Mathematics of Statistical Sequential Decision Making

Odalric-Ambrym Maillard

HABILITATION

à Diriger les recherches

manuscrit préparé au sein du laboratoire CRIStAL
et du centre de recherche Inria Lille - Nord Europe

…SequeL…

Mathematics of Statistical Sequential Decision Making

(Mathématique de la prise de décision séquentielle statistique)

par

Odalric-Ambrym Maillard

École doctorale des Sciences pour l'ingénieur

Discipline : Sciences Mathématiques.

Soutenue publiquement à l'**Université de Lille**, le **11 Février 2019** devant le jury composé de:

Pierre	Alquier	Ensae, Université Paris-Saclay, France	Examinateur (Président)
Peter	Grünwald	CWI and Leiden University, Pays-bas	Rapporteur
Joëlle	Pineau	McGill University, Canada	Examinateur
Vianney	Perchet	ENS Paris-Saclay, France	Rapporteur
Philippe	Preux	Université de Lille, France	Garant
Alexandre	Proutière	KTH Royal Institute of Technology, Suède	Rapporteur

Abstract

In this document, we give an overview of recent contributions to the mathematics of statistical sequential learning. Unlike research articles that start from a motivating example and provide little room to the mathematical tools in the main body of the article, we here give primary focus to these tools, in order to stress their potential as well as their role in the development of improved algorithms and proof techniques in the field. We revisit in particular properties of the log Laplace transform of a random variable, the handling of random stopping time in concentration of measure of empirical distributions, and we highlight the fundamental role of the "change of measure" argument both in the construction of performance lower-bounds as well as near-optimal strategies. We then give focus to obtaining finite-time error guarantees on the parameter estimation in parametric models before highlighting the strength of Legendre-Fenchel duality in the design of risk-averse and robust strategies. Finally, we turn the setting of Markov decision processes where we present some key insights for the development of the next generation of decision strategies. We end this manuscript by providing a more focused presentation of three key contributions in bandit theory, stochastic automata, and aggregation of experts.

Keywords: Sequential Learning, Multi-armed bandits, Reinforcement Learning, Concentration of Measure, Mathematical Statistics.

Résumé

Ce document montre un tour d'horizon de quelques contributions récentes à la mathématique de l'apprentissage statistique séquentiel. Contrairement aux articles de recherches qui partent d'exemples et donnent peu de place aux outils mathématiques, souvent relayés en annexe, nous présentons ici ces outils en pleine lumière, afin de souligner leur rôle capital dans le développement de nouvelles stratégies de prise de décision séquentielle dans l'incertain. Nous revisitons en particulier les propriétés de la transformée de Laplace d'une variable aléatoire, la prise en compte des temps d'arrêt pour la concentration de distributions empiriques, avant de souligner le rôle fondamental du « changement de mesure » dans la construction à la fois des meilleures bornes de performances atteignables et des stratégies quasi-optimales. Nous nous tournons ensuite vers l'obtention de bornes d'erreur en temps fini pour l'estimation de paramètre dans différents modèles paramétriques, avant d'expliquer le rôle clé de la dualité de Legendre-Fenchel dans la construction de stratégies robustes et sensibles au risque. Enfin, nous présentons, dans le cadre des processus décisionnels de Markov, de nouveaux éléments de compréhension utiles à la découverte de nouvelles stratégies de prise de décision séquentielle. Ce manuscrit se termine par une présentation plus détaillée de trois contributions clés à la théorie de bandits, aux automates stochastiques ainsi qu'à l'agrégation d'experts.

Mots clés : Apprentissage Séquentiel, Bandits Manchots, Apprentissage par Renforcement, Concentration de la Mesure, Statistique Mathématique.

Remerciements.

Les personnes que l'on rencontre nous influencent et nous aident de multiples façons, sans forcément que l'on s'en aperçoive, sans forcément que l'on en soit conscient. C'est sans doute cette diversité de points de vue qui est le moteur le plus efficace pour imaginer de nouveaux concepts, trouver des solutions inédites, nourrir et générer un enthousiasme et une énergie constante.

J'aimerais remercier ici plusieurs personnes, en m'excusant d'avance très sincèrement pour les personnes que j'aurais pu oublier. Ceci inclut les membres du jury, mes co-auteurs, les étudiants avec qui j'ai pu interagir, et l'Inria, notamment à travers les services de soutien à la recherche. Enfin, j'ai la chance, au sein de l'équipe SequeL, d'avoir des collègues merveilleux – Émilie Kaufmann, Michal Valko et Philippe Preux – qui sont chacun des chercheurs et des personnes exceptionnels. J'aimerais également remercier mes amis, mes parents, mes frères et ma soeur, ceux dont on ne parlent pas dans les articles, et bien évidemment les lecteurs et lectrices de ce manuscrit.

Enfin, l'habilitation à diriger les recherches est peut-être le bon moment pour remercier mes anciens professeurs, celles et ceux qui ont su m'initier et me transmettre leur passion pour la mathématique, cet art, qui au même titre que la musique ou la danse possède son propre langage, ses propres aspirations, et sa propre quête de sens, à la portée universelle.

Foreword: To the layman reader.

Let us consider you receive, one by one, a sequence of observations $Y^{(1)}, Y^{(2)}, Y^{(3)}, \cdots \in \mathbb{R}^d$, where $Y^{(1)}$ is the first observation you receive, $Y^{(2)}$ the second one, and so on and so forth. These observations may be of different nature, think for instance of each $Y^{(i)}$ as coming from

- sensors of a robot, such as video, audio, touch, mention or battery sensors.
- time-geographic whether data such as temperature, wind, humidity, gathered from many different probes.
- user data such as in health-care (history of past diseases of a patient, genetic data, response to a specific medicine), in web-advertisement (navigation history and clicked ads), or in job-care (CV of a person, location, interest in a given job).
- node/edge observations from an electric, water, producer-consumer or ecological network.

"All models are wrong" After having received many data, it seems natural to start understanding this sequence, what are its regularities, and what could be the next observation. In full generality, there is however no reason that the $n \in \mathbb{N}$ first observations are any informative about observation $Y^{(n+1)}$. For instance, let us denote $Y_1^{(n')} \in \mathbb{R}$ the first component of $Y^{(n')} = (Y_1^{(n')}, \dots, Y_d^{(n')})$ for each $n' \in \mathbb{N}$. Let us consider also that there exists $n_0 \in \mathbb{N}$, $n_0 \leqslant n$ such that

$$\max_{n' \leqslant n_0} Y_1^{(n')} = \max_{n' \leqslant n} Y_1^{(n')} .$$

Now, even if, say, $n_0/n < 10^{-5}$, that is you haven't seen the maximal value of the sequence change for a large amount of time, nothing prevents the first component of the next observation $Y_1^{(n+1)}$ to exceed this maximum. Thus, from a worst case perspective, the task of trying to say anything more than $Y^{(n+1)} \in \mathbb{R}^d$ is prone to error. Note, even worse, that considering $Y^{(n+1)} \in \mathbb{R}^d$ should be considered as an assumption, and for an arbitrary sequence given to you, it could be that in fact $Y^{(n+1)} \in \mathbb{C}^d$ (for instance). We thus have to consider that whatever restricting assumption we put on the observed sequence, this assumption may be wrong and contradicted by the next observation. Let us formalize a little bit this concept before continuing: We say an assumption can be contradicted if it implies that the next observation $Y^{(n+1)}$ should belong to a specific set $\mathcal{S}$, in which case either $Y^{(n+1)} \in \mathcal{S}$ or not. We further ask that this property can be decided. In the sequel, a set of such assumptions that be contradicted will be called a model:

Definition 1 (Model) *A model on the sequences of observations is a set of assumptions such that for each $n \in \mathbb{N}$, there exists a set $\mathcal{S}_n$ enforcing these assumptions such that $(Y^{(1)}, \dots, Y^{(n)}) \in \mathcal{S}_n$ is decidable.*

Scoring and adding stuff Although this notion of model looks fairly generic, in many situations however, one may want to extend this notion beyond using a single (or countably many) set. For instance, one may want to consider that both sets $[0, 1]$ or $[0, 2]$ are valid, but value more $[0, 1]$ than the other for some reason. Hence if we observe that $Y_1^{(n+1)} = 0.7 \in [0, 1]$ we may give this property a score of 1 and if $Y_1^{(n+1)} = 1.2 \in [0, 2] \setminus [0, 1]$

we give this property a lower score, say of 0.4. This score means that we prefer having an observation in $[0,1]$ over having an observation in $[0,2]$. We may further weight differently each set $[-a,a]$ for all $a \in \mathbb{R}$, each with a score $w_a \in [0,1]$, or even give a different weight to each singleton set $\{y\}$ for $y \in \mathbb{R}$. More generally, for some set Ω of possible observations, we consider subsets $\mathcal{B} \subset \Omega$ and want to give them a score.

Let us introduce a mild assumption on this notion of score: Basically, we would like to be able to "**add stuff**", in the sense that the score given to $[0,1] \cup [2,3]$ equals the score given to $[0,1]$ plus the score of $[2,3]$. More generally, it is natural to ask that the score given to any countable union of disjoint sets is the sum of the score of each set (countably additive), that is we want the score to be a measure, which we denote μ. In order to define the score in a meaningful way, the traditional approach is to define it on a specific sub-collection of subsets of Ω. We consider what is called a σ-algebra of Ω, that is a collection Σ of subsets of Ω containing at least the empty subset, and such that all complement, countable intersection or countable unions of sets of Σ still belong to Σ. This construction is natural in view of the countably additive property of the measure $\mu : \Sigma \to \mathbb{R}$. The tuple (Ω, Σ, μ) is simply called a measure space.

Coming back to our observations, given a set of possible outcomes Ω for a sequence $Y^{(1)}, Y^{(2)}, \dots$ of arbitrary finite length, a measure μ and a set $\mathcal{S} \in \Sigma$, the quantity $\mu(\mathcal{S}) \in \mathbb{R}$ tells us how much we value that our observations fall into $\mathcal{S}$. This leads to the following notion of measure model (we define likewise a measure model for sequences of length n or even 1):

Definition 2 (Measure model) *A measure model for sequences is a measure space (Ω, Σ, μ) where Ω is a set of outcomes of sequences of observations. For any $\mathcal{S} \in \Sigma$, we denote its measure by $\mu(\mathcal{S})$ or more explicitly*

$$\mu(Y^{(1)}, Y^{(2)}, \dots \in \mathcal{S}) \in \mathbb{R}\,.$$

Measures with richer properties It is convenient to consider cases when the observation space Ω is not arbitrary, but has some richer structure. Hence we ask the measure to obey some additional properties:

- The first case is when Ω is equipped with a specific topology τ: We want to be able to talk about neighborhoods of each points (topological space), such that each point has a compact neighbourhood (locally compact) and different points have different neighbourhoods (separable aka Hausdorff). In this case, we can define the smallest σ-algebra containing the open sets, called the Borel algebra, and it is enough to define our scores on this Borel algebra. The score function μ then qualifies as a Borel measure. For better compatibility with the considered topology, we may further want that $\mu(A) = \sup\{\mu(K) : \text{compact } K \subset a\}$ (inner regular) and that every point y has a neighbourhood A of finite score $\mu(A)$ (locally finite). Under these additional mild conditions, the score is now called a Radon measure.

- A second situation is when Ω is equipped with a metric ℓ, and we further ask Ω to be complete (every Cauchy sequence of Ω converges in Ω). Note that a metric naturally enables to define neighbourhoods and thus an induced topology that we denote τ_ℓ. Such a space $(\Omega, \ell, \tau_\ell)$ when τ_ℓ satisfies the above properties, is called a Polish space. A typical example of Polish space is $\mathbb{R}$, with Euclidean metric $d(x,y) = |x-y|$ and associated topology. The Borel algebra is generated by the sets $\{(-\infty, r) : r \in \mathbb{R}\}$. A typical Radon measure for the topology τ_d is the Lebesgue measure, usually denoted by λ.

Likewise, we may add further constraints to the measure μ. If $\mu(\Omega) = 1$, we call it a stochastic measure, and if further $\mu(\mathcal{S}) \geqslant 0$ for each $\mathcal{S} \in \Sigma$, we call it a probability measure. We naturally give specific names to a measure model having specific properties: Hence a stochastic model is a measure model with a stochastic measure, a probability model is a measure model with a probability measure. In the sequel, we will mostly focus on models that combine all previous assumptions and are simply termed "Probabilistic":

Definition 3 (Probabilistic model) *A probabilistic model is a measure space (Ω, Σ, μ) where Ω, equipped with metric ℓ and topology τ_ℓ is a Polish space, Σ is the Borel algebra induced by τ_ℓ, and μ is a Radon probability measure.*

A first probabilistic model Intuitively, while a *model* enables to decide whether a sequence of observations belong or not to the model, a *probabilistic model* values certain parts of the space more than other parts in a quantified way, and in a sense always considers a sequence of observations to be possibly outside of the model. A typical example of probabilistic model for sequences of length 1 is built with $\Omega = \mathbb{R}$, using the Euclidean metric d. Unfortunately $(\Omega, \Sigma, \lambda)$ is not a probabilistic model as the Lebesgue measure is not a probability measure. We can use instead the Gaussian measure with mean $m \in \mathbb{R}$ and variance $\sigma^2 > 0$ defined for each Borel set S by

$$g_{m,\sigma^2}(S) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_S \exp(-\frac{1}{2\sigma^2}|y-m|^2) d\lambda(y)\,,$$

where we introduce the integral, defined for a non-negative function $f : \Omega \to \mathbb{R}_+$ as

$$\int_S f(y)d\lambda(y) = \sup\left\{ \sum_{k=1}^K a_k \lambda(A_k) : K \in \mathbb{N}, a_k \in \mathbb{R}_+ \text{ and } \forall y \in S,\ \sum_{k=1}^K a_k \mathbb{I}\{y \in A_k\} \leqslant f(y) \right\},$$

where $\mathbb{I}\{y \in A_k\}$ equals 1 if $y \in A_k$ and 0 else. This quantity may be infinite, and extends to any signed function f by $\int_S f(y)d\lambda(y) = \int_S \max(f(y),0)d\lambda(y) - \int_S \max(-f(y),0)d\lambda(y)$ provided that both terms are finite; we say in this case that f is measurable with respect to the measure λ. Now it can be checked that $g_{m,\sigma^2}(\mathbb{R}) = 1$. In order to complete this presentation, let us remark that by construction, the Gaussian measure is in the form $g_{m,\sigma^2}(S) = \int_S f(y)d\lambda(y)$ for all measurable set S, where we introduced the function $f(y) = \exp(-\frac{1}{2\sigma^2}|y-m|^2)/\sqrt{2\pi\sigma^2}$. This function f is called the Radon-Nikodym derivative of g_{m,σ^2} with respect to λ and is denoted $\frac{dg_{m,\sigma^2}}{d\lambda}$. When the reference measure λ is specified without ambiguity, we may simply denote it g_{m,σ^2}, without confusion since the measure is $g_{m,\sigma^2} : 2^\Omega \to [0,1]$ while the derivative is $g_{m:\sigma^2} : \Omega \to \mathbb{R}_+$. Note for instance that $g_{m,\sigma^2}(\{y\}) = 0 \neq g_{m,\sigma^2}(y)$.

The previous probabilistic model is defined for a single observation. However it easily extends to $\Omega = \mathbb{R}^\star \stackrel{\text{def}}{=} \bigcup_{n\in\mathbb{N}} \mathbb{R}^n$, that captures all real-valued sequences of finite length. The topology can be extended to Cartesian products $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ with the product topology whose sets are called the cylinder sets. This in turns generates the cylinder σ-algebra and Borel cylinder sets. Likewise, a product space $\mathbb{R}^n$ of probabilistic models $(\mathbb{R}, \Sigma_i, \mu_i)_{i\leqslant n}$ naturally inherits a measure from the measures $(\mu_i)_{i\leqslant n}$ by the product measure μ defined on the Borel cylinder sets: Such a measure satisfies, for each $n \in \mathbb{N}$ and Borel sets $\mathcal{S}_1, \ldots, \mathcal{S}_n \subset \mathbb{R}$ that

$$\mu(Y^{(1)}, \ldots, Y^{(n)} \in \prod_{i=1}^n \mathcal{S}_i) = \prod_{i=1}^n \mu_i(Y^{(i)} \in \mathcal{S}_i)\,.$$

Note that here, each $Y^{(i)}$ is scored by the measure μ only according to μ_i, thus **independently** on the other measures $(\mu_j)_{j\neq i}$. For this reason, we also say that this probabilistic model considers that the observations are independent. Using our Gaussian model for observations of length 1 can be extended to a product measure by considering the same Gaussian measure $\mu_i = \sigma_{m,\sigma^2}$ for each i. In this case, we further remark that all measures are **identical**. Hence we say that this probabilistic model considers observations that are identically and independently distributed, or for short i.i.d. The typical probabilistic models for i.i.d. data that we will encounter include Bernoulli, Gaussian variables, and more generally exponential families (described later).

We have seen how to build a measure on $\mathbb{R}^\star$ from measures on $\mathbb{R}$. Conversely, any measure μ on $\mathbb{R}^\star$ can be used to define measures on $\mathbb{R}$, using the following product decomposition

$$\begin{aligned}\mu(Y^{(1)},\dots,Y^{(n)} \in \prod_{i=1}^{n} \mathcal{S}_i) &= \prod_{i=1}^{n} \mu_{Y^{(1)},\mathcal{S}_1,\dots,Y^{(i-1)},\mathcal{S}_{i-1}}(Y^{(i)} \in \mathcal{S}_i),\\ \text{where } \mu_{Y^{(1)},\mathcal{S}_1,\dots,Y^{(i-1)},\mathcal{S}_{i-1}}(Y^{(i)} \in \mathcal{S}_i) &= \frac{\mu(Y^{(1)},\dots,Y^{(i)} \in \prod_{j=1}^{i} \mathcal{S}_j)}{\mu(Y^{(1)},\dots,Y^{(i-1)} \in \prod_{j=1}^{i-1} \mathcal{S}_j)}.\end{aligned}$$

Note that in this generic decomposition, the i^{th} measure depends on all observations before i, but none after i. An interesting case is when there exists some $m \in \mathbb{N}$, such that for each i, the i^{th} measure only depends on the value of the last m *observations* before $Y^{(i)}$, namely of $Y^{(\max\{i-m,1\})},\dots,Y^{(i-1)}$ (but not on i, for $i > m$). Indeed in this case, μ is fully determined by $m+1$ measures on $\mathbb{R}$ (one for each $i \leqslant m$, and one for all $i > m$). Such models are called Markov of order m. Further, we recover the i.i.d. probabilistic models for $m = 0$.

Random variables and processes We are now ready to introduce the concept of random variable, that will be used extensively in all this manuscript. Given a probabilistic model (Ω, Σ, μ) and a measurable space $(E, \mathcal{E})$ (possibly equal to (Ω, Σ)), a random variable X is a function (sic!) $X : \Omega \to E$ such that $\{w \in \Omega : X(w) \in \mathcal{S}\} \in \Sigma$ holds for each measurable set $\mathcal{S} \in \mathcal{E}$ (Σ-measurable function). If $E = \mathbb{R}$, we naturally say the random variable is real-valued, if $E = \mathbb{R}^d$, we say it is vector valued, etc. A crucial object linked to the random variable is its law: we introduce the notation $\mathbb{P}_\mu(X \in \mathcal{S}) = \mu(\{w \in \Omega : X(w) \in \mathcal{S}\})$ and call μ the law of the random variable and write $X \sim \mu$ to say that X has law μ. Hence we see from this expression that a random variable is completely determined by the probability measure. The notation is coherent with the one introduced for models in the sense that in the trivial case when $(E, \mathcal{E}) = (\Omega, \Sigma)$ and $X(w) = w$, then $\mathbb{P}_\mu(X \in \mathcal{S}) = \mu(\{w \in \Omega : w \in \mathcal{S}\}) = \mu(\mathcal{S})$. Finally, when X is real-valued, we define its expectation (or mean) by $\mathbb{E}_\mu(X) = \int_{\mathbb{R}} y d\mu(y)$, whenever the identity function $y \to y$ is integrable with respect to its law μ.

The concept of random variable becomes more powerful when considering several random variables defined on the same probability space. For instance, instead of considering the sequence of observations as being a single random variable defined on a product space, an alternative view point is to consider each single $Y^{(i)} = X_i(w)$ as being the value of a different random variable X_i taken at a same point $w \in \Omega$. In that case, each variable may have its own associated probability measure μ_i. Hence from this standpoint, we have a collection $(X_i)_{i\in\mathbb{N}}$ of random variables (equivalently, a collection of probability measures $(\mu_i)_{i\in\mathbb{N}}$) indexed by the set of integers. This leads to the more general notion of process, that is a collection of random variables indexed by a set $\mathcal{I}$, or a function $\mathbb{X} : \mathcal{I} \times \Omega \to E$. The index set $\mathcal{I}$ does not have to be countable, in can be $\mathbb{R}$, or even a function space (function process). When $\mathcal{I}$ is totally ordered an interesting σ-algebra associated to the process $\mathbb{X}$ can be defined for each $i \in \mathcal{I}$: the smallest σ-algebra containing $\left\{X_j^{-1}(A) : j \in \mathcal{I}, j \leqslant i, A \in \mathcal{E}\right\}$, which we denote $\mathcal{F}_i^{\mathbb{X}}$, or more explicitly $\mathcal{F}(X_1,\dots,X_i)$ when $\mathcal{I} = \mathbb{N}_\star$. The collection $\mathcal{F}^{\mathbb{X}} = (\mathcal{F}_i^{\mathbb{X}})_{i\in\mathcal{I}}$ is a filtration and is called the natural filtration associated to the process $\mathbb{X}$.

In the rest of this document, we will encounter several of these processes that are of special interest in the context of sequential learning. For instance Markov processes of finite order enable to model dynamics in the observations, while function process indexed by a set of functions are particularly interesting to capture various kind of structures in real-valued observations.

The concept of "Likelihood" We now have a first example of a probabilistic model on sequences of observations thanks to the Gaussian measure on $\mathbb{R}$ and its extension to a product measure on sequences. We can

continue our tour and ask what to do with $Y^{(1)}, \dots, Y^{(n)}$: Since the measure provides a score for each set, it is tempting to interpret the Radon-Nikodym derivative as a score given to each observation point (note however, that this is not correct), and then to ask what is the score given to the sequence of observations that we actually receive. Following this rationale, and the product measure property, people have thus introduced and studied the following quantity (here specified to the Gaussian measure)

$$\mathcal{L}_{g_{m,\sigma^2}}(Y^{(1)}, \dots, Y^{(n)}) = \prod_{i=1}^{n} \frac{dg_{m,\sigma^2}}{d\lambda}(Y^{(i)}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{(Y^{(i)} - m)^2}{2\sigma^2} \right).$$

Let us remark that the closer the observations from point m, the larger this quantity. Such observations are more valued by the model, or put differently, they are more likely under this model than observations that are far from m. For this reason, the function $\mathcal{L}$, that can be defined more generally for any given probabilistic model, is called the likelihood function of the observations for the given probabilistic model.

At this point, given all the measures $\{g_{m,\sigma^2} : m \in \mathbb{R}\}$, it is equally tempting to ask which one maximizes the likelihood of the observations, in other-words, to look for a solution to

$$\sup_{m \in \mathbb{R}} \mathcal{L}_{g_{m,\sigma^2}}(Y^{(1)}, \dots, Y^{(n)}).$$

A solution, when it exists, is called the maximal likelihood estimate. An optimal value for our example using Gaussian measures exists, is unique and is given explicitly by

$$\widehat{m}_n = \arg\max_{m \in \mathbb{R}} \exp\left(-\sum_{i=1}^{n} \frac{(Y^{(i)} - m)^2}{2\sigma^2} \right) = \arg\min_{m \in \mathbb{R}} \sum_{i=1}^{n} (Y^{(i)} - m)^2 = \frac{1}{n} \sum_{i=1}^{n} Y^{(i)}.$$

This can be considered as a first instance of a learning problem, that is an optimization task based on observations. Assuming the observations are i.i.d. from a g_{m_0,σ^2} model, with $m_0 \in \mathbb{R}$, it is natural to ask how far is $\widehat{m}_n = \widehat{m}_n(Y^{(1)}, \dots, Y^{(n)})$ from m_0. A natural way to do so is to look at the mass given to the values of $\widehat{m}_n$ that are far away from m_0, say, for a given $\varepsilon > 0$,

$$\mathbb{P}_{g_{m_0,\sigma^2}}\Big(Y^{(1)}, \dots, Y^{(n)} \in S(m_0;\varepsilon)\Big) \text{where } S(m_0;\varepsilon) = \Big\{y_1, \dots, y_n : \big|\widehat{m}_n(y_1, \dots, y_n) - m_0\big| \geqslant \varepsilon\Big\},$$

which we simply denote $\mathbb{P}_{g_{m_0,\sigma^2}}\Big(\big|\widehat{m}_n(Y^{(1)}, \dots, Y^{(n)}) - m_0\big| \geqslant \varepsilon\Big)$. In the case of Gaussian $i.i.d.$ models, we will soon see we have an easy control on this concentration inequality. In the general case, the answer can be delicate: the observations can be dependent, the probabilistic model may have complicated parameters, and the maximal likelihood estimate may be tricky to compute (or do not exist); Also, we naturally want bounds that are valid for each number of observations n. We will present in part I of this document powerful tools in order to derive concentration inequalities.

Before moving on to the next step, let us point out that there are however a few known problems with the maximal likelihood estimate in general, and we warn the reader that maximizing is not necessary the correct way to make use of the likelihood function Skilling (2015). To get a hint why, let us point something a little awkward: remark that the value of the maximum in the Gaussian model is now given by

$$\mathcal{L}_{g_{\widehat{m}_n,\sigma^2}}(Y^{(1)}, \dots, Y^{(n)}) = \exp\left(-\frac{n}{2}\Big[\frac{\widehat{\sigma}_n^2}{\sigma^2} + \log(2\pi\sigma^2)\Big] \right) \text{where } \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (Y^{(i)} - m_n)^2.$$

Written in this form, we can observe that the value converges to 0 exponentially fast in general, which seems counterintuitive in view of the name "likelihood " given to this object. Indeed one may expect the likelihood of the observations, if they are all close to m, to be always high. This counterintuitive behavior potentially comes from over-interpreting the Radon-Nikodym derivative of the measure as a point-wise score in the first place, and highlights that combining the likelihood with *maximization* is **not** a good notion of score in order to assess how likely the observations are for the considered model. We will discuss this point and provide an alternative way when given a collection of measures. On the other hand, we will see later that the likelihood is especially appropriate in order to **compare** two models (rather than *selecting* one).

Parametric setup Now that we have discussed a first estimation task, it is convenient to introduce some further terminology. A family of probability measures $\mathcal{M}$ is said to be parametric if it is indexed by a subset Θ of $\mathbb{R}^d$ for some finite $d \in \mathbb{N}$, that is, such that each $\mu \in \mathcal{M}$ can be described uniquely by a parameter $\theta \in \Theta$. In this definition, the restriction is on the dimension of the parameters θ, that must be the same for all measures. Hence the collection of Gaussian measures $\{g_{m,\sigma^2} : m \in \mathbb{R}\}$ on $\mathbb{R}$ is an example of parametric family of dimension 1 with parameter $\theta = m \in \mathbb{R}$. The family $\{g_{m,\sigma^2} : m \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ has dimension 2. Now regarding our observations, we say we are in a parametric setup, or that the observation process is parametric, when the laws $(\mu_i)_{i\in\mathbb{N}}$ of the random variables $(Y^{(i)})_{i\in\mathbb{N}}$ belong to the same family. Note that for an i.i.d. process, all the laws are identical ($\mu_i = \mu$), hence it is parametric in case μ can be defined using a finite dimensional parameter θ. When each law μ_i belongs to a different family with dimension d_i, such that $\lim_{i\to\infty} d_i = \infty$, we sometimes say we face a semi-parametric setup.

Identifiability When assuming a parametric process, with parameter set Θ, it is natural, given observations, to try to identify for each i the parameter $\theta_i \in \Theta \subset \mathbb{R}^d$ corresponding to μ_i, that is we want to build an estimate of $\widehat{\theta}_{i,j} = \widehat{\theta}_i(Y^{(1)}, \dots, Y^{(j-1)}) \in \Theta$ of θ_i from observations before time j. This task can be done using ideas similar to the maximal likelihood estimate that we saw earlier, or other approaches. Before trying to estimate anything, a crucial question is whether it is possible at all to estimate θ_i from observations. Indeed, there is no reason a priori that there exists an estimation procedure such that $\forall \theta_i \in \Theta, \lim_{j\to\infty} \widehat{\theta}_{i,j} = \theta_i$. Such situations are unfortunately not uncommon: A typical example is when observation $Y^{(i)}$ corresponds to what is seen by an agent at its current location $s_i \in \mathcal{S}$ moving in a two dimensional set $\mathcal{S}$ that is invariant by rotation by $\pi/2$. That is, denoting the rotation R, in this situation $\mu_i = \mu_{s_i}$ and $\mu_{s_i} = \mu_{R(s_i)}$. In such a situation, if we try to infer our position s_i from the observations only, we will at best be able to estimate that we are in the set $\{s_i, R(s_i)\}$, hence the location s_i is not an identifiable parameter. This simple example shows that when considering a parameter set and an identification task one must always be cautious that the set consists only of identifiable parameters (or be happy recovering only the set of parameters indistinguishable from the targeted one).

Cones and orderings To continue with warnings, we now want to shed light on an important, and often overlooked notion. When looking for an estimate $\widehat{\theta}_i$ of θ_i, we will naturally compare the two quantities in order to assess how good is the estimate, and more importantly, we may want to understand how to move our current estimate in the direction of θ_i. The natural way to do so is by considering an order relation $\leqslant$ on Θ, and it turns out the correct way to build such order relations is thanks to the notion of cones. A cone $\mathcal{C}$ in a vector space $\mathcal{X}$ is a set that is stable by non-negative multiplication, namely $\{\lambda x : x \in \mathcal{C}, \lambda \in \mathbb{R}^+\} \subset \mathcal{C}$. An interesting property is that the relation defined by $x \leqslant_{\mathcal{C}} y$ if and only if $y - x \in \mathcal{C}$ is a partial order whenever $\mathcal{C}$ is convex, pointed ($0 \in \mathcal{C}$) and salient ($\forall x \neq 0, \{x, -x\} \notin \mathcal{C}$). A typical way to build such a cone (hence a partial order) from any point $x \in \mathcal{X}$ is by defining $\mathcal{C}^\star(x, p) = \{y \in \mathcal{X} : \langle y, x\rangle \geqslant p\|y\|\|x\|\}$, where $p \in [0, 1]$.

It is pointed and convex by construction, and salient for any $p > 0$. The definition extends to any set $\mathcal{S} \subset \mathcal{X}$ by $\mathcal{C}^\star(\mathcal{S}, p) = \bigcap_{x \in_c S} \mathcal{C}^\star(x, p)$, and $\mathcal{C}^\star(\mathcal{S}, 0)$ is called the dual cone of $\mathcal{S}$. Finally for any point x and cone $\mathcal{C}$, it is convenient to define $\mathcal{C}_x = x + \mathcal{C}$ to be the affine cone rooted at x. Coming back to our unknown parameter θ_i, we thus look at cones $\mathcal{C}_{\theta_i}$ rooted at θ_i. Now we are ready to deliver our warning message.

When $\Theta \subset \mathbb{R}$, the notion of cones is often hidden since there are only two possible pointed salient convex cones, namely $\mathbb{R}^+ = \{x \in \mathbb{R} : \langle x, 1\rangle \geqslant 0\}$ (the positive cone, or dual cone of $\{1\}$) and $\mathbb{R}^- = \{x \in \mathbb{R} : \langle x, -1\rangle \geqslant 0\}$ (the negative cone, or dual cone of $\{-1\}$). Note that if $\leqslant$ denotes the usual ordering on $\mathbb{R}$, then $\leqslant_{\mathbb{R}^+}$ is $\leqslant$ while $\leqslant_{\mathbb{R}^-}$ is $\geqslant$. Yet, cones do appear in the control of the probability of error such as $\mathbb{P}_{g_{m_0,\sigma^2}}\left(\left|\widehat{m}_n - m_0\right| \geqslant \varepsilon\right)$, where $\widehat{m}_n$ is an estimate of the parameter $m_0 \in \mathbb{R}$. Indeed we classically decompose this quantity in two parts

$$\text{(Positive cone)} \qquad \mathbb{P}_{g_{m_0,\sigma^2}}\left(\widehat{m}_n - m_0 \geqslant \varepsilon\right) = \mathbb{P}_{g_{m_0,\sigma^2}}\left(\left|\widehat{m}_n - m_0\right| \geqslant \varepsilon \cap \widehat{m}_n \in \mathbb{R}^+_{m_0}\right)$$

$$\text{(Negative cone)} \qquad \mathbb{P}_{g_{m_0,\sigma^2}}\left(m_0 - \widehat{m}_n \geqslant \varepsilon\right) = \mathbb{P}_{g_{m_0,\sigma^2}}\left(\left|\widehat{m}_n - m_0\right| \geqslant \varepsilon \cap \widehat{m}_n \in \mathbb{R}^-_{m_0}\right),$$

and often provide a separate controls for each partial order, before combining the two results (Note, interestingly, that a similar decomposition on the positive and negative cones is used to define integration). When moving to higher dimension $d > 1$, there are now *infinitely many* pointed salient convex cones, and handling them properly requires some specific care. It thus natural that in many settings, going from dimension 1 to higher dimensions is highly non-trivial and gives rise to long-lasting open questions. Fortunately, in such a situation the notion of cone is often beneficial. We illustrate the power of cones by solving an example of such intricate question in the multi-armed bandits setup in Maillard (2018) and another one for stochastic weighted automata in Balle and Maillard (2017).

A few questions Up to now, we have considered that we receive a sequence of observations, and that we have at hand a collection of probabilistic models to describe them. A few natural questions can be considered:

- What are the probabilistic models that provide the best fit to the data (model selection) ? Note that any collection of measures can naturally be combined into a new one: indeed, we can give a score to each model by defining a (probability) measure on the set of probabilistic models, which in turn induces a measure on the observations. This is called "aggregation" and we can extend this question of model selection to find the best aggregation of probabilistic models.
- Given the n first received observations, can we find a probabilistic model that gives maximal score to the next observations (prediction)?

