (MAB) (Chap. 3) and Linear Contextual Bandits (LCB) (Chap. 4) in censored environments. In both settings, at each round $t \leq T$, the agent observes an action set $\mathcal{A}_t \subset \mathcal{A}$. She then selects an action $a_t \in \mathcal{A}_t$ (i.e. an *arm*) to which a noisy feedback $r(a_t) + \epsilon_t$ is associated, where $r(a_t)$ is a bounded reward and ϵ_t is an i.i.d. sub-Gaussian noise of pseudo-variance σ^2. For action a, the sub-optimality gap at time t is denoted $\Delta_t(a) \triangleq \max_{\tilde{a} \in \mathcal{A}_t} r(\tilde{a}) - r(a)$, and the maximal gap $\Delta_{max} \triangleq \max_{a,t} \Delta_t(a)$. We now recall the specifics of each model:

- **MAB**: There is a finite number of actions d, enumerated as $\mathcal{A} \triangleq [d]$, each having a scalar reward $\theta^{\star}_a$. Arms are *independent*: playing one arm gives no information about the others.

- **LCB**: The action set $\mathcal{A}_t$ is a subset of the unit ball $\mathbb{B}_d$, possibly infinite. Non-stochastic contexts are modeled by the fact that $\mathcal{A}_t$ is drawn by an oblivious adversary. Here one does not need to rely on the typical i.i.d assumption on their generating process [50, 18]. Unless explicitly mentioned, the reward is assumed to be linear with respect to a latent unknown vector $\theta^\star \in \mathbb{R}^d$, i.e., $r(a) = \langle a, \theta^\star \rangle$.

2.2 Information Structure and Censorship

In the classical uncensored setting, the noisy feedback is immediately observed post-decision and utilized to make decisions in the next round. We introduce the following **censorship** model: an independent Bernoulli random variable of parameter p_{a_t} denoted as x_{a_t} is drawn after each decision a_t and the feedback is observed, i.e., *realized*, if and only if $x_{a_t} = 1$; else the feedback is said to be *censored*. We recover the uncensored setting when $p_a \equiv 1$. As briefly mentioned above, such censorship can also be seen as an infinite delay. One of the main novelties of this work and a key factor of difficulty is the fact that we allow the censoring probability to depend on the action, i.e. we consider a heterogeneous censoring. We instantiate the precise relationship between the censorship probability and the action chosen a in Chap.3 and 4. While this has already been partially done for the MAB case [19], it is an open challenge in the LCB case. Note also that this generic censorship model slows down the learning but doesn't introduce any bias in the estimation of the latent parameter [27].

Owing to the online nature of the problem where the principal learns endogenously about the environment while acting, the design of the information structure is paramount. More formally, in the case of uncensored bandits, the latter is characterized by the filtration $(\mathcal{F}_t^{NC})_{t\leq T}$ where $\mathcal{F}_t^{NC} \subset \mathcal{F}$ is the sigma algebra generated by $(a_1, r(a_1) + \epsilon_1, \ldots, a_{t-1}, r(a_{t-1}) + \epsilon_{t-1})$. Such actions are in turn generated by a (possibly randomized) policy $\pi \triangleq (\pi_t)_{t\leq T}$ that is $\mathcal{F}_t^{NC}$-adapted. In other word, π_t can be seen as a distribution probability over actions conditioned by $\mathcal{F}_t^{NC}$. In the censored case, the complete information structure can be characterized by the filtration $(\mathcal{F}_t^{C})_{t\leq T}$ where $\mathcal{F}_t^{C} \subset \mathcal{F}$ is the sigma algebra generated by

$(a_1, r(a_1)+\epsilon_1, x_{a_1}, \ldots, a_{t-1}, r(a_{t-1})+\epsilon_{t-1}, x_{a_{t-1}})$. We introduce Π_{adapt} the set of policies time-measurable with respect to this filtration and refer to them as adaptive policies. Yet, as described above, our main focus is on the more realistic setting where a reward is observed conditionally on realization and absence of realization i.e. censorship doesn't convey any information. In other words, we introduce the filtration $(\mathcal{F}_t)_{t \leq T}$ where $\mathcal{F}_t \subset \mathcal{F}$ is the sigma algebra generated recursively by:

$$\mathcal{F}_{t+1} \begin{cases} \mathcal{F}_t, & \text{if } x_{a_t} = 0 \\ \mathcal{F}_t \cup \sigma(a_t, r(a_t) + \epsilon_t), & \text{otherwise} \end{cases}$$

and we shall note $\Pi_{off} \subset \Pi_{adapt}$ the subset of policies time-measurable with respect to this filtration.

2.3 Upper Confidence Bound Algorithms

To study the impact of censorship on bandit problems, we consider the class of high-probability index algorithms based on the *optimism under uncertainty* principle, commonly referred as **UCB**-algorithms. Beyond the fact that they are widely used in practice, their theoretical study has proven to be intimately linked to that of a larger class of algorithms (notably Thompson Sampling and Information Directed Sampling). Algorithm 1 summarizes the generic UCB design framework. We detail bellow the specific instances of UCB for MAB (resp. LCB) used in Chap.3 (resp. Chap.4).

Algorithm 1: Generic UCB

```
Input: Total time $T$, Regularization $\lambda$, Precision $\delta$
for $t = 1, \dots, T$ do
    Provide reward estimator $\tilde{r}_t^\lambda$ verifying w.p. $1-\delta$:
        $\forall a \in \mathcal{A}_t, r(a) \le \tilde{r}_t^\lambda(a)$;
    Play action $a_t = \mathrm{argmax}_{a \in \mathcal{A}_t} \tilde{r}_t^\lambda(a)$ ;
    if $(a_t, r(a_t) + \epsilon_t)$ is realized i.e. $x_{a_t} = 1$ then
        Update $\tilde{r}_t^\lambda$;
    end
end
```

- **UCB-MAB:** Following [29], the UCB algorithms for the MAB case with homogeneous regularization $\lambda > 0$ uses the following optimistic reward estimator at time t:

$$\tilde{r}_t^\lambda(a) \triangleq \hat{\theta}_{t-1}^\lambda(a) + \sqrt{\frac{6\sigma^2 \log(T)}{\lambda + N_a(t-1)}} + \frac{\lambda \|\theta^\star\|_\infty}{\lambda + N_a(t-1)}.$$

It is based on the use of the regularized empirical mean to estimate the reward of action a at the end of round t:

$$\begin{aligned}\hat{\theta}_t^\lambda(a) &\triangleq \frac{1}{N_a(t) + \lambda} \sum_{\tau=1}^{t} (r(a_\tau) + \tau) \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\} \\ &= \frac{N_a(t)}{N_a(t) + \lambda} \theta_a^\star + \frac{1}{N_a(t) + \lambda} \sum_{\tau=1}^{t} \epsilon_{a_\tau} \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}.\end{aligned}$$

The high-confidence property of this algorithm is proven in Lemma 3.5.1.[1] Under a-priori known heteroskedasticity, the reward estimator can be expressed as:

$$\tilde{r}_t^\lambda(a) \triangleq \hat{\theta}_{t-1}^\lambda(a) + \sqrt{\frac{6\sigma_a^2 \log(T)}{\lambda + N_a(t-1)}} + \frac{\lambda \|\theta^\star\|_\infty}{\lambda + N_a(t-1)}.$$

[1] Typically, an upper bound on $\|\theta^\star\|_\infty$ for MAB (resp. $\|\theta^\star\|_2$ for LCB) is used instead of this unknown quantity. We keep $\|\theta^\star\|_\infty$ (resp. $\|\theta^\star\|_2$) not to overload notations but our results immediately extends to the use of the latter.

- **UCB for LCB** Following [1, 29], the UCB algorithms for the LCB case with homogeneous regularization $\lambda > 0$ uses the following optimistic reward estimator at time t:

  $$\tilde{r}^{\lambda}_t(a) \triangleq \langle a, \hat{\theta}^{\lambda}_{t-1} \rangle + \beta_{t-1}(\delta) \|a\|_{\mathbb{W}^C_{t-1}},,$$

  where we introduced the random quantity:

  $$\beta_{t-1}(\delta) \triangleq \sqrt{\sigma^2 \log\left(\frac{\det(\mathbb{W}^C_{t-1})}{\det(\lambda \mathbb{I}_d)}\right) + 2\sigma^2 \log(\frac{1}{\delta})} + \sqrt{\lambda}\|\theta^{\star}\|_2.$$

  It is based on the use of the regularized least square estimator to estimate the vector $\theta^{\star}$ at the end of round t:

  $$\hat{\theta}^{\lambda}_t = (\mathbb{W}^C_t)^{-1} \sum_{\tau=1}^{t} (\epsilon_\tau + \langle a_\tau, \theta^{\star} \rangle) x_{a_\tau} a_\tau$$

  The high-confidence property of this estimator is proven in Lemma 4.6.1. Extension to Generalized Linear Contextual Bandits is discussed in Sec. 4.6.6, where a regularized MLE estimator is used instead of a regularized least square estimator.

2.4 Performance Criterion

The frequentist performance of the agent is measured by the notion of *pseudo regret*, i.e., the difference between the algorithm's cumulative reward and the best total reward. More formally, we introduce for any policy $\pi \in \Pi$:

$$R(T, \pi) = \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} r(a) - \sum_{t=1}^{T} r(a_t) = \sum_{t=1}^{T} \Delta_t(a_t).$$

We aim to provide guarantees on $\mathbb{E}[R(T, \pi)]$ with respect to the number of rounds T and quantities that govern the *complexity* of the problem (for example number of arms, ambient dimension d, parameters of censorship model or smoothness properties of the reward r). Here, the expectation is with respect to the noise induced by the feedback, the censorship and a possibly randomized policy. This work provides both a proof methodology and an an answer to the following open questions:

- How to assess the quantitative impact of the presence of censorship on the expected regret for the UCB class of algorithms ?

- Are the dynamics of learning similar in Multi-Armed and Contextual Bandits model ? More precisely, how does the multiplicity of information acquisition means interacts with the censorship during learning for the latter ?

- What is the normative relationship between the difficulty of the problem i.e. the scaling of the regret and the parameters of the censorship model ? Stated differently, is there a way to characterize the class of censorship models corresponding to a given level of *effective dimension* ?

2.5 Notations

Transpose of a vector u is denoted by $u^\top$, classical Euclidean inner product by $\langle ., . \rangle$ and trace operator by Tr. For positive semi-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ and for any vector $u \in \mathbb{R}^d$, notation $\|u\|_\Sigma$ refers to $\sqrt{u^\top \mathbf{\Sigma} u}$. We use notation $\mathbb{I}_d$ to denote the $d \times d$ identity matrix. $\mathbb{B}_d$ is the unit ball in $\mathbb{R}^d$. $[n]$ is the set of integers $\{1, 2, \cdots, n\}$. For a given function f, we note $f^{(i)}$ the i^{th} derivative of f. To avoid confusion with the dimension d, we use ∂x instead of dx to denote an infinitesimal increase of x. We use the asymptotic notations $\sim$, $\mathcal{O}$, Θ and $\tilde{\mathcal{O}}$ ($\mathcal{O}$ when log factors are removed). Finally, for an event $\mathcal{H}$, we use $\neg\mathcal{H}$ to denote its complement.

To help the reader, we recall the notations used throughout the paper in Tab.2.1,2.2 and 2.3.

Table 2.1: Summary of Notations: Bandit Problem Variables

T	$\triangleq$	Total number of rounds of the sequential decision-making problem.
d	$\triangleq$	Number of arms in Chap.3, Dimension of action feature vector in Chap.4.
$(\mathcal{A}_t, \mathcal{A})$	$\triangleq$	Action set at time t; Union of all action sets $\mathcal{A}_t$.
a_t	$\triangleq$	Action picked at time t; selected by policy π, seen as a function of previous history.
(ϵ_t, σ^2)	$\triangleq$	Stochastic feedback noise a time t. Sub-Gaussian with pseudo-variance parameter σ^2. If σ^2 depends on the action selected (heteroskedasticity), we use σ_a^2 instead.
$(r, \theta^\star)$	$\triangleq$	Unknown reward function, maps action to scalar reward. Parameterized by unknown latent state $\theta^\star$.
$\Delta_t(a)$	$\triangleq$	Sub-optimality gap of action a at time t, reward difference with optimal decision of clairvoyant policy
$(\Delta_a, \Delta_{\max})$	$\triangleq$	If $\Delta_t(a)$ is independent of t, we use $\Delta_a \equiv \Delta_t(a)$. $\Delta_{\max}$ is an upper bound of $\Delta_t(a)$ for all actions a and time t.
$R(T, \pi)$	$\triangleq$	Pseudo regret of policy π over T rounds.

Table 2.2: Summary of Notations: Censorship Variables

p_a	$\triangleq$	Probability that action a is censored if selected, used in Chap. 3. Notation $p(a)$ is used in Chap.4 to emphasize the dependency of p on action a.
(ϕ_j, u, p_j)	$\triangleq$	Parameters of the multi-threshold censorship model. Vector u defines the direction of censorship, $(\phi_j)_{j \leq k+1}$ define the censorship regions with fixed censorship probability and $(p_j)_{j \leq k}$ define the probability of being censored for each region j.
x_{a_t}	$\triangleq$	Random variable indicating if feedback is censored as round t. Follows i.i.d Bernoulli distribution of parameter $p(a_t)$.

Table 2.3: Summary of Notations: Algorithmic and Analysis Variables

λ	$\triangleq$	Regularization tuning parameter. λ_a is used if heterogeneous action-based regularization.
$\tilde{\Delta}_t^{\lambda}(a)$	$\triangleq$	High-probability upper bound on the sub-optimality gap, used in UCB algorithms.
$\mathbb{V}_\alpha(T, \pi)$	$\triangleq$	Random cumulative censored potential, seen as a function of policy π and number of rounds T. First introduced in Chap.3 and extended in Chap.4.
ψ_α	$\triangleq$	Primitive of the function $x \mapsto x^{-\alpha}$, for a given $\alpha > 0$.
$N_a(t)$	$\triangleq$	Total number of time action a is *realized* at the end of round t by policy π. Used in Chap.3.
$\tau_a(t)$	$\triangleq$	Total number of time action a is *played* at the end of round t by policy π. Used in Chap.3.
$\mathbb{W}_t^C$	$\triangleq$	Censored Design Matrix. Linear generalization of $(N_a(t))_{a\in[d]}$. Used in Chap.4.
$\mathbb{W}_t$	$\triangleq$	Expected Design Matrix. Linear generalization of $(p_a\tau_a(t))_{a\in[d]}$. Used in Chap.4.
$\mathbb{W}(t)$	$\triangleq$	Continuous generalization of the expected design matrix$\mathbb{W}_t$.

Chapter 3

Multi-Armed Bandits

In this chapter, we focus specifically on the seminal Multi-Armed Bandit model. To evaluate the impact of censoring on decision making, we instantiate a proof framework based on the study of the *Cumulative Censored Potential* which is naturally extended in Chap.4 for more complex parameters. Thanks to this new methodology, we derive a precise estimate of the performance loss induced and introduce the notion of *effective dimension* resulting from the characterization of this potential. Finally, we perform a statistical analysis of the adaptive nature of censoring and derive precise asymptotic guarantees on the *adaptivity power*.

3.1 Effective Dimension and Regret Bounds

The main result of this section is that censorship effectively enlarges the dimension of the problem. We define the effective dimension as $d_{\mathit{eff}} \triangleq \sum_{a\in[d]} \frac{1}{p_a}$ and our result (Thm. 3.1.1) shows that, at first order, the regret is guaranteed to be the same as the uncensored problem with d_{eff} arms instead of d.

Theorem 3.1.1. *Under censorship, the UCB algorithm with regularization λ has an instance-independent expected regret of:*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d_{\mathit{eff}} T}).$$

Furthermore, we obtain analogous regret guarantees for instance-dependent cases where, at first order, the uncensored dimension $\sum_{a\neq a^\star} \frac{\sigma^2}{\Delta_a}$ enlarges to $\sum_{a\neq a^\star} \frac{\sigma^2}{p_a\Delta_a}$:

Proposition 3.1.2. *For a fixed action set $\mathcal{A}_t \equiv [d]$ and for a-priori known action gap $\Delta_a \triangleq \max_{\tilde{a}} \theta^\star_{\tilde{a}} - \theta^\star_a$, the UCB algorithm with regularization λ has the instance-dependent expected regret:*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq \mathcal{O}\Big(\log(T) \sum_{a\neq a^\star} \frac{1}{p_a} \max(\frac{\sigma^2}{\Delta_a}, \Delta_a)\Big).$$

On one hand, a preliminary understanding of censorship posits an increase of the average "*regret per information gain*" [26] (as it takes longer on average to get the same amount of information) but does not change the underlying complexity of the problem. One the other hand, our results (Thm. 3.1.1 and Prop. 3.1.2) postulate that the censored problem is equivalent at first order to a higher dimensional problem but explored with the same "*regret per information gain*".

The abovementioned results extends to a-priori known heteroskedasticity (see SI). For this general setting, the effective dimension for instance-independent (resp. dependent) case is given by $\sum_a \frac{\sigma_a^2}{p_a}$ (resp. $\sum_{a\neq a^\star} \frac{\sigma_a^2}{p_a\Delta_a}$), where σ_a^2 is the variance proxy of arm a. Although the scaling in $\sum_a \frac{1}{\Delta_a p_a}$ was already mentioned in [27] for unregularized setting with homogeneous variance σ^2 and proven to be optimal, our results generalize these findings.