The seemingly simple tasks of model selection, model aggregation and prediction have been the object of an intensive research agenda since the early ages of statistics, and will certainly require further decades or more of investigation. One of the reasons is that the observations can be of different nature, they can be real values, vectors, matrices, graphs, functions, etc. Further the probabilistic models defined on sequences of observations may consider not only i.i.d. observations, but also independent or not independent observations, with various notions of dependencies. Last, whatever processes $(\rho_j)_{j \leqslant J}$ are considered for a sequence of observations, it is always possible to define a new process ρ that uses ρ_{j_1} for the first $n_1 \in \mathbb{N}$ observations, ρ_{j_2} for the next $n_2 \in \mathbb{N}$ ones, and so on and so forth. Thus, the considered process changes every now and then, which introduces the questions of detecting and identifying changes in the observations (change point).

From passive to active observation Let us put these challenging questions aside for now and ask a simpler question: what is the source of these observations? Answering this simple question leads to an interesting point. Indeed, there are many different sources of observations out there and we are only observing some observations from some specific sources. Hence it is natural to consider that these sources are chosen in some way amongst all the possible ones. The choice of sources can be made independently on us and what we do with the observations, but it could also be done depending on us, either by an other system (in which case we are passive) or even by us (active learning) or by a combination of us and another system. This is a fundamental shift of paradigm, as we now consider a possible **interaction** between the agent and the observations. This opens a large research agenda that is one of the primary focus of this manuscript.

Indeed if we can choose at each step the source from which we want to receive the next observation(s), it is natural to consider these decisions are taken in order to optimize some criterion. More importantly, in case another system also chooses the sources, perhaps based on our choice, our decisions may now have a consequence on the next observations we receive. That is, we may have to change from a probabilistic model to another one according to the dynamical change of the sources. From this perspective, it is crucial to understand how the sequence of decisions that we make affects our observations and the criterion we want to optimize. This generic task is termed sequential decision making, or more precisely sequential decision making under uncertainty to emphasize that we may only have probabilistic models of the observations and of way the sources change (dynamics).

From this standpoint, we now have a sequence of decisions (actions) and observations

$$d^{(1)}, Y^{(1)}, d^{(2)}, Y^{(2)}, \dots,$$

where $Y^{(1)}$ is the (first) observation after taking decision $d^{(1)}$, $Y^{(2)}$ is the observation after taking decision $d^{(2)}$, etc. Here is a short illustrative list of problems of this kind:

- Clinical trials: A new disease appears. At each time step, we receive a patient suffering from it, and we have a few possible drugs that can be tested from a finite set $\mathcal{D}$. Upon applying drug $d^{(i)} \in \mathcal{D}$ on patient i, we quantify the success of the drug at curing patient i, thus creating the next observation $Y^{(i)}$. We assume in first approximation that the success value is a random variable whose law depends only on the considered drug. Our goal is to maximize the success score on all patients.

- Land probing: we want to estimate the level of nutrients, pesticides and micro-life activity of a land, by probing (observing, measuring) at different locations. We thus decide at each time step where to probe ($d^{(i)} \in \mathbb{R}^2$), then acquire the corresponding observation $Y^{(i)}$ before proceeding to the choice of the next location. Our goal is to get an accurate estimation of all levels, at a given precision error, with as few probing points as possible (equivalently, as fast as possible).

- Planning on the highway: We drive an autonomous vehicle and our actions are the possible commands on board. $Y^{(0)}$ is a vector that represents the scene around the autonomous vehicle (positions, speeds of other vehicles, road information, etc.). We have a module that, given a scene description Y and an action a at any time t, generates a possible scene description Y' at next time. We use it in order to generate possible trajectories of observations and actions: starting from $Y^{(0)}$, choosing $d^{(1)}$ we simulate $Y^{(1)}$, we choose $d^{(2)}$, etc. and decide when to stop a trajectory. We want to identify as fast as possible an action $d^{(1)}$ that maximizes a criterion, such as enabling to reach a specific destination in a secured way. This action (and only this one) will be executed in the real-world.

- Gardening: We consider growing one edible plant in a garden (e.g. strawberries). At each time step, we choose between different possible actions $d^{(t)} \in \{$do nothing, watering, mulching, cutting, etc.$\}$ and our

observation $Y^{(t)}$ consists of measurements about the plant (health status, stage of development, taste, etc) and contextual information such as whether conditions. The plant grows and its health may change over time, until we can harvest it. The goal is (for instance) to maximize the gustatory quality of the harvest.

It is worth noticing that for each decision step t, the decision $d^{(t)}$ is chosen based on past observations and decisions, but (obviously) without having knowledge of the future. Hence the decision process is such that for each $t \in \mathbb{N}_\star$, $d^{(t)}$ is measurable with respect to the σ-algebra $\mathcal{F}(d^{(1)}, Y^{(1)}, \dots, d^{(t-1)}, Y^{(t-1)})$ associated to all random variables indexed before $d^{(t)}$; we say it is adapted to the (filtration of the) observation-decision process. We sometimes call the natural filtration of the observation-decision process (indexed by the set $\mathbb{N}_\star \times \{1,2\}$, with lexicographic order) the filtration of the past. While the observation process, in full generality, has no reason to be adapted to the observation-decision process, it is often convenient to assume it is. The following notion of totally-adapted processes captures this (though we won't use it much in the sequel):

Definition 4 (Totally-adapted processes) *A finite collection of J processes $\mathbb{X}_j : \mathcal{I} \times \Omega \to E_j$, $j \in \{1, \dots, J\}$ is totally-adapted if for each $j \in \{1, \dots, J\}$, $\mathbb{X}_j$ is adapted to the* natural filtration *of the compound process $\mathbb{X} : \mathcal{I}_J \times \Omega \to \bigcup_j E_j$ indexed by $\mathcal{I}_J = \mathcal{I} \times \{1, \dots, J\}$ with lexicographic order $(i,j) \leqslant_{\mathcal{I}_J} (i',j')$ if $i \leqslant_{\mathcal{I}} i'$ or $i = i'$ and $j \leqslant_{\mathbb{N}} j'$, in the sense that $X_j^{(i)}$ is measurable with respect to $\mathcal{F}((X_{j'}^{(i')})_{(i',j') \leqslant_{\mathcal{I}_J} (i,j)})$ for all i, j.*

We mostly focus on totally-adapted observation and decision processes only, hence restricting the observation process. It is convenient to study even stronger restrictions: Assuming $Y^{(t)}$ is measurable with respect to $\mathcal{F}(d^{(t)})$ for each t is the typical focus of active learning and multi-armed bandits, while assuming $Y^{(t)}$ is measurable with respect to $\mathcal{F}(Y^{(t-m)}, d^{(t-m+1)}, \dots, Y^{(t-1)}, d^{(t)})$ for some integer m leads to the notion of Markov decision process (MDP) of order m. The case $m = 1$ is the most standard case, and the focus of extensive research.

Multi-armed bandits In order to prepare our first sequential learning task with active observation gathering, let us consider a simple situation when each decision consists in picking one element a in a set $\mathcal{A}$. The simplest case when this leads to an active setup is by considering the next observation is directly generated by this choice of action. Hence, let us consider that with each action a, there is a corresponding process $\mathbb{X}_a$ indexed by $\mathbb{N}$. Let us further consider the scenario when the random variables $(\mathbb{X}_a(n))_{n \in \mathbb{N}}$ are i.i.d. with common distribution ν_a, and independent from $(\mathbb{X}_{a'}(n))_{a' \neq a, n \in \mathbb{N}}$ (stochastic multi-armed bandit). Hence, the observations are $Y^{(t)} \sim \nu_{a_t}$, where a_t is the decision taken from observations before time t. The collection of distributions $\nu = (\nu_a)_{a \in \mathcal{A}}$ is sometimes called a bandit configuration. Finally, in order to specify a learning problem, one must decide what should be optimized. A multi-armed bandit problem considers again a direct approach: assuming that the distributions $(\nu_a)_{a \in \mathcal{A}}$ are real-valued and unknown, the goal is simply to accumulate observations with highest value, say in expectation, over some period of time T, that is to maximize $\mathbb{E}[\sum_{t=1}^T Y_t]$. This problem is perhaps one of the most direct combination of optimization and estimation and for this reason is at the heart of most sequential decision making problems.

In this specific formulation, we have the property that $\mathbb{E}[\sum_{t=1}^T Y_t] = \mathbb{E}[\sum_{t=1}^T \mu_{a_t}]$, where for each arm $a \in \mathcal{A}$, μ_a denotes the expectation of the distribution ν_a. Hence, an optimal strategy (knowing the distributions) is to pull an arm with maximal mean at each step, that is some $a^\star \in \operatorname{Argmax}_{a \in \mathcal{A}} \mu_a$. Hence, the expected error we make by pulling action a instead of $a_\star$ is $\Delta_a = \mu^\star - \mu_a$, where $\mu^\star$ is a short-hand notation for $\mu_{a^\star}$. This fundamental quantity is called the gap of arm a. Summing the expected errors on each chosen action enables to quantify how suboptimal a strategy is. This is called the regret (of not knowing in advance an optimal strategy):

Definition 5 (Expected regret) *The quality of a decision making strategy is evaluated using the notion of expected regret (or simply, regret) at round $T \geqslant 1$, defined as*

$$\mathfrak{R}_T \stackrel{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{a_t}\right] = \sum_{a\in\mathcal{A}} \Delta_a \, \mathbb{E}\big[N_a(T)\big] \,, \text{ with } N_a(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{a_t=a\}} \qquad (1)$$

where we used the tower rule for the first equality. The expectation is with respect to the random draws of the Y_t according to the ν_{a_t} and to the possible auxiliary randomization introduced by the decision-making strategy.

The term multi-armed bandit problem was probably coined during the 60's in reference to the casino slot machines of the 19th century: A popular way to illustrate this problem is indeed by trying to maximize its expected outcome when playing on casino slots machines, where each machine is also called a "one-armed bandit". Due to this illustration, the general setup with many machines is called a "multi-armed" bandit problem. Now, the above formulation of this problem is due to Herbert Robbins – one of the most brilliant mind of his time, see Robbins (1952) and takes its origin in earlier questions about optimal stopping policies for clinical trials, see Thompson (1933, 1935), Wald (1945). We refer the interested reader to Robbins (2012) regarding the legacy of the immense work of H. Robbins in mathematical statistics for the sequential design of experiments, compiling his most outstanding research for his 70's birthday. Since then, the field of multi-armed bandits has grown large and bold.

The questions that one may naturally ask for this problem are: where is the difficulty? what is the best performance that one can hope to reach? and is there a way to design an algorithm that is provably optimal (or near optimal)? For now, we only provide a quick hint at the first question, as we investigate these questions in greater details later in this manuscript.

A fundamental difficulty of multi-armed bandit problems is called the Exploration-Exploitation trade-off. Namely, at each time step, the learning agent must decide what arm to pull, but the distributions of each arm, and more importantly their means, are unknown to the learner. Hence they have to be estimated based on the observations available until the current time. Now, in order to improve the quality of the mean estimate of arm a, one should get more samples from distribution ν_a: This is called exploration. On the other hand, if we have a lot of observations from all arms, then we may simply trust our empirical estimates and play an arm whose empirical mean estimate has maximum value: This is called exploitation. An agent should then balance these two objectives, making sure all arms are estimated with sufficient accuracy so that we can pull an optimal arm. It turns out that solving this difficulty is a challenging research question.

One reason for the popularity of bandits is its versatility. Let us mention three complementary extensions to the initial problem:

a) Continuous arms and optimization While in its basic formulation, the set $\mathcal{A}$ is a finite set of cardinality $A \in \mathbb{N}$, that is, we only have to choose between finitely many distributions, the setup extends to a full-blown optimization problem when $\mathcal{A} \subset \mathbb{R}^d$. Indeed in that case, if we introduce the mean function $f : \mathcal{A} \to \mathbb{R}, a \mapsto \mu_a$, the goal is to maximize the function f by sequentially sampling the set $\mathcal{A}$. If we

knew the distributions $(\nu_a)_{a\in\mathcal{A}}$, this would be just a (possibly complicated) optimization problem. But here the distributions are unknown, and sampling at point a costs us Δ_a, which adds a key difficulty to the problem. One fruitful example if the setup of linear bandits where observations are $Y_{a_t} = \langle \theta, a_t \rangle + \xi_t$, with vector-valued actions $a_t \in \mathcal{A} \subset \mathbb{R}^d$, an unknown parameter $\theta \in \mathbb{R}^d$, and a random noise variable ξ_t (on which we put some assumption, such as sub-Gaussianity).

b) Expected versus risk-averse criterion In real applications, some actions may have an important negative effect (e.g. when considering applications of bandits in clinical trials, a drug may endanger the life of some patients). In this situation, minimizing the regret in expectation is not satisfactory. For instance, we prefer a treatment that is safer and does not endanger patients even if less good on average than another one that performs better on average but has a high chance of endangering life of some patients. The same is true when considering actions on a garden, where we may avoid actions that pollute the groundwater, or destroys the ground micro-life. The notion of risk can be captured by different approaches. We explore a sound notion derived from statistical properties of random variables later in this manuscript.

c) Beyond i.i.d.: drift, changes, state machines Yet another natural extension is to relax the i.i.d. process assumption, in which case an optimal policy may no longer reduce to choosing the same arm at each step. For instance the reward process on one arm may change, either slowly (drift) or abruptly (change) as a function of time, or a function of the number of pulls of this arm, or as a function of yet another observable quantity. The example of Markov Decision Process detailed below corresponds to the latter case, when the observed quantity is a specific quantity called a state.

Markov decision processes An MDP is specified by a tuple $(\mathcal{S}, \mathcal{A}, I, R, P)$ such that $I \in \mathcal{P}(\mathcal{S})$, $P : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ and $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{P}(\mathbb{R})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on a set $\mathcal{X}$. $\mathcal{S}$ is called the state space of the MDP, $\mathcal{A}$ is the action space and I is called the initial state distribution. The two most important objects are the transition function P that assigns to each state-action pair a distribution of states, and the reward function R that assigns a real-value to each state-action-state tuple. The process specifies S_0 to be a random-variable with law I, then for any decision A_0, S_1 has law $P(S_0, A_0)$, and R_1 has law $R(S_0, A_0, S_1)$. More generally:

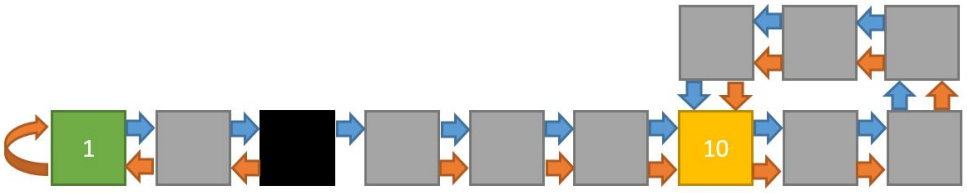
$$S_0 \sim I, \qquad \forall t \in \mathbb{N}_\star, \quad S_t \sim P(S_{t-1}, A_{t-1}), \quad R_t \sim R(S_{t-1}, A_{t-1}, S_t),$$

where for each $t \in \mathbb{N}$, A_t is adapted to $\mathcal{F}(S_0, A_0, S_1, R_1, A_1, \ldots, S_{t-1}, R_{t-1}, A_{t-1}, S_t)$. A typical example is when $A_t \sim \pi(S_t)$, for a function $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. Such functions are called (stochastic, stationary) policies. Note, importantly, that we see here MDPs from the perspective of statistics, when the transition function and reward functions are both unknown. This is in stark contrast with control theory that considers these functions are perfectly known. Intuitively, the transition models the dynamics of the state process (how the state is modified when choosing an action), while the reward function provides a notion of score to each transition that is done. Note that by construction, S_t and R_t are measurable with respect to $\mathcal{F}(S_{t-1}, A_{t-1})$. In an MDP, the observations are the states and rewards $Y_t = (S_t, R_t)$. Hence we say that the states are observed. This contrasts with the more general situation when we only observe a function of the state $Y_k = (f(S_t), R_t)$ but the states are unobserved (think of the previous example of navigation in a maze where S denotes the location and we only observe what can be seen from this location in the maze). When the state-space $\mathcal{S}$ is known, this is called a partially observable Markov decision process (POMDP), while some methods such as Predictive State Representations study the more general case when even $\mathcal{S}$ is unknown (and try to build a meaningful notion of states from the observations). On the other hand, the case when $\mathcal{S} = \{s\}$ is a singleton captures the stochastic

multi-armed bandits formulation, with $\nu_a = R(s,a,s)$ for each $a \in \mathcal{A}$. Similar to the multi-armed bandit setup, a natural goal is to sequentially chose the actions in order to maximize the cumulative sum of rewards in expectation over (possibly unknown) T time steps. for instance, when restricting to stochastic stationary policies, it is natural to introduce the quantity

$$V^{\pi}_{I,R,P,T} = \mathbb{E}\left[\sum_{t=1}^{T} R^{\pi}_t\right] \quad \text{where } \forall t \in \mathbb{N}_\star,\ S^{\pi}_t \sim P(S^{\pi}_{t-1}, A^{\pi}_{t-1}),\ R^{\pi}_t \sim R(S^{\pi}_{t-1}, A^{\pi}_{t-1}, S^{\pi}_t),\ A^{\pi}_t \sim \pi(S^{\pi}_t)\,,$$

which is sometimes called the (T-step) value of a policy π. Unfortunately, unlike in the multi-armed bandit setting, in full generality a stochastic stationary policy that maximizes this criterion has no reason to be optimal. Indeed, maximizing $\mathbb{E}\left[\sum_{t=1}^{T} R_t\right]$ over all actions generally corresponds to a policy $\pi_{I,R,P,T} : \mathbb{N} \times \mathcal{S} \to \mathcal{P}(\mathcal{A})$ that depends on the time horizon T, and that generates at step t an action $A_t \sim \pi_{I,R,P,T}(t, S_t)$ depending on both the current state and the current time step. This creates at least two fundamental difficulties. First, we generally don't want to specify the time horizon T of the problem, in the sense that we are looking for strategies $\pi_{I,R,P}$ that would be simultaneously optimal for all horizon T. But such uniformly horizon-optimal policies may not exist, even in very simple situations. Let us take for example the following example of an MDP with 2 actions $\mathcal{A} = \{a, b\} = \{\leftarrow, \rightarrow\}$, deterministic transitions and rewards, with initial state being the black state, and reward 1 (respectively 10) each time entering the green (respectively yellow) state.



In this simple example where everything is deterministic, the set of optimal sequences of actions can be made explicit for all time horizon T, as we summarize in the following table. Yet no sequence (hence policy) can be made optimal simultaneously for all time T (or even for all large enough T).

T	1	2	3	$4,\dots,7$	8	9	$10,\dots,13$	14	15	…
$\star$	$\mathcal{A}$	a^2	a^3	$b\mathcal{A}^{T-1}$	$a^2b^3\mathcal{A}^{T-5}$	$a^3b^3\mathcal{A}^{T-6}$	$b\mathcal{A}^{T-1}$	$a^2b^3\mathcal{A}^{T-5}$	$a^3b^3\mathcal{A}^{T-6}$	…
$V_{I,R,P,R}$	0	1	2	10	11	12	20	21	22	…

This is due to the fact that visiting the "small cycle" with rewards 1 costs 4 time steps only, which is less than the time to complete the "large cycle" of 6 steps starting from the gold state. Indeed, increasing the length of the small cycle from 4 to 6 by adding an intermediate gray state, ensures that $b\mathcal{A}^\star$ uniquely describes optimal sequences simultaneously for all T but $T = 3$ (for which it is a^3).

Hence one should either restrict the set of considered MDPs to make sure we don't have such cycles, look for near-optimal rather than optimal policies, or modify the performance criterion. For instance two main classes of criterion, called discounted and average reward criterion, have been introduced that do not make appear an explicit time horizon T; Actually even for these two main classes, several notions of optimality can be introduced, see Puterman (1994). We will focus in chapter 6 on the average-reward criterion. Now, a second difficulty is that, even assuming there exists a uniformly horizon-optimal policy $\pi_{I,R,P}$, learning in an MDP when the reward and transitions are unknown comes with a price, and it seems challenging to compete with a policy that depends on the current time t. These two difficulties have made people to consider restricted scenarii when *stationary* policies are shown to be optimal.

Rewards? In the two previous models (bandits and MDPs), we have encountered the notion of reward. Since a decision problem involves decisions, it is natural to search for decisions that maximize some criterion. To this end, people assume a score is given to each decision that is made. This score may be itself a random variable, as it is the case for MDPs. From this standpoint, a natural question is how to maximize the cumulative sum of rewards $\sum_{k=1}^{T} R_k$ in an MDP, say for a given number of steps T, that is what choice of the actions $(A_k)_{k\in\mathbb{N}}$ maximizes this sum. People generally focus on maximizing the expected value of this sum, while some works also consider other criterion, notably when dealing with a notion of risk. We will see some example in chapter 5.

Reinforcement learning considers the case when we get to observe this process but without knowing the transition function P or reward function R (we may know that they belong to some set of transition and reward functions). Hence the laws of the random variables are unknown to the agent. This is the main difference with control theory in which P and R are perfectly known. Thus reinforcement learning is primarily interested in maximizing the cumulative rewards *while* estimating the dynamics of a system, which yields an avenue of research questions.

No-rewards? Reinforcement learning heavily relies on the rewards that are received. However, as for the observation process, it is natural to ask where the reward process comes from. Classical reinforcement learning considers the reward process is defined by Nature, independently on the agent. However, note that the function R has no reason to behave nicely with respect to the transition function P. Hence maximizing the cumulative rewards may turn to be especially hard in full generality. But what if we consider the rewards are no longer defined by nature, but are defined actively (as we did for the observations)?

This leads to a second and fundamental shift of paradigm, that leads us from Reinforcement Learning to Artificial Intelligence (AI). Indeed, while Reinforcement Learning assumes the reward is generated by an external entity, Artificial Intelligence will typically build a reward function actively. Hence the general AI problem is related to some high level questions, such as which objective to focus on, what objective function and state space to specify, and when to stop solving an RL task and start another one. Further, as for POMDPs, the observations may not be the states directly but only a function of the states. Hence reward construction is part of AI, and one of the reasons that make AI such a challenging task is precisely this apparent absence of objective function: without clearly defined objective function, there is just no learning task. Notions such as intrinsic rewards have been introduced precisely in order to introduce some well-defined objective, but whether this is enough to encompass all the questions behind the quest for AI is unclear. Note that even if removing the rewards, the transitions are still unknown, and one should certainly focus on learning them.

In this manuscript, we will not discuss much this second paradigm shift that is still largely unexplored from a theoretical standpoint and focus instead on the mathematical foundations behind the first paradigm shift (when the agent influences the observation process). More precisely, we focus on the role of non-asymptotic statistics in the design of (near-)optimal strategies for sequential decision making, across various setups.

Where is this going?

The more applied you go, the stronger theory you need.

Before we dive into the mathematics of sequential learning, let us take a small step off and think about what all this can be useful for, beyond mathematics. Indeed the good thing when our object of study is a sequence of observations $Y^{(1)}, Y^{(2)}, \dots$ that is actively sampled is that this can be connected to a tremendous number of real-life situations. Hence, we can not only enjoy the beauty of the mathematics behind understanding a generic sequence, but further have a direct impact on the physical world as well. Moreover, since considering a specific application naturally brings novel challenges for the researcher, this often leads us to explore uncovered questions of the mathematical world. On the other hand, since a well-targeted answer generally applies to several real-world applications, this also means we should choose which application area to promote or explore.

Hence, as a researcher in the field of sequential learning, we should question the physical-world applications of our research, and provide guidance towards the ones we value most (instead of following the current trend). In this short section, we advocate for a thrilling application domain, that has been poorly explored by the community compared to the mainstream applications to marketing and medical health.

TOWARDS ECO-SUSTAINABLE SEQUENTIAL DECISION MAKING

Sequential learning shares a long history with medical health, since the first models of sequential learning are attributed to Thompson (1933) in the context of clinical trials. Hence, adaptive treatment strategies, dynamic drug dosage, and personalized health care have benefited from tremendous efforts of many brilliant researchers working on sequential learning for decades. We refer to the book Bartroff et al. (2012) that gives a detailed overview of the many questions surrounding modern medical trials. A number of research groups are thus actively, and successfully working on combining sequential learning and reinforcement learning techniques with medical health.

The last two decades have seen the explosion of a novel type of application: that of optimization of ad-placement strategies on web pages, as well as the more general task of recommender systems for the web-industry (Netflix, Amazon, Facebook, Criteo, etc.). Here the basic task for a web-company is to select one or a few items to present to a user, seen as a client, and the goal is to maximize the number of items sold (or clicked) to a user. Hence the traditional representation of multi-armed bandits applications (that is a first approximation of the decision making problem) has progressively shifted from

- Clinical trials: (Thompson, 1933)






:

where at each time step, a doctor faces a new patient, all having the same disease, and must choose one treatment from a list of possible treatments (or different dosage of the same treatment) whose effect is unknown.

to

- Ad-placement: (Nowadays...)


:


where at each time step, a new user comes to a web-page that has one (or many) slots for ads, and an algorithm must choose which add to display to the user in each slot.

We want to take the opportunity to deviate from these two mainstream applications, and illustrate the concepts of sequential learning on sustainable agriculture, ecological gardening and ecosystem preservation. We think this is not only of much higher societal and ethical value than optimizing web-selling strategies but also largely untouched compared to health care applications: Indeed while agriculture for instance is all about sequential decision making under uncertainty, it has been largely overlooked by the main community publishing at NIPS, ICML, COLT, JMLR, etc. This also means there is a tremendous potential for highly beneficial impact on such applications. Interestingly, all the standard questions and variants of sequential learning (bandits, MDPs, POMDPs, etc.) can be written in the context of health care, web-marketing, and sustainable farming. Hence, we strongly encourage the young researchers to explore this third avenue of research, as we believe it is especially exciting. Let us give a quick glance at a few potential applications.

Similarly too clinical trials or web-advertisement, multi-armed bandits can be considered in first approximation to address the following problems:

- Plant-health care:


:



  where at each time step, we encounter a new plant having a disease, and we want to find which action is best in a sense to cure it, such as watering, attracting a specific species, adding chemicals, doing nothing, etc.

- Ground-health care:


:



  where at each time step, we consider a piece of deteriorated land, and we want to decide what (combination of) actions to do such as encouraging micro-life activity, protecting the ground, planting plants that stabilize the ground or that clean it from pollutants, amongst other things.

- Bio-diversity/Bio-equilibrium care:

 :  

where at each time step, we study a destabilized ecosystem, and we need to choose amongst a limited number of actions to preserve it (promoting local species, creating awareness, fighting an invasive species, testing a new design, etc.)

Now, as for clinical trials and web-marketing, a multi-armed bandit formulation is limited to somewhat stringent assumptions, and a more realistic scenario is to consider more expressive setups, such as using contextual linear or combinatorial bandits, (partially observable) Markov decision processes and beyond. Indeed, short-term optimal policies have no reason to be long-term optimal (think about maximizing food production over a single year, versus other ten cumulative years). Further, it is natural to consider the distributions may be subject to changes, for instance due to unstable whether conditions, an invasive species, an incoming pollution, or any other not anticipated event. Also, as the plants adapt to their environment, the population evolves and thus the dynamics of the system progressively shifts, generation after generation. Further, weather conditions may also evolve so that actions that were efficient for a garden at some time may not be as efficient some years later. This poses the question of recommendation in a shifting context scenario.

While in clinical trials the main objective is so save life of patients and in web-marketing to sell as much as possible, potential goals (out of many) can be to maximize resilience of ecosystems, or maximize food production with minimal input, energy or fertilizer. We list below a short list of questions for illustration:

- Which technique is best to avoid potatoes from getting this specific disease, depending on the ground and whether conditions (contextual multi-armed bandit)?
- Which set of plants maximize butterfly biodiversity (combinatorial bandits)?
- Where to observe for best estimating the propagation of species outside their usual territory (active bandit)?
- What sequences of plant culture (culture rotation) create best germination conditions for an eggplant (goal-state MDP)?
- How to maximize health of the system while minimizing work-load/inputs? (multi-objective reinforcement learning)?
- Which path to patrol on in order to maximize probability of observing diverse animals (exploration, shortest path)?

We hope that this brief presentation of the sequential learning challenges of sustainable agriculture, ecological gardening and ecosystem preservation pictures an exciting application domain for sequential decision making under uncertainty, for which some existing learning strategies can already be applied immediately, and many others need to be developed before the challenges of the field can be addressed in a satisfactory way.

Contents

Foreword: To the layman reader. **v**

Where is this going? **xix**

Summary of Scientific Activity **xxv**

Part I. Mathematics of Statistical Sequential Decision Making **1**

Chapter 1 $\log \mathbb{E} \exp(\lambda X)$ **5**
1 Control of probability tails . . . 7
2 Legendre-Fenchel dual of the Log-Laplace . . . 12
3 Loss and noise . . . 13
4 General exponential families, properties . . . 17

Chapter 2 $\{\tau \leqslant t\} \in \mathcal{F}_t$ **19**
1 The peeling technique for random stopping times . . . 24
2 Uniform bounds and the Laplace method . . . 30
3 Numerical comparison of a few bounds . . . 34

Chapter 3 $\log\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)$ **37**
1 Change of measure and lower bounds . . . 39
2 Further lower-bounds and extensions . . . 44
3 From lower bounds to sampling strategies . . . 49
4 Change point detection . . . 53

Chapter 4 $\mathbb{P}(? \notin \widehat{\Theta}_{n,\delta}) \leqslant \delta$ **63**
1 Kernel regression and the Laplace method . . . 65
2 Markov concentration . . . 71
3 Forecasters of stationary processes over a finite alphabet . . . 75

Chapter 5 $\sup_q \langle q, f\rangle - \mathcal{B}^{\psi}(q,p)$ **79**
1 Risk-aversion in multi-armed bandits . . . 81
2 Aggregation of experts: insights from duality. . . . 86

Chapter 6 $f = T[f]$ **and** $\|.\|_?$ **95**
1 MDPs and average reward criterion . . . 97
2 Reinforcement learning in the average-reward criterion . . . 105

Part II. Focus on three contributions **115**

Chapter 7 Boundary Crossing Probabilities **117**
1 Multi-armed bandit setup and notations . . . 117
2 Boundary crossing probabilities for the generic `KL-ucb` strategy. . . . 118
3 Boundary crossing for K-dimensional exponential families . . . 122
4 Main analysis . . . 126

Chapter 8 Hankel matrices **131**
1 Weighted automata . . . 131
2 Hankel matrices and spectral learning . . . 134

Chapter 9 Aggregation of growing experts **139**
1 Introduction . . . 139
2 Preliminary: the exponential weights algorithm . . . 140
3 Overview of the results . . . 141

Part III. Thoughts and perspective **147**

Bibliography **153**

Index **161**

Summary of Scientific Activity

We provide below a list of the main contributions done since 2011 (after my PhD defence). They can be categorized into three main research areas. Out of clarity and conciseness, this manuscript reflects the work corresponding to a small selection of these works. We encourage the interested reader to read the other ones, in particular the work about the sub-sampling strategies in multi-armed bandits or about state model selection in Markov decision processes, that are not detailed in the sequel.

Statistical Theory

- Boundary Crossing Probabilities for General Exponential Families, O.-A. Maillard, Algorithmic Learning Theory (ALT), 2017. Extended version in Mathematical Methods of Statistics.
- Spectral Learning from a Single Trajectory under Finite-State Policies, B. Balle, O.-A. Maillard in International conference on Machine Learning, 2017.
- Self-normalization techniques for streaming confident regression, O.-A. Maillard, submitted to Bernoulli.
- Pliable rejection sampling, A. Erraqabi, M. Valko, A. Carpentier and O.-A. Maillard in International Conference on Machine Learning (ICML) 2016.
- A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets, R. Bardenet and O.-A. Maillard in 28th Neural Information Processing Systems (NIPS) workshop, 2015.
- Concentration inequalities for sampling without replacement, R. Bardenet and O.-A. Maillard, in Bernoulli, 2014.

Multi-armed bandits and sequential learning learning

- Memory Bandits: a Bayesian approach for the Switching Bandit Problem, R. Alami, O.-A. Maillard, R. Féraud, Neural Information Processing Systems (NIPS) workshop, 2017.
- Efficient tracking of a growing number of experts, J. Mourtada, O.-A. Maillard, Algorithmic Learning Theory (ALT), 2017.
- The non-stationary stochastic multi-armed bandit problem, R. Allesiardo, R. Féraud, O.-A. Maillard, in International Journal of Data Science and Analytics, 2016.
- Random Shuffling and Resets for the Non-stationary Stochastic Bandit Problem, R. Allesiardo, R. Féraud, O.-A. Maillard,
- Low-rank Bandits with Latent Mixtures, A. Gopalan, O.-A. Maillard, M. Zaki, submitted to JMLR, Sep 2016. Arxiv.

- Sub-sampling for multi-armed bandits, A. Baransi, O.-A. Maillard, S. Mannor, in Europeean conference on Machine Learning (ECML), 2014.
- Latent bandits, O.-A. Maillard and S. Mannor, in Proceedings of the International Conference on Machine Learning (ICML), 2013.
- Robust risk-averse stochastic multi-armed bandits, O.-A. Maillard, in Proceedings of the International Conference on Algorithmic Learning Theory (ALT), volume 8139 of Lecture Notes in Computer Science, pages 218–233. Springer Berlin Heidelberg, 2013.
- Kullback–leibler upper confidence bounds for optimal sequential allocation, O. Cappé, A. Garivier, O.-A. Maillard, R. Munos and G. Stoltz, in The Annals of Statistics, 41(3):1516–1541, 2013.