3.2 Cumulative Censored Potential

We now provide a proof sketch of Thm. 3.1.1, and in doing so, we instantiate an analysis framework that will be extended in Sec. 4. This proof consists in the successive elimination of the noise induced by the feedback and censorship. This leads to regret guarantees on a resulting deterministic quantity by characterizing worst-case learning conditions. The first step of the proof is a variant of the classical reduction of the UCB regret to another quantity we refer to as the *expected cumulative censored potential.* Before stating it, we define at the end of a round $t \in [T]$, the random number of times an arm a has been *pulled* as $\tau_a(t) \triangleq \sum_{l=1}^{t} \mathbf{1}\{a_l = a\}$. Similarly, the number of times an action a has been *realized* at the end of round t is denoted $N_a(t) \triangleq \sum_{l=1}^{t} \mathbf{1}\{a_l = a, x_{a_l} = 1\}$. We then have:

Lemma 3.2.1. *Given an uniform regularization of $\lambda > 0$, the UCB algorithm verifies:*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq 2\sqrt{6\sigma^2 \log(T)}\mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T, \pi_{UCB})] + 2\lambda\|\theta^\star\|_\infty \mathbb{E}[\mathbb{V}_1(T, \pi_{UCB})] + \frac{2d\Delta_{max}}{T}$$

where, for any $\alpha > 0$ and $\pi \in \Pi$, the cumulative potential under censorship is given by:

$$\mathbb{V}_\alpha(T, \pi) = \sum_{t=1}^{T} (N_{a_t}(t-1) + \lambda)^{-\alpha}.$$

In contrast to the classical non-regularized analysis or to the LCB case of Sec. 4, we observe two different orders of α ($1/2$ and 1) coming from the use of the L_∞-norm instead of the L_2-norm. Taken independently, they lead to respective contributions of $\mathcal{O}(d_{\mathit{eff}} \log(T))$ and $\mathcal{O}(\sqrt{d_{\mathit{eff}} T})$. To further study $\mathbb{V}_\alpha$, we introduce the following property:

Proposition 3.2.2. *For all $\alpha > 0$, $\delta \in]0, 1]$ and given ψ_α a primitive of $x \mapsto x^{-\alpha}$, we have:*

$$\max_{\pi \in \Pi} \mathbb{E}[\mathbb{V}_\alpha(T, \pi)] \leq \frac{d_{\mathit{eff}}}{(1-\delta)^\alpha} \left[\psi_\alpha(\frac{T}{d_{\mathit{eff}}} + \frac{\lambda}{1-\delta}) - \psi_\alpha(\frac{\lambda}{1-\delta})\right] + \frac{24 d_{\mathit{eff}} \log(T)}{\lambda^\alpha} + \frac{4 d_{\mathit{eff}}}{\lambda^\alpha \delta^2 T^{12\delta^2}}.$$

The proof of this proposition involves two steps: firstly, we remove the stochastic dependence induced by the censorship through concentration properties (See SI), and we then solve the resulting policy maximization problem (Lemma 3.2.3). In the first step, we consider for a given $\delta \in]0, 1]$ the event:

$$\mathcal{H}_{CEN}(\delta) = \left\{\exists a \in [d], t \in [T], N_a(t) < (1-\delta) p_a \tau_a(t) \quad \text{and} \quad \tau_a(t) \geq T_0(a)\right\},$$

where $T_0(a) \triangleq 24\log(T)/p_a$ and claim that $\mathbb{P}(\mathcal{H}_{CEN}(\delta)) \leq \frac{4d_{\mathit{eff}}}{\delta^2} T^{-12\delta^2}$, improving a result of [27]. Our second step makes use of the following lemma (also known as a *water-filling process* in information theory [13]):

Lemma 3.2.3. *For ψ_α a primitive of $x \mapsto x^{-\alpha}$ where $\alpha \in]0,1]$, regularization $(\lambda_a)_{a\in[d]} \in (\mathbb{R}_{>0})^d$ and censorship vector $(p_a)_{a\in[d]}$, the solution of the optimization problem:*

$$\max_{\tau_1\dots,\tau_d\geq 0} \quad \sum_{a\in[d]} \frac{1}{p_a}\Big(\psi_\alpha(p_a\tau_a+\lambda_a)-\psi_\alpha(\lambda_a)\Big) \quad s.t. \quad \sum_{a\in[d]} \tau_a = T$$

is given by $\tau_a^\star = \frac{1}{p_a}[C-\lambda_a]^+$, where C ensures the total budget constraint $\sum_{a\in[d]}\tau_a^\star = T$. In particular, with $\lambda_{\mathit{eff}} \triangleq \frac{1}{d_{\mathit{eff}}}\sum_{a\in[d]}\frac{\lambda_a}{p_a}$ and $\lambda_a^0 \triangleq d_{\mathit{eff}}(\lambda_a - \lambda_{\mathit{eff}})$, the optimal solution is given by $\tau_a^\star \triangleq \frac{1}{p_a d_{\mathit{eff}}}(T-\lambda_a^0)$ for $T \geq \max_a \lambda_a^0$ and the optimal value is $d_{\mathit{eff}}\psi_\alpha(\frac{T}{d_{\mathit{eff}}}+\lambda_{\mathit{eff}}) - \sum_{a\in[d]}\frac{1}{p_a}\psi_\alpha(\lambda_a)$.

Remark 1. *Note that by working with general α, our analysis naturally extends beyond sub-Gaussian noise to more general assumptions about the Laplace transform of noise (e.g., lighter or heavier tails). Indeed, we note that assuming tails distribution for the reward noise ϵ of the form:*

$$\mathbb{P}\left(\epsilon \geq x\right) \leq \exp\left\{\frac{-x^{1+q}}{2\sigma^2}\right\}$$

for a given $q > 0$, as suggested for instance in [50], would lead the use of the confidence interval:

$$\mathcal{H}_{UCB}^{\lambda,q} = \Big\{\exists a \in [d], t\in[T], |\hat\theta_t^\lambda(a)-\theta_a^\star| > \Big(6\sigma^2\log(T)\Big)^{\frac{1}{1+q}}\Big(\lambda+N_a(t)\Big)^{-\frac{q}{1+q}} + \frac{\lambda\|\theta^\star\|_\infty}{\lambda+N_a(t)}\Big\}.$$

The same reasoning as in the proof of Lemma 3.5.1 would then yield:

$$\begin{aligned}\mathbb{P}\Big(|\frac{\sum_{l=1}^k \epsilon_l}{k+\lambda}| > (6\sigma^2\log(T))^{\frac{1}{1+q}}(k+\lambda)^{-\frac{a}{1+q}}\Big) &= \mathbb{P}\Big(|\sum_{l=1}^k \epsilon_l| > (6\sigma^2(k+\lambda)\log(T))^{\frac{1}{1+q}}\Big)\\ &\leq 2\exp\{-\frac{6\sigma^2(k+\lambda)\log(T)}{2k\sigma^2})\} \leq \frac{2}{T^3}\end{aligned}$$

and therefore $\mathbb{P}(\mathcal{H}_{UCB}^{\lambda,q}) \leq \frac{2d}{T^2}$. For $q=1$, we recover the sub-Gaussian case, which in turns lead to the study of $\mathbb{V}_{1/2}$, as done in Lemma 3.2.1. For general $q>0$, we *would would then consider* $\mathbb{V}_{q/(1+q)}$*, which lead to the upper bound* $\mathcal{O}(d_{\mathit{eff}}^{q/(1+q)}T^{1/(1+q)})$ *through the use of Prop. 3.2.2.*

For unregularized algorithms, this framework can be easily applied to provide instances-dependent guarantees by adding constraints of type $\tau_a \leq f(\Delta_a)$ within Lemma 3.2.3. Optimal guarantees under regularization such as the ones given in Prop. 3.1.2 require however to consider both orders of $\mathbb{V}_\alpha$ ($^1/_2$ and 1) simultaneously and not independently, leading to slight variations as shown in the proof of Prop. 3.1.2 in SI. Next, we further discuss the properties of $\mathbb{V}_\alpha$ given its importance in our analysis and therefore provide additional insights to the main result of this section (Thm. 3.1.1).

3.3 Evaluating Adaptivity Gain

In particular, we seek to gain intuition about how the policies that are adaptive to the realization of censorship process would perform in expectation against a class of non-adaptive (i.e. offline) policies. In order to precisely derive asymptotic behavior of such policies, we introduce and study a continuous counterpart of the discrete original policy maximization problem $\max_{\pi\in\Pi}\mathbb{E}[\mathbb{V}_\alpha(T,\pi)]$. In fact, through the introduction of $\mathcal{H}_{CEN}(\delta)$ and for any $\alpha\in[0,1]$, $\delta\in]0,1]$, we showed in Prop. 3.2.2 the upper bound $\frac{d_{\mathit{eff}}}{(1-\delta)^\alpha}\psi_\alpha(\frac{T}{d_{\mathit{eff}}}+\frac{\lambda}{1-\delta})$ for $\max\mathbb{E}[\mathbb{V}_\alpha(T,\pi)]$ where the maximum is taken over the class of adaptive policies Π_{adapt}, i.e., measurable with respect to the censorship. Note that the exact value of such maximum is notoriously difficult to study due to the adaptive nature of censorship induced by the decision-making process. Interestingly, we obtain a surprising result that the gain due to adaptivity is not significant. Indeed, Lemma 3.3.1 provides the basis for continuous approach in the case of offline policies by leveraging concentration inequalities for inverse Binomial distribution. We then extend this approach in the proof of Prop. 3.3.2. This extension enables us to provide an exact expression for the asymptotic gain of a policy class that monitors the censorship at a single point in time, as well as estimate the gain from fully adaptive policies. More precisely, we introduce Π_{off}, the class of policies that are not adaptive with respect to the censorship and we prove that :

Lemma 3.3.1. *For $\alpha \in]0,1]$ and $\lambda > 0$, we have* $\max_{\pi \in \Pi_{off}} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] \sim d_{\mathit{eff}} \psi_\alpha(\frac{T}{d_{\mathit{eff}}} + \lambda)$.

In other words, restricting attention to offline policies is sufficient to obtain the correct scaling. The next step to complete our claim is the asymptotic expansion:

Proposition 3.3.2. *For $\alpha \in]0,1]$, by denoting* $\gamma_\alpha(\mathbf{p}) \triangleq \frac{\alpha}{2d_{\mathit{eff}}^{1-\alpha}} \sum_{a \in [d]} \frac{1}{p_a}\Big(\sum_{\tilde{a} \neq a} \frac{1-p_{\tilde{a}}}{p_{\tilde{a}}}\Big)$, *we have:*

$$\max_{\pi \in \Pi_{adapt}} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] - \max_{\pi \in \Pi_{off}} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] = \gamma_\alpha(\mathbf{p}) \frac{1}{T^\alpha} + o(\frac{1}{T^\alpha}). \tag{$\star$}$$

Moreover, if for a given $\beta \in]0,1[$, we introduce $\Pi_{single}(\beta T)$ the policy class whose censorship information set has a single updating at time $\lfloor \beta T \rfloor$, we have:

$$\max_{\pi \in \Pi_{single}(\beta T)} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] - \max_{\pi \in \Pi_{off}} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] = \gamma_\alpha(\mathbf{p}) \frac{\beta}{T^\alpha} + o(\frac{1}{T^\alpha}). \tag{$\star\star$}$$

Thus, we find that ($\star\star$), the power of a single monitoring is sufficient to ensure almost the same gain as adaptivity i.e. constant monitoring. The linear dependency in T_0 (due to the linear increase of variance in Binomial models) is also surprising. In non-asymptotic regime, it is still true but for β verifying $0 < \beta_- \leq \beta \leq \beta_+ < 1$ for given (β_-, β_+). We also observe a more general concave property of the single monitoring gain seen as a function of T_0, with limits equals to 0 on the borders on the interval. We conjecture that this concavity is likely to turn in a submodular dependency for several monitoring shots. Moreover, $\gamma_\alpha(\mathbf{p})$ can be viewed as an adaptivity gain resulting from the continuous correction of the cumulative variance induced by the action selection process. Essentially, it is closely related to the Jensen Gap of an appropriate random variable and the proof involves the study of the Taylor expansion of the potential function ψ_α. This shows that censorship in MAB can be treated in an *offline* manner at first order.

3.4 Conclusion of Chap. 3

In this chapter, we considered Multi-Armed Bandit (MAB) models and proved in particular that the regret scales as $\tilde{\mathcal{O}}(d_{\mathit{eff}}\sqrt{T})$ (Thm. 3.1.1), where d_{eff} is the effective dimension with value $\sum_{a\in[d]}\frac{1}{p_a}$, mirroring the $\tilde{\mathcal{O}}(d\sqrt{T})$ for uncensored case. In particular, we prove that the effective dimension results from characterizing the so-called censored cumulative potential $\mathbb{V}_\alpha$. Our proof methodology easily allows to extend this result to the instance-dependence case, as demonstrated in Prop.3.1.2.

In the second part of the chapter, we focused on a technical study of the adaptive nature of censorship for $\mathbb{V}_\alpha$. Interestingly, we show in Lemma 3.3.1 and Prop. 3.3.2 that the adaptivity only plays a second order role, that is, impact of censorship can be treated in an *offline* manner at first order.

3.5 Proof of Chap. 3 - Multi-Armed Bandits

In this section, we prove the results of Chap.3. We start by proving Lemmas 3.2.1, 3.5.1, 3.5.2, 3.2.3 and Prop. 3.2.2. Thanks to those results, we then tackle Thm. 3.1.1 and Prop. 3.1.2. To conclude the section, we further study the properties of the adaptivity gain, by proving Lemma 3.3.1 and Prop. 3.3.2. Recall that effective dimension d_{eff} is referring to $\sum_{a\in[d]}\frac{1}{p_a}$ in this section.

3.5.1 Proof of Lemma 3.2.1

Lemma 3.2.1. *Given an uniform regularization of $\lambda > 0$, the UCB algorithm verifies:*

$$\mathbb{E}[R(T,\pi_{UCB})] \leq 2\sqrt{6\sigma^2\log(T)}\mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T,\pi_{UCB})] + 2\lambda\|\theta^\star\|_\infty\mathbb{E}[\mathbb{V}_1(T,\pi_{UCB})] + \frac{2d\Delta_{max}}{T}$$

where, for any $\alpha > 0$ and $\pi \in \Pi$, the cumulative potential under censorship is given by:

$$\mathbb{V}_\alpha(T,\pi) = \sum_{t=1}^{T}(N_{a_t}(t-1)+\lambda)^{-\alpha}.$$

Proof. At a given round $t \in [T]$, we have under the event $\neg\mathcal{H}^\lambda_{\text{UCB}}$ introduced in Lemma 3.5.1:

$$\Delta_t(a) = \max_{a\in\mathcal{A}_t}\theta^\star_a - \theta^\star_{a_t} \leq 2\sqrt{6\sigma^2\frac{\log(T)}{N_{a_t}(t-1)+\lambda}} + 2\frac{\lambda\|\theta^\star\|_\infty}{\lambda + N_{a_t}(t-1)},$$

where the inequality comes from the definition of the UCB algorithm and the conditioning on $\neg\mathcal{H}^\lambda_{\text{UCB}}$. We find there the origin of the two different orders of N_a ($1/2$ and 1). Taken independently, those lead to a contribution of respectively $\mathcal{O}(d_{eff}\log(T))$ and $\mathcal{O}(\sqrt{d_{eff}T})$. More precisely, we have:

$$\begin{aligned}R(T,\pi_{\text{UCB}}|\neg\mathcal{H}^\lambda_{\text{UCB}}) &\leq 2\sqrt{6\sigma^2\log(T)}\sum_{t=1}^{T}\sqrt{\frac{1}{N_{a_t}(t-1)+\lambda}} + 2\lambda\|\theta^\star\|_\infty\sum_{t=1}^{T}\frac{1}{N_{a_t}(t-1)+\lambda}\\ &= 2\sqrt{6\sigma^2\log(T)}\mathbb{V}_{\frac{1}{2}}(T,\pi_{\text{UCB}}) + 2\lambda\|\theta^\star\|_\infty\mathbb{V}_1(T,\pi_{\text{UCB}}).\end{aligned}$$

Therefore, thanks to Lemma 3.5.1, we deduce that:

$$\begin{aligned}R(T,\pi_{\text{UCB}}) &\le (1-\mathbb{P}(\mathcal{H}^{\lambda}_{\text{UCB}}))R(T,\pi_{\text{UCB}}|\neg\mathcal{H}^{\lambda}_{\text{UCB}}) + \mathbb{P}(\mathcal{H}^{\lambda}_{\text{UCB}})\Delta_{max}T\\ &\le 2\sqrt{6\sigma^2\log(T)}\mathbb{V}_{\frac{1}{2}}(T,\pi_{\text{UCB}}) + 2\lambda\|\theta^{\star}\|_{\infty}\mathbb{V}_1(T,\pi_{\text{UCB}}) + \frac{2d\Delta_{max}}{T}.\end{aligned}$$

Finally, we conclude that:

$$\mathbb{E}[R(T,\pi_{\text{UCB}})] \le 2\sqrt{6\sigma^2\log(T)}\mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T,\pi_{\text{UCB}})] + 2\lambda\|\theta^{\star}\|_{\infty}\mathbb{E}[\mathbb{V}_1(T,\pi_{\text{UCB}})] + \frac{2d\Delta_{max}}{T}.$$

□

3.5.2 Statement and Proof of Lemma 3.5.1

The main step in this reduction from regret to cumulative censored potential is the study of the *failure of optimism* event thanks to the following result:

Lemma 3.5.1. *For a regularization $\lambda > 0$ and $\delta \in]0,1]$, we introduce the event:*

$$\mathcal{H}^{\lambda}_{UCB} = \left\{\exists a \in [d], t \in [T], |\hat{\theta}^{\lambda}_t(a) - \theta^{\star}_a| > \sqrt{\frac{6\sigma^2\log(T)}{\lambda + N_a(t)}} + \frac{\lambda\|\theta^{\star}\|_{\infty}}{\lambda + N_a(t)}\right\}.$$

We then have $\mathbb{P}(\mathcal{H}^{\lambda}_{UCB}) \le \frac{2d}{T^2}$.

Proof. Although this event is similar to the one introduced in the classical UCB proof idea, the subtlety comes from the randomness induced by the censorship as well as the impact of regularization. The main idea is adopt a worst-case agnostic approach. First, let's note that for a given $t \in [T], a \in [d]$, we have:

$$\begin{aligned}|\hat{\theta}^{\lambda}_t(a) - \theta^{\star}_a| &= |\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t}\epsilon_{\tau}\mathbf{1}\{a_{\tau} = a, x_{a_{\tau}} = 1\} - \frac{\lambda}{N_a(t)+\lambda}\theta^{\star}_a|\\ &\le |\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t}\epsilon_{\tau}\mathbf{1}\{a_{\tau} = a, x_{a_{\tau}} = 1\}| + \frac{\lambda}{N_a(t)+\lambda}\|\theta^{\star}\|_{\infty}.\end{aligned}$$

Therefore, for a given $a \in [d], t \in [T]$, by introducing the event $\mathcal{B}_{(t,a)} \triangleq \Big\{|\hat{\theta}^{\lambda}_t(a) - \theta^{\star}_a| >$

$\sqrt{\frac{6\sigma^2 \log(T)}{\lambda+N_a(t)}} + \frac{\lambda\|\theta^\star\|_\infty}{\lambda+N_a(t)}\Big\}$, we deduce:

$$\begin{aligned}
\mathcal{B}_{(t,a)} \subset \Big\{ &|\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}| + \frac{\lambda}{N_a(t)+\lambda}\|\theta^\star\|_\infty > \sqrt{\frac{6\sigma^2\log(T)}{\lambda+N_a(t)}} \\
&+ \frac{\lambda\|\theta^\star\|_\infty}{\lambda+N_a(t)}\Big\} \\
\subset \Big\{ &|\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}| > \sqrt{\frac{6\sigma^2\log(T)}{\lambda+N_a(t)}}\Big\}.
\end{aligned}$$

Then, we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{H}^\lambda_{\text{UCB}}) &= \mathbb{P}\Big(\bigcup_{a\in[d]}\bigcup_{t\in[T]} \mathcal{B}_{(t,a)}\Big) \\
&\le \mathbb{P}\Big(\bigcup_{a\in[d]}\bigcup_{t\in[T]} \Big\{|\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}| > \sqrt{\frac{6\sigma^2\log(T)}{\lambda+N_a(t)}}\Big\}\Big) \\
&\le \sum_{a\in[d]} \mathbb{P}\Big(\bigcup_{t\in[T]} \Big\{|\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}| > \sqrt{\frac{6\sigma^2\log(T)}{\lambda+N_a(t)}}\Big\}\Big) \\
&= \sum_{a\in[d]} \mathbb{P}\Big(\bigcup_{k\in[T],t\in[T]} \Big\{|\frac{1}{N_a(t)+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}|^2 > \frac{6\sigma^2\log(T)}{\lambda+N_a(t)}; N_a(t) = k\Big\}\Big) \\
&= \sum_{a\in[d]}\sum_{k\in[T]} \mathbb{P}(N_a(t) = k) \\
&\qquad \cdot \mathbb{P}\Big(\bigcup_{t\in[T]} \Big\{|\frac{1}{k+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}|^2 > \frac{6\sigma^2\log(T)}{k}\Big| N_a(t) = k\Big\}\Big) \\
&\le \sum_{a\in[d]}\sum_{k\in[T]} \mathbb{P}\Big(\bigcup_{t\in[T]} \Big\{|\frac{1}{k+\lambda}\sum_{\tau=1}^{t} \epsilon_\tau \mathbf{1}\{a_\tau = a, x_{a_\tau} = 1\}|^2 > \frac{6\sigma^2\log(T)}{\lambda+k}\Big| N_a(t) = k\Big\}\Big) \\
&= \sum_{a\in[d]}\sum_{k\in[T]} \mathbb{P}\Big(|\frac{\sum_{l=1}^{k}\epsilon_l}{k+\lambda}|^2 > \frac{6\sigma^2\log(T)}{\lambda+k}\Big),
\end{aligned}$$

where we successively used union bounds over the action set and number of realizations and conditioned over number of realizations k. We re-indexed the random sub-Gaussian variables (ϵ_t) for last expression thanks to the i.i.d property. Then, for a given k, using

Hoeffding inequality for sub-Gaussian variables, we have:

$$\mathbb{P}\Big(|\frac{\sum_{l=1}^k \epsilon_l}{k+\lambda}|^2 > \frac{6\sigma^2\log(T)}{k+\lambda}\Big) = \mathbb{P}\Big(|\sum_{l=1}^k \epsilon_l| > \sqrt{6\sigma^2(k+\lambda)\log(T)}\Big)$$

$$\leq 2\exp\{-\frac{6\sigma^2(k+\lambda)\log(T)}{2k\sigma^2}\} \leq \frac{2}{T^3},$$

where the used that fact that $\sum_{l=1}^k \epsilon_l$ is sub-Gaussian of pseudo-variance parameter $k\sigma^2$ Therefore, this yields:

$$\sum_{a\in[d]}\sum_{k\in[T]}\mathbb{P}\Big(|\frac{\sum_{l=1}^k \epsilon_l}{k+\lambda}|^2 > \frac{6\sigma^2\log(T)}{k+\lambda}\Big) \leq \frac{2d}{T^2}.$$

Finally, we conclude that $\mathbb{P}(\mathcal{H}_{\text{UCB}}^{\lambda}) \leq \frac{2d}{T^2}$. □

3.5.3 Statement and Proof of Lemma 3.5.2

Lemma 3.5.2. *For any $\delta\in]0,1]$, $\lambda>0$ and censorship model, let's introduce the event:*

$$\mathcal{H}_{CEN}^I(\delta) = \{\exists a\in[d], t\in[T], N_a(t) < (1-\delta)p_a\tau_a(t) \quad \textit{and} \quad \tau_a(t)\geq T_0(a)\},$$

where $T_0(a) \triangleq 24\log(T)/p_a$. We then have $\mathbb{P}(\mathcal{H}_{CEN}^I(\delta)) \leq \frac{4d_{\mathit{eff}}}{\delta^2}T^{-12\delta^2}$.

Proof. First, we apply successively two unions bounds over the action set and the number of realizations, mirroring the analysis of [27]:

$$\begin{aligned}\mathbb{P}(\mathcal{H}_{\text{CEN}}^I(\delta)) &\leq \sum_{a\in[d]}\mathbb{P}\Big(\Big\{\exists t\in[T], \tau_a(t)\geq T_0(a), N_a(t) < (1-\delta)p_a\tau_a(t)\Big\}\Big)\\ &= \sum_{a\in[d]}\mathbb{P}\Big(\bigcup_{k_a\in[T_0(a),T]}\bigcup_{t\in[T]}\Big\{\tau_a(t)\geq T_0(a), N_a(t) < (1-\delta)p_a\tau_a(t), \tau_a(t)=k_a\Big\}\Big)\\ &\leq \sum_{a\in[d]}\sum_{k_a\geq T_0(a)}\mathbb{P}\Big(\bigcup_{t\in[T]}\Big\{N_a(t) < (1-\delta)p_a\tau_a(t)\Big|\tau_a(t)=k_a\Big\}\Big).\end{aligned}$$

We then use a multiplicative Chernoff inequality for Binomial Distribution to deduce:

$$\sum_{a\in[d]}\sum_{k_a\geq T_0(a)} \mathbb{P}\Big(N_a(t) < (1-\delta)p_a\tau_a(t)\Big|\tau_a(t)=k_a\Big) \leq \sum_{a\in[d]}\sum_{k_a\geq T_0(a)} \exp\{-\frac{\delta^2 k_a p_a}{2}\}.$$

The novelty of our proof is to leverage a integral comparison to deduce the improved control:

$$\sum_{a\in[d]}\sum_{k_a\geq T_0(a)} \exp\{-\frac{\delta^2 k_a p_a}{2}\} \leq 2\sum_{a\in[d]}\left[-\frac{2}{\delta^2 p_a}\exp\{-\frac{\delta^2 k_a p_a}{2}\}\right]_{T_0(a)-1}^{\tau_a(t)}$$
$$\leq \frac{4}{\delta^2}d_{\mathit{eff}}\frac{1}{T^{12\delta^2}} - \frac{4}{\delta^2}\sum_{a\in[d]}\frac{1}{p_a}\exp\{-\frac{\delta^2\tau_a(t)p_a}{2}\} \leq \frac{4}{\delta^2}d_{\mathit{eff}}\frac{1}{T^{12\delta^2}}.$$

Picking for instance $\delta=\frac{1}{2}$ yields $\mathbb{P}(\mathcal{H}^I_{\text{CEN}}(\frac{1}{2})) \leq \frac{16 d_{\mathit{eff}}}{T^3}$. □

3.5.4 Proof of Lemma 3.2.3

Lemma 3.2.3. *For ψ_α a primitive of $x\mapsto x^{-\alpha}$ where $\alpha\in]0,1]$, regularization $(\lambda_a)_{a\in[d]}\in(\mathbb{R}_{>0})^d$ and censorship vector $(p_a)_{a\in[d]}$, the solution of the optimization problem:*

$$\max_{\tau_1\dots,\tau_d\geq 0}\quad \sum_{a\in[d]}\frac{1}{p_a}\Big(\psi_\alpha(p_a\tau_a+\lambda_a)-\psi_\alpha(\lambda_a)\Big)\quad s.t.\quad \sum_{a\in[d]}\tau_a = T$$

is given by $\tau_a^\star=\frac{1}{p_a}[C-\lambda_a]^+$, where C ensures the total budget constraint $\sum_{a\in[d]}\tau_a^\star = T$. In particular, with $\lambda_{\mathit{eff}}\triangleq\frac{1}{d_{\mathit{eff}}}\sum_{a\in[d]}\frac{\lambda_a}{p_a}$ and $\lambda_a^0\triangleq d_{\mathit{eff}}(\lambda_a-\lambda_{\mathit{eff}})$, the optimal solution is given by $\tau_a^\star\triangleq\frac{1}{p_a d_{\mathit{eff}}}(T-\lambda_a^0)$ for $T\geq\max_a\lambda_a^0$ and the optimal value is $d_{\mathit{eff}}\psi_\alpha(\frac{T}{d_{\mathit{eff}}}+\lambda_{\mathit{eff}})-\sum_{a\in[d]}\frac{1}{p_a}\psi_\alpha(\lambda_a)$.

Proof. We first introduce the Lagrangian of the problem $\mathcal{L}(\tau_1,\dots,\tau_d,\mu):=\sum_{a\in[d]}\frac{1}{p_a}\Big(\psi_\alpha(p_a\tau_a+\lambda_a)-\psi_\alpha(\lambda_a)\Big)+\mu(T-\sum_{a\in[d]}\tau_a)$. Differentiating with respect to τ_a for all $a\in[d]$ yields the equations:

$$\frac{1}{(p_a\tau_a+\lambda_a)^\alpha}-\mu=0.$$

We then write it equivalently as:

$$\tau_a = \frac{1}{p_a}[\mu^{-1/\alpha} - \lambda_a].$$

However, since (τ_a) must be nonnegative, it may not always be possible to find a solution of this form. We then verify using KKT conditions that the solution:

$$\tau_a = \frac{1}{p_a}[C - \lambda_a]^+,$$

where C ensures the total budget constraint $\sum_{a\in[d]} \tau_a^\star = T$, is optimal. In particular, whenever $T \geq \max_a \lambda_a^0$, we recover the solution provided in the second part the Lemma. □

3.5.5 Proof of Prop. 3.2.2

Proposition 3.2.2. *For all $\alpha > 0$, $\delta \in]0,1]$ and given ψ_α a primitive of $x \mapsto x^{-\alpha}$, we have:*

$$\max_{\pi\in\Pi} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] \leq \frac{d_{\mathit{eff}}}{(1-\delta)^\alpha}\left[\psi_\alpha(\frac{T}{d_{\mathit{eff}}} + \frac{\lambda}{1-\delta}) - \psi_\alpha(\frac{\lambda}{1-\delta})\right] + \frac{24 d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + \frac{4d_{\mathit{eff}}}{\lambda^\alpha\delta^2 T^{12\delta^2}}.$$

Proof. For a given $\alpha \in]0,1]$, we condition on the event $\mathcal{H}^I_{\mathrm{CEN}}(\delta)$ introduced in Lemma 3.5.2 and consider the cases $\tau_a(t) \geq T_0(a)$ and $\tau_a(t) < T_0(a)$. This yields for any policy $\pi \in \Pi$:

$$\begin{aligned}
\mathbb{V}_\alpha(T,\pi|\mathcal{H}^I_{\mathrm{CEN}}(\delta)) &\leq \frac{\sum_{a\in[d]} T_0(a)}{\lambda^\alpha} + \sum_{t=1}^{T}((1-\delta)p_{a_t}\tau_{a_t}(t-1)+\lambda)^{-\alpha} \\
&\leq \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + \frac{1}{(1-\delta)^\alpha}\sum_{t=1}^{T}\left(p_{a_t}\tau_{a_t}(t-1) + \frac{\lambda}{1-\delta}\right)^{-\alpha} \\
&\leq \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + \frac{1}{(1-\delta)^\alpha}\sum_{a\in[d]}\int_0^{\tau_a(T)}\left(p_a u + \frac{\lambda}{1-\delta}\right)^{-\alpha}\partial u \\
&= \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + \frac{1}{(1-\delta)^\alpha}\sum_{a\in[d]}\frac{1}{p_a}[\psi_\alpha(p_a\tau_a(T) + \frac{\lambda}{1-\delta}) - \psi_\alpha(\frac{\lambda}{1-\delta})].
\end{aligned}$$

We then apply the Lemma 3.2.3 with constant $\tilde{\lambda} \triangleq \lambda/(1-\delta)$ to deduce:

$$\max_{\pi\in\Pi} \mathbb{V}_\alpha(T,\pi|\neg\mathcal{H}^I_{\text{CEN}}(\delta)) \leq \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + \frac{d_{\mathit{eff}}}{(1-\delta)^\alpha}\Big[\psi_\alpha(\frac{T}{d_{\mathit{eff}}} + \frac{\lambda}{1-\delta}) - \psi_\alpha(\frac{\lambda}{1-\delta})\Big].$$

Then, we conclude thanks to Lemma 3.5.2 that:

$$\begin{aligned}\max_{\pi\in\Pi} \mathbb{E}[\mathbb{V}_\alpha(T,\pi)] &\leq \mathbb{P}(\neg\mathcal{H}^I_{\text{CEN}}(\delta)) \max_{\pi\in\Pi} \mathbb{V}_\alpha(T,\pi|\neg\mathcal{H}^I_{\text{CEN}}(\delta)) + (1-\mathbb{P}(\neg\mathcal{H}^I_{\text{CEN}}(\delta)))\frac{1}{\lambda^\alpha}\\ &\leq \frac{1}{(1-\delta)^\alpha} d_{\mathit{eff}}\left[\psi_\alpha(\frac{T}{d_{\mathit{eff}}} + \frac{\lambda}{1-\delta}) - \psi_\alpha(\frac{\lambda}{1-\delta})\right] + \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha}\\ &+ \frac{4}{\delta^2} d_{\mathit{eff}} \frac{1}{\lambda^\alpha T^{12\delta^2}}.\end{aligned}$$

In particular, for $\alpha = 1$ and $\delta = \frac{1}{2}$, this involves:

$$\max_{\pi\in\Pi} \mathbb{E}[\mathbb{V}_1(T,\pi)] \leq 2d_{\mathit{eff}}\log(\frac{T}{2\lambda}+1) + \frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha} + 16d_{\mathit{eff}}\frac{1}{\lambda T^2},$$

and for $\alpha = \frac{1}{2}$ and $\delta = \frac{1}{2}$, this yields:

$$\max_{\pi\in\Pi} \mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T,\pi)] \leq \sqrt{2}d_{\mathit{eff}}\left[\sqrt{\frac{T}{d_{\mathit{eff}}} + 2\lambda} - \sqrt{2\lambda}\right] + \frac{24d_{\mathit{eff}}\log(T)}{\sqrt{\lambda}} + 16d_{\mathit{eff}}\frac{1}{\sqrt{\lambda}T^2}.$$

□

3.5.6 Proof of Thm. 3.1.1

Theorem 3.1.1. *Under censorship, the UCB algorithm with regularization λ has an instance-independent expected regret of:*

$$\mathbb{E}[R(T,\pi_{UCB})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d_{\mathit{eff}}T}).$$

Proof. We first apply Lemma 3.2.1 to deduce:

$$\begin{aligned}\mathbb{E}[R(T,\pi_{\text{UCB}})] &\leq 2\sqrt{6\sigma^2\log(T)}\mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T,\pi_{\text{UCB}})] + 2\lambda\|\theta^\star\|_\infty\mathbb{E}[\mathbb{V}_1(T,\pi_{\text{UCB}})] + \frac{2d\Delta_{max}}{T}\\ &\leq 2\sqrt{6\sigma^2\log(T)} \max_{\pi\in\Pi}\mathbb{E}[\mathbb{V}_{\frac{1}{2}}(T,\pi)] + 2\lambda\|\theta^\star\|_\infty \max_{\pi\in\Pi}\mathbb{E}[\mathbb{V}_1(T,\pi)] + \frac{2d\Delta_{max}}{T}.\end{aligned}$$

We then apply proposition 3.2.2, with $\delta = 1/2$ in order to deduce:

$$\mathbb{E}[R(T,\pi_{\mathrm{UCB}})] \leq 2\sqrt{6\sigma^2\log(T)}\Big(\sqrt{2}d_{\mathit{eff}}\Big[\sqrt{\frac{T}{d_{\mathit{eff}}}+2\lambda}-\sqrt{2\lambda}\Big]+\frac{24d_{\mathit{eff}}\log(T)}{\sqrt{\lambda}}+16d_{\mathit{eff}}\frac{1}{\sqrt{\lambda}T^2}\Big)$$
$$+2\lambda\|\theta^\star\|_\infty\Big(2d_{\mathit{eff}}\log\Big(\frac{T}{2\lambda}+1\Big)+\frac{24d_{\mathit{eff}}\log(T)}{\lambda^\alpha}+16d_{\mathit{eff}}\frac{1}{\lambda T^2}\Big)+\frac{2d\Delta_{max}}{T}.$$

By taking $\lambda = o(\log(T))$ and considering only the leading order, we conclude that:

$$\mathbb{E}[R(T,\pi_{\mathrm{UCB}})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d_{\mathit{eff}}T}).$$

Note that our proof easily allows to get high-probability bounds on regret instead of bounds on its expected value. □

3.5.7 Proof of Prop. 3.1.2

Proposition 3.1.2. *For a fixed action set $\mathcal{A}_t \equiv [d]$ and for a-priori known action gap $\Delta_a \triangleq \max_{\tilde{a}}\theta^\star_{\tilde{a}}-\theta^\star_a$, the UCB algorithm with regularization λ has the instance-dependent expected regret:*

$$\mathbb{E}[R(T,\pi_{UCB})] \leq \mathcal{O}\Big(\log(T)\sum_{a\neq a^\star}\frac{1}{p_a}\max(\frac{\sigma^2}{\Delta_a},\Delta_a)\Big).$$

Proof. As in the proof of Lemma 3.2.2, for a given round $t\in[T]$, we have under the event $\neg\mathcal{H}^\lambda_{\mathrm{UCB}}$

$$\Delta_a = \max_{\tilde{a}\in\mathcal{A}}\theta^\star_{\tilde{a}}-\theta^\star_a \leq 2\sqrt{6\sigma^2\frac{\log(T)}{N_{a_t}(t-1)+\lambda}}+2\frac{\lambda\|\theta^\star\|_\infty}{\lambda+N_{a_t}(t-1)}.$$

It is as an inequality of the second degree and thus for any $t\in[T]$, $a\in[d]$:

$$x_1\left(\sqrt{\frac{1}{\lambda+N_a(t)}}\right)^2+x_2\sqrt{\frac{1}{\lambda+N_a(t)}}-\Delta_a\geq 0,$$

where $x_1 = 2\lambda\|\theta^\star\|_\infty$ and $x_2 = 2\sqrt{6\sigma^2\log(T)}$. Solving it yields:

$$\sqrt{\frac{1}{\lambda + N_a(t)}} \geq \frac{1}{2x_1}(-x_2 + \sqrt{x_2^2 + 4\Delta_a x_1}),$$

or equivalently:

$$N_a(t) \leq \Big(\frac{4\lambda\|\theta^\star\|_\infty}{\sqrt{24\sigma^2\log(T) + 8\Delta_a\lambda\|\theta^\star\|_\infty} - \sqrt{24\sigma^2\log(T)}}\Big)^2 - \lambda.$$

Therefore, under $\neg\mathcal{H}^I_{\text{CEN}}(\frac{1}{2})$, we have:

$$\tau_a(t) \leq \max(T_0(a), \frac{2}{p_a}\Big(\frac{4\lambda\|\theta^\star\|_\infty}{\sqrt{24\sigma^2\log(T) + 8\Delta_a\lambda\|\theta^\star\|_\infty} - \sqrt{24\sigma^2\log(T)}}\Big)^2 - \lambda).$$

This yields a conditional regret of:

$$\begin{aligned} R(T|\neg(\mathcal{H}^I_{\text{CEN}}(\frac{1}{2}) \cup \mathcal{H}^\lambda_{\text{UCB}})) &\leq \sum_{a\in[d], a\neq a^\star} \Delta_a \tau_a(t) \\ &= \sum_{a\in[d], a\neq a^\star} \frac{2\Delta_a}{p_a}\max(12\log(T), \Big(\frac{4\lambda\|\theta^\star\|_\infty}{\sqrt{24\sigma^2\log(T) + 8\Delta_a\lambda\|\theta^\star\|_\infty} - \sqrt{24\sigma^2\log(T)}}\Big)^2 - \lambda), \end{aligned}$$

where $a^\star \triangleq \operatorname{argmax}_{\tilde{a}\in\mathcal{A}} \theta^\star_{\tilde{a}}$ and an expected regret of:

$$\begin{aligned} \mathbb{E}[R(T, \pi_{UCB})] \leq \sum_{a\in[d], a\neq a^\star} \frac{2\Delta_a}{p_a}\max(12\log(T), \Big(\frac{4\lambda\|\theta^\star\|_\infty}{\sqrt{24\sigma^2\log(T) + 8\Delta_a\lambda\|\theta^\star\|_\infty} - \sqrt{24\sigma^2\log(T)}}\Big)^2 \\ - \lambda) + \frac{d\Delta_{max}}{T} + \frac{16 d_{\mathit{eff}}\Delta_{max}}{T^2}. \end{aligned}$$

In particular, for the regularization $\lambda = o(\log(T))$, we have the asymptotic:

$$\Big(\frac{4\lambda\|\theta^\star\|_\infty}{\sqrt{24\sigma^2\log(T) + 8\Delta_a\lambda\|\theta^\star\|_\infty} - \sqrt{24\sigma^2\log(T)}}\Big)^2 = \frac{24\sigma^2\log(T)}{\Delta_a^2} + \frac{8\lambda\|\theta^\star\|_\infty}{2\Delta_a} + o(1).$$

And thus, we conclude that:

$$\mathbb{E}[R(T, \pi_{UCB})] \leq \mathcal{O}\Big(\log(T)\sum_{a\in[d], a\neq a^\star} \frac{1}{p_a}\max(\frac{\sigma^2}{\Delta_a}, \Delta_a)\Big).$$

Again, note that our proof easily allows to get high-probability bounds on regret instead of bounds on its expected value. □

Remark 2. *As in the instance-independent case, previous reasoning immediately extends to a-priori known heteroskedasticity and yields the upper bound:*

$$\mathbb{E}[R(T,\pi_{UCB})] \leq \mathcal{O}\Big(\log(T)\sum_{a\in[d],a\neq a^\star}\frac{1}{p_a}\max(\frac{\sigma_a^2}{\Delta_a},\Delta_a)\Big).$$

Chapter 4

Contextual Bandits

In this chapter, we study Linear Contextual Bandits under censorship. The regret analysis for this general setting under censorship is significantly more complex than for the MAB model. This is due among other to the multiplicity of information acquisition means: observing the reward of a given action allows to partially learn the reward of some others. Henceforth, this leads to a complex trade-off between the information gain and censorship probability of an action. Nevertheless, by extending the analysis of Chap.3, we derive an equivalent notion of effective dimension for a broad class of censorship models and characterize new insights on the impact of censorship on sequential decision making.