Reinforcement Learning

- Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs, M. S. Talebi, O.-A. Maillard, Algorithmic Learning Theory (ALT), 2018.
- "How hard is my MDP?" Distribution-norm to the rescue, O.-A. Maillard, T.A. Mann and S. Mannor in Proceedings of the 27th conference on advances in Neural Information Processing Systems (NIPS), 2014.
- Selecting Near-Optimal Approximate State Representations in Reinforcement Learning, R.Ortner, O.-A. Maillard and D. Ryabko, in Proceedings of the International Conference on Algorithmic Learning Theory (ALT), 2014.
- Competing with an infinite set of models in reinforcement learning, P. Nguyen, O.-A. Maillard, D. Ryabko, and R. Ortner, in Proceedings of the International Conference on Artificial Intelligence and Statistics (AI&STATS), volume 31 of JMLR W&CP , pages 463–471, Arizona, USA, 2013.
- Optimal regret bounds for selecting the state representation in reinforcement learning, O.-A. Maillard, P. Nguyen, R. Ortner, and D. Ryabko, in Proceedings of the International conference on Machine Learning (ICML), volume 28 of JMLR W&CP, pages 543–551, Atlanta, USA, 2013.

Part I

Mathematics of Statistical Sequential Decision Making

In this part, our goal is, unlike what is typically done in research articles, to present the mathematical tools first, and only then show in which context they are used. We believe that proceeding this way sheds better light on the strength and generality of these tools, which is complementary with the more applied motivation that is presented in the research articles. We now briefly detail the content of the different chapters.

Estimation error Given a sequence of n real-valued observations assumed to come from an i.i.d. process ν, it is natural to form an empirical estimate $\widehat{\mu}_n$ of the mean μ. We first recall how to control its error by concentration of measure. One of the key tools for that is the control of the log-Laplace transform $\lambda \to \log \mathbb{E}_\nu \exp(\lambda X)$ of the random variable. We spend some time to detail some of the most useful properties of this fundamental object in Chapter 1.

Random number of observations When considering an active setup, estimation becomes trickier. Indeed, since our decisions at each time step are based on all past observations and influence the next observations, we are no longer in the i.i.d. setup. Hence the strong properties provided by the log-Laplace control may not apply a priori. For illustration, let us focus on the simplest setup when there is a finite set $\mathcal{A}$ of possible decisions, where each $a \in \mathcal{A}$ is attached to one i.i.d. process with distribution ν_a and mean μ_a, so that a decision consists in sampling one novel observation from the chosen action; This is typically the situation in multi-armed bandits. At time t, let $N_t(a)$ denote the total number of observations sampled from ν_a until time t, so that $\sum_{a\in\mathcal{A}} N_t(a) = t$, and let us consider the empirical estimate $\widehat{\mu}_{N_t(a)}$ of μ_a. This empirical estimate looks similar to the standard i.i.d. setup. However, $N_t(a)$ is not a deterministic quantity but a random variable depending on all past data. Obtaining concentration inequalities in this setup is a priori difficult, but is fortunately handled thanks to powerful techniques that we present in Chapter 2.

Uniform optimality In Chapter 3, we give greater focus to the set in which our i.i.d. processes belong to, that is on $\mathcal{D}$ such that $\nu \in \mathcal{D}$. Without knowing ν, it is natural to ask whether the observations are generated according to ν or a different $\nu' \in \mathcal{D}$. The key tool for this purpose is called a change of measure argument. We detail this change of measure, and show it has fundamental implications in sequential decision making theory. In particular, this enables to derive sharp lower bounds inequality on the achievable performance of virtually any decision procedure aiming at being good uniformly over all $\nu \in \mathcal{D}$, where good depends on the problem. We illustrate this on the multi-armed bandit problem, and recall how the change of measure naturally suggests the construction of sampling strategies, for a given set $\mathcal{D}$. An especially fruitful and challenging research question is to provide sharp regret minimization guarantees for these lower-bound inspired strategies. While in full generality a definitive answer is still not unknown, we provide key advances for the important class of exponential families in the dedicated chapter 7.

Confidence sets, ellipsoids While the initial chapters focus mostly on real-valued random variables, in many situations we face a more general parametric process. This happens for instance when observations sampled at point $x \in \mathcal{X}$ follow a Gaussian distribution $\nu_x = \mathcal{N}(f^\star(x), \sigma^2)$ for some unknown $f^\star \in \mathcal{F}$, and $\mathcal{F}$ is a function space such as $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathbb{R}, : f_\theta(x) = \langle \theta, \varphi(x)\rangle, \theta \in \Theta \subset \mathbb{R}^d\}$ or some feature function φ and parameter set Θ. Here one would like to estimate the vector parameter $\theta^\star$ corresponding to the unknown function $f^\star$. We explain how to extend the previous results to the general setup of reproducing kernel Hilbert spaces (RKHS). Another example is that of Markov models of order m on a finite set of symbols $\mathcal{S}$ for which we provide some simple finite-time estimation results. Hence, Chapter 4 provides a short zoology of the construction of confidence sets for common categories of processes.

Part I

Risk and duality In the four first chapters, different families of processes are studied, with essentially the same core idea: mean estimation. In Chapter 5, a different approach is considered. A first motivation is to optimize a risk-averse criterion rather than the mean: for instance when choosing different treatments, some may be successful on average, but have a higher risk of killing patients. A second motivation is also related to risk: when we have a set of learners, each corresponding to a different family of processes, we may naturally combine them (aggregate them) in order to build strategies that are near uniformly optimal over the union of all these classes, in order to face a larger set of situations. It turns out, perhaps surprisingly, that the same fundamental tool can be used for both approaches: (Bregman) duality. Hence chapter 5 shows some tools of risk-aversion and aggregation that are based on the use of Bregman duality.

Markov decision processes In Chapter 6, we turn to the setup of Markov decision processes. We focus here in presenting the contraction properties of the Bellman operator when considering an undiscounted reinforcement learning setup. This was known from Puterman (1994), but seems to be largely overlooked by the modern reinforcement learning literature. We suggest that many difficulties appearing today in the analysis of average-reward MDPs from an RL standpoint are due to a poor understanding of this quantity (and its estimates).

CHAPTER 1
$\log \mathbb{E} \exp(\lambda X)$

Contents

1 Control of probability tails 7

1.1 A first consequence 7

1.2 Two complementary results 9

1.3 Illustrative cases 11

2 Legendre-Fenchel dual of the Log-Laplace 12

3 Loss and noise 13

4 General exponential families, properties 17

Take-home message

Tail control of $\mathbb{R}$-valued random variables

$$\exists \lambda \in \mathbb{R}^+ : \log \mathbb{E} \exp(\lambda X) \leqslant \varphi(\lambda) \implies \forall t \in \mathbb{R}, \quad \mathbb{P}(X \geqslant t) \leqslant \exp(-\lambda t + \varphi(\lambda)) .$$

$$\forall t \in \mathbb{R}, \quad \mathbb{P}(X \geqslant t) \leqslant \alpha(t) \implies \forall \lambda \in \mathbb{R}^+, \quad \log \mathbb{E} \exp(\lambda X) \leqslant \log \int_{\mathbb{R}} \alpha(u/\lambda) e^u d\mu(u) .$$

The log-exp pair behaves nicely with respect to product measures and independent processes.

Duality formulas

($\mu \to \mathtt{KL}(\mu, \nu)$ and $f \to \log \mathbb{E}_\nu \exp(f(X))$ are dual to each other.)

$$\text{(Entropy formula)} \quad \log \mathbb{E}_\nu[\exp(f(X))] = \sup_{\mu \in \mathcal{P}(\mathcal{X})} \langle \mu, f \rangle - \mathtt{KL}(\mu, \nu)$$

$$\text{(Variational formula)} \quad \mathtt{KL}(\mu, \nu) = \sup_{f \in \mathcal{C}_b(\mathcal{X})} \langle \mu, f \rangle - \log \mathbb{E}_\nu[\exp(f(X))] ,$$

$$\text{(Gibbs distribution)} \quad \frac{d\mu}{d\nu}(x) = \frac{\exp(f(x))}{\mathbb{E}_\nu[\exp(f(X))]} \quad \text{achieves maximum in Entropy formula} .$$

This induces a natural duality between a loss function and a noise distribution: To any loss corresponds a (dual) notion of noise, and vice-versa. E.g. σ-sub-Gaussian noise is dual to quadratic loss $\ell(x, x') = \frac{(x-x')^2}{2\sigma^2}$.

Structural properties

$$-\log \mathbb{E} \exp(-f) \leqslant \mathbb{E} f \leqslant \log \mathbb{E} \exp f$$

$$\forall x \in \mathcal{X}, 0 \leqslant f(x) \implies \mathbb{E} f \leqslant -\log \mathbb{E} \exp(-f) + \frac{1}{2} \mathbb{E} f^2 .$$

$$\forall x \in \mathcal{X}, |f(x)| \leqslant b \implies \log \mathbb{E} \exp f - \frac{e^b - 1 - b}{b^2} \mathbb{E} f^2 \leqslant \mathbb{E} f .$$

Exponential families

The log-partition function $\psi(\theta) = \log \int_{\mathcal{X}} \exp(\langle \theta, F(x) \rangle) d\nu_0(x)$ of an exponential family with reference measure ν_0 and feature function $F : \mathcal{X} \to \mathbb{R}^K$ satisfies:

$$\text{(Bregman divergence)} \quad \mathtt{KL}(\nu_\theta, \nu_{\theta'}) = \mathcal{B}^\psi(\theta, \theta') = \psi(\theta') - \psi(\theta) - \langle \theta' - \theta, \nabla \psi(\theta) \rangle$$

$$\text{(Bregman duality)} \quad \mathcal{B}^\psi(\theta, \theta') = \sup_{\eta \in \mathbb{R}^K} \langle \eta, \nabla \psi(\theta) \rangle - \log \mathbb{E}_{\nu_{\theta'}} \exp(\langle \eta, F(X) \rangle) .$$

$$\text{(Maximum likelihood)} \quad \sup_{\theta \in \Theta} \sum_{i=1}^n \log \frac{d\nu_\theta}{d\nu_0}(X_i) = n \psi^\star \left(\frac{1}{n} \sum_{i=1}^n F(X_i) \right) .$$

In this chapter, we focus on arguingly one of the most powerful and oldest tool in statistics, namely the log-Laplace transform, or cumulant generative function of a random variable. We provide below a short list of powerful properties of this crucial quantity.

1 CONTROL OF PROBABILITY TAILS

Let us first start with a simple property of non-negative random variables.

Lemma 1.1 (*Non-negative random variables*) *Let X be a $\mathbb{R}^+$-valued random variable with $\mathbb{E}(X) < \infty$. Then*

$$\text{(Markov inequality)} \qquad \mathbb{E}(X) \geqslant \varepsilon \mathbb{P}(X \geqslant \varepsilon) \quad \text{for each } \varepsilon \in \mathbb{R}^+,$$

$$\text{(Fubini formula)} \qquad \mathbb{E}(X) = \int_{\mathbb{R}^+} \mathbb{P}(X \geqslant x) d\mu(x) \quad \text{where } \mu \text{ is the Lebesgue measure}.$$

1.1 A first consequence

We can apply this result immediately to real-valued random variables by remarking that for any random variable distributed according to ν (which we note $X \sim \nu$) and $\lambda \in \mathbb{R}$, the random variable $\exp(\lambda X)$ is non-negative. Thus if we now define the domain of ν by $\mathcal{D}_\nu = \{\lambda : \mathbb{E}[\exp(\lambda X)] < \infty\}$, we deduce by application of Markov's inequality that for all $t > 0$,

$$\begin{aligned} \forall \lambda \in \mathbb{R}_\star^+ \cap \mathcal{D}_\nu \quad \mathbb{P}(X \geqslant t) &= \mathbb{P}(\exp(\lambda X) \geqslant \exp(\lambda t)) \\ &\leqslant \exp(-\lambda t)\mathbb{E}[\exp(\lambda X)]\,. \qquad (1.1) \\ \forall \lambda \in \mathbb{R}_\star^- \cap \mathcal{D}_\nu \quad \mathbb{P}(X \leqslant t) &= \mathbb{P}(\exp(\lambda X) \geqslant \exp(\lambda t)) \\ &\leqslant \exp(-\lambda t)\mathbb{E}[\exp(\lambda X)]\,. \qquad (1.2) \end{aligned}$$

One first immediate result is the following:

Lemma 1.2 (*Chernoff's rule*) *Let $X \sim \nu$ be a real-valued random variable. Then*

$$\log \mathbb{E} \exp(X) \leqslant 0\,, \quad \textit{implies} \quad \forall \delta \in (0,1], \quad \mathbb{P}\Big(X \geqslant \ln(1/\delta)\Big) \leqslant \delta\,.$$

The proof is immediate by considering $t = \ln(1/\delta)$ and $\lambda = 1$ in (1.1). More generally, one can obtain a control of the tail of a random variable X from a control of the log-Laplace. The following result shows that conversely, a control of the tails of X induces a control of the log-Laplace:

Lemma 1.3 (*Tails and log-Laplace*) *Let X be a $\mathbb{R}$-valued random variable.*

$$\exists \lambda \in \mathbb{R}^+ : \log \mathbb{E}\exp(\lambda X) \leqslant \varphi(\lambda) \quad \Longrightarrow \quad \forall t \in \mathbb{R}, \quad \mathbb{P}(X \geqslant t) \leqslant \exp(\varphi(\lambda) - \lambda t)\,.$$

$$\forall t \in \mathbb{R}, \quad \mathbb{P}(X \geqslant t) \leqslant \alpha(t) \quad \Longrightarrow \quad \forall \lambda \in \mathbb{R}^+, \quad \log \mathbb{E}\exp(\lambda X) \leqslant \log \int_{\mathbb{R}} \alpha(u/\lambda) e^u d\mu(u)\,.$$

Proof :

Indeed, for any $\lambda > 0$

$$\mathbb{P}(X \geqslant t) \;=\; \mathbb{P}(\exp(\lambda X) \geqslant \exp(\lambda t)) \leqslant \mathbb{E}[\exp(\lambda X)]\exp(-\lambda t)\,,$$

where we applied Markov inequality to the $\mathbb{R}^+$-valued random variable $Z = \exp(\lambda X)$.

For the reverse inequality, we apply Fubini formula for Z, and conclude with a change of variable:

$$\log \mathbb{E}\exp(\lambda X) = \log \int_{\mathbb{R}^+} \mathbb{P}(\exp(\lambda X) \geqslant x) d\mu(x) = \log \int_{\mathbb{R}^+} \mathbb{P}(X \geqslant \log(x)/\lambda) d\mu(x)$$

□

Why logarithm? In these derivations, the $\exp$ transform may seem arbitrary, and one could indeed use more general transforms. Lemma 1.3 is stated using $\log$ and $\exp$ function, but there is nothing too specific about using the function $\log$ here. Indeed, let $(\underline{f}, \overline{f})$ be any pair of functions such that $\overline{f} : \mathbb{R} \to \mathbb{R}^+$ is increasing with $\overline{f}(\mathbb{R}) = \mathbb{R}^+$ and $\underline{f} \circ \overline{f} = \overline{f} \circ \underline{f}$ is the identity mapping. Then

$$\exists \lambda \in \mathbb{R}^+ : \underline{f}(\mathbb{E}\overline{f}(\lambda X)) \leqslant \varphi(\lambda) \quad \Longrightarrow \quad \forall t \in \mathbb{R}, \quad \mathbb{P}(X \geqslant t) \leqslant \frac{\overline{f}(\varphi(\lambda))}{\overline{f}(\lambda t)}\,.$$

However, using the pair $(\log, \exp)$ makes appear the quantity $\lambda t - \varphi(\lambda)$, which when optimized on λ corresponds to the Legendre-Fenchel dual of φ, another powerful mathematical tool.

Another natural explanation is that logarithms are most appropriate to deal with product measures: Let us say we have two measures p_1, p_2, and we form the product measure $p = p_1 \otimes p_2$. Since we are generally happier with summing things, let us look for a function such that $h(p_1 \otimes p_2) = h(p_1) + h(p_2)$ (additivity). It turns out that there are not too many choices. Indeed, let us first note that every (non zero) continuous morphism from $(\mathbb{R}_+^\star, \times)$ to $(\mathbb{R}, +)$ must be a logarithm function. The one that maps the neutral $e_\times = 1$ to the neutral $e_+ = 0$, is the classical logarithm $\log$. Now we want to build a function acting on probability measures, not just $\mathbb{R}^+$ A natural way to do so is by combining measures $p(S) \in \mathbb{R}$ of Borel sets S. Hence given two Borel sets S_1 and S_2, and measures p_1, p_2, we may want a function such that $f(p_1(S_1)p_2(S_2)) = f(p_1(S_1)) + f(p_2(S_2))$. While a logarithm function works, the remaining dependency on S_1, S_2 is not desirable. This can be done by replacing evaluation at a Borel set by integration over the space. For illustration, let us consider p_1 and p_2

are probability measures on a discrete set $\mathcal{X}$. Discrete integration (summing) on $\mathcal{X}$ reveals, using the fact that $p_1(\mathcal{X}) = p_2(\mathcal{X}) = 1$, that

$$\begin{aligned}\sum_{i,j\in\mathcal{X}^2} p_1(i)p_2(j)\log(p_1(i)p_2(j)) &= \sum_{i,j} p_1(i)p_2(j)\log(p_1(i)) + \sum_{i,j} p_1(i)p_2(j)\log(p_2(j)) \\ &= \sum_i p_1(i)\log(p_1(i)) + \sum_j p_2(j)\log(p_2(j)).\end{aligned}$$

Hence $p \to \sum_i p(i)\log(p(i))$ is a good candidate for h. Changing the sign then gives the entropy function $H(p) = -\sum_i p_i \log(p_i)$ (any multiplicative factor works as well). It turns out that we do not require much more to uniquely determine H. Indeed additivity for any p_1, p_2 plus assuming that when $\mathcal{X}$ is discrete, $h(p) = \sum_{x\in\mathcal{X}} g(p(x))$ for some measurable g null at 0 are enough to ensure unicity of H up to a constant factor, see Daróczy (1971), Csiszár (2008).

1.2 Two complementary results

Now one can consider two complementary points of view: The first one is to fix the value of t in (1.1) and (1.2) and minimize the probability level (the term on the right-hand side of the inequality). The second one is to fix the value of the probability level, and optimize the value of t. This leads to the following lemmas.

Lemma 1.4 (*Cramer-Chernoff*) *Let $X \sim \nu$ be a real-valued random variable. Let us introduce the log-Laplace transform and its Legendre transform:*

$$\begin{aligned}\forall\lambda\in\mathbb{R}, \quad \varphi_\nu(\lambda) &= \log\mathbb{E}[\exp(\lambda X)], \\ \forall t\in\mathbb{R}, \quad \varphi_\nu^\star(t) &= \sup_{\lambda\in\mathbb{R}}\Big(\lambda t - \varphi_\nu(\lambda)\Big),\end{aligned}$$

and let $\mathcal{D}_\nu = \{\lambda\in\mathbb{R} : \varphi_\nu(\lambda) < \infty\}$.
If $\mathcal{D}_\nu \cap \mathbb{R}_\star^+ \neq \emptyset$, then $\mathbb{E}[X] < \infty$ and for all $t \geqslant \mathbb{E}[X]$

$$\log\mathbb{P}(X \geqslant t) \leqslant -\varphi_\nu^\star(t)\,.$$

Likewise, if $\mathcal{D}_\nu \cap \mathbb{R}_\star^- \neq \emptyset$, $\mathbb{E}[X] > -\infty$ and for all $t \leqslant \mathbb{E}[X]$,

$$\log\mathbb{P}(X \leqslant t) \leqslant -\varphi_\nu^\star(t)\,.$$

Remark 1.1 *The log-Laplace transform φ_ν is also known as the cumulant generative function.*

Proof of Lemma 1.4:

First, note that $\{\lambda\in\mathbb{R} : \mathbb{E}[\exp(\lambda X)] < \infty\}$ coincides with $\{\lambda\in\mathbb{R} : \varphi_\nu(\lambda) < \infty\}$. Using equations (1.1) and (1.2), it holds:

$$\begin{aligned}\mathbb{P}(X \geqslant t) &\leqslant \inf_{\lambda\in\mathbb{R}_\star^+\cap\mathcal{D}_\nu} \exp(-\lambda t + \log\mathbb{E}[\exp(\lambda X)]) \\ \mathbb{P}(X \leqslant t) &\leqslant \inf_{\lambda\in\mathbb{R}_\star^-\cap\mathcal{D}_\nu} \exp(-\lambda t + \log\mathbb{E}[\exp(\lambda X)])\end{aligned}$$

The Legendre transform $\varphi_\nu^\star$ of the log-Laplace function φ_ν unifies these two cases. Indeed, a striking property of $\varphi_\nu^\star$ is that if $\lambda \in \mathcal{D}_\nu$ for some $\lambda > 0$, then $\mathbb{E}[X] < \infty$. This can be seen by Jensen's inequality applied to the function $\ln$: Indeed it holds $\lambda \mathbb{E}[X] = \mathbb{E}[\ln \exp(\lambda X)] \leqslant \varphi_\nu(\lambda)$. Further, for all $t \geqslant \mathbb{E}[X]$, it holds

$$\varphi_\nu^\star(t) = \sup_{\lambda \in \mathbb{R}^+ \cap \mathcal{D}_\nu} \Big(\lambda t - \varphi_\nu(\lambda) \Big).$$

Note that this also applies if $\mathbb{E}[X] = -\infty$. Likewise, if $\lambda \in \mathcal{D}_\nu$ for some $\lambda < 0$ then $\mathbb{E}[X] > -\infty$ and for all $t \leqslant \mathbb{E}[X]$, it holds

$$\varphi_\nu^\star(t) = \sup_{\lambda \in \mathbb{R}^- \cap \mathcal{D}_\nu} \Big(\lambda t - \varphi_\nu(\lambda) \Big).$$

□

Alternatively, the second point of view is to fix the confidence level $\delta \in (0, 1]$, and then to solve the equation $\exp(-\lambda t)\mathbb{E}[\exp(\lambda X)] = \delta$ in $t = t(\delta)$. We then optimize over t. This leads to:

Lemma 1.5 (*Alternative Cramer-Chernoff*) *Let $X \sim \nu$ be a real-valued random variable and let $\mathcal{D}_\nu = \{\lambda \in \mathbb{R} : \log \mathbb{E}\exp(\lambda X) < \infty\}$. It holds,*

$$\mathbb{P}\Big[X \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{ \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] + \frac{\log(1/\delta)}{\lambda} \Big\} \Big] \leqslant \delta$$
$$\mathbb{P}\Big[X \leqslant \sup_{\lambda \in (-\mathcal{D}_\nu) \cap \mathbb{R}_\star^+} \Big\{ -\frac{1}{\lambda} \log \mathbb{E}[\exp(-\lambda X)] - \frac{\log(1/\delta)}{\lambda} \Big\} \Big] \leqslant \delta \,.$$

Proof of Lemma 1.5:

Solving $\exp(-\lambda t)\mathbb{E}[\exp(\lambda X)] = \delta$ for $\delta \in (0, 1]$ and $\lambda \neq 0$, we obtain he following equivalence

$$\begin{aligned} -\lambda t + \log \mathbb{E}[\exp(\lambda X)] &= \log(\delta) \\ \lambda t &= -\log(\delta) + \log \mathbb{E}[\exp(\lambda X)] \\ t &= \frac{1}{\lambda} \log(1/\delta) + \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] \,. \end{aligned}$$

Thus, we deduce from (1.1) and (1.2) that

$$\begin{aligned} \forall \lambda > 0 \quad \mathbb{P}\Big[X \geqslant \frac{1}{\lambda} \log(1/\delta) + \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)] \Big] &\leqslant \delta \\ \forall \lambda > 0 \quad \mathbb{P}\Big[X \leqslant -\frac{1}{\lambda} \log(1/\delta) - \frac{1}{\lambda} \log \mathbb{E}[\exp(-\lambda X)] \Big] &\leqslant \delta \,. \end{aligned}$$

□

The rescaled Laplace transform $\lambda \to \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda X)]$ is sometimes called the *entropic risk* measure. Note that Lemma 1.4 and 1.5 involve slightly different quantities, depending on whether we focus on the probability level δ or the threshold on X.

1.3 Illustrative cases

An immediate corollary is the following:

Corollary 1.1 (*Sub-Gaussian Concentration Inequality*) *Let $\{X_i\}_{i\leqslant n}$ be independent R-sub-Gaussian random variables with mean μ, that is such that*

$$\forall \lambda \in \mathbb{R}, \quad \log \mathbb{E} \exp\Big(\lambda(X_i - \mu)\Big) \leqslant \frac{1}{2}\lambda^2 R^2 \,.$$

Then,

$$\forall \delta \in (0,1) \qquad \mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \sqrt{2R^2 n \log(1/\delta)}\Big] \leqslant \delta$$

Remark 1.2 *This corollary naturally applies to Gaussian random variables with variance σ^2, in which case $R = \sigma$. It also applies to bounded random variable. Indeed random variables $\{X_i\}_{i\leqslant n}$ bounded in $[0,1]$ are $1/2$-sub-Gaussian. This can be understood intuitively by remarking that distributions with the highest variance on $[0,1]$ are Bernoulli, and that the variance of a Bernoulli with parameter $\theta \in [0,1]$ is $\theta(1-\theta) \leqslant 1/4$, thus resulting in $R^2 = 1/4$. This is proved more formally via Hoeffding's lemma.*

Proof of Corollary 1.1:

Indeed, it holds that

$$\begin{aligned}
\frac{1}{\lambda}\log \mathbb{E}\big[\exp(\lambda \sum_{i=1}^n (X_i - \mu))\big] &= \frac{1}{\lambda}\log \mathbb{E}\Big[\prod_{i=1}^n \exp(\lambda(X_i - \mu))\Big] \\
&\overset{(a)}{=} \frac{1}{\lambda}\log \prod_{i=1}^n \mathbb{E}\Big[\exp(\lambda(X_i - \mu))\Big] \\
&= \frac{1}{\lambda}\sum_{i=1}^n \log \mathbb{E}\Big[\exp(\lambda(X_i - \mu))\Big] \\
&\overset{(b)}{\leqslant} \frac{n}{2}\lambda R^2 \,,
\end{aligned}$$

where (a) is by independence, and (b) holds by using the sub-Gaussian assumption. We deduce by Lemma 1.5 that

$$\begin{aligned}
&\mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{\lambda R^2 n/2 + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \\
&\overset{(a)}{\leqslant} \mathbb{P}\Big[\sum_{i=1}^n (X_i - \mu) \geqslant \inf_{\lambda \in \mathcal{D}_\nu \cap \mathbb{R}_\star^+} \Big\{\frac{1}{\lambda}\log \mathbb{E}\Big[\exp\Big(\lambda \sum_{i=1}^n (X_i - \mu)\Big)\Big] + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \\
&\leqslant \delta \,,
\end{aligned}$$

where in (a), we used that $x < y$ implies $\mathbb{P}(X \geqslant y) \leqslant \mathbb{P}(X \geqslant x)$.

Now we note that $\mathcal{D}_\nu = \mathbb{R}$ by the sub-Gaussian assumption, where ν is the distribution of $\sum_{i=1}^n X_i$. We conclude by noticing that $\lambda_\delta = \sqrt{\frac{2\log(1/\delta)}{R^2 n}}$ achieves the minimum in

$$\inf_{\lambda \in \mathbb{R}_\star^+} \left\{ \lambda R^2 n/2 + \frac{\log(1/\delta)}{\lambda} \right\} \;=\; \sqrt{2R^2 n \log(1/\delta)}\,.$$

□

Another interesting case is that of Bernoulli distributions. Let $\{X_i\}_{i \leqslant n}$ be independent Bernoulli variables with mean p and φ_p denote the log-Laplace transform of X_1. It can be checked that its Legendre-Fenchel dual satisfies for each $q \in [0,1], \varphi_p^\star(q) = \mathtt{kl}(q,p)$ where $\mathtt{kl}(q,p) = q\log(q/p) + (1-q)\log((1-q)/(1-p))$. Hence, we deduce that

$$\forall \varepsilon \geqslant 0, \quad \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n (X_i - p) \geqslant \varepsilon\right] \leqslant \exp\Big(- n\mathtt{kl}(p+\varepsilon, p)\Big)\,.$$

The inverse map of $\varepsilon \mapsto \mathtt{kl}(p+\varepsilon, p)$ is unfortunately not explicit, however it is not difficult to show by a local version of Pinsker inequality that

$$\begin{aligned} \mathtt{kl}(p+\varepsilon, p) &\geqslant \frac{\varepsilon^2}{2}\Big(\frac{1}{\tilde{q}} + \frac{1}{1-\tilde{q}}\Big) \text{ where } \tilde{q} \in [p, p+\varepsilon] \\ &\geqslant \frac{\varepsilon^2}{2}\Big(\frac{1}{p+\varepsilon} + \frac{1}{1-p}\Big) = \frac{\varepsilon^2}{2p(1-p)} - o(\varepsilon^2) \text{ as } \varepsilon \to 0\,. \end{aligned}$$

We refer to Berend and Kontorovich (2013), Kearns and Saul (1998), or to the monograph Raginsky et al. (2013), p. 25-26 for further details and existing approximations.

2 Legendre-Fenchel dual of the Log-Laplace

Lemma 1.4 provides a first example of control of the tail of a random variable using the Legendre-Fenchel dual of its log-Laplace transform. It turns out the log-Laplace can also be interpreted as a form of Legendre-Fenchel dual. In order to see this, let us first remark that so far, we have considered a $\mathbb{R}$-valued random variable X. However, what happens when X is $\mathcal{X}$ valued instead, where $\mathcal{X}$ can be in $\mathbb{R}^d$, or on the other hand a discrete set? In that case, we consider a bounded function $\lambda : \mathcal{X} \to \mathbb{R}$ and form the quantity $\log \mathbb{E}\exp(\lambda(X))$. This generalizes the previous case that is recovered when choosing a constant function.

Let μ and ν denote two measures on $\mathcal{X}$, and for a function $\lambda : \mathcal{X} \to \mathbb{R}$, let $\langle \mu, f\rangle = \int f(x) d\mu(x)$ be the duality product. Using such notations, it is possible to show the following two striking results

$$\begin{aligned} &\text{(Entropy formula)} \quad \log \mathbb{E}_\nu[\exp(\lambda(X))] \;=\; \sup_{\mu \in \mathcal{P}(\mathcal{X})} \langle \mu, \lambda\rangle - \mathtt{KL}(\mu, \nu) \\ &\text{(Variational formula)} \qquad\quad \mathtt{KL}(\mu,\nu) \;=\; \sup_{f \in \mathcal{C}_b(\mathcal{X})} \langle \mu, f\rangle - \log \mathbb{E}_\nu[\exp(f(X))]\,, \end{aligned}$$

where the first supremum is over the set of probability measures on $\mathcal{X}$ (denoted $\mathcal{P}(\mathcal{X})$), and the second over the set $\mathcal{C}_b(\mathcal{X})$ of continuous and bounded functions on $\mathcal{X}$. Hence, we see from the Entropy formula that the log-Laplace transform (seen as a function of λ) is the Legendre-Fenchel dual of the function $\mu \to \mathtt{KL}(\mu, \nu)$,

and by the variational formula that the function $\mu \to \mathtt{KL}(\mu, \nu)$ is the Legendre-Fenchel dual of function $\lambda \to \log \mathbb{E}_\nu[\exp(\lambda(X))]$. It is thus no wonder that $\mathtt{KL}$ appears in so many results of statistics. Note also that the supremum is reached for a distribution μ whose Radon-Nikodym derivative is given explicitly by

$$\text{(Gibbs distribution)} \quad \frac{d\mu}{d\nu}(x) = \frac{\exp(\lambda(x))}{\mathbb{E}_\nu[\exp(\lambda(X))]} \,.$$

The Gibbs distribution appears naturally in the setting of aggregation of experts. In this case, $\mathcal{X}$ is typically a finite set of experts, and $\lambda(x)$ captures the opposite of the loss of expert x. Thus the distribution μ gives to each expert a weight that is exponentially decreasing with its loss. When appropriately chosen, this enables to build an aggregate expert that behaves nearly as good as the best (convex) combination of the experts' decisions. Before we present the core of aggregation strategies, let us first discuss the notion of loss in greater details.

3 Loss and noise

The control of the tails probability induced by the log-Laplace transform is especially interesting when applied to the concept of noise and the construction of a loss.

Indeed, let us say that a value $c \in \mathbb{R}$ is perturbed by some random variable ξ with zero mean. You don't know c but you are asked to provide a proposed value c'. Then you observe $c + \xi$. We may want to give a score to a proposed value c'. In general this is captured by a notion of loss $\ell(c + \xi, c')$. Intuitively, we would like this loss to be small when $c' = c$. However, there are many possible notions of losses, and many possible laws for ξ. Hence, there is no reason that the loss and law of ξ lead to a reasonable value for $\ell(c + \xi, c)$. Can we relate the law of ξ to the loss in order to ensure that $\mathbb{P}(\ell(c + \xi, c) \geqslant t)$ is controlled for any $t \in \mathbb{R}^+$ and quickly vanishes as $t \to \infty$?

The control of the log-Laplace provides an interesting answer to this question:

Lemma 1.6 (*Loss-adapted noise*) *Let ℓ satisfy $\ell(y, y') = \psi(y - y')$ where ψ is a convex, non-negative function that null in 0. Let $\psi^\star$ be the Legendre-Fenchel dual of ψ (that is $\psi^\star(\lambda) = \sup_{y \in \mathcal{Y} - \mathcal{Y}} \langle \lambda, y \rangle - \psi(y)$). Then*

$$\forall \lambda \in \mathbb{R}, \quad \ln \mathbb{E}[\exp(\lambda \xi)] \leqslant \psi^\star(\lambda) \quad \Longrightarrow \quad \forall t \in \mathbb{R}^+, \mathbb{P}(\ell(c + \xi, c) \geqslant t) \leqslant 2 \exp(-t) \,,$$

Proof :

Indeed, by convexity of ψ, it holds $\psi^{\star\star} = \psi$. Now let ψ_+ be the restriction of ψ to the positive cone $\mathbb{R}^+$, and ψ_- its restriction to the negative cone $\mathbb{R}^-$. From Lemma 1.4, we deduce that

$$\mathbb{P}((c + \xi) - c \geqslant \psi_+^{-1}(t)) \exp(-\psi^{\star\star}(\psi_+^{-1}(t))) = \exp(-t) \,.$$

The same bound holds for $\mathbb{P}((c+\xi) - c \leqslant \psi_-^{-1}(t))$. Thus we conclude by combining the two results with a simple union bounds. □

Definition 1.3 (*Loss-adapted noise*) *The noise is adapted to the loss $\ell(y, y') = \psi(y - y')$ if the cumulative generative function of the noise is dominated by the Legendre-Fenchel dual of the potential function.*

This result enables to understand easily, from a given loss function, what noises are adapted to it. A typical example is the quadratic loss $\ell(y, y') = \frac{(y-y')^2}{2R^2}$, for which $\psi^\star(\lambda) = \frac{\lambda^2 R^2}{2}$. Hence, any R-sub-Gaussian random variable is adapted to this loss. Typical examples of potential functions are the following

$$\begin{array}{lll} \textit{(quadratic potential)} & \psi(y) = & \frac{1}{2}y^2 \\ \textit{(tolerance potential)} & \psi_\varepsilon(y) = & \chi_{(-\varepsilon,\varepsilon)}(y) \\ \textit{(risk-averse potential)} & \psi_{[a,b],\alpha,\beta}(y) = & \frac{1}{\alpha}(a-y)_+^\alpha + \frac{1}{\beta}(y-b)_+^\beta \,, \end{array}$$

where $\chi_{(-\varepsilon,\varepsilon)}(y) = 0$ if $y \in (-\varepsilon, \varepsilon)$, and is else $+\infty$, and where the risk-averse potential models asymmetric sensitivity to upper and lower estimation. The choice of the potential function ψ is generally application-driven, and often influences implicitly the type of allowed noise model; Definition 1.3 makes it explicit.

We have shown how the choice of a loss can induce a natural notion of noise by duality. Conversely, a noise measure naturally induces a notion of loss: In order to show this, we proceed backward, starting from a class of distributions in order to build a loss function. For an abstract space $\mathcal{Y}$, there is no necessarily natural notion of linearity giving a meaning to $y - y'$ or $\langle \lambda, y \rangle$. Think for instance of a discrete space $\mathcal{Y} = \{1, \ldots, S\}$. However, one can still consider a class of distributions, and functions on $\mathcal{Y}$. Thus, given a candidate distribution π for Y_n, we consider $g(\lambda) = \ln \mathbb{E}_\pi \exp \lambda(Y)$, for any function λ that is bounded, continuous. Interpreting g as the dual $\psi^\star$ of a convex loss, it is then natural to look at its dual $g^\star$ in order to recover the definition of the loss. Note that $g^\star$ acts on measures ν. It comes

$$g^\star(\nu) \;=\; \sup_{\lambda \in \mathcal{C}_B(\mathcal{Y})} (\nu, \lambda) - g(\lambda) = \sup_{\lambda \in \mathcal{C}_B(\mathcal{Y})} \mathbb{E}_\nu[\lambda(Y)] - \ln \mathbb{E}_\pi \exp \lambda(Y) = \mathtt{KL}(\nu, \pi) \,,$$

where $(\cdot, \cdot)$ is the duality product, and $\mathcal{C}_B(\mathcal{Y})$ are continuous bounded functions on $\mathcal{Y}$. Thus, the loss induced by π on probability measures coincides with the Kullback-Leibler divergence. In particular interpreting an observation Y_n as a Dirac distribution at point Y_n, the loss becomes for this choice of ν, $-\ln(p(Y_n))$ where p denotes the density of π with respect to the reference measure. This justifies the introduction of the following

Definition 1.6 (*Self-information loss*) *The loss of a distribution π on $\mathcal{Y}$ with density p is given by*

$$\ell_I(\pi, y) = -\ln p(y) \,.$$

Remark 1.3 *The self-information loss is a popular and standard loss in the literature on sequential prediction. Its expectation with respect to y coincides with the Kullback-Leibler of the distribution of y with π. The notion of loss-adapted noise, although less frequent, also appears in certain works. We refer to (Merhav and Feder, 1998) for an extended study of universal sequential prediction and further details.*

We conclude this section on losses with a useful property that directly connects a loss function to the log-Laplace transform and that is especially useful when designing aggregation strategies.

Definition 1.9 (*Mixable loss*) *The loss function ℓ is* η-mixable *for some $\eta > 0$, if*

$$\forall \boldsymbol{x} = (x_i)_{1 \leqslant i \leqslant M} \in \mathcal{X}^M \forall \boldsymbol{v} = (v_i)_{1 \leqslant i \leqslant M} \in \mathscr{P}_M, \exists x_{\boldsymbol{v}} \in \mathcal{X} \forall y \in \mathcal{Y}, \quad \ell\left(x_{\boldsymbol{v}}, y\right) \leqslant -\frac{1}{\eta} \ln \sum_{i=1}^{M} v_i \, e^{-\eta \, \ell(x_i, y)} \tag{1.5}$$

The mapping $\boldsymbol{x}, \boldsymbol{v} \to x_{\boldsymbol{v}}$ is called the substitution function.

Remark 1.4 *A important class of η-mixable losses is that of* η-exp-concave *losses, in the sense that the function $\exp(-\eta\, \ell(\cdot, y))$ is concave on $\mathcal{X}$ for every observation $y \in \mathcal{Y}$, with substitution function $x_{\boldsymbol{v}} = \sum_{i=1}^{M} v_i \, x_i$. One example in the case when $\mathcal{X}$ is the set of probability measures over $\mathcal{Y}$ is the* logarithmic *or* self-information *loss $\ell(x, y) = -\log x(\{y\})$ for which the inequality holds with $\eta = 1$, and is actually an equality. Another example of special interest is the quadratic loss on a bounded interval: indeed, for $\mathcal{X} = \mathcal{Y} = [a, b] \subset \mathbb{R}$, $\ell(x, y) = (x - y)^2$ is $\frac{1}{2(b-a)^2}$-exp-concave, and $\frac{2}{(b-a)^2}$-mixable.*

In order to better understand this definition, let us recall the following structural properties.

Lemma 1.7 (*Structural properties*) *For any function f, provided all the following terms are finite, it holds*

$$\begin{aligned} -\log \mathbb{E} \exp(-f) \leqslant \quad & \mathbb{E} f \quad \leqslant \log \mathbb{E} \exp f \\ \forall x \in \mathcal{X}, 0 \leqslant f(x) \quad \Longrightarrow \quad & \mathbb{E} f \leqslant -\log \mathbb{E} \exp(-f) + \frac{1}{2} \mathbb{E} f^2. \\ \forall x \in \mathcal{X}, |f(x)| \leqslant b \quad \Longrightarrow \quad & \log \mathbb{E} \exp f - \frac{e^b - 1 - b}{b^2} \mathbb{E} f^2 \leqslant \mathbb{E} f \, . \end{aligned}$$

Proof :

The first line is proved by Jensen's inequality. The second and third line by a Taylor expansion. □

Hence, applying this to the function $f(\cdot) = \eta \ell(\cdot, y)$ when ℓ is convex in its first argument, we obtain

$$\ell\Big(\sum_{i=1}^{M} v_i \, x_i, y\Big) \leqslant \mathbb{E}_{\boldsymbol{v}}[\ell(X, y)] \leqslant -\frac{1}{\eta} \ln \sum_{i=1}^{M} v_i \, e^{-\eta \, \ell(x_i, y)} + \frac{\eta}{2} \mathbb{E}[\ell(X, y)^2] \, .$$

Note that η-exp-concavity removes the quadratic term on the right hand size of the inequality.

A simple aggregation strategy To provide an illustration of what can be achieved with such losses, we introduce now the simple but fundamental *exponential weights* or *Hedge algorithm* (Vovk, 1998, Cesa-Bianchi and Lugosi, 2006), designed to control, for a fixed set of experts $\mathcal{M} = \{1, \ldots, M\}$, the regret $L_T - L_{i,T} = \sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i,t}$ for each $i \in \mathcal{M}$, where $\ell_{i,t} = \ell(x_{i,t}, y_t)$ is the loss of expert i at time t, and $\ell_t = \ell(x_t, y_t)$ the loss of the expert choosing x_t at time t. This algorithm is our first example of an aggregation strategy. Strikingly,

this is also a way to build complex forecasting strategies. The algorithm depends on a *prior distribution* $\boldsymbol{\pi} \in \mathscr{P}_M (= \mathcal{P}(\mathcal{M}))$ on the experts and predicts as $x_t = x_{\boldsymbol{v}_t}$, where the weights $\boldsymbol{v}_t \in \mathscr{P}_M$ are sequentially updated in the following way: $\boldsymbol{v}_1 = \boldsymbol{\pi}$ and, after each round $t \geqslant 1$, $\boldsymbol{v}_{t+1}$ is set to the *posterior* distribution $\boldsymbol{v}_t^m$ of $\boldsymbol{v}_t$ given the losses $(\ell_{i,t})_{1 \leqslant i \leqslant M}$, defined by

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,. \tag{1.6}$$

The following result that can be found in Cesa-Bianchi and Lugosi (2006, Corollary 3.1) enables to analyze a number of challenging situations, by reducing complex forecasting strategies to the aggregation of experts under a suitable prior.

Proposition 1.1 (*Regret of aggregation*) *Assume the loss function ℓ is η-mixable. Irrespective of the values of the signal and the experts' predictions, the exponential weights algorithm with prior $\boldsymbol{\pi}$ achieves*

$$L_T - L_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \tag{1.7}$$

for each $i = 1, \dots, M$ and $T \geqslant 1$. More generally, for each probability vector $\boldsymbol{u} \in \mathscr{P}_M$,

$$L_T - \sum_{i=1}^M u_i\, L_{i,T} \leqslant \frac{1}{\eta}\, KL(\boldsymbol{u}, \boldsymbol{\pi})\,. \tag{1.8}$$

Choosing a uniform prior $\boldsymbol{\pi} = \frac{1}{M}\mathbf{1}$ yields an at most $\frac{1}{\eta} \log M$ regret with respect to the best expert.

Proof :

Since the loss function is η-mixable and $x_t = x_{\boldsymbol{v}_t}$, we have

$$e^{-\eta\, \ell(x_t, y_t)} \geqslant \sum_{i=1}^M v_{i,t}\, e^{-\eta\, \ell(x_{i,t}, y_t)}, \qquad \textit{i.e.} \qquad \ell_t \leqslant -\frac{1}{\eta} \log \left(\sum_{i=1}^M v_{i,t}\, e^{-\eta\, \ell_{i,t}} \right).$$

This yields, introducing the posterior weights $v_{i,t}^m$ defined by (1.6),

$$\ell_t - \ell_{i,t} \leqslant -\frac{1}{\eta} \log \left(\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}} \right) - \ell_{i,t} = \frac{1}{\eta} \log \left(\frac{e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}} \right) = \frac{1}{\eta} \log \frac{v_{i,t}^m}{v_{i,t}}\,.$$

Now recalling that the exponentially weighted average forecaster uses $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t^m$, this writes: $\ell_t - \ell_{i,t} \leqslant \frac{1}{\eta} \ln \frac{v_{i,t+1}}{v_{i,t}}$ which, summing over $t = 1, \dots, T$, yields $L_T - L_{i,T} \leqslant \frac{1}{\eta} \ln \frac{v_{i,T+1}}{v_{i,1}}$. Since $v_{i,1} = \pi_i$ and $v_{i,T+1} \leqslant 1$, this proves (1.7); moreover, noting that $\ln \frac{v_{i,T+1}}{v_{i,1}} = \ln \frac{u_i}{v_{i,1}} - \ln \frac{u_i}{v_{i,T+1}}$, this implies

$$\sum_{i=1}^M u_i\, (L_T - L_{i,T}) \leqslant \frac{1}{\eta} \sum_{i=1}^M u_i\, \ln \frac{v_{i,T+1}}{v_{i,1}} = \frac{1}{\eta} \big(\mathtt{KL}(\boldsymbol{u}, \boldsymbol{v}_1) - \mathtt{KL}(\boldsymbol{u}, \boldsymbol{v}_{T+1}) \big)\,,$$

which establishes (1.8) since $\boldsymbol{v}_1 = \boldsymbol{\pi}$ and $\mathtt{KL}(\boldsymbol{u}, \boldsymbol{v}_{T+1}) \geqslant 0$. □

Remark 1.5 *We can recover the bound (1.7) from inequality (1.8) by considering* $\boldsymbol{u} = \delta_i$*. Conversely, inequality (1.7) implies, by convex combination,*

$$L_T - \sum_{i=1}^{M} u_i \, L_{i,T} \leqslant \frac{1}{\eta} \sum_{i=1}^{M} u_i \log \frac{1}{\pi_i} \, ;$$

inequality (1.8) improves on this bound, as it replaces the terms $\ln \frac{1}{\pi_i}$ *by* $\ln \frac{u_i}{\pi_i}$*. Following Koolen et al. (2012), we make use of such refinement in Chapter 9.*

4 General exponential families, properties

In this section, we study exponential families that are closely related to the log-Laplace object. For a set $\mathcal{X} \subset \mathbb{R}$, we consider a multivariate function $F : \mathcal{X} \to \mathbb{R}^K$ and denote $\mathcal{Y} = F(\mathcal{X}) \subset \mathbb{R}^K$.

Definition 1.12 (*Exponential families*) *The exponential family generated by the function* F *and the reference measure* ν_0 *on the set* $\mathcal{X}$ *is*

$$\mathcal{E}(F;\nu_0) = \Big\{ \nu_\theta \in \mathcal{P}(\mathcal{X}) \, ; \, \forall x \in \mathcal{X} \;\; \nu_\theta(dx) = \exp\big(\langle \theta, F(x) \rangle - \psi(\theta) \big) \nu_0(dx), \;\; \theta \in \mathbb{R}^K \Big\} \, ,$$

where $\psi(\theta) \stackrel{\text{def}}{=} \log \displaystyle\int_{\mathcal{X}} \exp\Big(\langle \theta, F(x) \rangle \Big) \nu_0(dx)$ *is the normalization function (aka log-partition function) of the exponential family. The vector* θ *is called the vector of canonical parameters. The parameter set of the family is the domain* $\Theta_{\mathcal{D}} \stackrel{\text{def}}{=} \Big\{ \theta \in \mathbb{R}^K \, ; \, \psi(\theta) < \infty \Big\}$*, and the invertible parameter set of the family is* $\Theta_I \stackrel{\text{def}}{=} \Big\{ \theta \in \mathbb{R}^K \, ; \, 0 < \lambda_{MIN}(\nabla^2 \psi(\theta)) \leqslant \lambda_{MAX}(\nabla^2 \psi(\theta)) < \infty \Big\} \subset \Theta_{\mathcal{D}}$*, where* $\lambda_{MIN}(M)$ *and* $\lambda_{MAX}(M)$ *denote the minimum and maximum eigenvalues of a semi-definite positive matrix* M*.*

Remark 1.6 *When* $\mathcal{X}$ *is compact, which is the usual assumption in multi-armed bandits (*$\mathcal{X} = [0,1]$*) and* F *is continuous, then we automatically get* $\Theta_{\mathcal{D}} = \mathbb{R}^K$*.*

When ν_0 is a probability measure, the log-partition of the family coincides with the log-Laplace transform of the random variable $F(X)$. In the sequel, we always assume that the family is regular, that is $\Theta_{\mathcal{D}}$ has non empty interior. Another key assumption is that the parameter $\theta^\star$ of the optimal arm belongs to the interior of Θ_I and is away from its boundary, which essentially avoids degenerate distributions, as we illustrate below.

Examples Bernoulli distributions form an exponential family with $K = 1$, $\mathcal{X} = \{0,1\}$, $F(x) = x$,$\psi(\theta) = \log(1 + e^\theta)$. The Bernoulli distribution with mean μ has parameter $\theta = \log(\mu/(1-\mu))$. Note that $\Theta_{\mathcal{D}} = \mathbb{R}$ and that degenerate distributions with mean 0 or 1 correspond to parameters $\pm\infty$.

Gaussian distributions on $\mathcal{X} = \mathbb{R}$ form an exponential family with $K = 2$, $F(x) = (x, x^2)$, and for each $\theta = (\theta_1, \theta_2)$, $\psi(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2} \log\Big(-\frac{\pi}{\theta_2} \Big)$. The Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ has parameter $\theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$. It is immediate to check that $\Theta_{\mathcal{D}} = \mathbb{R} \times \mathbb{R}_\star^-$. Degenerate distributions with variance 0 correspond to a parameter θ with both infinite components, while as θ approaches the boundary $\mathbb{R} \times \{0\}$, then the variance tends to infinity. It is natural to consider only parameters that correspond to a not too large variance.

Bregman divergence induced by the exponential family An interesting property of exponential families is the following straightforward rewriting of the Kullback-Leibler divergence:

$$\forall \theta, \theta' \in \Theta_{\mathcal{D}}, \quad \mathtt{KL}(\nu_\theta, \nu_{\theta'}) = \langle \theta - \theta', \mathbb{E}_{X \sim \nu_\theta}(F(X)) \rangle - \psi(\theta) + \psi(\theta') \,,$$

In particular, the vector $\mathbb{E}_{X \sim \nu_\theta}(F(X))$ is called the vector of *dual (or expectation) parameters*. It is equal to the vector $\nabla \psi(\theta)$. Now, it is interesting to note that $\mathtt{KL}(\nu_\theta, \nu_{\theta'}) = \mathcal{B}^\psi(\theta, \theta')$, where $\mathcal{B}^\psi$ is known as the Bregman divergence with potential function ψ and is defined (see Bregman (1967) for further details) by

$$\mathcal{B}^\psi(\theta, \theta') \stackrel{\text{def}}{=} \psi(\theta') - \psi(\theta) - \langle \theta' - \theta, \nabla \psi(\theta) \rangle \,.$$

We continue by providing a powerful rewriting of the Bregman divergence.

Lemma 1.8 (*Bregman duality*) *For all $\theta^\star \in \Theta_{\mathcal{D}}$ and $\eta \in \mathbb{R}^K$ such that $\theta^\star + \eta \in \Theta_{\mathcal{D}}$, let $\Phi(\eta) = \psi(\theta^\star + \eta) - \psi(\theta^\star)$. Further, let us introduce the Fenchel-Legendre dual of Φ defined by*

$$\Phi^\star(y) = \sup_{\eta \in \mathbb{R}^K} \langle \eta, y \rangle - \Phi(\eta) \,.$$

Then, it holds $\log \mathbb{E}_{X \sim \nu_{\theta^\star}} \exp\Big(\langle \eta, F(X) \rangle \Big) = \Phi(\eta) \,.$ *Further, for all F such that $F = \nabla \psi(\theta)$ holds for some $\theta \in \Theta_{\mathcal{D}}$, then the following equality holds true* $\Phi^\star(F) = \mathcal{B}^\psi(\theta, \theta^\star)$.

Proof of Lemma 1.8:

The second equality holds by simple algebra. Now the first equality is immediate, since

$$\begin{aligned} \log \mathbb{E}_{\theta^\star} \exp(\langle \eta, F(X) \rangle) &= \log \int \exp(\langle \eta, F(x) \rangle + \langle \theta^\star, F(x) \rangle - \psi(\theta^\star)) \nu_0(dy) \\ &= \psi(\eta + \theta^\star) - \psi(\theta^\star) \,. \end{aligned}$$

□

This result is especially useful in the analysis of boundary crossing probabilities for multi-armed bandits, as we explain in chapter 7. The main reason is that the function $\Phi^\star$ is increasing on each affine dual cone centered at $\nabla \psi(\theta^\star)$ (with respect to the cone ordering).

Likelihood and duality We finally mention a nice property of exponential families and its relation with the likelihood of n observations. Indeed, the value of the maximum likelihood coincides with n times the Legendre-Fenchel dual of the log-partition function, applied to the empirical mean of the feature function F:

$$\sup_{\theta \in \Theta} \sum_{i=1}^n \log \nu_\theta(X_i) = \sup_{\theta \in \Theta} \langle \theta, \sum_{i=1}^n F(X_i) \rangle - n\psi(\theta) = n \sup_{\theta \in \Theta} \langle \theta, \widehat{F}_n \rangle - \psi(\theta) = n\psi^\star(\widehat{F}_n) \,,$$

where we introduced the empirical mean $\widehat{F}_n = \frac{1}{N} \sum_{i=1}^n F(X_i)$.

CHAPTER 2
$\{\tau \leqslant t\} \in \mathcal{F}_t$

Contents

1 **The peeling technique for random stopping times** **24**

2 **Uniform bounds and the Laplace method** **30**

3 **Numerical comparison of a few bounds** **34**

Take-home message

Standard concentration inequalities (Hoeffding, Bersntein, etc.) are valid for an integer number of observations $t \in \mathbb{N}$: they do not apply when we have instead a random number of observations.
When the number of observations is a random variable τ, we can amend these results in two main ways:

Peeling Principle: localizing a random variable N by partitioning its domain in many sub-domains.
Result: Let $Z = \{Z_i\}_{i=1}^{\infty}$ be a process that is predictable, and let $\mathcal{F}^Z$ denotes its natural filtration. Let φ be an upper-envelope of the log-Laplace of the Z_i, and let $\psi_{\star,+}^{-1}$ denote the positive invert map of its Legendre-Fenchel dual.
Let N_n be a $\mathbb{N}$-valued random variable that is $\mathcal{F}^Z$-measurable and a.s. bounded by n. Then

$$\forall \alpha \in (1, n], \delta \in (0,1), \quad \mathbb{P}\left[\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\left(\frac{\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)\right)\right] \leqslant \delta$$

Now, if N is a (possibly unbounded) $\mathbb{N}$-valued random variable that is $\mathcal{F}^Z$-measurable,

$$\forall \alpha > 1, \delta \in (0,1) \quad \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varphi_{\star,+}^{-1}\left(\frac{\alpha}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta \ln^2(\alpha)}\right]\right)\right] \leqslant \delta$$

Advantage: Generic method. Extends to concentration of multivariate random variables and beyond.
Drawback: Using a geometric grid is not necessarily optimal and lacks adaptivity.

Laplace Principle: Replace the optimization appearing in the dual of $\lambda \to \log \mathbb{E}\exp(\lambda Z)$ by an integration.
Result: Assume $\varphi(\lambda) = \sigma^2\lambda^2/2$ is an envelop on the log-Laplace transform of Z. Let N be a (possibly unbounded) random stopping time for the natural filtration $\mathcal{F}^Z$. Then

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \sigma\sqrt{2\left(1+\frac{1}{N}\right)\frac{\ln\left(\sqrt{N+1}/\delta\right)}{N}}\right) \leqslant \delta$$

Advantage: Follows the distribution tail, improves on peeling when it applies. Extends to concentration of multivariate random variables and beyond.
Drawback: Restricted to specific distributions and explicit computations.

Application When applied to a bandit with $[0,1]$-bounded observations, this yields one of the best bandit strategies

$$\text{(UCB-Laplace)} \quad A_t = \arg\max_{a\in\mathcal{A}} \widehat{\mu}_{a,t} + \sqrt{\frac{\left(1+\frac{1}{N_t(a)}\right)\ln\left(A\sqrt{N_t(a)+1}/\delta\right)}{2N_t(a)}}\,.$$

In practice, it is often desirable to control not only a random variable such as an empirical mean at a single time step n, but also at multiple time steps $n = 1, \dots$. The naive approach to do so is by controlling the concentration at each different time step and then use a union-bound to deduce the final bound. However, this is generally sub-optimal as the empirical mean at time n and at time $n+1$ are close to each other and correlated. We study here two powerful methods that enable to improve on this naive strategy. Both methods heavily rely on the notion of random stopping time.

In order to illustrate the power of the random stopping times, let us start by recalling two standard and important inequalities:

Lemma 2.1 (*Doob's maximal inequality for non-negative sub-martingale*) *Let $\{W_t\}_{t\in\mathbb{N}}$ be a non-negative sub-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, that is*

$$\forall t \in \mathbb{N}, t' \geqslant t \qquad \mathbb{E}[W_{t'}|\mathcal{F}_t] \geqslant W_t\,, \text{ and } W_t \geqslant 0\,.$$

Then, for all $p \geqslant 1$ and $\varepsilon > 0$, it holds for all $T \in \mathbb{N}$

$$\mathbb{P}\Big(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\Big) \leqslant \frac{\mathbb{E}[W_T^p]}{\varepsilon^p}\,.$$

Proof :

Let us introduce the random variable $\tau_\varepsilon = \min\{t \in [0,T] : W_t \geqslant \varepsilon\}$. Using this variable, we get

$$\begin{aligned}\varepsilon\mathbb{P}\Big(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\Big) &= \mathbb{E}[\varepsilon\mathbb{I}\{\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\}] \\ &\leqslant \mathbb{E}[W_{\tau_\varepsilon}\mathbb{I}\{\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\}] \\ &\leqslant \mathbb{E}[W_{\tau_\varepsilon}]\,.\end{aligned}$$

Now, we observe that τ_ε is a random stopping time bounded by T. Further, it holds

$$\begin{aligned}\mathbb{E}[W_{\tau_\varepsilon}] &= \mathbb{E}[\liminf_{t\to\infty} W_{\min(t,\tau_\varepsilon)}] \\ &\leqslant \liminf_{t\to\infty} \mathbb{E}[W_{\min(t,\tau_\varepsilon)}]\,.\end{aligned}$$

Using the sub-martingale property, we deduce that $\mathbb{E}[W_{\min(t,\tau_\varepsilon)}] \leqslant \mathbb{E}[W_T]$. Indeed, for $\tilde{\tau}_\varepsilon = \min(t, \tau_\varepsilon)$,

$$\begin{aligned}\mathbb{E}[W_{\tilde{\tau}_\varepsilon}] &= \mathbb{E}\Big[\sum_{s=0}^{t} W_s\mathbb{I}\{\tilde{\tau}_\varepsilon = s\}\Big] \leqslant \mathbb{E}\Big[\sum_{s=0}^{t} \mathbb{E}[W_t|\mathcal{F}_s]\mathbb{I}\{\tilde{\tau}_\varepsilon = s\}\Big] \\ &= \mathbb{E}\Big[\sum_{s=0}^{t} \mathbb{E}[W_t\mathbb{I}\{\tilde{\tau}_\varepsilon = s\}|\mathcal{F}_s]\Big] = \mathbb{E}[W_t \sum_{s=0}^{t} \mathbb{I}\{\tilde{\tau}_\varepsilon = s\}] = \mathbb{E}[W_t]\,.\end{aligned}$$

Eventually, this shows that $\varepsilon\mathbb{P}\Big(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\Big) \leqslant \mathbb{E}[W_T]$. □

Lemma 2.2 (*Doob's maximal inequality for non-negative super-martingale*) *Let $\{W_t\}_{t\in\mathbb{N}}$ be a non-negative super-martingale with respect to the filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$, that is*

$$\forall t \in \mathbb{N}, \qquad \mathbb{E}[W_{t+1}|\mathcal{F}_t] \leqslant W_t\,, \text{ and } W_t \geqslant 0\,.$$

Then, for all $p \geqslant 1$ and $\varepsilon > 0$, it holds for all $T \in \mathbb{N}$

$$\mathbb{P}\left(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\right) \leqslant \frac{\mathbb{E}[W_0^p]}{\varepsilon^p}\,.$$

Proof :

In order to prove this result, we use a slightly different decomposition:

$$\begin{aligned}
\mathbb{E}[W_{\tilde{\tau}_\varepsilon}] &= \mathbb{E}\left[W_0 + \sum_{s=0}^{\tilde{\tau}_\varepsilon - 1}(W_{s+1} - W_s)\right] = \mathbb{E}\left[W_0 + \sum_{s'=0}^{t}\sum_{s=0}^{s'-1}(W_{s+1} - W_s)\mathbb{I}\{\tilde{\tau}_\varepsilon = s'\}\right] \\
&= \mathbb{E}[W_0] + \mathbb{E}\left[\sum_{s'=0}^{t}\sum_{s=0}^{t-1}(W_{s+1} - W_s)\mathbb{I}\{\tilde{\tau}_\varepsilon = s', s' > s\}\right] \\
&= \mathbb{E}[W_0] + \mathbb{E}\left[\sum_{s=0}^{t-1}(W_{s+1} - W_s)\mathbb{I}\{\tilde{\tau}_\varepsilon > s\}\right] \\
&= \mathbb{E}[W_0] + \mathbb{E}\left[\sum_{s=0}^{t-1}\mathbb{E}[W_{s+1} - W_s|\mathcal{F}_s](1 - \mathbb{I}\{\tilde{\tau}_\varepsilon \leqslant s\})\right] \\
&\leqslant \mathbb{E}[W_0]\,.
\end{aligned}$$

The last equality holds by $\{\tilde{\tau}_\varepsilon \leqslant s\} \subset \mathcal{F}_s$, and the last inequality by the super-martingale property. □

In particular, if $\mathbb{E}[W_0] \leqslant 1$, then for all $T \in \mathbb{N}$, $\mathbb{P}\left(\max_{0\leqslant t\leqslant T} W_t \geqslant \varepsilon\right) \leqslant \varepsilon^{-1}$.

As an example of fruitful application, we now provide the following uniform concentration inequality.

Lemma 2.3 (*Asymptotic Maximal Hoeffding inequality*) *Assume that X_i has positive mean μ and that $X_i - \mu$ is σ-sub-Gaussian. Then*

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\frac{\max_{s\leqslant n}\sum_{i=1}^{s} X_i}{n} \leqslant (1+\varepsilon)\mu\right) = 1.$$

One may wonder why to derive an *asymptotic* result instead of a result holding for all time steps. This result is useful in establishing regret lower-bounds for multi-armed bandits, see Chapter 3.

Proof of Lemma 2.3:

For any $x \in \mathbb{R}$ and $\lambda > 0$, it holds

$$\begin{aligned}
\mathbb{P}\Big(\max_{s\leqslant n} \sum_{i=1}^{s} Z_i \geqslant x \Big) &= \mathbb{P}\Big(\max_{s\leqslant n} \exp(\lambda \sum_{i=1}^{s} Z_i) \geqslant \exp(\lambda x) \Big) \\
&= \mathbb{P}\Big(\max_{s\leqslant n} \exp(\lambda \sum_{i=1}^{s} Z_i - \frac{\lambda^2\sigma^2}{2} n) \geqslant \exp(\lambda x - \frac{\lambda^2\sigma^2}{2} n) \Big) \\
&\leqslant \mathbb{P}\Big(\max_{s\leqslant n} \exp\Big(\lambda \sum_{i=1}^{s} Z_i - \frac{\lambda^2\sigma^2}{2} s\Big) \geqslant \exp(\lambda x - \frac{\lambda^2\sigma^2}{2} n) \Big),
\end{aligned}$$

where we introduced the quantity

$$W_s = \exp\Big(\lambda \sum_{i=1}^{s} Z_i - \frac{\lambda^2\sigma^2}{2} s\Big)$$

Note that W_s is non-negative super-martingale since $\mathbb{E}[W_{s+1}\mathcal{F}_s] \leqslant W_s \mathbb{E}[\lambda Z_{s+1} - \frac{\lambda^2\sigma^2}{2}] \leqslant W_s$. Thus, we can apply Doob's maximal inequality:

$$\mathbb{P}\Big(\max_{0\leqslant t\leqslant n} W_t \geqslant \varepsilon \Big) \leqslant \frac{\mathbb{E}[W_0]}{\varepsilon}.$$

In particular, it holds

$$\mathbb{P}\Big(\max_{s\leqslant n} \sum_{i=1}^{s} Z_i \geqslant x \Big) \leqslant \underbrace{\mathbb{E}[W_0]}_{1} \exp(-\lambda x + \frac{\lambda^2\sigma^2}{2} n) \leqslant \exp(-\frac{x^2}{2n\sigma^2})$$

where we optimize in $\lambda = x/n\sigma^2$.

We deduce that if $\mu > 0$, then on an en event of probability higher than $1-\delta$,

$$\begin{aligned}
\frac{\max_{s\leqslant n} \sum_{i=1}^{s} X_i}{n} &= \frac{\max_{s\leqslant n} \sum_{i=1}^{s} (X_i - \mu) + \mu s}{n} \\
&\leqslant \frac{\max_{s\leqslant n} \sum_{i=1}^{s} (X_i - \mu) + \mu n}{n} \\
&\leqslant \sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \mu
\end{aligned}$$

That is for any δ,

$$\mathbb{P}\left(\frac{\max_{s\leqslant n} \sum_{i=1}^{s} X_i}{n} \leqslant \sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \mu \right) \geqslant 1 - \delta$$

In particular, choosing δ_n such that $\delta_n \to 0$ and $\frac{\log(1/\delta)}{n} \to 0$ (for instance $\delta_n = 1/n$), then we deduce that $1 - \delta_n \to 1$, and $\sigma\sqrt{\frac{2\log(1/\delta)}{n}} + \mu \to \mu$. Since $\mu < (1+\varepsilon)\mu$ for any $\varepsilon > 0$, we get

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\frac{\max_{s\leqslant n} \sum_{i=1}^{s} X_i}{n} \leqslant (1+\varepsilon)\mu \right) = 1.$$

□

1 THE PEELING TECHNIQUE FOR RANDOM STOPPING TIMES

In this section, we provide a powerful result that is useful when dealing with generic real-valued distributions. We say a process generating a sequence of random variables $\{Z_i\}_{i=1}^{\infty}$ is predictable if there exists a filtration $\mathcal{H} = (\mathcal{H}_n)_{n\in\mathbb{N}}$ ("filtration of the past") such that Z_n is $\mathcal{H}_n$-measurable for all n. We say a random variable N is a random stopping time for $\mathcal{H}$ if $\forall m \in \mathbb{N}, \{N \leqslant m\}$ is $\mathcal{H}_{m-1}$-measurable.