4.1 Multi-threshold Censorship Models and Regret Bounds

We now introduce the class of multi-threshold censorship models defined as:

$$p : a \in \mathbb{B}_d \mapsto \sum_{j=0}^{k} \mathbf{1}\{\sin(\phi_j) \leq \langle a, u \rangle < \sin(\phi_{j+1})\} p_j, \tag{$\mathcal{MT}$}$$

where $(\phi_j)_{j \leq k+1}$ is an increasing sequence verifying $\phi_0 = -\frac{\pi}{2}$, $\phi_{k+1} = \frac{\pi}{2}$ and $u \in \mathbb{R}^d$ is a unit vector. We assume that $(p_j)_{j \leq k}$ is decreasing, i.e. the censorship is increasing

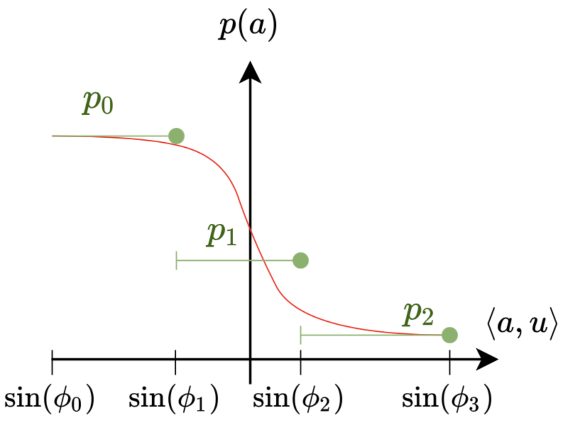

Figure 4-1: Example of a multi-threshold model for $k = 2$ (Green). Logistic censorship model (Red).

with j in direction u. Henceforth, we refer to the interval $[\sin(\phi_j), \sin(\phi_{j+1})[$ as *region* j. Note that simple models such as uniform censorship are subsumed by this family (for k equals 0). Furthermore, $\mathcal{MT}$ can be seen as a piecewise constant approximation of any generalized linear model [34]. Thus, we do not see this modeling abstraction as an inherently limiting factor on the generality of our subsequent results.

Moreover, $\mathcal{MT}$ admits a natural behavioral interpretation: Such a distribution can be seen as induced by a population model of heterogeneous random-utility maximizers agents. A single threshold model (i.e. k equals 1) corresponds to a given agent type, and the multi-threshold model naturally results from aggregate responses of heterogeneous population [4]. We now state the main result of this section:

Theorem 4.1.1. *For a given multi-threshold censorship model $\mathcal{MT}$, there exits $d_{\textit{eff}}$ such that the UCB algorithm with regularization λ has an instance-independent expected regret of:*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d \cdot d_{\textit{eff}}}\sqrt{T}).$$

Note that the mapping from the original dimension d to the enlarged $\sqrt{d \cdot d_{\textit{eff}}}$ is more surprising than previous dilation $d \mapsto d_{\textit{eff}}$ in the case of MAB problems. An extension to Generalized Linear Contextual Bandits is provided in the SI where we show that the dimension is governed by $\sqrt{d \cdot d_{eff}}/\kappa$, with κ corresponding to a minimum of the derivative of the link function (encompassing the smoothness of the generalized linear model at its maximum) [31, 18]. We conjecture that this result and the notion of *effective dimension* can be extended to more general censorship models as long as the radial property of censorship is verified i.e. p only depends on the action a through $\langle a, u \rangle$ for a given potentially time dependent vector u.

4.2 Generalized Cumulative Censored Potential

Analogous to the MAB case, we now introduce for LCB the random matrices corresponding to the effective realization $\mathbb{W}_t^C \triangleq \lambda \mathbb{I}_d + \sum_{n=1}^{t} x_{a_t} a_t a_t^\top$ and the expected realization $\mathbb{W}_t \triangleq \lambda \mathbb{I}_d + \sum_{n=1}^{t} p(a_t) a_t a_t^\top$. We also introduce the continuous counterpart of $\mathbb{W}_t$ defined as $\mathbb{W}(t) \triangleq \lambda \mathbb{I}_d + \int_{u=0}^{t} p(a(u)) a(u) a(u)^\top \partial u$, where $(a(u))_{u \leq T}$ is an integrable deterministic path.[1] We emphasise that the use of continuous counterpart is key in enabling our next results. As in the MAB case, we bound the regret although now using a generalization of $\mathbb{V}_\alpha$:

Lemma 4.2.1. *For all $\delta \in]0, 1]$, there exists a constant $\tilde{\beta}_\delta(T) = \Theta(\sqrt{d \log(T)})$ such that*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq 2\tilde{\beta}_\delta(T) \sqrt{T \mathbb{E}[\mathbb{V}_1(T, \pi_{UCB})]} + \delta T \Delta_{max},$$

where, for $\alpha > 0$ and $\pi \in \Pi$, the linear extension of the cumulative censored potential is given by:

$$\mathbb{V}_\alpha(T, \pi) \triangleq \sum_{t=1}^{T} \|a_t\|^2_{(\mathbb{W}_{t-1}^C)^{-\alpha}} = \sum_{t=1}^{T} \mathrm{Tr}((\mathbb{W}_{t-1}^C)^{-\alpha} a_t a_t^\top).$$

The proof idea is analogous (albeit more complex) than in the finite action case (see SI). In order to get a handle on $\mathbb{V}_\alpha$, we again leverage a two-step approach: first we eliminate the randomness due to censorship (here, we utilize matrix martingale

[1] In this section, the generic notation $X(t)$ is used for continuous time quantities and X_t for discrete time.

inequalities) and then optimize the resulting deterministic quantity seen through a continuous lens. The first step requires the following result:

Proposition 4.2.2. *For any $\delta \in]0,1]$, $\lambda > 0$, $\alpha > 0$ and policy $\pi \in \Pi$, we have:*

$$\mathbb{E}[V_\alpha(T,\pi)] \leq \frac{\delta}{\lambda^\alpha} + C(\delta)^\alpha \operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-\alpha} a(t)a(t)^\top \partial t\Big),$$

where $C(\delta) \triangleq 8(\lambda+1)\max(\log(d/\delta))/\lambda, 1)/\lambda$.

Remark 3. *The key idea of this result is to observe that the telescopic sum on which the classical Elliptical Potential lemma [1, 39, 7] heavily relies on is, in fact, the discrete approximation of an integral over a matrix path.*[2] *For the simpler case of classical uncensored environment, we obtain for $\alpha > 0, \alpha \neq 1$:*

$$\sum_{t=1}^T \|a_t\|^2_{\mathbb{W}_{t-1}^{-\alpha}} \leq \Big(\frac{\lambda+1}{\lambda}\Big)^\alpha \frac{\operatorname{Tr}\Big(\int_0^T \partial \mathbb{W}(t)^{1-\alpha}\Big)}{1-\alpha} = \Big(\frac{\lambda+1}{\lambda}\Big)^\alpha \frac{\operatorname{Tr}(\mathbb{W}_T^{1-\alpha} - \mathbb{W}_0^{1-\alpha})}{1-\alpha}.$$

For $\alpha = 1$, a similar reasoning is applied using the formula $\operatorname{Tr}(\log(A)) = \log(\det A)$:

$$\begin{aligned}\sum_{t=1}^T \|a_t\|^2_{\mathbb{W}_{t-1}^{-1}} &\leq \frac{\lambda+1}{\lambda}\int_0^T \frac{\partial \log\det(\mathbb{W}(t))}{\partial t}\partial t = \frac{\lambda+1}{\lambda}\operatorname{Tr}(\log \mathbb{W}_T - \log \mathbb{W}_0) \\ &\leq \frac{\lambda+1}{\lambda}\log\frac{\det \mathbb{W}_T}{\det \mathbb{W}_0}.\end{aligned}$$

A deeper study of the eigenvalues of $\mathbb{W}_T^{1-\alpha}$ then yields the worst-case upper bound $d^\alpha(d\lambda+T)^{1-\alpha}/(1-\alpha)$ for $\alpha < 1$ and $d\lambda^{1-\alpha}/(\alpha-1)$ for $\alpha > 1$, recovering more naturally and extending the results of [7]. Thus, analogous to the water filling process highlighted in the MAB case in Lemma 3.2.3, we now consider a spectral water-filling process [13] for the eigenvalues of $\psi_\alpha(\mathbb{W}_T)$ with a slight abuse of notations ($\mathbb{W}_T^{1-\alpha}$ and $\log \mathbb{W}_T$ in this discussion).

One of the main challenge introduced by the censorship is therefore to identify a suitable matrix operator on which the spectral maximization can be performed.

[2] Note that the rank 1 assumption is not needed in the continuous relaxation and therefore our results still hold whenever $a(t)a(t)^T$ is replaced by any positive semi-definite matrix $H(t)$.

Motivated by Lemma 4.2.1, we henceforth focus on the case of $\alpha = 1$ for which Prop. 4.2.2 implies that for any policy:

$$\mathrm{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1} a(t) a(t)^\top \partial t\Big) = \int_0^T \frac{1}{p(a(t))} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t.$$

4.3 Effective Dimension in Linear Settings

Again, the notion of effective dimension naturally appears. We now highlight intuition provided by this quantity, and then present its general study for the multi-threshold model $\mathcal{MT}$.

Remark 4. *Let us consider an uniform censorship model $p : a \mapsto \bar{p}$. By leveraging the case of equality in the Arithmetic-Geometric inequality applied to the eigenvalues of $\mathbb{W}_T$, we then simply deduce the associated effective dimension $d_{\mathit{eff}} \triangleq d/\bar{p}$:*

$$\max_{\pi \in \Pi} \int_0^T \frac{1}{\bar{p}} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t = d_{\mathit{eff}} \log(1 + \frac{T}{\lambda d_{\mathit{eff}}}).$$

We next illustrate the logarithmic scaling of this quantity in the general case as well as the importance of the leading dimension factor, crudely upper bounded by d/p_{min} in the next lemma:

Lemma 4.3.1. *For any censorship function p, by introducing lower and upper bounds (p_{min}, p_{max}) of p, we have:*

$$\frac{d}{p_{max}} \log(1 + \frac{p_{min} T}{d\lambda}) \leq \max_{\pi \in \Pi} \int_0^T \frac{1}{p(a(t))} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t \leq \frac{d}{p_{min}} \log(1 + \frac{p_{max} T}{d\lambda}).$$

Related problems in the generalized linear models literature [50, 31, 18] are implicitly solved in the spirit of Lemma 4.3.1, where a minimum of the derivative of the link function plays the role of p_{min} above. However, when the function p varies with action a, a more careful analysis is required to derive useful dimensional bounds. Our next major result addresses this gap in the literature by improving the bounds provided in Lemma 4.3.1:

Theorem 4.3.2. *For a multi-threshold censorship model $\mathcal{MT}$, we have:*

$$\max_{\pi\in\Pi} \int_0^T \frac{1}{p(a(t))} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t = d_{\mathit{eff}} \log(T) + o(\log(T)), \qquad (\mathcal{P})$$

where d_{eff} is the effective dimension. Furthermore, d_{eff} is characterized by two cases:

- ***Case 1:*** *Single region j effective dimension $d_{\mathit{eff}} = \frac{d}{p_j}$.*

- ***Case 2:*** *Bi-region (i,j) effective dimension, with $i < j$:*

$$d_{\mathit{eff}} = \frac{1}{p_j}\left[(d-1)\frac{1-l(i,j)}{\frac{p_i}{p_j} - l(i,j)} + \frac{u(i,j)-1}{u(i,j) - \frac{p_i}{p_j}}\right] < \frac{d}{p_j}. \qquad (\mathcal{D})$$

where $l(i,j) \triangleq \frac{\sin^2(\phi_i)}{\sin^2(\phi_j)}$ and $u(i,j) \triangleq \frac{\cos^2(\phi_i)}{\cos^2(\phi_j)}$.

Before moving to a discussion of the proof idea behind Thm. 4.3.2, we mention a few important remarks. First, a necessary condition for the bi-region (i,j) effective dimension to arise is the constraint on $\frac{p_i}{p_j}$:

$$\max(1, \underbrace{\frac{dl(i,j)u(i,j)}{u(i,j)+(d-1)l(i,j)}}_{\triangleq s^\star(i,j)}) < \frac{p_i}{p_j} < \underbrace{\frac{(d-1)u(i,j)+l(i,j)}{d}}_{\triangleq r^\star(i,j)}$$

As summarized in Figure 4-5, in the limit $\frac{p_i}{p_j} \to r^\star(i,j)$, d_{eff} goes again to d/p_j. We interpret this limiting case as *locally hard* in the sense that censorship in region j is sufficiently important in comparison to all other regions to impose a maximal effective dimension to the problem, irrespective of the values of p_i, matching Lemma 4.3.1. On the other hand, for the other limiting case (under additional mild assumptions), we find that d_{eff} also goes to d/p_j, but now for a *uniformly hard* reason: that is, censorship is approximately constant and equal to p_j, recovering the Remark 4. Note that in between these two extremes lies the *minimum effective dimension* for a given value of $\frac{p_i}{p_j}$.

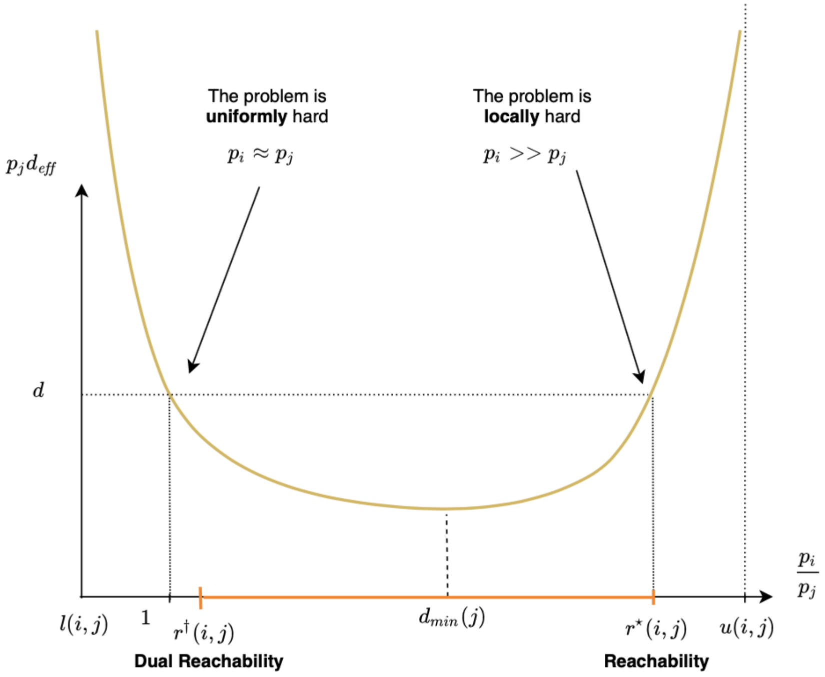

Figure 4-2: Sketch plot of normalized effective dimension $p_j d_{\text{eff}}$ with respect to $\frac{p_i}{p_j}$. We recover the uniform and local hardness conditions mentioned in the discussion of Thm. 4.3.2, as well as the existence of a *minimum effective dimension* for a certain value of $\frac{p_i}{p_j}$. The necessary conditions of reachability and dual reachability (Lemma 4.4.2 and 4.4.1) verified by $\frac{p_i}{p_j}$ impose that it belongs to the orange open interval.

4.4 Temporal dynamics of learning under censorship

We first summarize the dynamics of the optimal policy of ($\mathcal{P}$) through an algorithmic description in Alg. 2. Two key notions of our analysis are the concepts of reachability and dual reachability of a region i from a base region j, as described in Lemmas 4.4.1 and 4.4.2 and schematized in Fig.4-4,4-3 and 4-5. Formally, they can be written as two independent necessary constraints on the ratio $\frac{p_i}{p_j}$: $\frac{p_i}{p_j} < r^\star(i,j)$ for reachability and $\frac{p_i}{p_j} > r^\dagger(i,j)$ for dual reachability.

The categorization result provided in the statement of Thm. 4.3.2 follows from the two possible termination condition of the algorithm. We use as algorithmic invariant to ensure the termination the fact that the set of reachable regions is strictly decreasing

Algorithm 2: Algorithmic description of the dynamics of $\mathbb{W}(t)$

```
Initalization: Set current region $S \leftarrow k$
while a region is reachable from region $S$ do   /* Lemma 4.4.1,Fig.4-3 */
    play region $S$ optimal policy until first reachable region $i^\star$ is reached;
    if region $i^\star$ is dual reachable from region $S$ then   /* Lemma 4.4.2,
     Fig.4-4 */
        Bi-region $(i^\star, S)$ effective dimension (case 2);   /* Lemma 4.7.1 */
        play Bi-region $(i^\star, S)$ optimal policy;
        End;
    else
        Update current region $S \leftarrow i^\star$;   /* Lemma 4.4.2, Fig.4-5 */
    end
end
Single region $S$ effective dimension (case 1);   /* Lemma 4.4.1 */
play region $S$ optimal policy;
```

for inclusion and finite. Hence, the while loop will terminate either because a dual reachable region is reached or because no more regions are reachable. In order to not overload the presentation, time aspect is not present in the algorithmic description but is extensively covered in Lemmas 4.4.1, 4.4.2, 4.7.1 and Cor. 4.7.0.1, as well as in what follows. One of our main finding is that the dynamics of the optimal policy of ($\mathcal{P}$) are described through $\mathbb{W}(t)$ by two qualitatively different regimes. We emphasize that our continuous approach to analyzing cumulative censored potential is key to obtaining these results.

Transient Regime: From the **while** loop in the algorithmic description results a so-called transient regime. More precisely, there exists a decreasing sequence of censorship regions $\{i_1 = k, \ldots, i_l\}$ of length $l \in [k+1]$ and associated time sequence $\{t_0 \triangleq 0, t_1, \ldots, t_l\}$ such that whenever $t_j \leq t \leq t_{j+1}$ for a given index $j \leq l-1$, the evolution of $\mathbb{W}(t)$ is given by:

$$\mathbb{W}(t) = p_{i_{j+1}}(t - t_j)\mathbb{W}_{i_{j+1}} + \mathbb{W}(t_j) = p_{i_{j+1}}(t - t_j)\mathbb{W}_{i_{j+1}} + \sum_{n=1}^{j} p_{i_n}(t_n - t_{n-1})\mathbb{W}_{i_n} + \lambda \mathbb{I}_d.$$

This result follows from a simple induction with repeated use of Lemma 4.4.1, giving the exact sequence of censorship regions, Moreover, closed-formed formula for the time sequence is provided in Cor. 4.7.0.1. We interpret this transient step as an adversarial self-correction of the initial misspecification of censorship at an extra cost. This characterization of transient regime highlights an important consequence of using classical algorithms in censored environments.

Steady State Regime: Post-transient regime, the dynamics of $\mathbb{W}(t)$ enter a steady state regime, where one of the two cases necessarily arise:

- **Case 1: Single region i_l.** This case arises when the **while** loop ends because no other regions are reachable. It is equivalent to have last element of the time sequence t_l is equal to $+\infty$ and we have the single region evolution for all $t \geq t_{l-1}$ thanks to Lemma 4.4.1:

  $$\mathbb{W}(t) = p_{i_l}(t - t_{l-1})\mathbb{W}_{i_l} + \mathbb{W}(t_{l-1}) = p_{i_l}(t - t_{l-1})\mathbb{W}_{i_l} + \sum_{n=1}^{l-1} p_{i_n}(t_n - t_{n-1})\mathbb{W}_{i_n} + \lambda \mathbb{I}_d.$$

  The effective dimension corresponding to this dynamics is d/p_{i_l}, with the following equality for $T \geq t_{l-1}$:

  $$\int_0^T \frac{1}{p(a(t))} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t = \frac{1}{p_{i_l}} \log\det(\mathbb{W}(T)) + \sum_{n=1}^{l-1} (\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}}) \log\det \mathbb{W}(t_n),$$

  where the closed-form formula for $\mathbb{W}(t_n)$ is provided in Cor. 4.7.0.1 for all $n \leq l-1$.

- **Case 2: Bi-region (i_{l+1}, i_l).** This case arises when the **while** loop ends because dual reachable region i_{l+1} is reached from region i_l, with $i_{l+1} < i_l$. For all $t \geq t_l$, Lemma 4.4.2 yields the evolution:

  $$\mathbb{W}(t) \propto p_{i_{l+1}}(t + \lambda^\star) \begin{pmatrix} \cos^2(\phi_{i_l})(u(i_{l+1}, i_l) - \frac{p_{i_{l+1}}}{p_{i_l}})\mathbb{I}_{d-1} & (0) \\ (0) & \sin^2(\phi_{i_l})(\frac{p_{i_{l+1}}}{p_j} - l(i_{l+1}, i_l)) \end{pmatrix}.$$

  where $\lambda^\star$ and the proportionality factor are specified in the proof. The corresponding effective dimension is given by ($\mathcal{D}$) and the following equality holds for all $T \geq t_l$ thanks to Lemma 4.7.1:

$$\int_0^T \frac{1}{p(a(t))} \frac{\partial \log\det(\mathbb{W}(t))}{\partial t} \partial t = d_{\mathit{eff}} \log(1 + \frac{T - t_l}{t_l + \lambda^\star}) + \sum_{n=1}^{l} (\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}}) \log\det \mathbb{W}(t_n),$$

where the closed-form formula for $\mathbb{W}(t_n)$ is provided in Cor. 4.7.0.1 for all $n \leq l$.

Remark 5. *Fig.4-3 and 4-5 provide further insights on formula ($\mathcal{D}$) for d_{eff}. Throughout the proof and as illustrated on Fig.4-3, we see that for ($\mathcal{D}$) to arise, $\frac{p_i}{p_j}$ must belong to a certain interval $J \triangleq]\max(1, r^\dagger(i,j)), r^\star(i,j)[$. As $r^\star(i,j) < u(i,j)$ and $r^\dagger(i,j) > l(i,j)$, we see ($\mathcal{D}$) as a weighted average of the relative distance of $\frac{p_i}{p_j}$ to $u(i,j)$ and $l(i,j)$. Fig.4-5 provides a sketch of the variations of d_{eff} as $\frac{p_i}{p_j}$ evolves in this interval.*

Then, we formally present the key notions of reachability and dual reachability, used in the study of ($\mathcal{P}$) and illustrate possible behaviors thanks to sketch plots. In what follows, we show that the optimal policy adopts a greedy-like behavior by selecting in a specific way actions in a sequence of decreasingly censored regions. Reachability is a binary directed relation from a highly censored region i to a less censored region j. The latter is said to be reachable from the former if the playing only the highly censored region i leads to a point in time where region i and j are equally attractive from a greedy perspective. This happens whenever playing only region i create a potential gap between different actions that can be exploited to compensate the gap in censorship. Dual reachability is the symmetric relation: a highly censored region i is said to be dual reachable from a less censored region j if the playing only the weakly censored region i leads to a point in time where region i and j are equally attractive from a greedy perspective. For effective dimension of case 2 to arise, it is necessary to have two regions i and j such that j is both reachable and dual reachable from i.

Lemma 4.4.1. *[Reachability Analysis] Let's assume we start at a given time t_1 in* *transient censored region j, with a matrix*

$$\mathbb{W}(t_1) = \begin{pmatrix} \lambda_a \mathbb{I}_{d-1} & (0) \\ (0) & \lambda_b \end{pmatrix},$$

where $\lambda_a \geq \lambda_b$. We introduce $I_j \triangleq \{i; i < j \quad \text{and} \quad \frac{p_i}{p_j} < r^\star(i,j)\}$, the set of reachable regions from region j and affirm that we have the two possible cases:

- *If $I_j = \varnothing$, i.e. no region is reachable from region j, we switch to a steady state regime with single region j effective dimension $d_{\mathit{eff}} = d/p_j$.*
- *Otherwise, next region added to the transient sequence is $i^\star \triangleq \operatorname{argmin}_{i \in I_j} \mu^\star(i, j, \lambda_a, \lambda_b)$, at time $t_2 \triangleq t_1 + \frac{1}{p_j}\mu^\star(i^\star, j, \lambda_a, \lambda_b)$ and we have:*

$$\mathbb{W}(t_2) = \frac{(d-1)\sin^2(\phi_j)\lambda_a - \cos^2(\phi_j)\lambda_b}{d\cos^2(\phi_j)\sin^2(\phi_j)(r^\star(i^\star, j) - \frac{p_i}{p_j})}\mathbb{W}(i^\star, j).$$

Lemma 4.4.2. *[Dual Reachability Analysis] Let's assume we are currently playing transient region j and we reach the region i at time t_l. We then have the following two possible cases:*

- *If $\frac{p_i}{p_j} > r^\dagger(i,j)$, we say that regions i is dual reachable from region j, leading to a steady state regime with bi-region (i,j) effective dimension. In such case, for $t \geq t_l$, the potential increase is of the form:*

$$\mathbb{W}(t) = \frac{1}{p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j - r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_i(t + \lambda^\star)\mathbb{W}(i,j).$$

- *Otherwise, we switch from base region j to base region i and continue in the transient regime.*

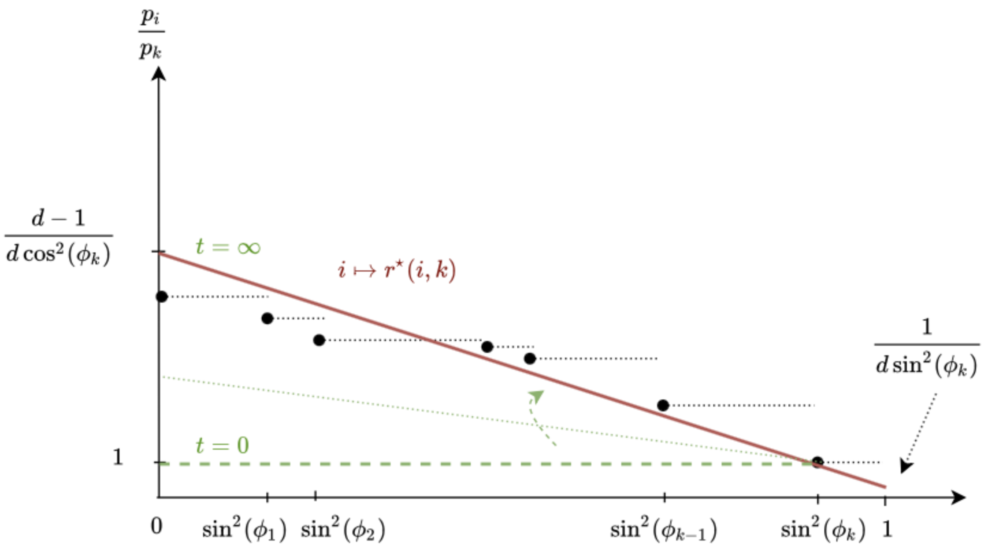

Figure 4-3: Illustration of the set of reachable regions from a base region k, as a function of $\frac{p_i}{p_k}$. Black dots and lines correspond to censorship regions defined by $\mathcal{MT}$. In this figure, we see that a region is reachable if and only if the black dot is below the red reachability line. As time increases, the green line rotates with region k as pivot and asymptotically approaches to the red line. Hence, the first reachable region is the one first *reached* by the green line.

To conclude this section, we present our results in the context of the canonical single-threshold model and effectively witness the full range of variation of the effective dimension between the local and uniform hardness edge scenarios. This variation is parameterized by the ratio of the censorship values and the value of the threshold.

Corollary 4.4.2.1. *For the single threshold model with two regions* 0 *and* 1 *and associated censorship probabilities* $p_0 < p_1$*, our main theorem yields:*

- *If* $\frac{p_0}{p_1} < \frac{d-1}{d\cos^2(\phi_1)}$*, then we reach bi-region steady state regime and have the effective dimension:*

$$d_{\mathit{eff}} = \frac{d-1}{p_0} + \frac{1}{p_0}\frac{\sin^2(\phi_1)}{\frac{p_1}{p_0} - \cos^2(\phi_1)} \in [\frac{d}{p_0}, \frac{d}{p_1}].$$

- *Otherwise, we are from* $t = 0$ *in single-region steady state regime and have the*

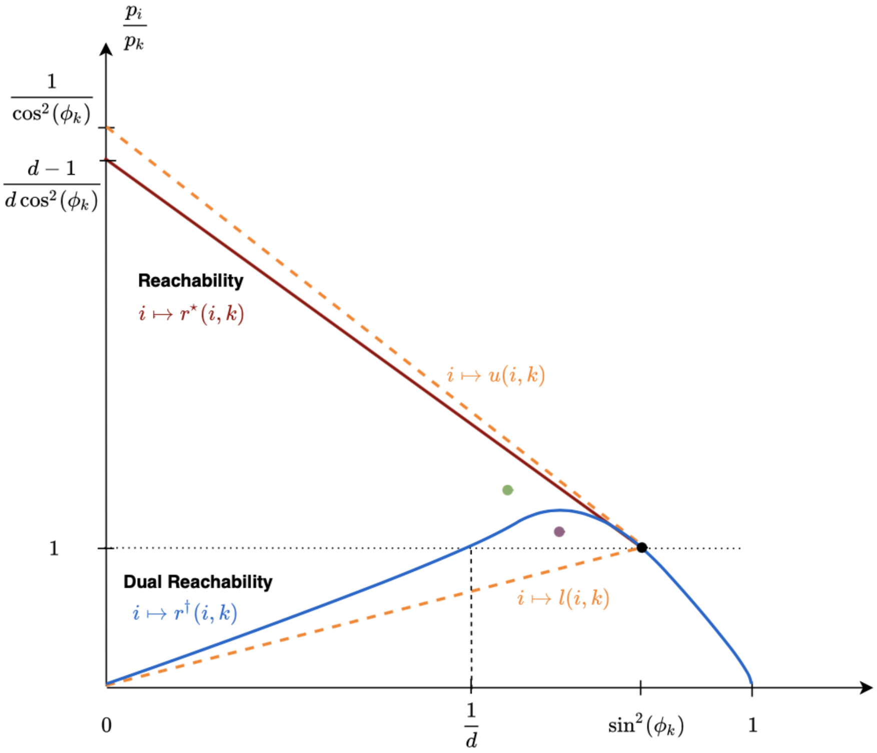

Figure 4-4: Sketch plot of reachability and dual reachability conditions from base region k associated with the black dot (Lemma 4.4.2 and 4.4.1) as a function of $\frac{p_i}{p_j}$. For a region i to be reachable, $\frac{p_i}{p_j}$ has to be below the red line. For a region i to be dual reachable, $\frac{p_i}{p_j}$ has to be above the blue line. Henceforth, the red dot here is a censorship region that is both reachable and dual reachable whereas the purple dot is a reachable but not dual reachable region. Orange lines represent the functions $u(i,k)$ and $l(i,k)$ introduced above in Sec.4.7.1.

effective dimension $d_{\mathit{eff}} = d/p_1$.

4.5 Conclusion of Chap. 4

The main result of this chapter yields that regret of LCB with censorship is still governed by the effective dimension, but now with a dependency of $\tilde{\mathcal{O}}(\sqrt{d \cdot d_{\mathit{eff}}}\sqrt{T})$ (Thm. 4.1.1). To the best of our knowledge, these regret bounds provide the first theoretical characterization in LCB with censorship, and contribute to the literature by evaluating the impact of censorship on the performance of UCB-type algorithms. In proving this result, we derive the value of the effective dimension for a broad class

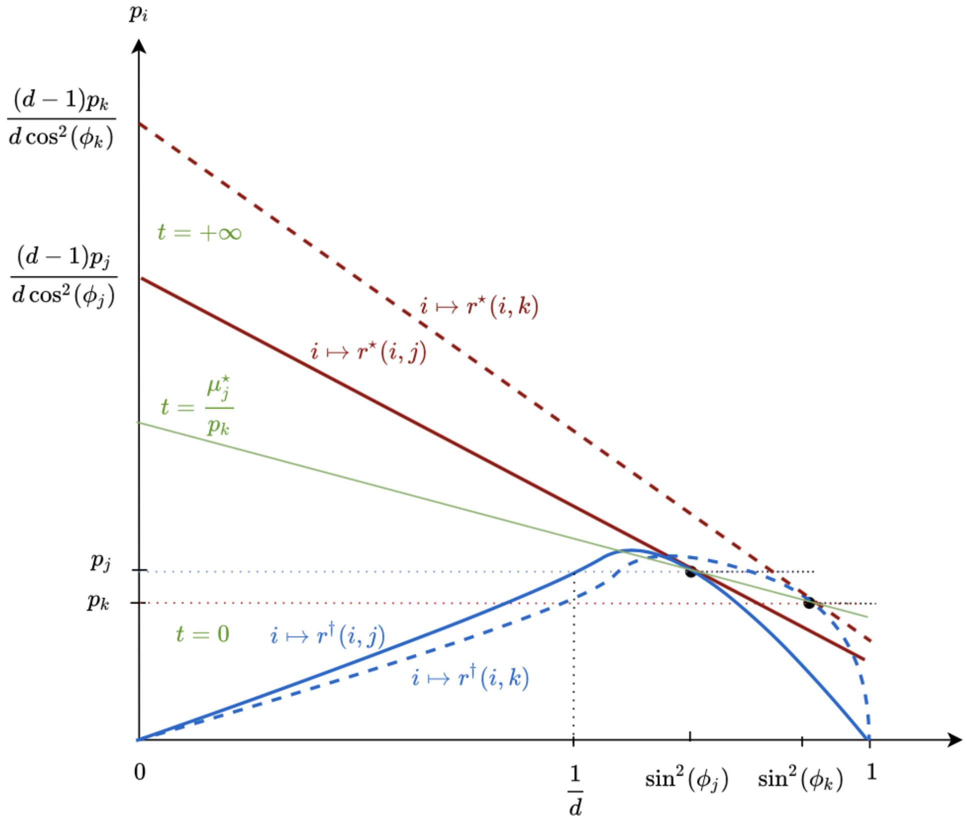

Figure 4-5: Sketch plot of the evolution of reachability and dual reachability conditions after a region j is reached from region k but is not dual reachable (Else condition in Alg. 2). Doted red (resp. blue) line is reachability (resp. dual reachability) condition for previous region k and full red (resp. blue) lines is reachability (resp. dual reachability) condition for new region j. Instead of starting from horizontal line at $t = 0$ to find new reachable state, rotation with region j as pivot is initialized at the green line associated with $t = \frac{\mu_j^\star}{p_k}$. Note that the y-axis is not normalized here.

of multi-threshold models $\mathcal{MT}$ as well as a precise understanding of the dynamic behavior induced by these models (Thm. 4.3.2). In particular, we find that censorship introduces a two-phase behavior: a transient phase during which the initial censoring misspecification is self-corrected at an additional cost; followed by a stationary phase that reflects the inherent slowdown of learning governed by the effective dimension. Moreover, in extending our analysis from MAB to LCB, we also develop a continuous generalization of the widely used Elliptical Potential Inequality (Prop. 4.2.2), which we believe is also of independent interest.

4.6 Proof of Chap. 4 - Contextual Bandits

In this section, we prove Thm. 4.1.1 of Chap.4, extending the results of MAB to LCB. To do so, we prove Lemmas 4.2.1, 4.6.1 and Prop. 4.2.2. Note that the proof of Thm. 4.3.2 is differed to next section. We conclude the section by discussing the extension of our analysis to Generalized Linear Contextual Bandits.