Lemma 2.4 (*Concentration inequality for predictable processes*) *Let $Z = \{Z_i\}_{i=1}^{\infty}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F}^Z$ be its natural filtration. Let $\varphi : \mathbb{R} \to \mathbb{R}^+$ be a convex upper-envelope of the cumulant generative function of the conditional distributions with $\varphi(0) = 0$, and $\varphi_\star$ its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathcal{D}, \forall i, \qquad & \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathcal{H}_{i-1}\Big] \leqslant \varphi(\lambda)\,, \\ \forall x \in \mathbb{R} \qquad & \varphi_\star(x) = \sup_{\lambda\in\mathbb{R}} \big(\lambda x - \varphi(\lambda)\big), \end{aligned}$$

where $\mathcal{D} = \{\lambda \in \mathbb{R} : \forall i, \ln \mathbb{E}\Big[\exp\Big(\lambda Z_i\Big) \Big| \mathcal{H}_{i-1}\Big] < \infty\}$. Assume that $\mathcal{D}$ contains an open neighborhood of 0. Then, $\forall c \in \mathbb{R}^+$, there exists a unique x_c such that for all i, $x_c > \mathbb{E}\Big[Z_i \Big| \mathcal{H}_{i-1}\Big]$, and $\varphi_\star(x_c) = c$, and a unique x'_c such that for all i, $x'_c < \mathbb{E}\Big[Z_i \Big| \mathcal{H}_{i-1}\Big]$ and $\varphi_\star(x'_c) = c$. We define $\varphi^{-1}_{\star,+} : c \mapsto x_c$, $\varphi^{-1}_{\star,-} : c \mapsto x'_c$. Then $\varphi^{-1}_{\star,+}$ is not decreasing and $\varphi^{-1}_{\star,-}$ is not increasing.
Let N_n be a $\mathbb{N}$-valued random variable that is $\mathcal{F}^Z$-measurable and a.s. bounded by n. Then for all $\alpha \in (1, n]$, and $\delta \in (0, 1)$,

$$\begin{aligned} \mathbb{P}\left[\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)\right)\right] &\leqslant \delta \\ \mathbb{P}\left[\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)\right)\right] &\leqslant \delta \end{aligned}$$

In particular, one can take α to be the minimal solution to $\ln(\alpha)e^{1/\ln(\alpha)} = \ln(n)/\delta$.
Now, if N is a (possibly unbounded) $\mathbb{N}$-valued random variable that is $\mathcal{F}^Z$-measurable, it holds for all deterministic $\alpha > 1$ and $\delta \in (0, 1)$,

$$\begin{aligned} \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varphi^{-1}_{\star,+}\left(\frac{\alpha}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\right]\right)\right] &\leqslant \delta \\ \mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i \leqslant \varphi^{-1}_{\star,-}\left(\frac{\alpha}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\right]\right)\right] &\leqslant \delta \end{aligned}$$

Remark 2.1 *In the case of i.i.d. random variables following a Gaussian law $\mathcal{N}(0,\sigma^2)$, then it holds*

$$\varphi(\lambda)=\tfrac{\lambda^2\sigma^2}{2},\quad \varphi^\star(x)=\tfrac{x^2}{2\sigma^2},\quad \varphi_{\star,+}^{-1}(y)=\sigma\sqrt{2y},\quad \varphi_{\star,-}^{-1}(y)=-\sigma\sqrt{2y}\,.$$

Proof of Lemma 2.4:

First, one easily derives the following properties, from properties of the Legendre-Fenchel transform.

- $\varphi_\star(0)=0$, $\varphi_\star(x)\overset{x\to+\infty}{\to}\infty$, $\varphi_\star$ is convex, increasing on $\mathbb{R}^+$.
- $\forall x$ such that $\varphi_\star(x)<\infty$, there exists a unique $\lambda_x\in\mathcal{D}_\nu$ such that $\varphi_\star(x)=\lambda_x x-\varphi(\lambda_x)$.
- $\forall c\in\mathbb{R}^+$, there exists a unique $x_c>\mathbb{E}[Z]$ such that $\varphi_\star(x_c)=c$. We write it $\varphi_{\star,+}^{-1}(c)$. $\varphi_{\star,+}^{-1}$ is not decreasing.

---

1. A peeling argument We start with a peeling argument. Let us choose some $\eta>0$ and define $t_k=(1+\eta)^k$, for $k=0,\dots,K$, with $K=\lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil$ (thus $n\leqslant t_K$).

Let $\varepsilon_t\in\mathbb{R}^+$ be a sequence that is non-increasing in t.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n}Z_i\geqslant\varepsilon_{N_n}\Big)\\
&\quad\leqslant\ \mathbb{P}\Big(\bigcup_{k=1}^{K}\{t_{k-1}<N_n\leqslant t_k\}\cap\{\sum_{i=1}^{N_n}Z_i\geqslant N_n\varepsilon_{N_n}\}\Big)\\
&\quad\leqslant\ \sum_{k=1}^{K}\mathbb{P}\Big(\exists t\in(t_{k-1},t_k]:\sum_{i=1}^{t}Z_i\geqslant t\varepsilon_t\Big)
\end{aligned}$$

Let $\lambda_k > 0$, for $k = 1, \dots, K$.

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] : \sum_{i=1}^{t} Z_i \geqslant t\varepsilon_t\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] : \exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big)\Big) \geqslant \exp(\lambda_k t\varepsilon_t)\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] : \underbrace{\exp\Big(\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big) - t\varphi(\lambda_k)\Big)}_{W_{k,t}} \\
&\qquad\qquad \geqslant \exp\Big(t\big(\lambda_k\varepsilon_t - \varphi(\lambda_k)\big)\Big)\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] : W_{k,t} \geqslant \exp\Big(t\big(\lambda_k\varepsilon_{t_k} - \varphi(\lambda_k)\big)\Big)\Big).
\end{aligned}$$

Since $\varepsilon_{t_k} > 0$, we can choose a $\lambda_k > 0$ such that $\varphi^\star(\varepsilon_{t_k}) = \lambda_k\varepsilon_{t_k} - \varphi(\lambda_k)$.

---

2. Doob's maximal inequality At this, point, we show that the sequence $\{W_{k,t}\}_t$ is a non-negative super-martingale, where $W_{k,t} = \exp\Big(\lambda_k\big(\sum_{i=1}^{t} Z_i\big) - t\varphi(\lambda_k)\Big)$. Indeed, note that:

$$\begin{aligned}
\mathbb{E}[W_{k,t+1}|\mathcal{F}_t] &= W_{k,t}\mathbb{E}[\exp\big(\lambda_k Z_{t+1}\big)|\mathcal{F}_t]\exp(-\varphi(\lambda_k)) \\
&\leqslant W_{k,t}\,.
\end{aligned}$$

Thus, using that $t_{k-1} \geqslant t_k/(1+\eta)$, we find

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big) \\
&\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big) \\
&\overset{(a)}{\leqslant} \sum_{k=1}^{K} \exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big),
\end{aligned}$$

where (a) holds by application of Doob's maximal inequality for non-negative super-martingales, using that $\max_{t\in(t_{k-1},t_k]} W_{k,t} \leqslant \max_{t\in(0,t_k]} W_{k,t}$ and $W_{k,0} \leqslant 1$.

---

3. Parameter tuning for bounded N_n Now, let us choose ε_t such that $t\varphi_\star(\varepsilon_t) = c > 1$ is a constant, that is $\varepsilon_t = \varphi_{\star,+}^{-1}(c/t)$ (non increasing with t). Thus, we get for all $\eta \in (0, n-1)$:

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varepsilon_{N_n}\Big) &\leqslant \sum_{k=1}^{\lceil \frac{\ln(n)}{\ln(1+\eta)}\rceil} \exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\\ &\leqslant \lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil \exp\Big(-\frac{c}{1+\eta}\Big)\,,\end{aligned}$$

which suggest to set $c = (1+\eta)\ln\Big(\lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil\frac{1}{\delta}\Big)$. We thus obtain for all $\eta \in [0, n-1]$,

$$\mathbb{P}\Big(\frac{1}{N_n}\sum_{i=1}^{N_n} Z_i \geqslant \varphi_{\star,+}^{-1}\Big[\frac{1+\eta}{N_n}\ln\Big(\Big\lceil\frac{\ln(n)}{\ln(1+\eta)}\Big\rceil\frac{1}{\delta}\Big)\Big]\Big) \leqslant \delta\,.$$

Then, it makes sense to find the minimum value of $f : x \to x\ln(\frac{a}{\ln(x)})$, for $x > 1$. An optimal point $x_\star > 1$ satisfies

$$f'(x) = \ln\Big(\frac{a}{\ln(x)}\Big) + x\frac{-(1/x)/\ln^2(x)}{1/\ln(x)} = \ln\Big(\frac{a}{\ln(x)}\Big) - \frac{1}{\ln(x)} = 0\,,$$

that is $x_\star$ satisfies $a = \ln(x_\star)e^{1/\ln(x_\star)}$. We may thus choose the (slightly suboptimal) minimal value x that satisfies $\ln(x)e^{1/\ln(x)} = \ln(n)/\delta$.

---

4. Parameter tuning for unbounded N
We restart from

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varepsilon_N\Big) \leqslant \sum_{k=1}^{K}\exp\Big(-\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\,,$$

where $t_k = (1+\eta)^k$ and $K = \lceil\frac{\ln(n)}{\ln(1+\eta)}\rceil$, and choose a different tuning for ε_t in order to handle an infinite sum (with $K = \infty$). Let us choose ε_t that satisfies $t\varphi_\star(\varepsilon_t) = c(t)$, where $c(t)$ is chosen such that

$$\sum_{k=1}^{\infty}\exp\Big(-\frac{c(t_k)}{1+\eta}\Big) < \infty\,.$$

Choosing $c(t) = (1+\eta)\ln\Big(\frac{\ln(t)}{\delta\ln(1+\eta)}[\frac{\ln(t)}{\ln(1+\eta)}+1]\Big)$, it comes $c(t_k) = (1+\eta)\ln(k(k+1)\delta)$ and thus

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varepsilon_N\Big) \leqslant \sum_{k=1}^{\infty}\frac{\delta}{k(k+1)} = \delta\,,$$

With this choice, we thus deduce

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geqslant \varphi_{\star,+}^{-1}\Big(\frac{(1+\eta)}{N}\ln\Big(\frac{\ln(N)\ln(N(1+\eta))}{\delta\ln^2(1+\eta)}\Big)\Big) \leqslant \delta\,.$$

5. Reverse bounds. We now provide a similar result for the reverse bound. Let $\varepsilon_t \in \mathbb{R}$ be a sequence that is non-decreasing with t, and $\lambda_k > 0$, for $k = 1, \dots, K$. Then

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \leqslant \varepsilon_N\Big) &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k] \exp\Big(-\lambda_k\Big(\sum_{i=1}^{t} Z_i\Big) - t\varphi(-\lambda_k)\Big) \\ &\qquad \geqslant \exp\Big(t(-\lambda_k\varepsilon_t - \varphi(-\lambda_k))\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t(-\lambda_k\varepsilon_{t_k} - \varphi(-\lambda_k))\Big)\Big)\end{aligned}$$

If $\varepsilon_{t_k} < \mathbb{E}[Z_{t_k}]$, we can choose $\lambda_k = \lambda_{\varepsilon_{t_k}} > 0$ such that $\varphi^\star(\varepsilon_{t_k}) = -\lambda_k\varepsilon_{t_k} - \varphi(-\lambda_k) \geqslant 0$. Thus, using that $t_{k-1} > t_k/(1+\eta)$, it comes

$$\begin{aligned}\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \leqslant \varepsilon_N\Big) &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\exists t \in (t_{k-1}, t_k], W_{k,t} \geqslant \exp\Big(t\varphi^\star(\varepsilon_{t_k})\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \mathbb{P}\Big(\max_{t\in(t_{k-1},t_k]} W_{k,t} \geqslant \exp\Big(\frac{t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\Big) \\ &\leqslant \sum_{k=1}^{K} \exp\Big(\frac{-t_k\varphi^\star(\varepsilon_{t_k})}{1+\eta}\Big)\end{aligned}$$

Now, let us choose $\varepsilon_t < \mathbb{E}[Z_t]$ such that $t\varphi_\star(\varepsilon_t) = c > 1$, that is $\varepsilon_t = \varphi_{\star,-}^{-1}(c/t)$ (non decreasing with t). For $\eta = 1/(c-1)$ and $c = \ln(e/\delta)$, we obtain

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \leqslant \varphi_{\star,-}^{-1}(\ln(e/\delta)/N)\Big) \leqslant \lceil \ln(n)\ln(e/\delta)\rceil\delta\,.$$

□

Application to the control of variance We conclude this section by applying Lemma 2.4 to the concentration of the quadratic sum of a noise term ξ_i. We believe that this illustrates the power of this method. Assume that the noise terms are second-order sub-Gaussian in the sense that

$$\forall\lambda \in \mathcal{D}_\nu, \forall i \quad \log\mathbb{E}[\exp(\lambda\xi_i^2)|\mathcal{H}_{i-1}] \leqslant \varphi(\lambda)$$

where $\varphi(\lambda) = -\frac{1}{2}\log(1-2\lambda R^2)$. Note that this is the cumulant generative function of the square of a centered Gaussian. Then we can prove the following result:

Lemma 2.5 (*Birge-Massart concentration for predictable process*) *Assume that N_n is an $\mathcal{F}^Z$-measurable $\mathbb{N}$-valued random variable that satisfies $N_n \leqslant n$ almost surely, then it holds for all $\alpha > 1$*

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)} + \frac{2\alpha R^2}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)\right] \leqslant \delta$$

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{i=1}^{N_n}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\alpha}{N_n}\ln\left(\left\lceil\frac{\ln(n)}{\ln(\alpha)}\right\rceil\frac{1}{\delta}\right)}\right] \leqslant \delta$$

Further, for an $\mathcal{F}^Z$-measurable $\mathbb{N}$-valued random variable N, then it holds for all $\alpha > 1$,

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \geqslant R^2 + 2R^2\sqrt{\frac{2\alpha}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\right]} + \frac{2\alpha R^2}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\right]\right] \leqslant \delta$$

$$\mathbb{P}\left[\frac{1}{N}\sum_{i=1}^{N}\xi_i^2 \leqslant R^2 - 2R^2\sqrt{\frac{\alpha}{N}\ln\left[\frac{\ln(N)\ln(\alpha N)}{\delta\ln^2(\alpha)}\right]}\right] \leqslant \delta$$

Proof of Lemma 2.5:

According to Lemma 2.4 applied to $Z_i = \xi_i^2$, all we have to do is to compute an upper bound on the quantity $\varphi_{\star,+}^{-1}(c)$, first for the value $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1 + \frac{2}{\ln(1/\delta)}\ln\left(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\right)) \leqslant \frac{\ln(e/\delta)}{N}(1 + c_N/\ln(1/\delta))$. We proceed in the following way. First, the envelope function is given by

$$\varphi(\lambda) = -\frac{1}{2}\ln(1-2\lambda R^2) \leqslant \frac{\lambda R^2}{1-2\lambda R^2}\,.$$

for $\lambda \in (0, \frac{1}{2R^2})$. Let $x > R^2$. It holds that $\varphi^\star(x) \geqslant \sup_\lambda[\lambda x - \frac{\lambda R^2}{1-2\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = (1 - \frac{R}{\sqrt{x}})\frac{1}{2R^2} \in (0, \frac{1}{2R^2})$, with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= (1-\frac{R}{\sqrt{x}})\frac{x}{2R^2} - (1-\frac{R}{\sqrt{x}})\frac{\sqrt{x}}{2R}\\ &= \frac{x}{2R^2} - \frac{\sqrt{x}}{R} + \frac{1}{2}\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1+\sqrt{2c})^2$. As a result, we deduce that $\varphi_{\star,+}^{-1}(c) \leqslant R^2(1+\sqrt{2c})^2 = R^2 + 2R^2c + 2R^2\sqrt{2c}$.

Now, for the reverse inequality, we have to compute a lower bound on the quantity $\varphi_{\star,-}^{-1}(c)$, first for $c = \frac{\ln(e/\delta)}{N_n}$, then for $c = \frac{\ln(e/\delta)}{N}(1 + \frac{2}{\ln(1/\delta)}\ln\left(\frac{\pi\ln(N)\ln(1/\delta)}{6^{1/2}(1+\ln(1/\delta))}\right)) \leqslant \frac{\ln(e/\delta)}{N}(1 + c_N/\ln(1/\delta))$. We proceed in

the following way. First, the envelope function is given for $\lambda > 0$ by

$$\varphi(-\lambda) = -\frac{1}{2}\ln(1+2\lambda R^2) \geqslant -\frac{\lambda R^2}{1+\lambda R^2}\,.$$

Thus, for $0 < x < R^2$ it holds $\varphi^\star(x) \geqslant \sup_{\lambda>0}[-\lambda x + \frac{\lambda R^2}{1+\lambda R^2}] = 1 + \sup_{\lambda>0}[-\lambda x - \frac{1}{1+\lambda R^2}]$. Solving this optimization by differentiating over λ, the supremum is reached for $\lambda = \frac{1}{R^2}(\frac{R}{\sqrt{x}} - 1) > 0$ with corresponding value given by

$$\begin{aligned}\tilde{\varphi}^\star(x) &= 1 - \frac{x}{R^2}\big(\frac{R}{\sqrt{x}} - 1\big) - \frac{\sqrt{x}}{R} \\ &= \frac{x}{R^2} - 2R\frac{\sqrt{x}}{R} + 1\,.\end{aligned}$$

Now, for $c > 0$, it is easily checked that $\tilde{\varphi}^\star(x) = c$ holds for $x_c = R^2(1-\sqrt{c})^2$, and $x_c < R^2$ if $c < 1$. As a result, we deduce that if $c \in (0,1)$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c} + R^2 c$. On the other hand, for all $c > 0$, choosing $\lambda = \frac{1}{R^2}\sqrt{c}$, and using the inequality $\frac{1}{1+v} \geqslant 1 - v$ for $v > 0$, then

$$\begin{aligned}\varphi^\star(x) &\geqslant -\frac{x}{R^2}\sqrt{c} + 1 - \frac{1}{1+\sqrt{c}} = \sqrt{c}\big(-\frac{x}{R^2} + \frac{1}{1+\sqrt{c}}\big) \\ &\geqslant \tilde{\varphi}^\star(x) \stackrel{\text{def}}{=} \sqrt{c}\big(-\frac{x}{R^2} + 1 - \sqrt{c}\big)\end{aligned}$$

Thus, $\tilde{\varphi}^\star(x) = c$ for $x_c = R^2 - 2R^2\sqrt{c} < R^2$. As a result, we deduce that if $c > 0$, then $\varphi^{-1}_{\star,-}(c) \geqslant R^2 - 2R^2\sqrt{c}$. □

2 UNIFORM BOUNDS AND THE LAPLACE METHOD

In this section, we present another very powerful tool, that is the Laplace method (method of mixtures for sub-Gaussian random variables, see Peña et al. (2008)). Like the peeling method, the Laplace method belongs to the set of methods making use of a martingale construction. Unlike the peeling method that makes use of interval with geometrically increasing size, regardless of the distribution, the Laplace method considers infinitesimal intervals that closely follow the shape of the considered distribution. For this reason, it is more powerful, but applies to a more restricted set of situations. We provide the illustrative following result here, for real-valued random variables. The result however extends naturally to dimension d, and even, to some extent, to infinite dimension, as we explain in chapter 4.

Lemma 2.6 (*Time-uniform concentration inequalities*) *Let $Y_1, \dots Y_t$ be a sequence of t i.i.d. real-valued random variables bounded in $[0,1]$, with mean μ. Let $\mu_t = \frac{1}{t}\sum_{s=1}^t Y_s$ be the empirical mean estimate. Then, for all $\delta \in (0,1)$, it holds*

$$\begin{aligned}\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu_t - \mu \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{t+1}/\delta\big)}{2t}}\right) &\leqslant \delta \\ \mathbb{P}\left(\exists t \in \mathbb{N}, \quad \mu - \mu_t \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{t+1}/\delta\big)}{2t}}\right) &\leqslant \delta\,.\end{aligned}$$

Proof of Lemma 2.6:

In order to prove this result, we introduce for a fixed $\delta \in [0,1]$ the random variable

$$\tau = \min\left\{ t \in \mathbb{N} : \mu_t - \mu \geqslant \sqrt{\Big(1+\frac{1}{t}\Big)\frac{\ln\big(\sqrt{1+t}/\delta\big)}{2t}} \right\}.$$

This quantity is a random stopping time for the filtration $\mathcal{F} = (\mathcal{F}_t)_t$, where $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$, since $\{\tau \leqslant m\}$ is $\mathcal{F}_m$-measurable for all m. We want to show that $\mathbb{P}(\tau < \infty) \leqslant \delta$. To this end, for any λ, and t, we introduce the following quantity

$$M_t^\lambda = \exp\bigg(\sum_{s=1}^{t} (\lambda(Y_s - \mu) - \frac{\lambda^2}{8}) \bigg).$$

By the i.i.d. bounded assumption, the random variables are $1/2$-sub-Gaussian and it is immediate to show that $\{M_t^\lambda\}_{t\in\mathbb{N}}$ is a non-negative super-martingale that satisfies $\ln \mathbb{E}[M_t^\lambda] \leqslant 0$ for all t. It then follows that $M_\infty^\lambda = \lim_{t\to\infty} M_t^\lambda$ is almost surely well-defined (this is a consequence of Doob's upcrossing lemma for super-martingales). Hence, M_τ^λ is well-defined as well. Further, let us introduce the stopped version $Q_t^\lambda = M_{\min\{\tau,t\}}^\lambda$. An application of Fatou's lemma shows that $\mathbb{E}[M_\tau^\lambda] = \mathbb{E}[\liminf_{t\to\infty} Q_t^\lambda] \leqslant \liminf_{t\to\infty} \mathbb{E}[Q_t^\lambda] \leqslant 1$. Thus, $\mathbb{E}[M_\tau^\lambda] \leqslant 1$.

The next step is to introduce the auxiliary variable $\Lambda = \mathcal{N}(0,4)$, independent of all other variables, and study the quantity $M_t = \mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty]$. Note that the standard deviation of Λ is $(1/2)^{-1}$ due to the fact we consider $1/2$-sub-Gaussian random variables. We immediately get $\mathbb{E}[M_\tau] = \mathbb{E}[M_\tau^\Lambda | \Lambda]] \leqslant 1$. For convenience, let $S_t = t(\mu_t - \mu)$. By construction of M_t, we have

$$\begin{aligned}
M_t &= \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\bigg(\lambda S_t - \frac{\lambda^2 t}{8} - \frac{\lambda^2}{8} \bigg) d\lambda \\
&= \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\bigg(-\Big[\lambda\sqrt{\frac{t+1}{8}} - \frac{\sqrt{2}S_t}{\sqrt{t+1}}\Big]^2 + \frac{2S_t^2}{t+1} \bigg) d\lambda \\
&= \exp\bigg(\frac{2S_t^2}{t+1}\bigg) \frac{1}{\sqrt{8\pi}} \int_{\mathbb{R}} \exp\Big(-\lambda^2 \frac{t+1}{8} \Big) d\lambda \\
&= \exp\bigg(\frac{2S_t^2}{t+1}\bigg) \frac{\sqrt{8\pi/(t+1)}}{\sqrt{8\pi}}.
\end{aligned}$$

Thus, we deduce that

$$S_t = \sqrt{\frac{t+1}{2} \ln\Big(\sqrt{t+1} M_t\Big)}.$$

We conclude by applying a simple Markov inequality:

$$\mathbb{P}\bigg(\tau(\mu_\tau - \mu) \geqslant \sqrt{\frac{\tau+1}{2}\ln\Big(\sqrt{\tau+1}/\delta\Big)}\bigg) = \mathbb{P}(M_\tau \geqslant 1/\delta) \leqslant \mathbb{E}[M_\tau]\delta\,.$$

□

Proceeding with similar steps, more generally we obtain the following result for sums of sub-Gaussian random variables.

Lemma 2.7 (*Time-uniform sub-Gaussian concentration*) *Let $Y_1, \dots Y_t$ be a sequence of t independent real-valued random variables where for each $s \leqslant t$, Y_s has mean μ_s and is σ_s-sub-Gaussian. Then for all $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \sum_{s=1}^{t}(Y_s - \mu_s) \geqslant \sqrt{2\sum_{s=1}^{t}\sigma_s^2\left(1+\frac{1}{t}\right)\ln\left(\sqrt{t+1}/\delta\right)}\right) \leqslant \delta$$

$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \sum_{s=1}^{t}(\mu_s - Y_s) \geqslant \sqrt{2\sum_{s=1}^{t}\sigma_s^2\left(1+\frac{1}{t}\right)\ln\left(\sqrt{t+1}/\delta\right)}\right) \leqslant \delta\,.$$

An immediate application of this result is to derive a version of the Upper confidence Bound (UCB) algorithm Auer et al. (2002) for multi-armed bandits, in the context of bounded (or sub-Gaussian) distributions. The first appearance of the strategy involving the Laplace method is due to Abbasi-Yadkori et al. (2011), although the bound that is reported in the pseudo-code of the algorithm contains a mistake (not present in their analysis). The correct version of the strategy, is the following one,

$$\text{(UCB-Laplace)} \quad A_t = \arg\max_{a\in\mathcal{A}} \widehat{\mu}_{a,t} + \sqrt{\frac{\left(1+\frac{1}{N_t(a)}\right)\ln\left(A\sqrt{N_t(a)+1}/\delta\right)}{2N_t(a)}}\,,$$

where $\delta \in (0,1)$ is the input confidence level of the algorithm, and where ties are solved by choosing uniformly amongst the least pulled arms.

Extension to exponential families One may wonder whether the previous method applies more generally to any exponential family, and not only to Gaussian (or sub-Gaussian) distributions. The following result provides such a generalization, stated in a way similar to Lemma 2.4; However this requires some condition and terms that may not be easy to make explicit beyond specific distributions.

Lemma 2.8 (*Laplace concentration for Exponential families*) *For a sample $X_1, \dots, X_n \sim p_\theta$ generated from a parametric distribution in an (F, ψ, ν_0)-exponential family (that is $p_\theta(dx) = \exp(\langle\theta, F(x)\rangle - \psi(\theta))\nu_0(dx)$), let us define the variables $Z_i = F(X_i) - \nabla\psi(\theta)$, for $i = 1..n$.*
Let us introduce, for some constant $c \in \mathbb{R}^+$ and function f, the following function

$$G_{n,\theta}^{c,f} : u \mapsto \frac{\int_{\mathbb{R}} \exp(-(n+c)\mathcal{B}^\psi(\theta+u, \theta+u+\lambda) - f(\lambda))d\lambda}{\int_{\mathbb{R}} \exp(-c\mathcal{B}^\psi(\theta, \theta+\lambda) - f(\lambda))d\lambda}\,.$$

Assume that for some choice of f, c, $\exists C_{n,\theta}^{c,f} > 0$ such that $\inf_u G_{n,\theta}^{c,f}(u) \geqslant \frac{1}{C_{n,\theta}^{c,f}}$. Then, for any random stopping time N, it holds

$$\mathbb{P}_{p_\theta}\left[\varphi_\theta^\star\left(\frac{1}{N+c}\sum_{i=1}^N Z_i\right) \geqslant \frac{\ln(C_{N,\theta}^{c,f}/\delta)}{N+c}\right] \leqslant \delta\,,$$

where we introduced $\varphi_\theta^\star(s) = \max_\lambda \langle\lambda, s\rangle - \mathcal{B}^\psi(\theta, \theta+\lambda)$.

Note that in the Gaussian case (with known variance), $G_{n,\theta}^{c,f}$ is a constant function. Now, in the Bernoulli case where $p_\theta = \mathcal{B}(\mu)$, it can be checked that $\varphi_\theta^\star\left(\frac{1}{N+c}\sum_{i=1}^N Z_i\right) = \mathtt{kl}\left(\frac{N}{N+c}\widehat{\mu}_N + \frac{c}{c+N}\mu, \mu\right) \simeq \frac{1}{(1+c/N)^2}\mathtt{kl}(\widehat{\mu}_N, \mu)$.

Proof :

Let $S_n = \sum_{i=1}^n Z_i$, and let us introduce the following quantity

$$M_n^\lambda = \exp(\langle\lambda, S_n\rangle - n\mathcal{B}^\psi(\theta, \theta+\lambda))$$

Since by construction $\log \mathbb{E}\exp(\langle\lambda, S_n\rangle) = n\mathcal{B}^\psi(\theta, \theta+\lambda)$, this is a martingale such that $\mathbb{E}[M_n^\lambda] = 1$. Further, it is not difficult to show that for any random stopping time N, $\mathbb{E}[M_N^\lambda] \leqslant 1$.

We apply the method of mixture by integrating over λ. To this end, let us define for some $c \in \mathbb{R}$ the function $g(\lambda) = c\mathcal{B}^\psi(\theta, \theta+\lambda) + f(\lambda)$. We also introduce for convenience $s_n = S_n/(n+c)$. Then, it holds

$$M_n = \frac{\int_{\mathbb{R}} M_n^\lambda \exp(-g(\lambda))d\lambda}{\int_{\mathbb{R}} \exp(-g(\lambda))d\lambda} = \frac{\int_{\mathbb{R}} \exp\Big((n+c)(\langle\lambda, s_n\rangle - \mathcal{B}^\psi(\theta, \theta+\lambda)) - f(\lambda)\Big)d\lambda}{\int_{\mathbb{R}} \exp(-g(\lambda))d\lambda}$$

It is thus natural to introduce $\quad u^\star(s_n, \theta) = \arg\max_\lambda \langle\lambda, s_n\rangle - \mathcal{B}^\psi(\theta, \theta+\lambda)$.

Note that by construction $u^\star = u^\star(s_n, \theta)$ satisfies $s_n - \nabla\psi(\theta+u^\star) + \nabla\psi(\theta) = 0$. Hence, we deduce that

$$\begin{aligned}
&\langle\lambda, s_n\rangle - \mathcal{B}^\psi(\theta, \theta+\lambda) - \langle u^\star, s_n\rangle + \mathcal{B}^\psi(\theta, \theta+u^\star) \\
&= \langle\lambda - u^\star, s_n\rangle + \psi(\theta+\lambda) - \langle\lambda, \nabla\psi(\theta)\rangle - \psi(\theta+u^\star) + \langle u^\star, \nabla\psi(\theta)\rangle \\
&= \langle\lambda - u^\star, s_n - \nabla\psi(\theta)\rangle + \psi(\theta+\lambda) - \psi(\theta+u^\star) \\
&= \langle\lambda - u^\star, s_n - \nabla\psi(\theta)\rangle - \mathcal{B}^\psi(\theta+u^\star, \theta+\lambda) + \langle\lambda - u^\star, \nabla\psi(\theta+u^\star)\rangle \\
&= -\mathcal{B}^\psi(\theta+u^\star, \theta+\lambda)\,.
\end{aligned}$$

Thus, plugging-in this equality in the definition of M_n, we get for each $c \in \mathbb{R}$ and f,

$$\begin{aligned} M_n &= \exp\Big((n+c)(\langle u^\star, s_n\rangle - \mathcal{B}^\psi(\theta,\theta+u^\star)\Big) \frac{\int_{\mathbb{R}} \exp(-(n+c)\mathcal{B}^\psi(\theta+u^\star,\theta+\lambda)-f(\lambda))d\lambda}{\int_{\mathbb{R}} \exp(-c\mathcal{B}^\psi(\theta,\theta+\lambda)-f(\lambda))d\lambda}\,. \\ &= \exp\Big((n+c)(\langle u^\star, s_n\rangle - \mathcal{B}^\psi(\theta,\theta+u^\star)\Big) \frac{\int_{\mathbb{R}} \exp(-(n+c)\mathcal{B}^\psi(\theta+u^\star,\theta+u^\star+\tilde{\lambda})-f(\tilde{\lambda}))d\tilde{\lambda}}{\int_{\mathbb{R}} \exp(-c\mathcal{B}^\psi(\theta,\theta+\lambda)-f(\lambda))d\lambda} \\ &= \exp((n+c)\psi_\theta^\star(s_n))G_{n,\theta}^{c,f}(u^\star) \end{aligned}$$

where $u^\star$ satisfies $s_n - \nabla\psi(\theta+u^\star) + \nabla\psi(\theta) = 0$. Now by assumption, there exists some $C_{n,\theta}^{c,f} > 0$ such that

$$M_n \geqslant \tilde{M}_n = \exp((n+c)\psi_\theta^\star(s_n))/C_{n,\theta}^{c,f}\,.$$

Thus, we deduce that

$$\begin{aligned} \mathbb{P}\left(\psi_\theta^\star(s_n) \geqslant \frac{\ln(C_{n,\theta}^{c,f}/\delta)}{n+c}\right) &= \mathbb{P}(\tilde{M}_n \geqslant 1/\delta) \leqslant \mathbb{P}(M_n \geqslant 1/\delta) \\ &\leqslant \delta\mathbb{E}[M_n] \leqslant \delta\,. \end{aligned}$$

By properties of M_n, this holds also for any random stopping time N. □

3 NUMERICAL COMPARISON OF A FEW BOUNDS

In this section, we compare the behavior of the bounds obtained by the Laplace and Peeling method on the illustrative example of sub-Gaussian random variables. Let $\mu_t = \frac{1}{t}\sum_{t'=1}^t Y_{t'}$ denote the empirical mean with t observations. We first recall the following uniform confidence bound that is obtained by an application of the Laplace method (method of mixtures for sub-Gaussian variables) in the i.i.d σ^2-sub-Gaussian case.

$$\text{(Laplace method)} \quad \mathbb{P}\left(\exists t\in\mathbb{N},\ \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma\sqrt{\frac{2(1+\frac{1}{t})\ln\left(\sqrt{t+1}/\delta\right)}{t}}\right) \leqslant \delta\,,$$

Note that this holds simultaneously over all t. For comparison, the peeling method yields

$$\text{(Peeling method)} \qquad \mathbb{P}\left(\exists t\in\mathbb{N},\quad \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma\sqrt{\frac{2(1+\eta)}{t}\ln\left(\frac{\ln(t)\ln(t(1+\eta))}{\delta\ln^2(1+\eta)}\right)}\right) \leqslant \delta\,,$$

where $\eta > 0$ is any fixed constant not depending on t. For reference, a simple union bound gives

$$\text{(Union bound)} \qquad \mathbb{P}\left(\exists t\in\mathbb{N},\quad \mu_t - \mathbb{E}[\mu_t] \geqslant \sigma\sqrt{\frac{2\ln\left(t(t+1)/\delta\right)}{t}}\right) \leqslant \delta\,,$$

The bound obtain by a union bound is very crude, and even the *a priori appealing* $\ln\ln(t)$ scaling of the bound obtained by the peeling method is however not better than the one derived by the Laplace method, unless for huge times t ($t \geqslant 10^6$, for $\delta = 0.05$ and any η, see also Figure 3). This should not be surprising, since neither methods make use of the fact that the variables are sub-Gaussian, contrary to the Laplace method. This is illustrated in Figure 2.1, where we choose various values for η (even values $\eta = \eta(t)$ depending on t, for which the peeling bound is no longer valid a priori). We hope this illustrates how powerful the Laplace method can be, and why it should be used instead of more naive approaches.