4.6.1 Proof of Lemma 4.2.1

Lemma 4.2.1. *For all $\delta \in]0,1]$, there exists a constant $\tilde{\beta}_\delta(T) = \Theta(\sqrt{d\log(T)})$ such that*

$$\mathbb{E}[R(T,\pi_{UCB})] \le 2\tilde{\beta}_\delta(T)\sqrt{T\mathbb{E}[\mathbb{V}_1(T,\pi_{UCB})]} + \delta T\Delta_{max},$$

where, for $\alpha > 0$ and $\pi \in \Pi$, the linear extension of the cumulative censored potential is given by:

$$\mathbb{V}_\alpha(T,\pi) \triangleq \sum_{t=1}^{T} \|a_t\|^2_{(\mathbb{W}^C_{t-1})^{-\alpha}} = \sum_{t=1}^{T} \mathrm{Tr}((\mathbb{W}^C_{t-1})^{-\alpha} a_t a_t^\top).$$

Proof. We have under the event $\neg\mathcal{H}^{II}_{\text{UCB}}(\delta)$ introduced in Lemma 4.6.1 and thanks to Holder inequality:

$$\Delta_t(a) \triangleq \max_{\tilde{a}\in\mathcal{A}_t}\langle\theta^\star,\tilde{a}\rangle - \langle\theta^\star,a_t\rangle \le 2\beta_\delta(t-1)\|a_t\|_{(\mathbb{W}^C(t-1))^{-1}}.$$

Therefore, the conditional regret is upper-bounded by:

$$R(T|\neg\mathcal{H}^{II}_{\text{UCB}}(\delta)) \le \beta_\delta(T)\sum_{t=1}^{T}\|a_t\|_{(\mathbb{W}^C(t-1))^{-1}} = \beta_\delta(T)\tilde{\mathbb{V}}_{\frac{1}{2}}(T,\pi),$$

where we introduced $\tilde{\mathbb{V}}_{\frac{1}{2}}(T,\pi) \triangleq \sum_{t=1}^{T}\|a_t\|_{\mathbb{W}^C(t-1)^{-1}}$. Cauchy Schwartz inequality then allows to make the junction $\tilde{\mathbb{V}}_{\frac{1}{2}}(T,\pi) \le \sqrt{T}\sqrt{\mathbb{V}_1(T,\pi)}$. We then introduce $\tilde{\beta}_\delta(T)$

a deterministic upper bound on $\beta_\delta(T)$:

$$\begin{aligned}\beta_\delta(T) &= \sqrt{\sigma^2 \log\left(\frac{\det(\mathbb{W}_T^C)}{\det(\lambda \mathbb{I}_d)}\right) + 2\sigma^2\log(\frac{1}{\delta})} + \sqrt{\lambda}\|\theta^\star\|_2 \\ &\leq \underbrace{\sqrt{\sigma^2 d \log(1+\frac{T}{d\lambda}) + 2\sigma^2\log(\frac{1}{\delta})} + \sqrt{\lambda}\|\theta^\star\|_2}_{\triangleq \tilde{\beta}_\delta(T)} \\ &= \Theta(\sqrt{d\log(T)}).\end{aligned}$$

Using the concavity of square root and Jensen's inequality, we have $\mathbb{E}[\sqrt{\mathbb{V}_1(T,\pi)}] \leq \sqrt{\mathbb{E}[\mathbb{V}_1(T,\pi)]}$. Finally, thanks to Lemma 4.6.1, we conclude that:

$$\mathbb{E}[R(T,\pi_{\text{UCB}})] \leq 2\tilde{\beta}_\delta(T)\sqrt{T\mathbb{E}[\mathbb{V}_1(T,\pi_{UCB})]} + \delta T \Delta_{max}.$$

□

4.6.2 Statement and Proof of Lemma 4.6.1

Analogous to Lemma 3.5.1 for the MAB case, one key step in the proof is introduction of the failure of optimism event. Nevertheless, note the difference with the choice of norm.

Lemma 4.6.1. *For any $\delta \in]0,1]$, uniform regularization $\lambda > 0$ and censored action generating process $(\mathbb{W}_t^C)_{t\leq T}$, let's introduce the event:*

$$\mathcal{H}_{UCB}^{II}(\delta) \triangleq \left\{\exists t \geq 0, \|\hat{\theta}_t^\lambda - \theta^\star\|_{\mathbb{W}_t^C} > \underbrace{\sqrt{\sigma^2 \log\left(\frac{\det(\mathbb{W}_t^C)}{\det(\lambda \mathbb{I}_d)}\right) + 2\sigma^2\log(\frac{1}{\delta})} + \sqrt{\lambda}\|\theta^\star\|_2}_{\triangleq \beta_\delta(t)}\right\}.$$

We then have $\mathbb{P}(\mathcal{H}_{UCB}^{II}(\delta)) \leq \delta$.

Proof. The proof closely mirrors the self-normalized bound for vector-valued martingales of Thm.1 from [1]. The main subtlety is to apply the results to the censored measurable vectors $(x_{a_t}a_t)$ instead of classically (a_t). This yields that with probability

$1-\delta$, for all $t \geq 0$:

$$\| \sum_{n=1}^{t} \epsilon_n x_{a_n} a_n \|^2_{\mathbb{W}_t^C} \leq \sigma^2 \log \frac{\det(\mathbb{W}_t^C)}{\det(\lambda \mathbb{I}_d)} + 2 \log(\frac{1}{\delta}).$$

Thus, still on this event, for any $t \geq 0$ and action $a \in \mathbb{R}^d$, we have by definition of $\hat{\theta}_t^\lambda$:

$$\langle a, \hat{\theta}_t^\lambda \rangle - \langle a, \theta^\star \rangle = \langle a, (\mathbb{W}_t^C)^{-1} \sum_{n=1}^{t} \epsilon_n x_{a_n} a_n \rangle - \lambda \langle a, (\mathbb{W}_t^C)^{-1} \theta^\star \rangle,$$

and therefore, thanks to Cauchy-Schwartz inequality:

$$|\langle a, \hat{\theta}_t^\lambda \rangle - \langle a, \theta^\star \rangle| \leq \|a\|_{(\mathbb{W}_t^C)^{-1}} \Big(\| \sum_{n=1}^{t} \epsilon_t x_{a_t} a_t \|_{\mathbb{W}_t^C} + \lambda^{1/2} \|\theta^\star\|_2 \Big)$$

Using previous result, for all $a \in \mathbb{B}_d, t \geq 0$, with probability $1-\delta$, we have:

$$|\langle a, \hat{\theta}_t^\lambda \rangle - \langle a, \theta^\star \rangle| \leq \sigma \sqrt{\log \Big(\frac{\det(\mathbb{W}_t^C)}{\det(\lambda \mathbb{I}_d)} \Big) + 2 \log(\frac{1}{\delta})} + \lambda^{1/2} \|\theta^\star\|_2$$

To conclude, we classically plug-in the value $a = \mathbb{W}_t^C(\hat{\theta}_t^\lambda - \theta^\star)$ and divide both sides by $\|\hat{\theta}_t^\lambda - \theta^\star\|_{\mathbb{W}_t^C}$ to get that for all $t \geq 0$, with probability $1-\delta$, we have:

$$\|\hat{\theta}_t^\lambda - \theta^\star\|_{\mathbb{W}_t^C} \leq \sigma \sqrt{\log \Big(\frac{\det(\mathbb{W}_t^C)}{\det(\lambda \mathbb{I}_d)} \Big) + 2 \log(\frac{1}{\delta})} + \lambda^{1/2} \|\theta^\star\|_2$$

and therefore, by definition $\mathbb{P}(\mathcal{H}^{II}_{\text{UCB}}(\delta)) \leq \delta$.

□

4.6.3 Proof of Prop. 4.2.2

Proposition 4.2.2. *For any $\delta \in]0,1]$, $\lambda > 0$, $\alpha > 0$ and policy $\pi \in \Pi$, we have:*

$$\mathbb{E}[V_\alpha(T, \pi)] \leq \frac{\delta}{\lambda^\alpha} + C(\delta)^\alpha \operatorname{Tr} \Big(\int_0^T \mathbb{W}(t)^{-\alpha} a(t) a(t)^\top \partial t \Big),$$

where $C(\delta) \triangleq 8(\lambda + 1) \max(\log(d/\delta))/\lambda, 1)/\lambda$.

Proof. First, we use Lemma 4.6.2 to deduce that under $\mathcal{H}_{\mathrm{CEN}}^{II}(\delta)$:

$$\mathbb{V}_\alpha(T,\pi|\mathcal{H}_{\mathrm{CEN}}^{II}(\delta)) = \sum_{t=1}^{T} \mathrm{Tr}((\mathbb{W}_{t-1}^{C})^{-\alpha} a_t a_t^\top) \leq c_\delta^\alpha \sum_{t=1}^{T} \mathrm{Tr}(\mathbb{W}_{t-1}^{-\alpha} a_t a_t^\top).$$

For all $t \geq 1$, we then use the fact that $W_t \preceq (1+\frac{1}{\lambda})W_{t-1}$ to deduce $\mathrm{Tr}(\mathbb{W}_{t-1}^{-\alpha} a_t a_t^\top) \leq (1+\frac{1}{\lambda})^\alpha \,\mathrm{Tr}(\mathbb{W}_t^{-\alpha} a_t a_t^\top)$. The last and most important step is the integral comparison:

$$\sum_{t=1}^{T} \mathrm{Tr}(\mathbb{W}_t^{-\alpha} a_t a_t^\top) \leq \int_0^T \mathrm{Tr}(\mathbb{W}(t)^{-\alpha} a(t) a(t)^\top) \partial t = \mathrm{Tr}\Big(\int_0^T \mathbb{W}(t)^{-\alpha} a(t) a(t)^\top \partial t\Big).$$

In the previous result, the continuous extension $(a(t), \mathbb{W}(t))_{t\leq T}$ of $(a_t, \mathbb{W}_t)_{t\in[T]}$ for a given policy π is defined for any time $t \geq 1$ as:

$$a(t) \triangleq a_{\lfloor t \rfloor} \quad \text{and} \quad \mathbb{W}(t) \triangleq \int_{u=1}^{t} p_{a(u)} a(u) a(u)^\top \partial u = \mathbb{W}_{\lfloor t \rfloor} + (t - \lfloor t \rfloor) p(a_{\lceil t \rceil}) a_{\lceil t \rceil} a_{\lceil t \rceil}^\top.$$

This yields the result:

$$\mathbb{V}_\alpha(T,\pi|\mathcal{H}_{\mathrm{CEN}}^{II}(\delta)) \leq c_\delta^\alpha (1 + \frac{1}{\lambda})^\alpha \,\mathrm{Tr}\Big(\int_0^T \mathbb{W}(t)^{-\alpha} a(t) a(t)^\top \partial t\Big).$$

Finally, we conclude thanks to Lemma 4.6.2 that:

$$\mathbb{E}[\mathbb{V}_\alpha(T,\pi)] \leq \frac{\delta}{\lambda^\alpha} + C(\delta)^\alpha \,\mathrm{Tr}\Big(\int_0^T \mathbb{W}(t)^{-\alpha} a(t) a(t)^\top \partial t\Big).$$

□

Remark 6. *The main tour de force of the continuous approximation we employ is to relax the maximization problem by considering the class of continuous deterministic integrable policies, which is considerably more tractable from an analysis perspective. On the one hand, it allows to get closed-form solution for the maximization problem whereas the discrete approach can only deal with approximations and upper bounds. On the other hand, it clearly reveals the underlying matrix function the discrete approach is approximating and henceforth allows to leverage powerful integration results. We leverage again this idea in the context of Chap.4 to tackle impact of censorship.*

To illustrate the abovementioned points, we remark that for the simpler case of classical uncensored environment, we obtain for $\alpha > 0, \alpha \neq 1$:

$$\sum_{t=1}^{T} \|a_t\|^2_{\mathbb{W}_{t-1}^{-\alpha}} \leq \Big(\frac{\lambda+1}{\lambda}\Big)^{\alpha} \frac{\mathrm{Tr}\left(\int_0^T \partial \mathbb{W}(t)^{1-\alpha}\right)}{1-\alpha} = \Big(\frac{\lambda+1}{\lambda}\Big)^{\alpha} \frac{\mathrm{Tr}(\mathbb{W}_T^{1-\alpha} - \mathbb{W}_0^{1-\alpha})}{1-\alpha}.$$

For $\alpha < 1$, we then have thanks to Lemma 3.2.3 the worst case bound $\mathrm{Tr}(\mathbb{W}_T^{1-\alpha}) \leq d^{\alpha}(d\lambda + T)^{1-\alpha}$ and henceforth:

$$\sum_{t=1}^{T} \|a_t\|^2_{\mathbb{W}_{t-1}^{-\alpha}} \leq \Big(\frac{\lambda+1}{\lambda}\Big)^{\alpha} \frac{d^{\alpha}(d\lambda+T)^{1-\alpha} - d\lambda^{1-\alpha}}{1-\alpha}$$

On the other hand, for $\alpha > 1$, we deduce:

$$\sum_{t=1}^{T} \|a_t\|^2_{\mathbb{W}_{t-1}^{-\alpha}} \leq \Big(\frac{\lambda+1}{\lambda}\Big)^{\alpha} \frac{d\lambda^{1-\alpha}}{\alpha-1}.$$

Finally, for $\alpha = 1$, we use the formula $\mathrm{Tr}(\log(A)) = \log(\det A)$ to deduce:

$$\sum_{t=1}^{T} \|a_t\|^2_{\mathbb{W}_{t-1}^{-1}} \leq \frac{\lambda+1}{\lambda} \int_0^T \frac{\partial \log \det(\mathbb{W}(t))}{\partial t} \partial t = \frac{\lambda+1}{\lambda} \mathrm{Tr}(\log \mathbb{W}_T - \log \mathbb{W}_0)$$
$$= \frac{\lambda+1}{\lambda} \log \frac{\det \mathbb{W}_T}{\det \mathbb{W}_0} \leq \frac{\lambda+1}{\lambda} \log(1 + \frac{T}{\lambda d}),$$

where we used again Lemma 3.2.3 to obtain the last (worst-case) upper bound. In doing so, we recover and extend the recent results of [7] in a more natural way.[3]

4.6.4 Statement of Lemma 4.6.2

In order to prove previous property on $\mathbb{V}_{\alpha}$, a key step mirroring the MAB case is the use of high confidence lower bound on the censorship process, proven using anytime matrix martingale inequalities:

[3] Yet, we conjecture that the preliminary use of Cauchy Schwartz inequality in the case $\alpha > 1$ to affirm $\sum_{t=1}^{T} \|a_t\|_{\mathbb{W}_{t-1}^{-\alpha}} \leq \sqrt{T \sum_{t=1}^{T} \|a_t\|^2_{\mathbb{W}_{t-1}^{-\alpha}}}$ is suboptimal in this case as it imposes a $\mathcal{O}(\sqrt{T})$ scaling.

Lemma 4.6.2. *([39]) For any $\delta \in]0,1]$, $\lambda > 0$ and policy π, let's introduce the event:*

$$\mathcal{H}^{II}_{CEN}(\delta) \triangleq \Big\{\exists t \geq 0, \mathbb{W}^C_t \prec \frac{1}{c_\delta}\mathbb{W}_t\Big\},$$

where $c_\delta \triangleq 8\max(\frac{\log(d/\delta))}{\lambda}, 1)$. We then have $\mathbb{P}(\mathcal{H}^{II}_{CEN}(\delta)) \leq \delta$.

Note that picking as in the MAB case $\delta \sim d/T^2$ would lead to a constant $c_\delta = \Theta(\log(T))$, that is a worsening confidence interval, except if we manage to control the initialization. One interesting technical question for future work would be to allow an initialization condition as in Lemma 3.5.2 ensuring $\mathbb{W}(T_0)$ counterbalance $\log(d/\delta)$.

4.6.5 Proof of Thm. 4.1.1

Theorem 4.1.1. *For a given multi-threshold censorship model $\mathcal{MT}$, there exits d_{eff} such that the UCB algorithm with regularization λ has an instance-independent expected regret of:*

$$\mathbb{E}[R(T, \pi_{UCB})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d \cdot d_{\mathit{eff}}}\sqrt{T}).$$

Proof. Analogous to the MAB case, we use Lemma 4.2.1 to deduce:

$$\mathbb{E}[R(T, \pi_{\mathrm{UCB}})] \leq 2\tilde{\beta}_\delta(T)\sqrt{T\mathbb{E}[\mathbb{V}_1(T, \pi_{UCB})]} + \delta T \Delta_{max},$$

where we have:

$$\tilde{\beta}_\delta(T) = \sqrt{\sigma^2 d\log(1 + \frac{T}{d\lambda}) + 2\sigma^2\log(\frac{1}{\delta})} + \sqrt{\lambda}\|\theta^\star\|_2.$$

We then pick $\delta = \frac{d}{T^2}$, which yields:

$$\mathbb{E}[R(T, \pi_{\mathrm{UCB}})] \leq 2\Big(\sqrt{\sigma^2 d\log(1 + \frac{T}{d\lambda}) + 2\sigma^2\log(\frac{T^2}{d})} + \sqrt{\lambda}\|\theta^\star\|_2\Big)\sqrt{T\mathbb{E}[\mathbb{V}_1(T, \pi_{UCB})]}$$
$$+ \frac{d\Delta_{max}}{T}.$$

We then apply Lemma 4.2.2 with $\alpha = 1$ and $\delta = \frac{d}{T^2}$ to deduce:

$$\begin{aligned}\mathbb{E}[\mathbb{V}_\alpha(T,\pi)] &\leq \frac{d}{\lambda T^2} + 8\frac{\lambda+1}{\lambda}\max(\frac{2\log(T)}{\lambda}, 1)\operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big) \\ &\leq \frac{d}{\lambda T^2} + 8\frac{\lambda+1}{\lambda}\max(\frac{2\log(T)}{\lambda}, 1)\max_{\pi\in\Pi}\operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big).\end{aligned}$$

By applying Thm. 4.3.2, we deduce the two possibilities:

- **Case 1: Single region i_l.** The effective dimension corresponding to this dynamics is d/p_{i_l}, with the following equality for $T \geq t_{l-1}$:

  $$\max_{\pi\in\Pi}\operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big) = \frac{1}{p_{i_l}}\log\det(\mathbb{W}(T)) + \sum_{n=1}^{l-1}(\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}})\log\det\mathbb{W}(t_n),$$

  where we have for $T \geq t_{l-1}$ $\mathbb{W}(T) = p_{i_l}(T - t_{l-1})\mathbb{W}_{i_l} + \mathbb{W}(t_{l-1})$. Explicit formula of $(t_n, \mathbb{W}(t_n))$ are given for all $n \leq l$ in Cor. 4.7.0.1. We then note that:

  $$\begin{aligned}\frac{1}{p_{i_l}}\log\det(\mathbb{W}(T)) &= \frac{1}{p_{i_l}}\log\det(p_{i_l}(T - t_{l-1})\mathbb{W}_{i_l} + \mathbb{W}(t_{l-1})) \\ &= d_{\mathit{eff}}\log(T) + \frac{1}{p_{i_l}}\log\det(p_{i_l}(1 - \frac{t_{l-1}}{T})\mathbb{W}_{i_l} + \frac{1}{T}\mathbb{W}(t_{l-1})).\end{aligned}$$

  For $T \geq t_{l-1}$, we then write this in the form:

  $$\max_{\pi\in\Pi}\operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big) = d_{\mathit{eff}}\log(T) + f(T),$$

  where $f(T) = o(\log(T))$.

- **Case 2: Bi-region** (i_{l+1}, i_l)**.** Similarly, for $T \geq t_l$, we have:

$$\begin{aligned}\max_{\pi\in\Pi} \operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big) &= d_{\mathit{eff}} \log(1 + \frac{T - t_l}{t_l + \lambda^\star}) \\ &\quad + \sum_{n=1}^{l}(\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}}) \log\det \mathbb{W}(t_n) \\ &= d_{\mathit{eff}} \log(T) + d_{\mathit{eff}} \log(\frac{1}{T} + \frac{1 - \frac{t_l}{T}}{t_l + \lambda^\star}) \\ &\quad + \sum_{n=1}^{l}(\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}}) \log\det \mathbb{W}(t_n) \\ &= d_{\mathit{eff}} \log(T) + f(T),\end{aligned}$$

where $f(T) = o(\log(T))$.

Therefore, for given d_{eff}, f and t_0, we know that the following holds for all $T \geq t_0$:

$$\begin{aligned}\mathbb{E}[\mathbb{V}_\alpha(T,\pi)] &\leq \frac{d}{\lambda T^2} + 8\frac{\lambda+1}{\lambda}\max(\frac{2\log(T)}{\lambda}, 1) \operatorname{Tr}\Big(\int_0^T \mathbb{W}(t)^{-1}a(t)a(t)^\top \partial t\Big) \\ &\leq \frac{d}{\lambda T^2} + 8\frac{\lambda+1}{\lambda}\max(\frac{2\log(T)}{\lambda}, 1)(d_{\mathit{eff}} \log(T) + f(T)).\end{aligned}$$

Putting the pieces together yields for $T \geq t_0$:

$$\begin{aligned}\mathbb{E}[R(T,\pi_{\text{UCB}})] \leq 2\Big(\sqrt{\sigma^2 d\log(1 + \frac{T}{d\lambda}) + 2\sigma^2\log(\frac{T^2}{d})} + \sqrt{\lambda}\|\theta^\star\|_2\Big)\sqrt{T}\Big(\frac{d}{\lambda T^2} \\ + 8\frac{\lambda+1}{\lambda}\max(\frac{2\log(T)}{\lambda}, 1)(d_{\mathit{eff}}\log(T) + f(T))\Big)^{1/2} + \frac{d\Delta_{max}}{T}.\end{aligned}$$

By imposing regularization of order $\lambda = o(\log(T))$ only considering the leading order, this yields:

$$\mathbb{E}[R(T,\pi_{\text{UCB}})] \leq \tilde{\mathcal{O}}(\sqrt{(d+4)\sigma^2}\sqrt{d_{\mathit{eff}}}\sqrt{T}).$$

Finally, by working in large d regime, we finally conclude that:

$$\mathbb{E}[R(T,\pi_{\text{UCB}})] \leq \tilde{\mathcal{O}}(\sigma\sqrt{d\cdot d_{\mathit{eff}}}\sqrt{T}).$$

Again, we note that our proof easily allows to get high-probability bounds on regret instead of bounds on its expected value.

□

4.6.6 Extension to Generalized Linear Contextual Bandits

On what follows, we provide a sketch of the extension our results to Generalized Linear Contextual Bandits (GLCB) but differ the complete treatment to future work. In this model, the reward of a given action a is assumed to be of the form:

$$r(a) = \mu(\langle a, \theta^\star \rangle)$$

for a given function μ strictly increasing, continuously differentiable and real-valued. Notable instances of such a problem include the Logistic bandit and the Poisson bandit. Of particular importance in the dimensionality study of the problem are the constants:

$$L_\mu = \sup_{a \in \cup \mathcal{A}_t} \mu^{(1)}(\langle a, \theta^\star \rangle) \quad \text{and} \quad \kappa = \inf_{a \in \cup \mathcal{A}_t} \mu^{(1)}(\langle a, \theta^\star \rangle).$$

An important requirement of GLCB is the assumption $\kappa > 0$ needed to ensure identifiability of $\theta^\star$ and asymptotic normality. Given this, the suited definition of pseudo-regret considered is:

$$R(T, \pi) \triangleq \sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} \mu(\langle a, \theta^\star \rangle) - \mu(\langle a_t, \theta^\star \rangle)$$

Note that this regret can be easily mapped to the one studied above thanks to the fact that L_μ is a Lipschitz constant for μ: for all $a, \tilde{a} \in \cup \mathcal{A}_t$, $|\mu(\langle a, \theta^\star \rangle) - \mu(\langle \tilde{a}, \theta^\star \rangle)| \leq L_\mu |\langle a, \theta^\star \rangle - \langle \tilde{a}, \theta^\star \rangle|$. Mirroring the proof of [31], we use a Maximum Likelihood Estimator (MLE) instead of a Least-Square Estimator for $\theta^\star$. More precisely, we define

$\hat{\theta}_t^{MLE}$ as the solution of the equation:

$$\sum_{n=1}^{t} \langle a_n, \epsilon_t + \mu(\langle a_n, \theta^\star \rangle) - \mu(\langle a_n, \theta \rangle) = 0$$

A minor difference between the approach of [31] and what precedes is the use of a period of initial random sampling (e.g. *exploration*) instead of the regularization to ensure inversibility of the design matrix $\mathbb{W}_t^C$. More precisely, the initial sampling ensures that with high-probability, $\lambda_{\min}(\mathbb{W}_t^C) > 0$ in a finite time T_{init}. To be possible, this requires the assumption that there exists $\sigma_0^2 > 0$ such that for all $t \geq 1$, we have $\lambda_{min}\left(\mathbf{E}_{a \in \mathcal{A}_t}\left[aa^\top\right]\right) \geq \sigma_0^2$, where the expectation $\mathbf{E}$ is associated with an uniform sampling of actions. Under the same assumption, the impact of censorship on this initialization step is at worst an increase of the sampling time to $\tilde{T}_{\text{init}} \triangleq T_{\text{init}}/p_{\min}$, which is still constant. Following Lemma 9 of [31], we then consider the censored high-probability confidence set for any $\delta \in [\frac{1}{T}, 1]$:

$$\mathcal{H}_{\text{UCB}}^{III}(\delta) \triangleq \Big\{ \exists t \geq 0, \|\hat{\theta}_t^{MLE} - \theta^\star\|_{\mathbb{W}_t^C} > \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2\frac{t}{d}) + \log(1/\delta)} \quad \text{and} \quad \lambda_{\min}(\mathbb{W}_t^C) > 1 \Big\}.$$

and a direct extension of their results allows us to conclude $\mathbb{P}(\mathcal{H}_{\text{UCB}}^{III}(\delta)) \leq \delta$. Note that the constant κ appears when upper bounding in the Loewner order the Fischer Information Matrix of the MLE by the matrix $\mathbb{W}_t^C$. Post-initialization, the conditional regret is then upper bounded by:

$$\begin{aligned} R(T, \pi_{\text{UCB}} | \neg \mathcal{H}_{\text{UCB}}^{III}(\delta)) &\leq \tilde{T}_{\text{init}} \Delta_{max} + \sum_{t=T_{\text{init}}}^{T} L_\mu \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2\frac{t}{d}) + \log(1/\delta)} \|a_t\|_{(\mathbb{W}_t^C)^{-1}} \\ &\leq \tilde{T}_{\text{init}} \Delta_{max} + L_\mu \frac{\sigma}{\kappa}\sqrt{\frac{d}{2}\log(1 + 2T/d) + \log(1/\delta)} \sqrt{T \mathbb{V}_1(\pi_{\text{UCB}}, T)}, \end{aligned}$$

Combining these elements and taking $\delta = \frac{1}{T}$, we conclude that:

$$\mathbb{E}[R(T, \pi_{\text{UCB}})] \leq \tilde{\mathcal{O}}\Big(\frac{L_\mu}{\kappa}\sqrt{d}\sqrt{T\mathbb{E}[\mathbb{V}_1(\pi_{\text{UCB}}, T)]}\Big) \leq \tilde{\mathcal{O}}\Big(L_\mu \frac{\sqrt{d \cdot d_{\textit{eff}}}}{\kappa}\sqrt{T}\Big),$$

where we used Thm. 4.3.2 to control $\mathbb{E}[\mathbb{V}_1(\pi_{\text{UCB}}, T)]$ as done in the proof of Th.4.1.1.

4.7 Proof of Chap. 4 - Temporal Dynamics for Multi-Threshold Models

In this section, we prove Thm. 4.3.2 and discuss its implications. In doing so, we introduce and prove Lemmas 4.4.1, 4.4.2, 4.7.1 and Cor. 4.7.0.1. We conclude the section by illustrating results for the single-threshold model, through Cor. 4.4.2.1.

4.7.1 Supplementary Notations

Without loss of generality (i.e. up to an orthogonal transformation), we can consider that $u \equiv e_d$, the d^{th} basis vector. Given this, for two regions $i < j$, we introduce the notations:

$$l(i,j) \triangleq \frac{\sin^2(\rho_i)}{\sin^2(\rho_j)} \quad \textit{and} \quad u(i,j) \triangleq \frac{\cos^2(\rho_i)}{\cos^2(\rho_j)}$$

$$r^\star(i,j) \triangleq \frac{(d-1)u(i,j)+l(i,j)}{d} \quad \textit{and} \quad r^\dagger(i,j) \triangleq \frac{1}{r^\star(j,i)} = \frac{dl(i,j)u(i,j)}{u(i,j)+(d-1)l(i,j)}$$

$$\mathbb{W}_i \triangleq \begin{pmatrix} \frac{\cos^2(\rho_i)}{d-1}\mathbb{I}_{d-1} & (0) \\ (0) & \sin^2(\rho_i) \end{pmatrix}$$

$$\mathbb{W}(i,j) \triangleq \begin{pmatrix} \cos^2(\rho_j)(u(i,j) - \frac{p_i}{p_j})\mathbb{I}_{d-1} & (0) \\ (0) & \sin^2(\rho_j)(\frac{p_i}{p_j} - l(i,j)) \end{pmatrix}.$$

Whenever i and j are clear from context, we use in u (resp. l) as abbreviation for $u(i,j)$ (resp. $l(i,j)$).

4.7.2 Proof of Lemma 4.4.1

Lemma 4.4.1. *[Reachability Analysis] Let's assume we start at a given time t_1 in transient censored region j, with a matrix*

$$\mathbb{W}(t_1) = \begin{pmatrix} \lambda_a\mathbb{I}_{d-1} & (0) \\ (0) & \lambda_b \end{pmatrix},$$

where $\lambda_a \geq \lambda_b$. We introduce $I_j \triangleq \{i; i < j \quad \text{and} \quad \frac{p_i}{p_j} < r^\star(i,j)\}$, the set of reachable regions from region j and affirm that we have the two possible cases:

- *If $I_j = \varnothing$, i.e. no region is reachable from region j, we switch to a steady state regime with single region j effective dimension $d_{\mathit{eff}} = d/p_j$.*

- *Otherwise, next region added to the transient sequence is $i^\star \triangleq \operatorname{argmin}_{i \in I_j} \mu^\star(i, j, \lambda_a, \lambda_b)$, at time $t_2 \triangleq t_1 + \frac{1}{p_j}\mu^\star(i^\star, j, \lambda_a, \lambda_b)$ and we have:*

$$\mathbb{W}(t_2) = \frac{(d-1)\sin^2(\phi_j)\lambda_a - \cos^2(\phi_j)\lambda_b}{d\cos^2(\phi_j)\sin^2(\phi_j)(r^\star(i^\star, j) - \frac{p_i}{p_j})}\mathbb{W}(i^\star, j).$$

Proof. First, we note that the initial starting point is recovered for $t_1 = 0$, base censored state k and $\lambda_a = \lambda_b = \lambda$ but this Lemma allows to go beyond the first step in the study of the behavior of the system. We know the temporal evolution for normalized budget $\mu \triangleq p_1(t - t_1)$ is of the form:

$$\mathbb{W}(t) = \begin{pmatrix} (\mu\frac{\cos^2(\phi_j)}{d-1} + \lambda_a)\mathbb{I}_d & (0) \\ (0) & \mu\sin^2(\phi_j) + \lambda_b \end{pmatrix} = \mu\mathbb{W}_j + \mathbb{W}(t_1).$$

We recall that the set of actions associated with region j is $\{a \in \mathbb{B}_d, \sin(\phi_j) \leq \langle a, e_d\rangle < \sin(\phi_{j+1})\}$. Therefore, the use of Kiefer-Wolfowitz theorem [29] combined with the fact $\lambda_a \geq \lambda_b$ yields that the optimal policy while evolving in region j only plays unit action vector $v_j \equiv (\cos(\phi_j)/(d-1)^{1/2}, \ldots, \cos(\phi_j)/(d-1)^{1/2}, \sin(\phi_j))$. By noting that $v_j v_j^\top = \mathbb{W}_j$, we obtain the formula announced. Reachability of a given state $i < j$ from state j after time t_1 is then defined as:

$$\exists t \geq t_t, \quad \frac{1}{p_i}\operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_i) = \frac{1}{p_j}\operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_j).$$

We interpret this as a classical a first-order optimally condition for convex maximization problems, where the matrix $\mathbb{W}_j$ is weighted by the censorship probability representing the speed of increase in region j. We then rewrite this condition as:

$$\exists \mu \geq 0, \quad \frac{1+f(\mu)\cos^2(\phi_i)}{1+f(\mu)\cos^2(\phi_j)} = \frac{p_i}{p_j} \quad where \quad f(\mu) \triangleq \frac{\mu \sin^2(\phi_j)+\lambda_b}{\mu \frac{\cos^2(\phi_j)}{d-1}+\lambda_a} - 1.$$

We know that f is increasing in μ and the LHS of the equation above is decreasing in $f(\mu)$ as $i < j$. Hence, the reachability condition than be stated by looking at the limit of f in $+\infty$. By using the fact that $\lim_{\mu \to +\infty} f(\mu) = \frac{d\sin^2(\phi_j)-1}{\cos^2(\phi_j)}$, we deduce that the reachability condition is equivalent to looking at the position of $\frac{p_i}{p_j}$ with respect to:

$$r^\star(i,j) \triangleq \frac{1+ud[\sin^2(\phi_j)-\frac{1}{d}]}{d\sin^2(\phi_j)} = \frac{(d-1)u+l}{d} = \frac{1}{d}\operatorname{Tr}(\mathbb{W}_j^{-1}\mathbb{W}_i).$$

On the one hand, if $\frac{p_i}{p_j} \geq r^\star(i,j)$, the state in never reachable in a finite time. On the other hand, whenever $\frac{p_i}{p_j} < r^\star(i,j)$, the state is reachable by investing a budget $\mu^\star(i,j,\lambda_a,\lambda_b)$ such that:

$$f(\mu^\star(i,j,\lambda_a,\lambda_b)) = \frac{1}{\cos^2(\phi_j)} \frac{\frac{p_i}{p_j}-1}{u-\frac{p_i}{p_j}},$$

which in turn involves:

$$\mu^\star(i,j,\lambda_a,\lambda_b) = \frac{d-1}{d\sin^2(\phi_j)\cos^2(\phi_j)} \frac{(\sin^2(\phi_j)\lambda_a+\cos^2(\phi_j)\lambda_b)\frac{p_i}{p_j} - (\sin^2(\phi_i)\lambda_a+\cos^2(\phi_i)\lambda_b)}{r^\star(i,j)-\frac{p_i}{p_j}}.$$

In particular, at $t_1 = 0$ whenever $\lambda_b = \lambda_a = \lambda$ and $j = k$, this gives:

$$\mu^\star(i,k,\lambda,\lambda) = \frac{(d-1)\lambda}{d\sin^2(\phi_k)\cos^2(\phi_k)} \frac{\frac{p_i}{p_k}-1}{r^\star(i,k)-\frac{p_i}{p_k}}.$$

The first reachable region from region j is then defined as $i^\star \triangleq \operatorname{argmin}_{i\in I} \mu^\star(i,j,\lambda_a,\lambda_b)$, where $I \triangleq \{i; i < j \quad and \quad \frac{p_i}{p_j} < r^\star(i,j)\}$. Note that at the moment $t_2 \triangleq t_1 + \frac{1}{p_j}\mu^\star(i^\star,j,\lambda_a,\lambda_b)$ when this region is reached, we have:

$$\mathbb{W}(t_2) = \frac{(d-1)\sin^2(\phi_j)\lambda_a - \cos^2(\phi_j)\lambda_b}{d\cos^2(\phi_j)\sin^2(\phi_j)(r^\star(i^\star,j)-\frac{p_i}{p_j})}\mathbb{W}(i,j).$$

On the other hand, whenever the set I is empty, by definition, the process reaches case 1 steady-state regime and only plays optimal policy of region j for remaining budget. To be fully general, we note that two or more regions can be reached simultaneously. In this case, the optimal policy tie-breaks by taking the region with maximal index i.e. higher censorship, as further described in Lemma 4.4.2. □

4.7.3 Statement and Proof of Cor. 4.7.0.1

More generally, this allows us to deduce the next technical corollary:

Corollary 4.7.0.1. *For a sequence of censored regions $\{i_1 = k, \dots, i_l, i_{l+1}, \dots\}$, we have for the l^{th} region of the transient sequence, with starting time t_{l-1} and ending time t_l:*

$$\begin{aligned}\mathbb{W}(t_l) &= \lambda \mathbb{I}_d + \sum_{n=1}^{l} \mu^\star(i_{n+1}, i_n, \lambda_a^{\mathbb{W}(t_{n-1})}, \lambda_b^{\mathbb{W}(t_{n-1})}) \mathbb{W}_{i_n} \\ &= \frac{\lambda^{\frac{(d-1)\sin^2(\phi_k) - \cos^2(\phi_k)}{\cos^2(\phi_{i_l})\sin^2(\phi_{i_l})}} \prod_{n=1}^{l-1} \left(r^\dagger(i_{n+1}, i_n) - \frac{p_{i_{n+1}}}{p_{i_n}} \right)}{d^l \prod_{n=1}^{l} \left(r^\star(i_{n+1}, i_n) - \frac{p_{i_{n+1}}}{p_{i_n}} \right) \prod_{n=1}^{l-1} \left(u(i_{n+1}, i_n) + dl(i_{n+1}, i_n) \right)} \mathbb{W}(i_{l+1}, i_l),\end{aligned}$$

where t_l is characterized by:

$$t_l = \sum_{n=1}^{l} \frac{1}{p_{i_n}} \mu^\star(i_{n+1}, i_n, \lambda_a^{\mathbb{W}(t_{n-1})}, \lambda_b^{\mathbb{W}(t_{n-1})}),$$

and where $\lambda_a^{\mathbb{W}(t_n)}$ and $\lambda_b^{\mathbb{W}(t_n)}$ refer respectively to the upper and lower coefficient of the diagonal matrix $\mathbb{W}(t_n)$.