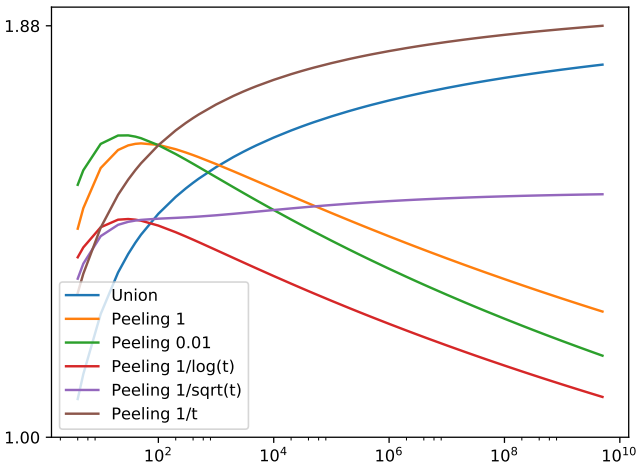


Figure 2.1: Ratio of different time-uniform concentration bounds over that of the Laplace method, as a function of t, for a confidence level $\delta = 0.01$ and various choice of $\eta = \eta(t)$. This indicates that all other bounds are larger than the Laplace bound by a multiplicative factor up to 1.88 here, and none is smaller until at least time $t = 10^{10}$. Notice the logarithmic scale.

CHAPTER 3
$\log\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)$

Contents

1 Change of measure and lower bounds 39

2 Further lower-bounds and extensions 44

3 From lower bounds to sampling strategies 49

4 Change point detection 53

4.1 Doubly time uniform concentration of scan-statistics 54

4.2 Non-asymptotic detection delay of sub-Gaussian GLR 59

Take-home message

Fundamental lemma

(Change of measure) $\forall \Omega, \forall c \in \mathbb{R},\ \mathbb{P}_\nu\left(\Omega \cap \left\{\log\left(\frac{d\nu}{d\tilde{\nu}}(X)\right) \leqslant c\right\}\right) \leqslant \exp(c)\mathbb{P}_{\tilde{\nu}}(\Omega).$

(Fundamental lemma) $\mathbb{E}_\nu\left[\log\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)\right] \geqslant \sup_{g:\mathcal{X}\to[0,1]} \mathtt{kl}\left(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\tilde{\nu}}[g(X)]\right).$

Usage: derive **performance lower bounds** for sequential sampling strategies with ν the distribution of the observations, $\tilde{\nu}$ another distribution, and g specified by a "uniformly good" property requirement.
Example in **multi-armed bandits**: Let $\mathcal{D} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_A$, where $\mathcal{D}_a \subset \mathcal{P}(\mathcal{X})$ for each $a \in \mathcal{A}$ be any (unstructured) set of configurations, let $\nu \in \mathcal{D}$. Then any uniformly-good strategy must pull arms such that

$$\forall a \in \mathcal{A}, \mu_a(\nu) < \mu_\star(\nu) \quad \Longrightarrow \quad \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\log(T)} \geqslant \frac{1}{\inf\{\mathtt{KL}(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D}_a, \mu_a(\nu') > \mu_\star(\nu)\}}.$$

Usage: Lower bound inspired **sampling strategies**, e.g. `KL-ucb`.
Key step: Identify the set of **maximally confusing distributions**

$$\tilde{\mathcal{D}}(\nu) = \left\{\nu' \in \mathcal{D} : \mathcal{A}^\star(\nu') \cap \mathcal{A}^\star(\nu) = \emptyset, \forall a \in \mathcal{A}^\star(\nu), \mathtt{KL}(\nu_a, \nu'_a) = 0\right\}.$$

Uniformly-good price principle

"Uniform guarantee (over a set $\mathcal{D}$) comes with a sampling price."

Optimistic principle (revisited)

"Pull the arm that enables to rule-out the seemingly best environment from the plausible ones."

Change point detection

(GLR stopping rule) $\min\left\{t \in [1,n] : \max_{s\in[0,t)} G^{\mathcal{E}}_{1:s:t} \geqslant c\right\}$ where

$$G^{\mathcal{E}}_{t_0:s:t} = \sup_{\theta_1,\theta_2} \sum_{t'=t_0}^{s} \log p_{\theta_1}(Y_{t'}) + \sum_{t'=s+1}^{t} \log p_{\theta_2}(Y_{t'}) - \sup_{\theta} \sum_{t'=t_0}^{t} \log p_\theta(Y_{t'})$$

Tuning of threshold c: by concentration of measure.

In this chapter we focus on the log likelihood ratio of two measures, and the key properties that result from its control. This object is at the heart of the change of measure argument, which is a fundamental tool in order to provide lower bound guarantees on a learning problem, and at least in a few cases, give hints at how to design an optimal decision strategy.

We illustrate the strength of this tool below, by showing two proofs strategy for the lower bound in multi-armed bandits. The first proof follows the original proof strategy from Lai and Robbins (1985b), and sees the log likelihood ratio as a random variable that must be controlled by concentration of measure. The second proof technique looks directly at the expectation of this quantity, thus making appear the Kullback-Leibler divergence. Both paths are interesting: the first one can naturally yield a finite-time regret lower bound, while the second one nicely extends to setups with a specific structure.

1 Change of measure and lower bounds

In this section, we present what is called the change of measure argument together with some powerful results, and apply them for illustration to the multi-armed bandit setup.

Change of measure In its most basic form, the change of measure argument simply relates the expectation of a function under a distribution ν to its expectation under another one $\tilde{\nu}$

Lemma 3.1 (*Change of measure*) *For each measurable f with respect to ν and $\tilde{\nu}$, it holds*

$$\mathbb{E}_\nu\big[f(X)\big] = \mathbb{E}_{\tilde{\nu}}\left[\frac{d\nu}{d\tilde{\nu}}(X) f(X)\right].$$

In particular, for every measurable event Ω with respect to ν and $\tilde{\nu}$,

$$\forall c \in \mathbb{R},\ \mathbb{P}_\nu(\Omega \cap \mathcal{C}_c) \leqslant \exp(c)\mathbb{P}_{\tilde{\nu}}(\Omega) \text{ where } \mathcal{C}_c = \left\{\log\left(\frac{d\nu}{d\tilde{\nu}}(X)\right) \leqslant c\right\}$$

Perhaps one of the most direct application of the change of measure is to consider two non-foreign distributions $\nu, \tilde{\nu} \in \mathcal{P}(\mathcal{X})$ on a discrete set $\mathcal{X}$. Then we have that

$$\mathbb{E}_\nu\left[\frac{d\tilde{\nu}}{d\nu}(X)\right] = \mathbb{E}_{\tilde{\nu}}[1] = 1, \quad \text{that is} \quad \log \mathbb{E}_\nu \exp\left(\log(\tilde{\nu}(\{X\})) - \log(\nu(\{X\}))\right) = 0\,.$$

In particular, the log-Laplace of the log-likelihood ratio at value 1 is less than 0, so that we deduce by Markov's inequality that for all $\delta \in [0,1]$ then

$$\mathbb{P}_\nu\Big[-\log(\tilde{\nu}(\{X\})) \leqslant -\log(\nu(\{X\})) - \log(1/\delta)\Big] \leqslant \delta\,,$$

which is precisely the core inequality of compression theory (with $K = \log(1/\delta)$ bits).

Lemma 3.1 can be applied to two finite collection of independent measures (bandit configurations) $(\nu_a)_{a\in\mathcal{A}}$, $(\tilde{\nu}_a)_{a\in\mathcal{A}}$ on $\mathcal{X}$. Indeed let us consider for instance a deterministic sequence $(a_i)_{i\leqslant n}$ of index in $\mathcal{A}$, and corresponding random variable $X = (X_i)_{i\leqslant n}$, where $X_i \sim \nu_{a_i}$. By forming the product measures $\nu = \otimes_{i=1}^n \nu_{a_i}$ and

$\tilde{\nu} = \otimes_{i=1}^{n} \tilde{\nu}_{a_i}$ on $\mathcal{X}^n$, we obtain that the event $\mathcal{C}_c$ bounds the log-Likelihood ratio of the observations as follows.

$$\mathcal{C}_c = \left\{ \sum_{i=1}^{n} \log\left(\frac{d\nu_{a_i}}{d\tilde{\nu}_{a_i}}(X_i)\right) \leqslant c \right\}.$$

A regret lower bound for the multi-armed bandit problem Combining this change of measure with concentration inequalities, Lai and Robbins (1985a) was able to provide one of the first regret lower bound on the achievable regret in this setup. To this end, one should first specify the considered set of sampling strategies.

Definition 3.3 (*Uniformly-good strategy for bandits*) *Let $\mathcal{D}$ be a set of bandit configurations on $\mathcal{X} \subset \mathbb{R}$. For a configuration $\nu = (\nu_a)_{a\in\mathcal{A}} \in \mathcal{D}$, we denote $\mu_a(\nu)$ the mean of ν_a and $\mu_\star(\nu) = \max_{a\in\mathcal{A}} \mu_a(\nu)$ its maximal mean. A bandit strategy is uniformly-good on $\mathcal{D}$ if*

$$\forall \nu \in \mathcal{D}, \forall a \in \mathcal{A} : \mu_a(\nu) < \mu_\star(\nu) \implies \mathbb{E}[N_a(T)] = o(T^\alpha) \text{ for all } \alpha \in (0,1].$$

Intuitively, a uniformly-good strategy is just a strategy that "pulls sub-optimal arms not too often", when facing any bandit configuration from a set $\mathcal{D}$. Note that this is an asymptotic notion (for historical reasons). Lemma 3.4 below shows a fundamental barrier to the regret achievable by any such strategy. Namely if one wants to be uniformly good on $\mathcal{D}$, the regret against any $\nu \in \mathcal{D}$ must be *lower-bounded*, thus has to be high. We provide it below for the special case of Bernoulli distributions, that is when ν_a is a Bernoulli distribution $\mathcal{B}(\theta_a)$, with mean parameter $\theta_a \in [0,1]$. Since the notion of uniformly good strategy is asymptotic, so is the lower bound:

Lemma 3.2 (*Regret lower bound for uniformly good strategies*) *Let $\mathcal{B}$ denotes the set of all possible Bernoulli configurations, and $\nu \in \mathcal{B}$. Then any uniformly-good strategy on $\mathcal{B}$ must satisfy that*

$$\forall a \in \mathcal{A}, \mu_a(\nu) < \mu_\star(\nu) \quad \implies \quad \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\log(T)} \geqslant \frac{1}{\mathtt{kl}(\mu_a(\nu), \mu_\star(\nu))}.$$

where $\mathtt{kl}(\mu,\mu') = \mathtt{KL}(\mathcal{B}(\mu), \mathcal{B}(\mu'))$. In particular, it must incur a regret

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_T}{\log(T)} \geqslant \sum_{a\in\mathcal{A}} \frac{\mu_\star(\nu) - \mu_a(\nu)}{\mathtt{kl}(\mu_a(\nu), \mu_\star(\nu))}.$$

Proof of Lemma 3.2:

The first step is an application of Markov inequality.

$$\forall c \in \mathbb{R}^+, \quad \frac{\mathbb{E}[N_T(a)]}{\log(T)} \geqslant c\mathbb{P}_\theta(N_T(a) \geqslant c\log(T)) \quad \text{(Markov inequality)},$$

which motives to study the event $\Omega = \{N_T(a) < c\log(T)\}$.

The second step is to introduce a bandit configuration $\tilde{\nu} \in \mathcal{B}$ with parameters $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_A)$, that we specify as $\begin{cases} \tilde{\theta}_{a'} = \theta_{a'} & \text{if } a' \neq a \\ \tilde{\theta}_a = \lambda & \text{where } \lambda > \mu_\star \end{cases}$ for some λ. We call it a **maximally confusing** instance for a.

Let us now introduce the event to control the log-Likelihood ratio. For any $\alpha \in (0,1]$,

$$\begin{aligned} \text{let } \mathcal{E} &= \left\{ \sum_{t=1}^{T} \log\left(\frac{d\nu_{A_t}}{d\tilde{\nu}_{A_t}}(X_t)\right) \leqslant (1-\alpha)\log(T) \right\} \\ &= \left\{ \sum_{t=1}^{T} \mathbb{I}\{A_t = a\} \log\left(\frac{d\nu_a}{d\tilde{\nu}_a}(X_t)\right) \leqslant (1-\alpha)\log(T) \right\} \text{where } \frac{d\nu_a}{d\tilde{\nu}_a}(x) = \frac{\theta_a^x(1-\theta_a)^{1-x}}{\lambda^x(1-\lambda)^{1-x}}. \end{aligned}$$

The third step is to control the probability of the event $\Omega \cap \mathcal{E}$ by

$$\begin{aligned} \mathbb{P}_\nu(\Omega \cap \mathcal{E}) &\leqslant T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}(\Omega) && \text{(Change of measure)} \\ &= T^{1-\alpha}\mathbb{P}_{\tilde{\nu}}\Big(\textstyle\sum_{a' \neq a} N_T(a') > T - c\log(T) \Big) && \Big(\textstyle\sum_{a'} N_T(a') = T \Big) \\ &\leqslant T^{1-\alpha} \frac{\sum_{a' \neq a} \mathbb{E}_{\tilde{\nu}}[N_T(a')]}{T - c\log(T)} && \text{(Markov inequality)} \\ &= o(1) && \text{(Consistency for } \tilde{\theta}) \end{aligned},$$

where the last line follows by the assumption that the considered strategy is uniformly-good on $\mathcal{B}$ and thus in particular it pulls sub-optimal arms of $\tilde{\nu} \in \mathcal{B}$ not too often.

The fourth and last step is to control the remaining event $\Omega \cap \mathcal{E}^c$. Let us introduce $X_{a,j} = X_{\tau_{a,j}}$ with $\tau_{a,j} = \min\{t \in \mathbb{N} : N_a(t) = j\}$. Note that the random variables $\tau_{a,j}$ are predictable stopping times, since $\{\tau_{a,j} = t\}$ is a measurable with respect to the filtration generated by $A_1, X_1, \dots, A_{t-1}, X_{t-1}$. Hence we obtain

$$\begin{aligned} \mathbb{P}_\nu(\Omega \cap \mathcal{E}^c) &\leqslant \mathbb{P}_\nu\left(\exists m < c\log(T) : \sum_{j=1}^{m} \underbrace{\log\left(\frac{d\nu_a}{d\tilde{\nu}_a}(X_{a,j})\right)}_{Z_j} > (1-\alpha)\log(T) \right). \\ &= \mathbb{P}_\nu\left(\frac{\max_{m < c\log(T)} \sum_{j=1}^{m} Z_j}{c\log(T)} > \frac{1-\alpha}{c\mathtt{kl}(\theta_a, \lambda)} \underbrace{\mathtt{kl}(\theta_a, \lambda)}_{\mathbb{E}_\theta[Z_j]} \right) \end{aligned}$$

Remarking that the Z_j are i.i.d. bounded, with **positive** mean $\mu = \mathtt{kl}(\theta_a, \lambda)$ we can now apply the Asymptotic maximal Hoeffding inequality (see Lemma 2.3, that we recall below):

$$\forall \eta > 0, \lim_{n \to \infty} \mathbb{P}_\theta\left(\frac{\max_{m<n} \sum_{j=1}^{m} Z_j}{n} > (1+\eta)\mu \right) = 0.$$

It remains to choose e.g. $c = \dfrac{1-2\alpha}{\mathtt{kl}(\theta_a, \lambda)}$ to ensure that $\dfrac{1-\alpha}{c\mathtt{kl}(\theta_a, \lambda)} > 1$ and conclude by letting $\alpha \to 0$. □

Let us note that there is nothing too specific about using a family of Bernoulli distributions. The lower bound can actually be extended much beyond this case. Instead of following the same proof, let us show this with an alternative proof technique, that replaces the use of concentration inequality of the log-Likelihood with a control on its expectation. Both techniques yield interesting developments that we discuss later. Before we present it, we introduce a stronger version of the change of measure inequality, whose original proof technique can be traced at least back to Wald (1945), that we state below in a generic form. This inequality yields a variety of results in hypothesis testing, sequential decision making and beyond.

Lemma 3.3 (*Fundamental change of measure inequality*) *Let $(\underline{f}, \overline{f})$ be any conjugate pair of functions such that $\overline{f} : \mathbb{R} \to \mathbb{R}^+$ is convex and increasing with $\overline{f}(\mathbb{R}) = \mathbb{R}^+$ and $\underline{f} \circ \overline{f} = \overline{f} \circ \underline{f} = \mathbf{1}$. Then, for any measures ν, $\tilde{\nu}$ that admit a Radon-Nikodym derivative $d\nu/d\tilde{\nu}$, it holds*

$$\mathbb{E}_{\tilde{\nu}}\left[\underline{f}\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)\right] \leqslant \inf_{g:\mathcal{X}\to[0,1]} \mathbb{E}_{\tilde{\nu}}[g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[g(X)]}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right) + \mathbb{E}_{\tilde{\nu}}[1-g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[1-g(X)]}{\mathbb{E}_{\tilde{\nu}}[1-g(X)]}\right).$$

In particular, for the conjugate pair $(\underline{f}, \overline{f}) = (\log, \exp)$, we obtain the following form

$$\mathtt{KL}(\tilde{\nu}, \nu) = \mathbb{E}_{\tilde{\nu}}\left[\log\left(\frac{d\tilde{\nu}}{d\nu}(X)\right)\right] \geqslant \sup_{g:\mathcal{X}\to[0,1]} \mathtt{kl}\left(\mathbb{E}_{\tilde{\nu}}[g(X)], \mathbb{E}_{\nu}[g(X)]\right).$$

where we introduced for convenience the function $\mathtt{kl}(x, y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

Proof :

Let us consider a function $g : \mathcal{X} \to [0, a]$. Then it holds

$$\begin{aligned}
\mathbb{E}_{\nu}[g(X)] &= \mathbb{E}_{\tilde{\nu}}\left[\frac{d\nu}{d\tilde{\nu}}(X)g(X)\right] \\
&= \mathbb{E}_{\tilde{\nu}}\left[\overline{f}(\underline{f}(\frac{d\nu}{d\tilde{\nu}}(X)))\frac{g(X)}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right]\mathbb{E}_{\tilde{\nu}}[g(X)] \\
&\geqslant \overline{f}\left(\mathbb{E}_{\tilde{\nu}}\left[\underline{f}(\frac{d\nu}{d\tilde{\nu}}(X))\frac{g(X)}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right]\right)\mathbb{E}_{\tilde{\nu}}[g(X)]
\end{aligned}$$

The second line holds thanks to $\overline{f} \circ \underline{f} = \mathbf{1}$ and linearity of the expecation. The inequality follows by Jensen's inequality applied to the convex function $\overline{f}$, using the fact that $dq(x) = \frac{g(x)}{\mathbb{E}_{\tilde{\nu}}[g(x)]}d\tilde{\nu}(x)$ is a probability measure (because $g(x) \geqslant 0$ for all $x \in \mathcal{X}$). Hence, using the fact that $\underline{f} \circ \overline{f} = \mathbf{1}$ and $\underline{f}$ is increasing, we deduce that

$$\mathbb{E}_{\tilde{\nu}}\left[\underline{f}\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)g(X)\right] \leqslant \mathbb{E}_{\tilde{\nu}}[g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[g(X)]}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right).$$

Since $\tilde{g}(x) = a - g(x)$ is also non-negative, the same bound applies replacing g with $\tilde{g}$. Thus, using the key property that $a = g(x) + \tilde{g}(x)$ for all $x \in \mathcal{X}$, we deduce that

$$a\mathbb{E}_{\tilde{\nu}}\left[\underline{f}\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)\right] \leqslant \mathbb{E}_{\tilde{\nu}}[g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[g(X)]}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right) + \mathbb{E}_{\tilde{\nu}}[\tilde{g}(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[\tilde{g}(X)]}{\mathbb{E}_{\tilde{\nu}}[\tilde{g}(X)]}\right).$$

Thus, we deduce that

$$\begin{aligned}
\mathbb{E}_{\tilde{\nu}}\left[\underline{f}\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)\right] &\leqslant \inf\left\{\frac{1}{a}\inf_{g:\mathcal{X}\to[0,a]} \mathbb{E}_{\tilde{\nu}}[g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[g(X)]}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right) + \mathbb{E}_{\tilde{\nu}}[a-g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[a-g(X)]}{\mathbb{E}_{\tilde{\nu}}[a-g(X)]}\right) : a > 0\right\} \\
&= \inf_{g:\mathcal{X}\to[0,1]} \mathbb{E}_{\tilde{\nu}}[g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[g(X)]}{\mathbb{E}_{\tilde{\nu}}[g(X)]}\right) + \mathbb{E}_{\tilde{\nu}}[1-g(X)]\underline{f}\left(\frac{\mathbb{E}_{\nu}[1-g(X)]}{\mathbb{E}_{\tilde{\nu}}[1-g(X)]}\right)
\end{aligned}$$

□

Product measures As for Lemma 3.1, the previous result becomes especially interesting when considering two finite collection of independent measures $(\nu_a)_{a\in\mathcal{A}}$, $(\tilde{\nu}_a)_{a\in\mathcal{A}}$ on $\mathcal{X}$ and a deterministic sequence $(a_i)_{i\leqslant n}$ of index in $\mathcal{A}$. For instance if we form the product measures $\nu = \otimes_{i=1}^n \nu_{a_i}$ and $\tilde{\nu} = \otimes_{i=1}^n \tilde{\nu}_{a_i}$ on $\mathcal{X}^n$, and consider the random variable $X = (X_i)_{i\leqslant n}$, we obtain, thanks to the properties of the logarithm and the independence of the random variables,

$$\begin{aligned}\mathbb{E}_{\tilde{\nu}}\left[\log\left(\frac{d\tilde{\nu}}{d\nu}(X)\right)\right] &= \mathbb{E}_{\tilde{\nu}}\left[\sum_{i=1}^n \log\left(\frac{d\tilde{\nu}_{a_i}}{d\nu_{a_i}}(X_i)\right)\right] = \mathbb{E}_{\tilde{\nu}}\left[\sum_{a\in\mathcal{A}}\sum_{i=1}^n \log\left(\frac{d\tilde{\nu}_a}{d\nu_a}(X_i)\right)\mathbb{I}\{a_i = a\}\right] \\ &= \sum_{a\in\mathcal{A}}\sum_{i=1}^n \mathbb{I}\{a_i = a\}\mathbb{E}_{\tilde{\nu}_a}\left[\log\left(\frac{d\tilde{\nu}_a}{d\nu_a}(X_i)\right)\right] = \sum_{a\in\mathcal{A}} n_a \mathtt{KL}(\tilde{\nu}_a, \nu_a),\end{aligned}$$

where we introduced the integers $n_a = \sum_{i=1}^n \mathbb{I}\{a_i = a\}$. Note that using another pair than $(\log, \exp)$ would yield much more complicated expression when dealing with product measures. Interestingly, this result can be extended to the case when the sequence a_i is not deterministic but adapted to the filtration of the observations. In that case, the n_a become random variables N_a, and we get instead

$$\mathbb{E}_{\tilde{\nu}}\left[\log\left(\frac{d\tilde{\nu}}{d\nu}(X)\right)\right] = \mathbb{E}\left[\sum_{a\in\mathcal{A}} N_a \mathtt{KL}(\tilde{\nu}_a, \nu_a)\right],$$

where the expectation is over the law of all random variables. For further details, we refer to the manuscript of Emilie Kaufmann (see Kaufmann (2014)), who rediscovered this result independently. See also Garivier et al. (2016) for an extensive use of this result.

Another regret lower bound for the multi-armed bandit problem We provide below a regret lower bound for multi-armed bandits, whose proof follows the fundamental change of measure argument. We provide it in a slightly more general setup than the Bernoulli case (yet still restricted to product set of distributions).

Lemma 3.4 (*Regret lower bound for uniformly good strategies*) *Let $\mathcal{D} = \mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_A$, where $\mathcal{D}_a \subset \mathcal{P}(\mathcal{X})$ for each $a \in \mathcal{A}$ be any (unstructured) set of configurations, and let $\nu \in \mathcal{D}$. Then any uniformly-good strategy must pull arms such that*

$$\forall a \in \mathcal{A}, \mu_a(\nu) < \mu_\star(\nu) \quad \Longrightarrow \quad \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\log(T)} \geqslant \frac{1}{\mathcal{K}_a(\nu_a, \mu_\star(\nu))}.$$

where $\mathcal{K}_a(\nu_a, \mu_\star(\nu)) = \inf\{\mathtt{KL}(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D}_a, \mu_a(\nu') > \mu_\star(\nu)\}$. In particular, it must incur a regret

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_T}{\log(T)} \geqslant \sum_{a\in\mathcal{A}} \frac{\mu_\star(\nu) - \mu_a(\nu)}{\mathcal{K}_a(\nu_a, \mu_\star(\nu))}.$$

Proof of Lemma 3.4:

Let $\nu \in \mathcal{D}$ denote the configuration from which observations are sampled. First, from the fundamental lower bound using the $(\log, \exp)$ pair and isolating the number of pull of an arm $N_T(a)$, it comes for all sub-optimal $a \in \mathcal{A}$ ($\mu_a(\nu) < \mu_\star(\nu)$) that

$$\mathbb{E}_\nu[N_T(a)] \geqslant \sup_{\nu' \in \mathcal{D}, g:\mathcal{X}^T \to [0,1]} \frac{\mathtt{kl}(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\nu'}[g(X)]) - \sum_{a' \in \mathcal{A}\setminus\{a\}} \mathbb{E}[N_T(a')]\mathtt{KL}(\nu_{a'}, \nu'_{a'})}{\mathtt{KL}(\nu_a, \nu'_a)} .$$

Now, from the definition of uniformly-good strategy, is is natural to choose $g(X_1, \dots, X_T) = \mathbb{I}\{\Omega_\alpha\}$ where $\Omega_\alpha = \{N_T(a) > T^\alpha\}$. Indeed, for each a that is sub-optimal, one get by Markov inequality that for each $\alpha \in (0,1)$

$$\mathbb{E}_\nu[g(X)] = \mathbb{P}_\nu(N_T(a) > T^\alpha) \leqslant \mathbb{E}_\nu[N_T(a)]T^{-\alpha} = o(1) .$$

Hence, we deduce that $\mathtt{kl}(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\nu'}[g(X)]) \simeq -\log\big(\mathbb{P}_{\nu'}(N_T(a) \leqslant T^\alpha)\big)$. We now use the structural property that for all T, $\sum_{a\in\mathcal{A}} N_T(a) = T$ together with a second application of Markov inequality to get

$$\mathtt{kl}(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\nu'}[g(X)]) \simeq -\log\big(\mathbb{P}_{\nu'}(N_T(a) \leqslant T^\alpha)\big) \geqslant \log\Big(T - T^\alpha\Big) - \log\Big(\sum_{a'\neq a} \mathbb{E}_{\nu'}[N_T(a')]\Big) .$$

Finally, it remains to choose ν'. We choose $\nu' = \nu^\alpha$ such that a is the unique optimal arm: $\forall a' \neq a, \mu_{a'}(\nu^\alpha) < \mu_a(\nu^\alpha)$. This ensures that $\log\Big(\sum_{a'\neq a} \mathbb{E}_{\nu^\alpha}[N_T(a')]\Big) = o(\log(T))$, by uniform consistency. We further choose ν^α such that $\nu^\alpha_{a'} = \nu_{a'}$ for all $a' \neq a$, to ensure that $\mathtt{KL}(\nu_{a'}, \nu^\alpha_{a'}) = 0$. Such a choice exists since $\mathcal{D}$ is unstructured, in the sense that we can freely modify a distribution on one arm without having to modify the distribution of any other arm to ensure we stay in the family. This is also valid for any α. Note that due to these two conditions, we must have $\mu_a(\nu^\alpha) > \mu_\star(\nu)$. This ensures that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\log(T)} \geqslant \sup\left\{\frac{1-\alpha}{\mathtt{KL}(\nu_a, \nu^\alpha_a)} : \alpha \in (0,1], \nu^\alpha_a \in \mathcal{D}_a, \mu(\nu^\alpha_a) > \mu_\star(\nu)\right\} .$$

We conclude by letting $\alpha \to 0$. □

2 FURTHER LOWER-BOUNDS AND EXTENSIONS

In the previous section, we introduced two proof techniques for providing a regret lower bounds in multi-armed bandit sampling strategies. It turns out that both techniques yield interesting extensions. We first present a fully non-asymptotic lower bound on the regret, as a simple extension of Lemma 3.2. We then present two extensions of the second proof technique involving the fundamental change of measure: A first one to obtain regret bounds for a structured set of bandit configurations, a second one to obtain regret bounds in sequential hypothesis testing.

A non-asymptotic lower bound The direct proof of the lower bound from Lemma 3.2 that is based on concentration inequalities is especially interesting: Indeed by replacing the maximal asymptotic concentration inequality with a finite-time version, and making the notion of uniformly good strategy non-asymptotic, it gives the possibility of deriving a non-asymptotic lower bound for the regret of strategies. More precisely, we get

Lemma 3.5 (*Price for uniform optimality*) *Let $\mathcal{B}$ be the set of all Bernoulli bandit configurations. For any constants $C = (c_a)_{a\in\mathcal{A}}$ and function $f : \mathbb{N} \to \mathbb{R}^+$, it is not possible for an algorithm to achieve simultaneously $\forall \nu \in \mathcal{B}, \forall a \in \mathcal{A} : \mu_a(\nu) < \mu_\star(\nu), \forall T \in \mathbb{N} \quad \mathbb{E}_\theta[N_T(a)] \leqslant c_a f(T)$ and*

$$\mathbb{E}_\theta[N_T(a)] \leqslant \sup_{\alpha,\varepsilon\in(0,1),\lambda>\mu^\star} \frac{(1-\varepsilon)(1-\alpha)}{\mathtt{kl}(\mu_a,\lambda)} \log(T)\Big[1-\delta_{C,f,T}(\mu,\alpha,\varepsilon,\lambda)\Big] \quad \textit{where}$$

$$\delta_{C,f,T}(\mu,\alpha,\varepsilon,\lambda) = \frac{\sum_{a'\neq a} c_{a'} T^{1-\alpha} f(T)}{T - \frac{(1-\varepsilon)(1-\alpha)\log(T)}{\mathtt{kl}(\mu,\lambda)}} + 1 \wedge \sqrt{\frac{(1-\varepsilon)(1-\alpha)}{\mathtt{kl}(\mu,\lambda)}\log(T)} \exp\left(-\frac{2\varepsilon^2 \mathtt{kl}(\mu,\lambda)(1-\alpha)\log(T)}{(1-\varepsilon)|\log(\frac{\mu(1-\lambda)}{(1-\mu)\lambda})|^2}\right).$$

We may call (C, f)-uniformly-good a pulling strategy that always satisfies the first of the two lines, in reference to the corresponding asymptotic notion. Now denoting the bound in the second inequality $L_{c,f,T}(\mu_a(\nu), \mu_\star(\nu))$, the result shows that either a pulling strategy is not (C, f)-uniformly-good (for at least one $\nu \in \mathcal{B}$), or it is but the expected number of pulls of a suboptimal arm a must then be at least $L_{c,f,T}(\mu_a(\nu), \mu_\star(\nu))$. From this result, it is natural to ask: What is a smallest set of values and function C, f such that

$$\forall \nu \in \mathcal{B}, \forall a \in \mathcal{A} : \mu_a(\nu) < \mu_\star(\nu), \forall T \in \mathbb{N} \quad L_{c,f,T}(\mu_a(\nu), \mu_\star(\nu)) \leqslant c_a f(T)\,?$$

The choice $c_a = \frac{1}{\mathtt{kl}(\mu_a,\mu_\star)}$ with $f(T) = \log(T)$ is admissible, which enables to recover the asymptotic lower bound from below, hence showing the behavior of the lower-bound given by the asymptotic result can be beaten non-asymptotically. Whether we can get smaller values is currently an open question.

Proof of Lemma 3.5:

We first replace the Asymptotic maximal Hoeffding inequality with a non-asymptotic result. For instance the Laplace concentration inequality specialized to σ-sub-Gaussian random variables Z_j with mean μ gives

$$\mathbb{P}_\nu\left(\exists m \in \mathbb{N}, \quad \sum_{j=1}^{m} Z_j > m\mu + \sigma\sqrt{2(m+1)\log(\sqrt{m+1}/\delta)}\right) \leqslant \delta\,.$$

In our setup, the random variables are $Z_j = \log(\frac{\theta_a}{\lambda})$ if $X_{a,j} = 1$, and $\log(\frac{1-\theta_a}{1-\lambda})$ if $X_{a,j} = 0$. Hence $Z_j \in [A, B]$ where $A = \min(\log(\frac{\theta_a}{\lambda}), \log(\frac{1-\theta_a}{1-\lambda}))$ and $B = \max(\log(\frac{\theta_a}{\lambda}), \log(\frac{1-\theta_a}{1-\lambda}))$, thus we deduce that $\sigma_\lambda = (B-A)/2 = |\log(\frac{\theta_a(1-\lambda)}{(1-\theta_a)\lambda})|/2$ is a suitable value of σ (although in the case of Bernoulli distributions, this approach is a little crude).