Proof. We leverage a simple induction reasoning using for $l \geq 1$ the formula given within the proof of lemma 4.4.1:

$$\begin{aligned} t_l &= t_{l-1} + \frac{1}{p_{i_l}} \mu^\star(i_{l+1}, i_l, \lambda_a^{\mathbb{W}(t_{l-1})}, \lambda_b^{\mathbb{W}(t_{l-1})}) \\ \mathbb{W}(t_l) &= \frac{(d-1)\sin^2(\phi_{i_l})\lambda_a^{\mathbb{W}(t_{l-1})} - \cos^2(\phi_{i_l})\lambda_b^{\mathbb{W}(t_{l-1})}}{d\cos^2(\phi_{i_l})\sin^2(\phi_{i_l})(r^\star(i_{l+1}, i_l) - \frac{p_{i_{l+1}}}{p_{i_l}})} \mathbb{W}(i_{l+1}, i_l), \end{aligned}$$

and the initialization conditions $t_0 = 0$ and $\mathbb{W}(0) = \lambda \mathbb{I}_d$. □

4.7.4 Proof of Lemma 4.4.2

Lemma 4.4.2. *[Dual Reachability Analysis] Let's assume we are currently playing transient region j and we reach the region i at time t_l. We then have the following two possible cases:*

- *If $\frac{p_i}{p_j} > r^\dagger(i,j)$, we say that regions i is dual reachable from region j, leading to a steady state regime with bi-region (i,j) effective dimension. In such case, for $t \geq t_l$, the potential increase is of the form:*

$$\mathbb{W}(t) = \frac{1}{p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j) - p_i/p_j}{p_i/p_j - r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_i(t+\lambda^\star)\mathbb{W}(i,j).$$

- *Otherwise, we switch from base region j to base region i and continue in the transient regime.*

Proof. Using previous section, we know that $\mathbb{W}(t_l) \propto \mathbb{W}(i,j)$ where we recall that the matrix $\mathbb{W}(i,j)$ has the strong property that the gains in regions i and j are equal i.e.:

$$\frac{1}{p_i}\operatorname{Tr}(\mathbb{W}(i,j)^{-1}\mathbb{W}_i) = \frac{1}{p_j}\operatorname{Tr}(\mathbb{W}(i,j)^{-1}\mathbb{W}_j).$$

One of the main result we show in the multi-threshold censorship model is that for $t \geq t_l$, we have:

$$\mathbb{W}(t) - \mathbb{W}(t_l) \propto (t - t_l)\mathbb{W}(i,j),$$

which involves in particular that for $t \geq t_l, \mathbb{W}(t) \propto \mathbb{W}(i,j)$. This is possible thanks to the fact that the optimal policy produces a combination of $p_i\mathbb{W}_i$ and $p_j\mathbb{W}_j$ proportional to $\mathbb{W}(i,j)$ so that optimally of both regions i and j is maintained while maximal first-order gain is simultaneously ensured. The proportionality condition is then written as the existence of $\mu_i, \mu_j > 0$ such that $p_i\mu_i\mathbb{W}_i + p_j\mu_j\mathbb{W}_j \propto \mathbb{W}(i,j)$ or equivalently as:

$$\exists \mu_i, \mu_j > 0, \quad \frac{\frac{1}{d-1}[p_i\mu_i \cos^2(\phi_i) + p_j\mu_j \cos^2(\phi_j)]}{p_i\mu_i \sin^2(\phi_i) + p_j\mu_j \sin^2(\phi_j)} = \frac{\cos^2(\phi_j)(u(i,j) - \frac{p_i}{p_j})}{\sin^2(\phi_j)(\frac{p_i}{p_j} - l(i,j))} \triangleq R,$$

where μ_i and μ_j are the infinitesimal time increase in regions i and j. It leads in turn to the ratio equality:

$$\frac{p_i\mu_i}{p_j\mu_j} = \frac{\sin^2(\phi_j)(d-1)R - \cos^2(\phi_j)}{\cos^2(\phi_i) - \sin^2(\phi_i)(d-1)R} = \frac{(d-1)u + l - d\frac{p_i}{p_j}}{(u + (d-1)l)\frac{p_i}{p_j} - dlu} = \frac{d}{u + (d-1)l}\frac{r^\star(i,j) - \frac{p_i}{p_j}}{\frac{p_i}{p_j} - r^\dagger(i,j)}.$$

Thus, we see that bi-region stationarity is possible if and only if $\frac{p_i}{p_j} > r^\dagger(i,j)$ where we introduced the dual reachability condition:

$$r^\dagger(i,j) \triangleq \frac{dl(i,j)u(i,j)}{u(i,j) + (d-1)l(i,j)} = \Big(\frac{\frac{d-1}{u(i,j)} + \frac{1}{l(i,j)}}{d}\Big)^{-1} = \Big(\frac{1}{d}\operatorname{Tr}(\mathbb{W}_i^{-1}\mathbb{W}_j)\Big)^{-1} = \frac{1}{r^\star(j,i)}.$$

Hence, the use of the term dual reachability comes from the fact that region i is dual reachable from region j if and only if region j is reachable from region j. In such case, further algebraic calculation then lead to the instantaneous potential increase ∂W for infinitesimal time $\partial t \triangleq \mu_j + \mu_j$:

$$\partial W(\partial t) \triangleq p_j\mu_j\mathbb{W}_j + p_i\mu_i\mathbb{W}_i = \frac{u - l}{u + (d-1)l}\frac{1}{\frac{p_i}{p_j} - r^\dagger(i,j)}p_j\mu_j\mathbb{W}(i,j).$$

We then note that:

$$\frac{\mu_i + \mu_j}{\mu_j} = 1 + \frac{1}{\frac{p_i}{p_j}}\frac{d}{u + (d-1)l}\frac{r^\star(i,j) - \frac{p_i}{p_j}}{\frac{p_i}{p_j} - r^\dagger(i,j)}.$$

Therefore, we conclude that:

$$
\begin{aligned}
\partial W(\partial t) &= \frac{1}{p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j-r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_i(\mu_j+\mu_i)\mathbb{W}(i,j) \\
&= \frac{1}{p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j-r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_i \partial t \mathbb{W}(i,j).
\end{aligned}
$$

We then introduce $\lambda^\star$ defined such that:

$$
(t_l+\lambda^\star)\mathbb{W}(i,j) \triangleq \frac{1}{p_i}\frac{(u+(d-1)l)(p_i/p_j - r^\dagger(i,j))}{u-l}\left(p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{\frac{p_i}{p_j}-r^\dagger(i,j)}\right)\mathbb{W}(t_l).
$$

Given the previous two results, we conclude that for all $t \geq t_l$:

$$
\mathbb{W}(t) = \frac{1}{p_i/p_j + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j-r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_i(t+\lambda^\star)\mathbb{W}(i,j).
$$

Note that entering the bi-region stationary regime impedes new regions to be reachable. Indeed, going back to the initial definition of reachability, region n is said to be reachable from region j after time t_l if and only if:

$$
\exists t \geq t_l, \quad \frac{1}{p_n}\operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_n) = \frac{1}{p_j}\operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_j).
$$

Yet, using previous result on the evolution of $\mathbb{W}(t)$, we know that the ratio of those two quantities remain equal for any $t \geq t_l$ i.e. no new regions can be reached.

Moreover, using the optimality criterion of Lemma 4.4.1, when several regions are reached simultaneously, the tie-breaking is performed by considering the most censored region, i.e. the one with the highest i index. If the chosen region is not dual reachable, then the next one is considered. In the case where none of them is dual reachable, the base region becomes the maximally censored region and we immediately reiterate the procedure described in Lemma 4.4.2.

□

4.7.5 Proof of Lemma 4.7.1

Lemma 4.7.1 (Bi-Region Effective Dimension)**.** *Let's assume we reach a bi-region (i,j) steady state regime at time $t_l \leq T$. Then, we have:*

$$\int_{t_l}^{T} \frac{1}{p(a(t))} \frac{\partial \log \det(\mathbb{W}(t))}{\partial t} \partial t = d_{\mathit{eff}} \log(1 + \frac{T - t_l}{t_l + \lambda^\star}) \sim d_{\mathit{eff}} \log(T),$$

where $d_{\mathit{eff}} = \frac{1}{p_j}\left[(d-1)\frac{1-l(i,j)}{p_i/p_j - l(i,j)} + \frac{u(i,j)-1}{u(i,j)-p_i/p_j}\right]$ and $\lambda^\star$ is given in the proof of Lemma 4.4.2. Moreover, we have the cumulative transient potential:

$$\begin{aligned}\int_{0}^{t_l} \frac{1}{p(a(t))} \frac{\partial \log \det(\mathbb{W}(t))}{\partial t} \partial t &= \sum_{n=1}^{l} \frac{1}{p_{i_n}} \int_{t_{n-1}}^{t_n} \partial \log \det(\mathbb{W}(t)) = \sum_{n=1}^{l} \frac{1}{p_{i_n}} \log \frac{\det(\mathbb{W}(t_n))}{\det(\mathbb{W}(t_{n-1}))} \\ &= \sum_{n=1}^{l} (\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}}) \log \det \mathbb{W}(t_n).\end{aligned}$$

Proof. For $t \geq t_l$, we have the infinitesimal two-step increase ∂G during the infinitesimal time $\partial t \triangleq \mu_i + \mu_j$:

$$\begin{aligned}\partial G(\partial t) &\triangleq \mu_i \operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_i) + \mu_j \operatorname{Tr}((\mathbb{W}(t) + \mu_i p_i \mathbb{W}_i)^{-1}\mathbb{W}_j) \\ &= \mu_i \operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_i) + \mu_j \operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_j) + o(\partial t) \\ &= \frac{p_i\mu_i + p_j\mu_j}{p_j} \operatorname{Tr}(\mathbb{W}(t)^{-1}\mathbb{W}_j) + o(\partial t),\end{aligned}$$

where we used the property of $\mathbb{W}(i,j)$. Invoking lemma 4.4.2, we know the evolution of $\mathbb{W}(t)$ for $t \geq t_l$:

$$\mathbb{W}(t) = \frac{1}{1 + \frac{1}{p_i/p_j}\frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j - r^\dagger(i,j)}} \frac{u-l}{u+(d-1)l} \frac{1}{p_i/p_j - r^\dagger(i,j)} p_j(t+\lambda^\star)\mathbb{W}(i,j),$$

as well as the relations between μ_i and μ_j:

$$\begin{cases} \frac{p_i\mu_i + p_j\mu_j}{p_j} &= \mu_j(1 + \frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j - r^\dagger(i,j)}) \\ \frac{\mu_i+\mu_j}{\mu_j} &= 1 + \frac{1}{p_i/p_j}\frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{\frac{p_i}{p_j} - r^\dagger(i,j)}. \end{cases}$$

We invoke the fact that $\mathrm{Tr}(\mathbb{W}(i,j)^{-1}\mathbb{W}_j) = \frac{1}{u-p_i/p_j} + \frac{1}{p_i/p_j-l}$ to conclude that:

$$\begin{aligned}
\partial G(\partial t) &= \frac{1}{p_j}\frac{[(d-1)l+u-d]\frac{p_i}{p_j} - [dlu-((d-1)u+l)]}{(u-\frac{p_i}{p_j})(\frac{p_i}{p_j}-l)}\frac{(1+\frac{1}{p_i/p_j}\frac{d}{u+(d-1)l}\frac{r^\star(i,j)-p_i/p_j}{p_i/p_j-r^\dagger(i,j)})\mu_j}{t+\lambda^\star} \\
&= \frac{1}{p_j}\left[(d-1)\frac{1-l}{\frac{p_i}{p_j}-l} + \frac{u-1}{u-\frac{p_i}{p_j}}\right]\frac{\partial t}{t+\lambda^\star} \\
&= d_{\mathit{eff}}\frac{\partial t}{t+\lambda^\star}.
\end{aligned}$$

Given that ∂t is an infinitesimal time increase, we have in the steady state regime:

$$\int_{t_l}^{T}\partial G = d_{\mathit{eff}}\int_{t_l}^{T}\frac{\partial t}{t+\lambda^\star} = d_{\mathit{eff}}\log(\frac{T+\lambda^\star}{t_l+\lambda^\star}) = d_{\mathit{eff}}\log(1+\frac{T-t_l}{t_l+\lambda^\star}).$$

We finally note that the cumulative potential coming from the transient period is equal to:

$$\begin{aligned}
\int_0^{t_l}\partial G &= \sum_{n=1}^{l}\frac{1}{p_{i_n}}\int_{t_{n-1}}^{t_n}\partial\log\det(\mathbb{W}(t)) = \sum_{n=1}^{l}\frac{1}{p_{i_n}}\log\frac{\det(\mathbb{W}(t_n))}{\det(\mathbb{W}(t_{n-1}))} \\
&= \sum_{n=1}^{l}(\frac{1}{p_{i_n}} - \frac{1}{p_{i_{n+1}}})\log\det\mathbb{W}(t_n),
\end{aligned}$$

where the closed-form expression of $\mathbb{W}(t_n)$ is given in Corollary 4.7.0.1. □

4.7.6 Special case: Single-threshold model

Corollary 4.4.2.1. *For the single threshold model with two regions* 0 *and* 1 *and associated censorship probabilities* $p_0 < p_1$*, our main theorem yields:*

- *If* $\frac{p_0}{p_1} < \frac{d-1}{d\cos^2(\phi_1)}$*, then we reach bi-region steady state regime and have the effective dimension:*

$$d_{\mathit{eff}} = \frac{d-1}{p_0} + \frac{1}{p_0}\frac{\sin^2(\phi_1)}{\frac{p_1}{p_0}-\cos^2(\phi_1)} \in [\frac{d}{p_0},\frac{d}{p_1}].$$

- *Otherwise, we are from* $t=0$ *in single-region steady state regime and have the* *effective dimension* $d_{\mathit{eff}} = d/p_1$.

Proof. Using Lemma 4.4.2 in the case of the single threshold model, we note that if region 0 is reachable, it is necessarily dual reachable given that $r^{\dagger}(0,1) = 0$ and henceforth, we always have $p_0/p_1 > r^{\dagger}(0,1)$. Thanks to the results of Lemma 4.4.1, we also note that $r^{\star}(0,1) = \frac{p_0}{p_1} < \frac{d-1}{d\cos^2(\phi_1)}$ and that if region 0 is reachable, it is done in a time:

$$t_1 = \frac{1}{p_1} \frac{(d-1)\lambda}{d\sin^2(\phi_1)\cos^2(\phi_1)} \frac{\frac{p_0}{p_1} - 1}{\frac{d-1}{d\cos^2(\phi_1)} - \frac{p_0}{p_1}}$$

□

Chapter 5

Conclusion and Future Directions

In this work, we demonstrate that the complexity of bandit learning under censorship is governed by the notion of effective dimension. To do so, we developed a novel analysis framework which enables us to precisely estimate this quantity first for general censored MAB bandits and then for a broad class of censorship models in the CB case. From a methodological viewpoint, the key contributions are the analysis of the dynamics of learning through a continuous lens, the characterization of the adaptivity gain and the fruitful introduction of the multi-threshold models. An important future work would be to extend our approach to Bayesian settings, which will likely provide us with useful insights on the cumulative censored potential $\mathbb{V}_\alpha$, as initiated by [22]. Future work also includes the study of time-dependent censorship models such as exponential counting processes or Markov Decision Processes. Moreover, we believe that our new perspective can be exploited to design robust algorithms operating efficiently in unknown (and possibly adversarial) environments. To conclude this work, we further detail two directions deemed of particular relevance with respect to our initial questioning:

- **Towards a generic assessment of the value of information in sequential settings**: More generally, modelling relations between user preferences (i.e. latent state) and perceived feedback(s) as a Dynamic Choice Structural Models [4] allows to extend our approach to more complex behavioral processes. Building blocks of this structure include sequences of nested interdependent decisions (censorship process in what precedes) and of feedback (realized reward). On the one hand, regarding decision modeling, a first degree of variability is to be introduced in terms of amount of randomness present in the decision making process, ranging in term of difficulty from deterministic utility maximization to uniformly random decisions. Another one degree deals with identifiability of the underlying latent state driving the output. For instance, the multi-threshold decision model doesn't allow direct identifiability of the latent state in the neighborhood of the optimal action as the derivative with respect to such latent state is null. On the other hand, feedback analysis also presents similar complexity of randomness and identifiability issues but usually under a different class of parametric models. A complete and normative study of the resulting difficulty of decision making under uncertainty arising from the Graphical structure of the Dynamic Choice Model is to the best of our knowledge an open and fascinating challenge.

- **Some connections with Information Directed Sampling and Partial Monitoring**: Finally, we believe that this work sketches some deeper connections between online statistical learning literature and information theory. Indeed, it is possible to see the effective dimension as a form of measure of the communication capacity between the principal and the environment and this raises several fascinating questions. Firstly, it would seem very relevant to further investigate the links between adaptativity in the MAB framework and the impact of feedback in non-asymptotic channel coding [38]. Secondly, and more related to the recent literature on Information Directed Sampling (IDS) and partial monitoring, our work allows us to study in a tractable way the evolution of information flows through a continuous lens. An open challenge would be to extend the analysis of IDS to allow for the adoption of this methodology as well, thus unlocking more subtle dimensional estimates for a larger class of problems.

Bibliography

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[2] Jacob Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandit, with and without censored feedback. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4896–4904, Red Hook, NY, USA, 2016. Curran Associates Inc.

[3] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[4] Victor Aguirregabiria and Pedro mira. Dynamic Discrete Choice Structural Models: A Survey. Technical Report tecipa-297, University of Toronto, Department of Economics, July 2007.

[5] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

[6] Xueying Bai, Jian Guan, and Hongning Wang. A model-based reinforcement learning with adversarial training for online recommendation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[7] Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited, 2020.

[8] Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 605–622, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[9] Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The assistive multi-armed bandit. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, page 354–363. IEEE Press, 2019.

[10] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016.

[11] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017.

[12] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[14] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Computationally and statistically efficient truncated regression, 2020.

[15] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples, 2020.

[16] Constantinos Daskalakis, Dhruv Rohatgi, and Manolis Zampetakis. Truncated linear regression in high dimensions, 2020.

[17] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. *arXiv preprint arXiv:1106.2369*, 2011.

[18] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[19] Manegueu Anne Gael, Claire Vernade, Alexandra Carpentier, and Michal Valko. Stochastic bandits with arm-dependent delays. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3348–3356. PMLR, 13–18 Jul 2020.

[20] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning, 2016.

[21] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design, 2017.

[22] Nima Hamidi and Mohsen Bayati. The randomized elliptical potential lemma with an application to linear thompson sampling, 2021.

[23] James Honaker and Gary King. What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(3):561–581, 2010 2010.

[24] Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Delay and cooperation in nonstochastic linear bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4872–4883. Curran Associates, Inc., 2020.

[25] Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1453–1461, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[26] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 358–384. PMLR, 06–09 Jul 2018.

[27] Tal Lancewicki, Shahar Segal, Tomer Koren, and Y. Mansour. Stochastic multi-armed bandits with unrestricted delay distributions. In *ICML*, 2021.

[28] Tor Lattimore and Csaba Szepesvari. An information-theoretic approach to minimax regret in partial monitoring. In *COLT*, 2019.

[29] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, 2020.

[30] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[31] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.

[32] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data.* Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2002.

[33] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2849–2856. AAAI Press, 2015.

[34] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

[35] Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[36] Therese D. Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

[37] Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grunewalder. Bandits with delayed, aggregated anonymous feedback. In *International Conference on Machine Learning*, pages 4105–4113. PMLR, 2018.

[38] Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. Feedback in the non-asymptotic regime. *IEEE Transactions on Information Theory*, 57(8):4903–4925, 2011.

[39] Chao Qin and Daniel Russo. Adaptivity and confounding in multi-armed bandit experiments, 2022.

[40] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

[41] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

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

[43] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 1015–1022, Madison, WI, USA, 2010. Omnipress.

[44] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory.*

[45] Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brückner. Linear bandits with stochastic delayed feedback. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9712–9721. PMLR, 13–18 Jul 2020.

[46] Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[47] Han Wu and Stefan Wager. Thompson sampling with unrestricted delays. *arXiv preprint arXiv:2202.12431*, 2022.

[48] Fuwen Yang and Yongmin Li. Set-membership filtering for systems with sensor saturation. *Automatica*, 45(8):1896–1902, 2009.

[49] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[50] Zhengyuan Zhou, Renyuan Xu, and Jose Blanchet. Learning in generalized linear contextual bandits with stochastic delays. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.