Then we solve the following constraints in δ, for each α, T

$$\forall m < c\log(T), \quad m\mu + \sigma_\lambda\sqrt{2(m+1)\log(\sqrt{m+1}/\delta)} \leqslant (1-\alpha)\log(T)\,.$$

This leads, since the most constraining value of m is for $\overline{m} = \lfloor c\log(T)\rfloor$ to

$$\delta_{c,\alpha} \;=\; \min\left\{1, \sqrt{\lfloor c\log(T)\rfloor}\exp\left(-\frac{1}{2(\lfloor c\log(T)\rfloor+1)}\left(\frac{(1-\alpha)\log(T)-\lfloor c\log(T)\rfloor \mathtt{kl}(\mu_a,\lambda)}{\sigma_\lambda}\right)_+^2\right)\right\}$$

Hence, reproducing the same steps as for the proof of Lemma 3.2, for all c,α such that $1-\alpha > c\log(T)\mathtt{kl}(\mu_a,\lambda) \geqslant \lfloor c\log(T)\rfloor \mathtt{kl}(\mu_a,\lambda)$, it holds

$$\begin{aligned}\frac{\mathbb{E}_\theta[N_T(a)]}{\log(T)} &\geqslant c\left[1-e^{(1-\alpha)\log(T)}\frac{\sum_{a'\neq a}\mathbb{E}_{\tilde{\theta}}[N_T(a')]}{T-c\log(T)} - \delta_{c,\alpha}\right]\\ &\geqslant c\left[1-e^{(1-\alpha)\log(T)}\frac{\sum_{a'\neq a}\mathbb{E}_{\tilde{\theta}}[N_T(a')]}{T-c\log(T)} - 1\wedge\sqrt{c\log(T)}\exp\left(-\frac{\log(T)}{2c}\left(\frac{(1-\alpha)-c\mathtt{kl}(\mu_a,\lambda)}{\sigma_\lambda}\right)^2\right)\right].\end{aligned}$$

Choosing $c = \frac{(1-\varepsilon)(1-\alpha)}{\mathtt{kl}(\mu_a,\lambda)}$ for some $\varepsilon<1$ yields,

$$\begin{aligned}\frac{\mathbb{E}_\theta[N_T(a)]}{\log(T)} \;\geqslant\; \sup_{\alpha,\varepsilon\in(0,1),\lambda>\mu^\star} \frac{(1-\varepsilon)(1-\alpha)}{\mathtt{kl}(\mu_a,\lambda)}\Bigg[1-e^{(1-\alpha)\log(T)}\frac{\sum_{a'\neq a}\mathbb{E}_{\tilde{\theta}}[N_T(a')]}{T-\frac{(1-\varepsilon)(1-\alpha)\log(T)}{\mathtt{kl}(\mu_a,\lambda)}}\\ -1\wedge\sqrt{\frac{(1-\varepsilon)(1-\alpha)}{\mathtt{kl}(\mu_a,\lambda)}\log(T)}\exp\left(-\frac{\mathtt{kl}(\mu_a,\lambda)\varepsilon^2}{2(1-\varepsilon)\sigma_\lambda^2}(1-\alpha)\log(T)\right)\Bigg]\end{aligned}$$

Now the second and third term converge to 0 as T goes to ∞, for all $\alpha,\varepsilon\in(0,1)$. We finally use the non-asymptotic consistency property stating that $\mathbb{E}_{\tilde{\nu}}[N_T(a')] \leqslant c_{a'} f(T)$ for each $a'\neq a$, to conclude. □

Regret lower bounds for $\mathcal{D}$-constrained configuration sets Following the same proof steps as for Lemma 3.4 (or Lemma 3.2) one can obtain the following much more general result:

Lemma 3.6 ($\mathcal{D}$-*constrained regret lower bound*) *Let $\mathcal{D}$ be any set of bandit configurations,and let $\nu\in\mathcal{D}$. Then any uniformly-good strategy on $\mathcal{D}$ must incur a regret*

$$\liminf_{T\to\infty}\frac{\mathfrak{R}_T}{\log(T)} \geqslant \inf\left\{\sum_{a\in\mathcal{A}} c_a(\mu_\star(\nu)-\mu_a(\nu)) : \forall a\in\mathcal{A}, c_a\geqslant 0, \inf_{\nu'\in\tilde{\mathcal{D}}(\nu)}\sum_{a\in\mathcal{A}} c_a KL(\nu_a,\nu'_a)\geqslant 1\right\}.$$

where we introduced the set of maximally confusing distributions

$$\tilde{\mathcal{D}}(\nu) = \left\{\nu'\in\mathcal{D} : \mathcal{A}^\star(\nu')\cap\mathcal{A}^\star(\nu)=\emptyset, \forall a\in\mathcal{A}^\star(\nu), KL(\nu_a,\nu'_a)=0\right\}.$$

This result can be seen as a specialization to the multi-armed bandit setup of an even more general result obtained by Graves and Lai (1997) (extending the work of Agrawal et al. (1989)).

Proof :

Using the fundamental change of measure argument and a similar construction than for the proof of Lemma 3.4, we get that, for each sub-optimal arm a and any $\nu' \in \mathcal{D}$ such that a is its unique optimal arm, then asymptotically as $T \to \infty$,

$$\sum_{a' \in \mathcal{A}} \mathbb{E}[N_T(a')] \mathtt{KL}(\nu_{a'}, \nu'_{a'}) \geqslant \log\Big(T - T^{\alpha}\Big) - \log\Big(\sum_{a' \neq a} \mathbb{E}_{\nu'}[N_T(a')]\Big)\,,$$

in the sense that $\liminf_T \frac{A}{\log(T)} \geqslant \liminf_T \frac{B}{\log(T)}$, with A and B being the two terms of the inequality. Since by uniformly-good assumption, it must be that $\liminf_T \frac{\log\Big(\sum_{a' \neq a} \mathbb{E}_{\nu'}[N_T(a')]\Big)}{\log(T)} = 0$, we deduce that for any $\nu' \in \mathcal{D}$ that has no optimal arm in common with an optimal arm of ν, then

$$\liminf_T \sum_{a' \in \mathcal{A}} \frac{\mathbb{E}[N_T(a')]}{\log(T)} \mathtt{KL}(\nu_{a'}, \nu'_{a'}) = \sum_{a' \in \mathcal{A}} \left(\liminf_T \frac{\mathbb{E}[N_T(a')]}{\log(T)}\right) \mathtt{KL}(\nu_{a'}, \nu'_{a'}) \geqslant 1\,.$$

This holds in particular choosing ν' such that $\mathtt{KL}(\nu_{a'}, \nu'_{a'}) = 0$ whenever a' is optimal for ν. We conclude by remarking that

$$\liminf_{T \to \infty} \frac{\mathfrak{R}_T}{\log(T)} = \sum_{a \in \mathcal{A}} \left(\liminf_{T \to \infty} \frac{\mathbb{E}[N_T(a)]}{\log(T)}\right) (\mu_\star(\nu) - \mu_a(\nu))\,.$$

□

When considering a generic set of bandit configurations $\mathcal{D}$, another related that is useful to identify is the number of times a sub-optimal arm needs to be pulled. A direct application of the fundamental change of measure inequality together with simple reordering of the terms shows that

$$\mathbb{E}_\nu[N_T(a)] \geqslant \sup_{\nu' \in \mathcal{D}} \frac{\sup_{g:\mathcal{X} \to [0,1]} \mathtt{kl}\Big(\mathbb{E}_{\tilde{\nu}}[g(X)], \mathbb{E}_\nu[g(X)]\Big) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] \mathtt{KL}(\nu_{a'}, \nu'_{a'})}{\mathtt{KL}(\nu_a, \nu'_a)}\,.$$

When specified to the quest of uniformly-good strategies on $\mathcal{D}$, this motivates the following definition

Definition 3.6 (*Asymptotic price for uniformly-good strategies*) *For $\nu \in \mathcal{D}$ and $a \notin \mathcal{A}_\star(\nu)$, we define the asymptotic price to pay on arm a for being uniformly-good on $\mathcal{D}$ by*

$$n_T(a, \nu, \mathcal{D}) = \sup_{\nu' \in \mathcal{D} : a \in \mathcal{A}_\star(\nu)} \frac{\log(T) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] KL(\nu_{a'}, \nu'_{a'})}{KL(\nu_a, \nu'_a)}\,.$$

Indeed, the number of pulls of a by any uniformly-good strategies should satisfy $\mathbb{E}_\nu[N_T(a)] \geqslant n_T(a, \nu, \mathcal{D})$ asymptotically, when $T \to \infty$.

Detection lower bounds for hypothesis testing We now present another simple application of the fundamental change of measure result to an hypothesis-testing problem. Let us say we are facing a configuration of K distributions, and we have the possibility, at each time step to choose one of these distributions, and receive one observation sampled from it. We consider two disjoint sets of configurations $\mathcal{D}_0, \mathcal{D}_1 \subset \mathcal{P}(\mathcal{X})^K$, and our goal is to decide, as fast as possible, to which set the distributions we sample belong, while ensuring a low probability of error. Hence, a strategy will decide at each time step what distribution to sample, until some (random) time when it stops and output its decision (the set belongs to $\mathcal{D}_0$, or to $\mathcal{D}_1$). To avoid "lucky" sampling strategies, we restrict to the following strategies:

Definition 3.9 ($(\mathcal{D}_0, \mathcal{D}_1)$-*uniformly*-$\delta$-*correct detection strategy*) *A uniformly-δ-correct detection strategy to separate $\mathcal{D}_0$ from $\mathcal{D}_1$ (where $\mathcal{D}_0 \cap \mathcal{D}_1 = \emptyset$) ensures that, if c_τ denotes the class $\mathcal{D}_0$ or $\mathcal{D}_1$ chosen by the strategy at its random stopping time τ, then*

$$\forall \nu \in \mathcal{D}_0,\ \mathbb{P}\Big(c_\tau = \mathcal{D}_1\Big) \leqslant \delta, \quad \forall \nu \in \mathcal{D}_1,\ \mathbb{P}\Big(c_\tau = \mathcal{D}_0\Big) \leqslant \delta.$$

Applying the fundamental change of measure Lemma 3.3 is here direct, using $g(X_1, \dots, X_\tau) = \mathbb{I}\{\Omega\}$ where $\Omega = \{c_\tau = \mathcal{D}_1\}$, so that $\mathbb{P}_{\nu'}[\Omega] \geqslant 1-\delta$ and $\mathbb{P}_\nu[\Omega] \leqslant \delta$. In this case, we deduce that $\mathtt{kl}\Big(\mathbb{P}_{\nu'}[\Omega], \mathbb{P}_\nu[\Omega]\Big) \geqslant (1-\delta)\log\frac{1-\delta}{\delta}$, and thus any $(\mathcal{D}_0, \mathcal{D}_1)$-uniformly-$\delta$-correct detection strategy must satisfy

$$\sum_a \mathbb{E}_{\nu'}[N_\tau(a)]\mathtt{KL}(\nu'_a, \nu_a) \geqslant (1-\delta)\log\frac{1-\delta}{\delta}.$$

In particular, the random stopping time τ when the algorithm outputs a decision must satisfy

$$\mathbb{E}[\tau] \geqslant \frac{(1-\delta)\log\big((1-\delta)/\delta\big)}{\sup\{\sum_a w_a \mathtt{KL}(\nu'_a, \nu_a) : \sum_a w_a = 1, w_a \geqslant 0\}}\,.$$

Isolating $N_\tau(a)$ in the previous expression also gives a lower bound on the expected number of pulls $\mathbb{E}[N_\tau(a)]$ of ν'_a. Generalizing this simple idea from 2 to M decision sets yields the more general following result:

Definition 3.12 (*Uniformly-good strategy for separation*) *Let $\mathcal{D}_1, \dots, \mathcal{D}_M \subset \mathcal{D}$ be M sets of configurations included in a reference set $\mathcal{D} \subset \mathcal{P}(\mathcal{X})^{\mathcal{A}}$. A separation strategy samples the arms, until some stopping time τ decided by the strategy, where it outputs a score[a] $s_\tau : [\![0, M]\!] \to \mathbb{R}$ such that $s_\tau(m)$ is the score given by the strategy to the hypothesis $\nu \in \mathcal{D}_m$ for each $m \in [\![1 : M]\!]$ and $s_\tau(0)$ is the score given to the hypothesis $\nu \notin \cup_m \mathcal{D}_m$. For convenience, let $\mathcal{D}_0 = \mathcal{D} \setminus \{\cup_m \mathcal{D}_m\}$. A $(\boldsymbol{\delta}, \mathcal{D})$-uniformly good separation strategy, where $\boldsymbol{\delta} = (\delta_m)_{m \in [\![0,M]\!]}$ ensures that when it stops,*

$$\forall m \in [\![0, M]\!], \forall \nu \in \mathcal{D}_m, \quad \mathbb{P}_\nu\Big(s_\tau(m) < \max_{m'} s_\tau(m')\Big) \leqslant \delta_m\,.$$

[a] For any two integers $m < M \in \mathbb{Z}$, we denote $[\![m, M]\!] = \{m, \dots, M\} \subset \mathbb{Z}$.

Lemma 3.7 (*Detection lower bounds*) *Any $(\boldsymbol{\delta}, \mathcal{D})$-uniformly good separation strategy for $(\mathcal{D}_m)_{m\in[\![0,M]\!]}$ must sample the arms in a way such that*

$$\forall m \in [\![0, M]\!], \forall \nu \in \mathcal{D}_m, \quad \mathbb{E}_\nu[N_\tau(a)] \geqslant \kappa_{m,a}(\nu; \boldsymbol{\delta}) \, .$$

where we introduce the following term

$$\kappa_{m,a}(\nu; \boldsymbol{\delta}) = \max_{m'\in[\![0,M]\!]} \sup_{\nu'\in\mathcal{D}_{m'}\setminus\mathcal{D}_m} \frac{\mathtt{kl}(\delta_m, 1-\delta_{m'}) - \sum_{a'\in\mathcal{A}\setminus\{a\}} \mathbb{E}[N_\tau(a')]\mathtt{KL}(\nu_{a'}, \nu'_{a'})}{\mathtt{KL}(\nu_a, \nu'_a)} \, .$$

In the case of unstructured configurations of arms such that $\forall m, m' \in [\![1, M]\!], \quad \mathcal{D}_m \subset \otimes_{a\in\mathcal{A}}\mathcal{P}(\mathcal{X})$ and $\mathcal{D}_m \cap \mathcal{D}_{m'} = \emptyset$, we obtain the following simplifications for each $m \in [\![1, M]\!]$,

$$\kappa_{m,a}(\nu; \boldsymbol{\delta}) = \max_{m'\in[\![0,M]\!]} \frac{\mathtt{kl}(\delta_m, 1-\delta_{m'})}{\mathtt{KL}(\nu_a, \mathcal{D}_{m'})} \, , \qquad \textit{with } \mathtt{KL}(\nu, \mathcal{D}) \stackrel{\text{def}}{=} \inf\{\mathtt{KL}(\nu, \nu') : \nu' \in \mathcal{D}\} \, .$$

The prof of this result is a direct application of the change of measure argument, together with the definition of uniformly-good strategies to specify the functions g used in the argument.

3 FROM LOWER BOUNDS TO SAMPLING STRATEGIES

The previous paragraph illustrates that change of measure can yield powerful lower bounds in sequential decision making. We now want to point that the result gives more than a lower bound: it actually suggests a sampling strategy.

Empirical distributions Before proceeding, let us remind that that at time t, we only have access to an empirical distribution $\widehat{\nu}_a(t)$ of ν_a for each $a \in \mathcal{A}$. We denote empirical distributions in two related ways, depending on whether random averages indexed by the global time t or averages of given numbers t of pulls of a given arms are considered. The first series of averages will be referred to by using a functional notation for the indexation in the global time: $\widehat{\nu}_a(t)$, while the second series will be indexed with the local times t in subscripts: $\widehat{\nu}_{a,t}$. These two related indexations, functional for global times and random averages versus subscript indexes for local times, will be (hopefully) consistent throughout the manuscript for all quantities at hand, not only empirical averages.

Definition 3.15 (*Empirical distributions*) *For each $m \geqslant 1$, we denote by $\tau_{a,m}$ the round at which arm a was pulled for the m–th time, that is*

$$\tau_{a,m} = \min\big\{t \in \mathbb{N} : \ \ N_a(t) = m\big\}\,.$$

For each round t such that $N_a(t) \geqslant 1$, we then define the following two empirical distributions

$$\widehat{\nu}_a(t) = \frac{1}{N_a(t)}\sum_{s=1}^{t}\delta_{Y_s}\,\mathbb{I}_{\{a_s=a\}} \quad \textit{and} \quad \widehat{\nu}_{a,n} = \frac{1}{n}\sum_{m=1}^{n}\delta_{X_{a,m}}, \quad \textit{where} \quad X_{a,m} \overset{\text{def}}{=} Y_{\tau_{a,m}}\,.$$

where δ_x denotes the Dirac distribution on $x \in \mathbb{R}$.

Lemma 3.8 (*Empirical distributions*) *The random variables $X_{a,m} = Y_{\tau_{a,m}}$, where $m = 1, 2, \ldots$, are independent and identically distributed according to ν_a. Moreover, we have the rewriting $\widehat{\nu}_a(t) = \widehat{\nu}_{a,N_a(t)}$.*

Proof of Lemma 3.8:

For means based on local times we consider the filtration $(\mathcal{F}_t)$, where for all $t \geqslant 1$, the σ–algebra $\mathcal{F}_t$ is generated by $a_1, Y_1, \ldots,\ a_t, Y_t$. In particular, a_{t+1} and all $N_a(t+1)$ are $\mathcal{F}_t$–measurable. Likewise, $\big\{\tau_{a,m} = t\big\}$ is $\mathcal{F}_{t-1}$–measurable. That is, each random variable $\tau_{a,m}$ is a (predictable) stopping time. Hence, the result follows by a standard result in probability theory (see, e.g., Chow and Teicher 1988, Section 5.3). □

In practice when $\mathcal{D}$ is a given set of bandit configurations, we consider an operator $\Pi_{\mathcal{D}} : \mathcal{P}(\mathcal{X}) \to \mathcal{D}$ (in spirit, a projection operator) that we apply to the empirical distribution $\widehat{\nu}(t)$. A typical example is that of a parametric family of distributions, where $\widehat{\nu}(t)$ is mapped to a parametric distribution $\nu_{\widehat{\theta}(t)}$ with parameter $\widehat{\theta}(t)$.

A lower-bound inspired strategy We are now ready to illustrate the construction of a strategy from lower bounds on the hypothesis testing task.

Since our goal is to get a uniformly good strategy, it is natural to take a look at the probability of mistake (for each class $\mathcal{D}_m$), and more precisely the lowest probability of mistake compatible with the current observations. As long as the probability of mistake is too large compared to the threshold δ, then we should continue sampling, otherwise we should stop. Hence the strategy presented in Algorithm 1 tries to *minimize the probability of mistake* as fast as possible, and is inspired from the previous lower bounds. More precisely, it computes the minimal probability of error compatible with the constraint on the number of observations provided by the lower bound, defined as $\delta_{t,a,m}$ for each m and a. Note that for small values of δ_m, $\mathtt{kl}(\delta_m, 1-\delta_{m'}) \simeq \log(1/\delta_{m'})$. This means for instance that the value $\delta_{t,a,m} = 0$ is achievable provided that

$$\exp(-N_t(a)\mathtt{KL}(\widehat{\nu}_{m,a}(t), \mathcal{D}_{m'})) \leqslant \delta_{m'}, \forall m'\,,$$

where $\widehat{\nu}_{m,a}(t) = (\Pi_{\mathcal{D}_m}(\widehat{\nu}(t)))_a$. We also note that we no longer need to pull an arm for hypothesis class m if $\delta_{t,a,m} \leqslant \delta_m$. The algorithm then simply pulls an arm corresponding to the smallest $\delta_{t,a,m}$ that is larger than δ_m among all a, m if such a pair exist. When no such pair exists, the algorithm can stop pulling new observations.

Algorithm 1 `KL-Separation`$((\mathcal{D}_m)_m, \mathcal{D}, \boldsymbol{\delta})$

$t = 0, \mathcal{M}_t = [\![0, M]\!]$.
while $\mathcal{M}_t \neq \emptyset$ **do**
  $t = t + 1$
  Compute for each $a \in \mathcal{A}, m \in [\![0, M]\!]$: $\delta_{t,a,m} = \inf\left\{\delta \in [0,1] : N_t(a) \leqslant \kappa_{m,a}(\Pi_{\mathcal{D}_m}(\widehat{\nu}(t)); \boldsymbol{\delta})\right\}$.
  Let $\mathcal{A}_{m,t} = \{a \in \mathcal{A} : \delta_{t,a,m} > \delta_m\}$ and $\mathcal{M}_t = \{m \in [\![0, M]\!] : \mathcal{A}_{m,t} \neq \emptyset\}$
  if $\mathcal{M}_t \neq \emptyset$ **then**
    Pull arm $a_t \in \mathcal{A}$ defined by

$$a_t = \operatorname*{arg\,min}_{a \in \mathcal{A}_{m_t,t}} \delta_{t,a,m_t} \quad \text{where } m_t = \operatorname*{arg\,min}_{m \in \mathcal{M}_t} \min_{a \in \mathcal{A}_{m,t}} \delta_{t,a,m}$$

output $s_t : m \mapsto -\mathtt{KL}(\widehat{\nu}(t), \mathcal{D}_m)$.

Remark 3.1 *We provide this algorithm for illustration purpose. It is currently unknown whether this strategy is provably optimal in some sense and what is a bound on its performance.*

`KL-ucb` a lower-bound inspired strategy for multi-armed bandits Since situations when sampling can be done at no cost are not very common, the setup of multi-armed bandit where each decision step may yield an instantaneous loss and the goal is to minimize the cumulative error is often more appealing in practice than pure hypothesis testing. We now provide the construction of the `KL-ucb` strategy for a set of bandit configurations $\mathcal{D}$, that can be traced at least back to Lai (1987).

The generic form of the algorithm is described as Algorithm 4, that relies a parameter that is non-decreasing function f, typically chosen such that $f(t) \approx \log(t)$.

At each round $t \geqslant K + 1$, an upper bound $U_a(t)$ is associated with the expectation μ_a of the distribution ν_a of each arm, then an arm a_{t+1} with highest upper bound is played.

Algorithm 2 The `KL-ucb` algorithm for unstructured $\mathcal{D}$.

Parameters: A set $\mathcal{D}$ of bandit configurations, a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\{1, \dots, K\}$ once

for each round $t + 1$, where $t \geqslant K$:

compute for each arm a the quantity

$$U_a(t) \;=\; \sup\left\{\mu_a(\nu) :\; \nu \in \mathcal{D},\; \forall a' \in \mathcal{A}\backslash\{a\},\, \nu_{a'} = \widehat{\nu}_{\mathcal{D},a'}(t) \quad \text{and} \quad N_a(t) \leqslant \frac{f(t)}{\mathtt{KL}\Big(\widehat{\nu}_{\mathcal{D},a}(t),\, \nu_a\Big)}\right\}$$

where $\widehat{\nu}_{\mathcal{D},a}(t) = \big(\Pi_{\mathcal{D}}\big(\widehat{\nu}(t)\big)\big)_a$

Pull an arm $a_{t+1} \in \operatorname*{arg\,max}_{a \in \mathcal{A}} U_a(t)$.

In the literature, another a variant of `KL-ucb` is introduced where the term $f(t)$ is replaced with $f(t/N_a(t))$. We refer to this algorithm as `KL-ucb+`. While `KL-ucb` has been analyzed and shown to be provably near-optimal for some specific sets $\mathcal{D}$ in Cappé et al. (2013) (dimension 1 exponential families), the variant

KL-ucb+ has not been analyzed. In chapter 7 corresponding to Maillard (2018) we provide the required tools to obtain near-optimal regret bounds for both KL-ucb and KL-ucb+ in the context of general exponential families of arbitrary finite dimension d.

The KL-ucb strategy is introduced in the case of an unstructured set of bandit configurations $\mathcal{D}$, in the sense that a most confusing instance for an arm a and configuration ν can always be found without modifying the distributions on the other arms. In view of the price to pay in the case of structured sets $\mathcal{D}$ (See definition 3.6), it is natural to introduce an alternative version of KL-ucb, that uses the following modified index:

$$U_a(t) = \sup\left\{\mu_a(\nu) : \ \nu \in \mathcal{D} \quad \text{and} \quad N_a(t) \leqslant \frac{f(t) - \sum_{a' \in \mathcal{A}\setminus\{a\}} N_{a'}(t)\mathtt{KL}\Big(\widehat{\nu}_{\mathcal{D},a'}(t),\, \nu_{a'}\Big)}{\mathtt{KL}\Big(\widehat{\nu}_{\mathcal{D},a}(t),\, \nu_a\Big)}\right\}.$$

Whether this strategy is indeed provably optimal is however an open question. Other variants have been introduced in Magureanu (2018), but there is currently no definitive answer.

The optimistic principle, revisited We now want to revisit a popular principle in multi-armed bandit theory that we think is stated in a slightly misleading way. This popular principle is the "optimism in face of uncertainty". The generic idea is as follows: let us say you want to optimize a criterion $\max_a g_a(\nu)$, over actions $a \in \mathcal{A}$, that depends on the distributions $\nu = (\nu_a)_a \in \mathcal{A}$ that you do not known. Further, say that, based on your hypothesis on the problem and the past observations up to time t,, you can build a high-probability confidence set $\mathcal{D}_{t,\delta}$ such that $\mathbb{P}(\nu \in \mathcal{D}_{t,\delta}) \geqslant 1-\delta$. Then, the optimistic principle tells you to sample, at time $t+1$ you should choose arm in

$$\arg\max_{a\in\mathcal{A}} \max\{g_a(\nu') : \nu' \in \mathcal{D}_{t,\delta}\}.$$

When the $g_a(\nu)$ is the mean μ_a of ν_a, and we assumed bounded distributions, we recover using some simple Hoeffding concentration inequality the basis for the popular UCB algorithm. This principle has been applied and extended to other situations, generally with success. However, we want to point out that this principle does not really comply with what suggests the lower bounds (Note also that the above rule is a little myopic). Indeed, the lower bounds are based on the construction of a most confusing instance $\tilde{\nu}$ that is statistically indistinguishable from ν given the past observations. They also quantity the number of pulls of each arm, or better the information that should be gathered for each distribution in order to rull-out a bad instance. It is important since the only thing we know is that by pulling arm a, then we receive one new observations from ν_a. Hence the lower bound construction thus suggests something a little different than the optimistic principle: It suggests to pull an arm so that we can rull-out the environment that seems to give largest gain. If after receiving the new observation, the environment is rulled-out, then we made indeed the right decision. Otherwise, this means we may have indeed made the optimal decision. This gives rise to a slightly different principle:

"Pull the arm that enables to rull-out the seemingly best environment from the plausible ones."

This may look like a subtle modification of the principle. However, in typical situations when we have a structured set of distributions, and for which sampling the distribution of an arm gives information about another one, this may yield strategies that differ from the naive application of the optimistic principle.

4 CHANGE POINT DETECTION

In this last section on the log-likelihood ratio, we now deviate from sequential sampling theory to consider a change-point setup. Indeed, change of measure is intimately linked with change point detection, and we want to provide below a quick overview of the basics of sequential change point detection, showing what can be achieved when combining tools such as the Laplace method with change-of-measure. In a change point detection problem, we receive a sequence of observations $Y_1, Y_2, \dots$ one by one, assumed to be generated from some process $\rho \in \mathcal{D}$, where $\mathcal{D}$ is a known family of processes. However, from some unknown time τ on, the observations $Y_{\tau+1}, Y_{\tau+2}, \dots$ are generated from a different process $\rho' \in \mathcal{D}$. The task of change-point detection is to detect the change, that is to raise an alarm at a time $t > \tau$, as early as possible after time the change occurred. We recall below the most emblematic change-point detection strategies.

CUSUM One of the most famous change-point detection algorithm is the CUSUM strategy from Page (1954) that is based on likelihood ratio thresholding: Assuming that $Y_1, \dots, Y_\tau$ is i.i.d. from a distribution p_0 and $Y_{\tau+1}, \dots, Y_n$ is i.i.d. from the distribution p_1, where both p_0 and p_1 are perfectly known and $\tau \in \mathbb{N}$ is the unknown change point, the original CUSUM change-point detection procedure takes a positive constant $c \in \mathbb{R}^+$ as input parameter and builds the following quantity:

$$\text{(CUSUM)} \quad \tau(c; p_0, p_1) = \min\Big\{t \in [1,n] : \max_{s\in[0,t)} L_{s:t} \geqslant c\Big\} \quad \text{where } L_{s:t} = \sum_{t'=s+1}^{t} \log \frac{p_1(Y_{t'})}{p_0(Y_{t'})}\,. \tag{3.1}$$

This quantity is a stopping time and enjoys nice theoretical properties: Let $\mathbb{E}_\tau$ and $\mathbb{P}_\tau$ denote the expectation and probability with respect to the process that changes from p_0 to p_1 at change-point $\tau + 1$. CUSUM minimizes the worst-case delay $\max_\tau \mathbb{E}_\tau(\widehat{\tau} - \tau | \widehat{\tau} \geqslant \tau)$ amongst all algorithms outputting $\widehat{\tau}$ for which $\mathbb{E}_0(\widehat{\tau}) = \mathbb{E}_0(\tau(c; p_0, p_1))$, see e.g. Blazek et al. (2001). On the other hand, this procedure is restricted to the case when p_0 and p_1 are known. The same criticism applies to the Shiryaev-Pollak stopping time $\min\{t \in [1,n] : \log \sum_{s=0}^{t-1} \exp(L_{s:t}) \geqslant c\}$.

GLR When p_0, p_1 are unknown, it is natural replace the log-likelihood ratios with a generalized likelihood ratio (GLR). While initially introduced for the case when p_0 is known and p_1 is not, Lai and Xing (2010) extends the GLR to the case when both distributions are unknown, assuming they come from the same canonical exponential family. Namely, for the density model $p_\theta(y) = \exp(\theta^\top y - \psi(\theta))$ with log-partition function ψ defining the exponential family $\mathcal{E} = \{p_\theta : \psi(\theta) < \infty\}$ it writes

$$\begin{aligned} \text{(GLR)} \quad \tau_n(c; \mathcal{E}) &= \min\Big\{t \in [1,n] : \max_{s\in[0,t)} G^{\mathcal{E}}_{1:s:t} \geqslant c\Big\} \quad \text{where} \qquad (3.2)\\ G^{\mathcal{E}}_{t_0:s:t} &= \sup_{\theta_1,\theta_2} \sum_{t'=t_0}^{s} \log p_{\theta_1}(Y_{t'}) + \sum_{t'=s+1}^{t} \log p_{\theta_2}(Y_{t'}) - \sup_{\theta} \sum_{t'=t_0}^{t} \log p_\theta(Y_{t'}) \\ &= (s - t_0 + 1)\psi_\star(\mu_{t_0:s}) + (t-s)\psi_\star(\mu_{s+1:t}) - (t - t_0 + 1)\psi_\star(\mu_{t_0:t})\,, \end{aligned}$$

in which we introduced the empirical means and Fenchel-Legendre dual following notations

$$\mu_{t':t} = \frac{1}{t - t' + 1} \sum_{s=t'}^{t} Y_s, \quad \psi_\star(\mu) = \sup_\theta \{\theta^\top \mu - \psi(\theta)\}\,.$$

Example 3.1 *For the family $\mathcal{N}_1$ of standard univariate Gaussian distributions $\{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$, the GLR statistics simplifies to*

$$G^{\mathcal{N}_1}_{t_0:s:t} = (s - t_0 + 1)(t - s)(\mu_{t_0,s} - \mu_{s+1,t})^2/(t - t_0 + 1)\,,$$

thus leading to the stopping time

$$(\mathcal{N}_1\text{-GLR}) \quad \tau_n(c; \mathcal{N}_1) = \min\Big\{t \in [1,n] : \max_{s\in[t_0,t)} \frac{(s - t_0 + 1)(t-s)}{t - t_0 + 1}(\mu_{t_0,s} - \mu_{s+1,t})^2 \geqslant c\Big\}\,. \tag{3.3}$$

Sequential setup The previous formulation is in the batch setup, however both CUSUM and GLR (and their variants) can be phrased in the sequential setup as well (see Downey (2008)). In this case, at time t, upon having observed $Y_{t_0+1}, \dots, Y_t$, an alert is raised according to the boolean test

$$\text{CUSUM}(t_0, t) = \mathbb{I}\{\max_{s\in[t_0,t)} L_{s:t} \geqslant c\}, \quad \text{or to} \quad \text{GLR}^{\mathcal{E}}(t_0, t) = \mathbb{I}\{\max_{s\in[t_0,t)} G^{\mathcal{E}}_{t_0:s:t} \geqslant c\}$$

where c may depend on t. Note that the observations $Y_{t'}$ for $t' \in [t+1, n]$ are not available at time t.

Delay and false alarms We measure the quality of a detection algorithm using the two following notions: First the *probability of false alarm*, that is of detecting a change at some time t while there is no change: For GLR this quantity is $\mathbb{P}_\infty(\exists t \in \mathbb{N} : \max_{s\in[t_0,t)} G^{\mathcal{E}}_{t_0:s:t} \geqslant c)$. Second the *detection delay*, that is the difference between the first time step when an algorithm detects a change and $\tau + 1$. For GLR, this is the random variable $\tau_t(c, \mathcal{E}) - \tau - 1$ for $t > \tau$, that can be studied in expectation or high probability. A natural question is then how to choose the threshold c.

While the classical literature only studies an asymptotic control of these and related quantities (e.g. expressed for the limiting case when the probability of false alarm tends to 0), we show we can be more precise, by building sequential change-point detection procedures that are uniformly-good in the following sense:

Definition 3.18 (*Uniformly-good change-point detection strategies*) *A change-point detection strategy is called uniformly-good on a class of processes $\mathcal{D}$ if for each $\nu \in \mathcal{D}$ generating the observations, for any given $\delta \in [0, 1]$*
i) *with probability higher than $1 - \delta$, uniformly over all $t \in [t_0, \tau]$, no alarm is raised at time t and*
ii) *its detection delay is controlled with probability $1-\delta$ and expressed in terms of the magnitude of the change. We thus request non-asymptotic results that hold for each t, each δ and each τ.*

We now consider a sequential change-point detection problem and generalize the GLR analysis from $\mathcal{N}_1$ to the class $\mathcal{D}^{\sigma\text{-sub}}$ of processes with σ-sub-Gaussian observation noise, that is we make the following mild assumption on the sequence $(Y_t)_t$ of real-valued observations

Assumption 3.1 (Sub-Gaussian observation noise) *A sequence $(Y_t)_t$ has σ-sub-Gaussian noise if*

$$\forall t, \forall \lambda \in \mathbb{R}, \quad \log \mathbb{E}[\lambda(Y_t - \mathbb{E}[Y_t])] \leqslant \frac{\lambda^2\sigma^2}{2}\,.$$

We further restrict to the case of change in the mean only (change of variance could be considered as well) and assume (piecewise) i.i.d. data.

We provide below a refined concentration inequality on the scan-statistics of the GLR test (Lemma **??**, Theorem 3.1) that improves on naive bounds derived from applications of Bonferroni inequality (aka union bound) thanks to the Laplace method. This result is used to derive Theorem 3.3, showing that given a confidence level $\delta \in (0, 1)$, the threshold $c = (1+\frac{1}{t-t_0+1})2\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}}{\delta}\right]$ enforces properties i) and ii), with an explicit detection delay that improves over state-of-the-art analysis.

4.1 Doubly time uniform concentration of scan-statistics

In order to handle changes of the mean, it is natural to study the concentration of $\mu_{1:s} - \mu_{s+1:t}$. A simple way to achieve time-uniform confidence bounds for such quantities is to make use of uniform concentration inequalities for $\mu_{1:s}$ and $\mu_{s+1:t}$ separately, and combine them with a simple union bound. This leads to the

bound $b_{t_0}^{\text{disjoint}}(s,t,\delta)$ given in the following Theorem 3.1. This however requires to extend the Laplace method, which we do now:

Lemma 3.9 (*Doubly-time-uniform concentration*) *Let $Y_1, \dots Y_t$ be a sequence of t independent real-valued random variables satisfying Assumption 3.1. Let $\mu_{t_1+1:t_2} = \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} Y_s$ be the empirical mean estimate on the time interval $[t_1+1, t_2]$. Then, for all $\delta \in (0,1)$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N}_\star, \exists s \in [0,t), \quad |\mu_{s+1:t} - \mathbb{E}[\mu_{s+1:t}]| \geqslant \sqrt{\frac{2\sigma^2(1+\frac{1}{t-s})}{t-s}\ln\left(\frac{t\ln^2(t)\sqrt{t+1-s}}{\ln(2)\delta}\right)}\right) \leqslant \delta\,.$$

Proof of Lemma 3.9:

Step 1. Let us start with the disjoint case, and consider that $t_0 = 1$. Let $\overline{z}_{s+1:t} = \mu_{s+1:t} - \mathbb{E}[\mu_{s+1:t}]$ be the centered empirical mean using observations from $s+1$ to t. We first introduce for each $\lambda \in \mathbb{R}$ and each $s \leqslant t$ the following quantity:

$$B_{s,t}^{\lambda} = \exp\left(\lambda(t-s)\overline{z}_{s+1:t} - \frac{\lambda^2\sigma^2(t-s)}{2}\right).$$

Note that $(B_{s,t}^{\lambda})_{t\in[s,\infty]\cap\mathbb{N}}$ is a non-negative supermartingale. Let us introduce $B_{s,t} = \mathbb{E}[B_{s,t}^{\Lambda}]$, where $\Lambda \sim \mathcal{N}(0, \frac{1}{\sigma^2(t-s)c})$, for some $c > 0$. We note that by simple algebra,

$$|\overline{z}_{s+1:t}| = \sqrt{\frac{2\sigma^2(1+c)}{t-s}\ln\left(B_{s,t}\sqrt{1+1/c}\right)}\,.$$

In particular, choosing $c = 1/(t-s)$, it comes for all deterministic $g(t) > 0$, that

$$\begin{aligned}
\mathbb{P}\left(\exists t, \exists s<t,\ |\overline{z}_{s+1:t}| \geqslant \sqrt{\frac{2\sigma^2(1+\frac{1}{t-s})}{t-s}\ln\left(\frac{g(t)\sqrt{1+t-s}}{\delta}\right)}\right) &= \mathbb{P}\left(\exists t, \exists s<t,\ B_{s,t} \geqslant g(t)/\delta\right) \\
&= \mathbb{P}\left(\exists t, \max_{s<t} B_{s,t} \geqslant g(t)/\delta\right) \\
&\leqslant \delta\mathbb{E}\left[\max_t \frac{\max_{s<t} B_{s,t}}{g(t)}\right].
\end{aligned}$$

Step 2. This leads to study the quantity $\frac{\max_{s<t} B_{s,t}}{g(t)}$. To this end, it is convenient to introduce $\overline{B}_t = \frac{\sum_{s<t} B_{s,t}}{g(t)}$ for $t > 1$. Indeed, for every random stopping time $\tau > 1$,

$$\mathbb{E}\left[\frac{\max_{s<\tau} B_{s,\tau}}{g(\tau)}\right] \leqslant \mathbb{E}\left[\overline{B}_\tau\right] = \mathbb{E}\left[\overline{B}_2 + \sum_{t=2}^{\infty}(\overline{B}_{t+1} - \overline{B}_t)\mathbb{I}\{\tau > t\}\right].$$

Further, we note that, conveniently

$$\overline{B}_{t+1} - \overline{B}_t = \frac{B_{t,t+1}}{g(t+1)} + \sum_{s<t}\left(\frac{B_{s,t+1}}{g(t+1)} - \frac{B_{s,t}}{g(t)}\right).$$

Next, by construction, we note that

$$\mathbb{E}[B_{s,t+1}|\mathcal{F}_t] = \frac{\sigma}{\sqrt{2\pi}}\int_{\mathbb{R}} \mathbb{E}[B^{\lambda}_{s,t+1}|\mathcal{F}_t] e^{-\frac{\lambda^2\sigma^2}{2}} d\lambda \;\leqslant\; \frac{\sigma}{\sqrt{2\pi}}\int_{\mathbb{R}} B^{\lambda}_{s,t} e^{-\frac{\lambda^2\sigma^2}{2}} d\lambda = B_{s,t}\,.$$

Thus, since $\mathbb{I}\{\tau > t\} \in \mathcal{F}_t$, we deduce that

$$\begin{aligned}\mathbb{E}\left[\frac{\max_{s<\tau} B_{s,\tau}}{g(\tau)}\right] &\leqslant \mathbb{E}\big[\overline{B}_2\big] + \sum_{t=2}^{\infty}\frac{\mathbb{E}[B_{t,t+1}]}{g(t+1)} + \sum_{t=1}^{\infty}\sum_{s<t}\mathbb{E}\left[\left(\frac{1}{g(t+1)} - \frac{1}{g(t)}\right)B_{s,t}\mathbb{I}\{\tau > t\}\right] \\ &= \mathbb{E}\big[\overline{B}_2\big] + \sum_{t=2}^{\infty}\frac{\mathbb{E}[B_{t,t+1}]}{g(t+1)} + \sum_{t=1}^{\infty}\sum_{s<t}\left(\frac{1}{g(t+1)} - \frac{1}{g(t)}\right)\underbrace{\mathbb{E}\Big[B_{s,t}\mathbb{I}\{\tau > t\}\Big]}_{\geqslant 0}\,.\end{aligned}$$

Hence, choosing g as an increasing function of t ensures that the last sum is upper bounded by 0. Since on the other hand $\mathbb{E}[B_{t,t+1}] \leqslant 1$ and $\mathbb{E}\big[\overline{B}_2\big] \leqslant 1/g(2)$, we deduce that

$$\mathbb{E}\left[\frac{\max_{s<\tau} B_{s,\tau}}{g(\tau)}\right] \;\leqslant\; \frac{1}{g(2)} + \sum_{t=2}^{\infty}\frac{1}{g(t+1)} = \sum_{t=2}^{\infty}\frac{1}{g(t)}\,.$$

Choosing $g(t) = Ct\ln^2(t)$ for $t > 1$ yields

$$\mathbb{E}\left[\frac{\max_{s<\tau} B_{s,\tau}}{g(\tau)}\right] \;\leqslant\; \frac{1}{C\ln(2)}\,.$$

Plugging-in this in the control of the deviation and choosing $C = 1/\ln(2)$ thus gives

$$\mathbb{P}\left(\exists t, \exists s < t \quad |\overline{z}_{s+1:t}| \geqslant \sqrt{\frac{2\sigma^2(1+\frac{1}{t-s})}{t-s}\ln\left(\frac{t\ln^2(t)\sqrt{t+1-s}}{\ln(2)\delta}\right)}\right) \;\leqslant\; \delta\,.$$

□

We obtain the following Theorem 3.1 upon using similar arguments and an additional union bound over s.

Theorem 3.1 (*Doubly-time-uniform concentration*) *Let $Y_1, \dots Y_t$ be a sequence of t independent real-valued random variables satisfying Assumption 3.1. Let $\mu_{t_1+1:t_2} = \frac{1}{t_2-t_1}\sum_{s=t_1+1}^{t_2} Y_s$ be the empirical mean estimate on the time interval $[t_1+1, t_2]$. Then, for each $t_0 \in \mathbb{N}_\star$, for all $\delta \in (0,1)$,*

$$\mathbb{P}\Big(\exists t \in \mathbb{N}_\star, s \in [t_0 : t), \big|\mu_{t_0:s} - \mu_{s+1:t} - \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}]\big| \geqslant b^{\textit{disjoint}}_{t_0}(s,t,\delta)\Big) \leqslant \delta \quad \textit{where}$$

$$b^{\textit{disjoint}}_{t_0}(s,t,\delta) = \sqrt{2}\sigma\left[\sqrt{\frac{1+\frac{1}{s-t_0+1}}{s-t_0+1}\ln\left(\frac{2\sqrt{s-t_0+2}}{\delta}\right)} + \sqrt{\frac{1+\frac{1}{t-s}}{t-s}\ln\left(\frac{2(t-t_0+1)\sqrt{t-s+1}\ln^2(t-t_0+1)}{\ln(2)\delta}\right)}\right].$$

Proof of Theorem 3.1:

We consider that $t_0 = 1$ without loss of generality. On the one hand, by Lemma 3.9, it holds

$$\mathbb{P}\left(\exists t, \exists s < t \quad |\overline{z}_{s+1:t}| \geqslant \sqrt{\frac{2\sigma^2(1+\frac{1}{t-s})}{t-s} \ln\left(\frac{t\ln^2(t)\sqrt{t+1-s}}{\ln(2)\delta}\right)}\right) \quad \leqslant \quad \delta\,.$$

Since on the other hand, by the classical Laplace method,

$$\mathbb{P}\left(\exists s, \quad |\overline{z}_{1:s}| \geqslant \sqrt{\frac{2\sigma^2(1+\frac{1}{s})}{s} \ln\left(\frac{\sqrt{s+1}}{\delta}\right)}\right) \quad \leqslant \quad \delta\,,$$

we conclude by using the triangular inequality $|\overline{z}_{1:s} - \overline{z}_{s+1:t}| \leqslant |\overline{z}_{1:s}| + |\overline{z}_{s+1:t}|$ together with a union bound argument. □

Instead of controlling each term in a disjoint way. A better approach may be to handle the concentration of the terms in $z_{1:s:t} := \mu_{1:s} - \mu_{s+1:t}$ jointly, since it is a sum of t independent random variables, s of which are σ/s-sub-Gaussian, and the others are $\sigma/(t-s)$-sub-Gaussian. We conjecture it might be possible to replace $b_{t_0}^{\text{disjoint}}(s,t,\delta)$ with the tighter bound

$$b_{t_0}^{\text{joint}}(s,t,\delta) \;=\; \sigma\sqrt{\Big(\frac{1}{s-t_0+1}+\frac{1}{t-s}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)2\ln\left[\frac{2(t-t_0+1)\sqrt{t-t_0+2}\ln(t-t_0+2)}{\delta}\right]}\,.$$

Lemma 3.9 generalizes beyond the Gaussian case, thanks to the following key result.

Lemma 3.10 (*Doubly-time uniform martingale control*) *For all positive integer $t \in \mathbb{N}_\star$ and integer $s \in \{0,\dots,t-1\}$, let us consider $M_{s+1:t}$ to be a non-negative random variable. Assume that for each fixed s, $(M_{s+1:t})_t$ is a super-Martingale with respect to the same filtration $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{N}_\star}$, and that $\forall t \in \mathbb{N}_\star, \mathbb{E}[M_{t,t}] \leqslant 1$. Then, for any non-decreasing positive function g, and any random stopping time τ with respect to the filtration $\mathcal{F}$, it holds*

$$\mathbb{E}\left[\frac{\max_{s\in\{0,\dots,\tau-1\}} M_{s+1:\tau}}{g(\tau)}\right] \leqslant \sum_{t=1}^{\infty}\frac{1}{g(t)}\,.$$

Remark 3.2 (Choice of function g). *Possible choices of function g to ensure that $\sum_{t=1}^{\infty}\frac{1}{g(t)} \leqslant 1$, include for instance $g(t) = t(t+1)$, $g(t) = 3t^{3/2}$, $g(t) = \frac{(t+1)\ln^2(t+1)}{\ln(2)}$ or $g(t) = \frac{(t+2)\ln(t+2)(\ln\ln(t+2))^2}{\ln\ln(3)}$.*

Proof of Lemma 3.10:

It is convenient to introduce the quantity $\overline{M}_t = \frac{\sum_{s\in\{0,\dots,t-1\}} M_{s+1,t}}{g(t)}$ for each $t \in \mathbb{N}_\star$. Since each $M_{s+1,t}$ and $g(t)$ is non-negative, we first get that for every random stopping time $\tau \in \mathbb{N}_\star$, it holds

$$\mathbb{E}\left[\frac{\max_{s<\tau} M_{s+1,\tau}}{g(\tau)}\right] \quad \leqslant \quad \mathbb{E}\left[\overline{M}_\tau\right] = \mathbb{E}\left[\overline{M}_1 + \sum_{t=1}^{\infty}(\overline{M}_{t+1} - \overline{M}_t)\mathbb{I}\{\tau > t\}\right].$$

Further, we note that, conveniently

$$\overline{M}_{t+1} - \overline{M}_t \;=\; \frac{M_{t+1,t+1}}{g(t+1)} + \sum_{s=0}^{t-1}\left(\frac{M_{s+1,t+1}}{g(t+1)} - \frac{M_{s+1,t}}{g(t)}\right).$$

Next, by assumption, we note that $\mathbb{E}[M_{s+1,t+1}|\mathcal{F}_t] \leqslant M_{s+1,t}$. Thus, since $\mathbb{I}\{\tau > t\} \in \mathcal{F}_t$, we deduce that

$$\begin{aligned}\mathbb{E}\left[\frac{\max_{s<\tau} M_{s+1,\tau}}{g(\tau)}\right] &\leqslant \mathbb{E}\big[\overline{M}_1\big] + \sum_{t=1}^{\infty}\frac{\mathbb{E}[M_{t+1,t+1}]}{g(t+1)} + \sum_{t=1}^{\infty}\sum_{s<t}\mathbb{E}\left[\left(\frac{1}{g(t+1)} - \frac{1}{g(t)}\right)M_{s+1,t}\mathbb{I}\{\tau > t\}\right] \\ &= \mathbb{E}\big[\overline{M}_1\big] + \sum_{t=1}^{\infty}\frac{\mathbb{E}[M_{t+1,t+1}]}{g(t+1)} + \sum_{t=1}^{\infty}\sum_{s<t}\left(\frac{1}{g(t+1)} - \frac{1}{g(t)}\right)\underbrace{\mathbb{E}\Big[M_{s+1,t}\mathbb{I}\{\tau > t\}\Big]}_{\geqslant 0}.\end{aligned}$$

Hence, the assumption that g is non-decreasing ensures that the last sum is upper bounded by 0. Since on the other hand $\mathbb{E}[M_{t+1,t+1}] \leqslant 1$ holds for all t (and thus $\mathbb{E}\big[\overline{M}_1\big] \leqslant 1/g(1)$), we deduce that

$$\mathbb{E}\left[\frac{\max_{s<\tau} M_{s+1,\tau}}{g(\tau)}\right] \;\leqslant\; \frac{1}{g(1)} + \sum_{t=1}^{\infty}\frac{1}{g(t+1)} = \sum_{t=1}^{\infty}\frac{1}{g(t)}.$$

□

Theorem 3.2 (*Doubly-time-uniform concentration*) *Let $Y_1, \dots, Y_t$ be a sequence of independent random variables, generating some filtration $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{N}_\star}$. Assume that for some random variables $(M_{s+1:t})_{s,t}$ satisfying the conditions of Lemma 3.10, and for some deterministic real-valued functions $(F_{s,t})_{s,t}$, the following inequalities hold*

$$\forall t \in \mathbb{N}_\star, s \in [0, t-1], \quad F_{s,t}(Y_{s+1}, \dots, Y_t) \leqslant M_{s+1:t}.$$

Then, for all $\delta \in (0,1)$ and any non-decreasing positive g, it holds

$$\mathbb{P}\left(\exists t \in \mathbb{N}_\star, \exists s \in [0, t-1], \quad F_{s,t}(Y_{s+1}, \dots, Y_t) \geqslant \frac{g(t)}{\delta}\right) \leqslant \delta \sum_{t=1}^{\infty}\frac{1}{g(t)}.$$

Theorem 3.2 directly generalizes Lemma 3.9, that is recovered using the following functions

$$F_{s,t}(Y_{s+1}, \dots, Y_t) = \frac{1}{\sqrt{t+1-s}}\exp\left(\frac{(t-s)\overline{z}_{s+1:t}^2}{2\sigma^2(1+\frac{1}{t-s})}\right) \quad \text{and } g(t) = \frac{(t+1)\ln^2(t+1)}{\ln(2)}.$$

Proof of Theorem 3.2:

Indeed, by assumption, we get

$$\begin{aligned}\mathbb{P}\left(\exists t\in\mathbb{N}_\star, \exists s\in[0,t-1],\quad F_{s,t}(Y_{s+1},\dots,Y_t)\geqslant\frac{g(t)}{\delta}\right) &\leqslant \mathbb{P}\left(\exists t\in\mathbb{N}_\star, \exists s\in[0,t-1],\ M_{s+1:t}\geqslant\frac{g(t)}{\delta}\right)\\ &= \mathbb{P}\left(\max_{t\in\mathbb{N}_\star}\max_{s\in[0,t-1]}\frac{M_{s+1:t}}{g(t)}\geqslant\frac{1}{\delta}\right)\\ &\leqslant \delta\mathbb{E}\left[\max_{t\in\mathbb{N}_\star}\max_{s\in[0,t-1]}\frac{M_{s+1:t}}{g(t)}\right].\end{aligned}$$

We then conclude by Lemma 3.10. □

4.2 Non-asymptotic detection delay of sub-Gaussian GLR

We now make use of the confidence bounds in order to tune the GLR change-point detection procedure in the sub-Gaussian setting, that we define now. The next result bounds its detection delay.

$$\mathrm{GLR}^{\text{sub-}\sigma}(t_0,t)=\mathbb{I}\{\exists s\in[t_0:t):\left|\mu_{t_0:s}-\mu_{s+1:t}\right|\geqslant b_{t_0}(s,t,\delta)\}$$

Theorem 3.3 (*Detection delay*) *Let $Y_{t_0},\dots Y_\tau$ be a sequence of τ i.i.d. real-valued random variables with mean μ_1. Let $Y_{\tau+1},\dots Y_t$ be a sequence of $t-\tau$ i.i.d. real-valued random variables with mean μ_2. Consider the procedure GLR$^{\text{sub-}\sigma}$ started at time t_0, run for each subsequent time using $b^{joint}_{t_0}(s,t,\delta)$. Then the following holds under Assumption 3.1:*
(i) With probability higher than $1-\delta$, no change point is detected on the whole time interval $[t_0,\tau]$.
(ii) If the change point that occurs at $\tau+1$ has magnitude $\Delta=|\mu_2-\mu_1|$, it is detected with probability higher than $1-\delta$, with a delay not exceeding $d(t_0,\tau+1,\Delta)$ (that is at time $\leqslant\tau+1+d(t_0,\tau+1,\Delta)$), with

$$\textit{(Delay)}\quad d(t_0,\tau+1,\Delta)=\min\left\{d'\in\mathbb{N}:\, d'>\frac{8\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2x_{d'}}{\delta}\right]}{\left(\Delta^2-\frac{8\sigma^2}{\tau-t_0+1}\ln\left[\frac{2x_{d'}}{\delta}\right]\right)_+}-1\right\},$$

where we introduced the notation $x_d=f(d+\tau-t_0+1)$ with $f(x)=x\sqrt{x+2}\ln(x+2)$ and $(x)_+=\max\{x,0\}$.
(iii) if $\tau=\tau_c$ is undetectable in the sense that no algorithm can detect the change before time τ_{c+1} using only data from time $[\tau_{c-1}+1,\tau_{c+1}]$, where $t_0-1=\tau_{c-1}<\tau_c<\tau_{c+1}$, then the gap must be of magnitude $\Delta\leqslant\overline{\Delta}(\tau_{c-1}+1,\tau_c+1,\tau_{c+1})$ where

$$\textit{(Gap)}\quad \overline{\Delta}(t_0,\tau+1,t)=\sigma\sqrt{\frac{(t-t_0+2)}{(t-\tau)(\tau-t_0+1)}8\ln\left[\frac{2(t-t_0)\sqrt{t-t_0+2}\ln(t-t_0+2)}{\delta}\right]}.$$

Remark 3.3 (Scaling) *It is intuitive that the detection delay $d(t_0,\tau+1,\Delta)$ may not be bounded for change points of too small magnitude. Actually taking $t_0=1$ in Theorem 3.3 shows that when the number of observations τ before the change point and its magnitude Δ satisfy $\Delta<\sigma\sqrt{\frac{8}{\tau}\ln\left(2\tau\sqrt{\tau+2}\ln(\tau+2)/\delta\right)}=\tilde{O}(\frac{\sigma}{\sqrt{\tau}})$, then no change is detected (the detection delay is infinite). Now for larger Δ, Theorem 3.3 shows that the delay of the detection scales essentially as $O(\frac{\sigma^2}{\Delta^2}\ln(\tau+\frac{\sigma}{\Delta}))$. This scaling is order-optimal, by comparison with the*

asymptotic results for the parametric setup, see e.g. Theorem 3.1 in Lai and Xing (2010), after using a Pinsker inequality to bound the Kullback-Leibler divergence. Recall that we ask a bounded time-uniform false alarm probability.

Remark 3.4 (Other work) *Up to our knowledge, it is surprisingly the first time such precise bounds on the detection delay are obtained. This result is coherent with the analysis from (Garreau and Arlot, 2016, Theorem 3.1) in the slightly different batch setting with kernels. Now Theorem 3.3 improves the constants and log scalings. This result contrasts with existing asymptotic results presented in the limit when δ goes to 0 see e.g. Theorem 3.1 in Lai and Xing (2010).*

Remark 3.5 (Extensions) *If we further know that when there is a gap, this gap must be at least of magnitude Δ_0, or that we do not care about detecting gaps of smaller magnitude, then we may add Δ_0 to b_{t_0} on the right hand side of the test, and replace Δ with $\Delta - \Delta_0$ in the definition of d.*

Proof of Theorem 3.3:

---

i) False detection By definition of the detection procedure, a detection occurs a time t if $\exists s \in [t_0 : t)$ such that

$$|\mu_{t_0:s} - \mu_{s+1:t}| \quad > \quad b_{t_0}(s,t,\delta)\,.$$

In the first case (i), since there is no change point before τ, then for all $s, t \leqslant \tau$, $\mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}] = 0$. Now, we observe that, thanks to the uniform concentration inequality, it holds

$$\mathbb{P}\left(\exists t \in \mathbb{N}_\star, s \in [t_0:t), \left|\mu_{t_0:s} - \mu_{s+1:t} - \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}]\right| \geqslant b_{t_0}(s,t,\delta)\right) \leqslant \delta\,.$$

We deduce that on an event of probability higher than $1-\delta$, no detection occurs for any $t \leqslant \tau$.
ii) Detection delay We now turn to the second case (ii). In the sequel, we consider that $t_0 = 1$. On the same event, it holds for all $t > \tau$ and $s < t$,

$$\begin{aligned}
\mu_{t_0:s} - \mu_{s+1:t} &\geqslant \mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}] - b_{t_0}(s,t,\delta) \\
\mu_{s+1:t} - \mu_{t_0:s} &\geqslant \mathbb{E}[\mu_{s+1:t} - \mu_{t_0:s}] - b_{t_0}(s,t,\delta)\,,
\end{aligned}$$

which implies that

$$|\mu_{t_0:s} - \mu_{s+1:t}| \quad \geqslant \quad \left|\mathbb{E}[\mu_{t_0:s} - \mu_{s+1:t}]\right| - b_{t_0}(s,t,\delta)\,.$$

At this point, note that we have the relations

$$\begin{aligned}
\forall t' > \tau, \quad \mu_{t_0:t'} &= \frac{\tau - t_0 + 1}{t' - t_0 + 1}\mu_{t_0:\tau} + \frac{t' - \tau}{t' - t_0 + 1}\mu_{\tau+1:t'} \\
\forall t' < \tau < t, \quad \mu_{t'+1:t} &= \frac{\tau - t'}{t - t'}\mu_{t'+1:\tau} + \frac{t - \tau}{t - t'}\mu_{\tau+1:t}\,.
\end{aligned}$$

Thus, taking the expectation on each side, this means that

$$\mathbb{E}[\mu_{t_0:t'} - \mu_{t'+1:t}] = \begin{cases} \frac{t-\tau}{t-t'}(\mu_1 - \mu_2) & \text{if } t' \leqslant \tau \leqslant t \\ \frac{\tau - t_0 + 1}{t' - t_0 + 1}(\mu_1 - \mu_2) & \text{if } \tau \leqslant t' \leqslant t\,. \end{cases}$$

Using this expressions, we deduce that with probability higher than $1-\delta$, a detection occurs at most at the first time $t > \tau$ such that for some $s < t$,

$$\Big(\frac{t-\tau}{t-s}\mathbb{I}\{s \leqslant \tau\} + \frac{\tau-t_0+1}{s-t_0+1}\mathbb{I}\{s > \tau\}\Big)\Delta > 2b_{t_0}(s,t,\delta)$$

$$\text{where} \quad b_{t_0}(s,t,\delta) = \sigma\sqrt{\Big(\frac{1}{s-t_0+1}+\frac{1}{t-s}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)2\ln\left[\frac{2f(t-t_0)}{\delta}\right]}\,.$$

For $s > \tau$, this corresponds to the condition

$$(\tau-t_0+1)\Delta > 2\sigma\sqrt{\min_{s\in[\tau+1:t-1]}\Big(s-t_0+1+\frac{(s-t_0+1)^2}{t-s}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)2\ln\left[\frac{2f(t-t_0)}{\delta}\right]}$$

which can be simplified into

$$\frac{(\tau-t_0+1)^2\Delta^2}{(\tau-t_0+2)8\sigma^2} > \Big(1+\frac{\tau-t_0+2}{t-\tau-1}\Big)\Big(1+\frac{1}{t-t_0+1}\Big)\ln\left[\frac{2f(t-t_0)}{\delta}\right]$$

Introducing the delay $d = t-(\tau+1)$, it comes

$$\frac{(\tau-t_0+1)^2\Delta^2}{(\tau-t_0+2)8\sigma^2} > \Big(\frac{d+\tau-t_0+3}{d}\Big)\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right]$$

Thus, a detection occurs for the minimal delay $d = t-(\tau+1) \in \mathbb{N}$ (if any) that satisfies

$$d > \frac{8\sigma^2(1+\frac{2}{\tau-t_0+1})\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right]}{\frac{(\tau-t_0+1)\Delta^2}{(\tau-t_0+2)} - \frac{8\sigma^2}{\tau-t_0+1}\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right]}\,. \tag{3.4}$$

We now detail the case of $s \leqslant \tau$ that corresponds to the condition

$$(t-\tau)^2\Delta^2 > 8\sigma^2 \min_{s\in[t_0:\tau]}(t-s)\frac{t-t_0+2}{s-t_0+1}\ln\left[\frac{2f(t-t_0)}{\delta}\right].$$

which shows a detection occurs for the minimal t (if any) that satisfies

$$\frac{(t-\tau)\Delta^2}{8\sigma^2} > \frac{t-t_0+2}{\tau-t_0+1}\ln\left[\frac{2f(t-t_0)}{\delta}\right].$$

Thus, the detection delay must satisfy in this case

$$\frac{(\tau-t_0+1)\Delta^2}{8\sigma^2} > \Big(1+\frac{\tau-t_0+2}{d+1}\Big)\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right],$$

That is

$$d > \frac{8\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right]}{\Delta^2 - \frac{8\sigma^2}{\tau-t_0+1}\ln\left[\frac{2f(d+\tau-t_0+1)}{\delta}\right]} - 1\,. \tag{3.5}$$

Combining inequalities (3.4) and (3.5), the detection delay $d = t - (\tau + 1)$, is not larger than

$$\min\left\{d' \in \mathbb{N}\colon d' \text{ satisfies (3.4) or (3.5)}\right\} \leqslant \min\left\{d' \in \mathbb{N}\colon d' > \frac{8\sigma^2(1+\frac{1}{\tau-t_0+1})\ln\left[\frac{2x_{d'}}{\delta}\right]}{\Delta^2 - \frac{8\sigma^2}{\tau-t_0+1}\ln\left[\frac{2x_{d'}}{\delta}\right]} - 1\right\}.$$

where $x_d = f(d+\tau-t_0+1)$.

iii) Maximal no-detection gap It remains to handle the maximal not-detectable gap. Proceeding with similar steps, we have obtained that with probability higher than $1-\delta$, if a change occurs at $\tau+1$, then a detection occurs at most at the first time $t > \tau$ such that

$$\begin{aligned} \text{either} \quad \frac{\Delta^2}{8\sigma^2} &> \frac{(\tau-t_0+2)}{(\tau-t_0+1)^2}\frac{t-t_0+2}{t-\tau-1}\ln\left[\frac{2f(t-t_0)}{\delta}\right] \\ \text{or} \quad \frac{\Delta^2}{8\sigma^2} &> \frac{t-t_0+2}{(t-\tau)(\tau-t_0+1)}\ln\left[\frac{2f(t-t_0)}{\delta}\right]. \end{aligned}$$

Looking at the minimum of the left-hand side quantities, we deduce that if a change occurring at $\tau+1$ is not detectable, where $\tau = \tau_c$, then the change must be of magnitude $\Delta \leqslant \min\{\overline{\Delta}(\tau_{c-1}+1, \tau_c+1, t), t \in [\tau_c+1, \tau_{c+1}]\}$, where we introduced the quantity

$$\overline{\Delta}(t_0, \tau+1, t) = \sigma\sqrt{\frac{(t-t_0+2)}{(t-\tau)(\tau-t_0+1)}8\ln\left[\frac{2f(t-t_0)}{\delta}\right]}.$$

□

CHAPTER 4
$\mathbb{P}(? \notin \widehat{\Theta}_{n,\delta}) \leqslant \delta$

Contents

1 Kernel regression and the Laplace method 65

2 Markov concentration 71

2.1 Entry-wise concentration 71

2.2 Trajectory-wise concentration 72

3 Forecasters of stationary processes over a finite alphabet 75

Take-home message

Kernel regression in RKHS $\mathcal{K}$

Predictable sequence $(x_t, y_t)_t$ with $y_t = f_\star(x_t) + \xi_t$, ξ_t is sub-Gaussian conditioned on past, $f_\star \in \mathcal{F} \subset \mathcal{K}$.

$$\forall \delta \in [0,1] \quad \mathbb{P}\left(\exists t \in \mathbb{N}, f_\star \notin \widehat{\Theta}_{t,\delta} \right) \leqslant \delta \text{ where}$$

$$\widehat{\Theta}_{t,\delta} = \left\{ f \in \mathcal{F} : \forall x \in \mathcal{X}, \forall t' \leqslant t, |f_\star(x) - f_{\lambda,t'}(x)| \leqslant \sqrt{k_{\lambda,t'}(x,x)} \left[\|f_\star\|_{\mathcal{K}} + \frac{\sigma}{\sqrt{\lambda}} \sqrt{2\ln(1/\delta) + 2\gamma_{t'}(\lambda)} \right] \right\},$$

where we introduced the quantities:

$$\text{(Mean estimate)} \quad f_{\lambda,t}(x) = k_t(x)^\top (\mathbf{K}_t + \lambda I_t)^{-1} Y_t, \quad k_{\lambda,t}(x,x) = k(x,x) - k_t(x)^\top (\mathbf{K}_t + \lambda I_t)^{-1} k_t(x)$$

$$\text{(Information gain)} \quad \gamma_t(\lambda) = \frac{1}{2} \sum_{t'=1}^{t} \ln\left(1 + \frac{1}{\lambda} k_{\lambda,t'-1}(x_{t'}, x_{t'})\right).$$

and notations: $k_t(x) = (k(x, x_{t'}))_{t' \leqslant t}$ is a $t \times 1$ (column) vector and $\mathbf{K}_t = (k(x_s, x_{s'}))_{s,s' \leqslant t}$.
Extension: Estimation of σ, see Maillard (2016).

In this chapter, we focus on the construction of confidence sets, built from finitely many samples of a distribution. The previous chapters have focused on estimating a real-valued quantity such as the mean of a Bernoulli distribution. We consider in this chapter richer distributions, for instance when one unknown vector (or matrix) parameter $\theta \in \Theta$ specifies a process. We deal with the estimation of unknown parameters and build a set $\widehat{\Theta}_{t,\delta}$ from the observations obtained until time t, such that $\theta \in \widehat{\Theta}_{t,\delta}$ holds with probability higher than $1-\delta$. In other words, we want "the unknown" to belong to our empirical set of parameters.

1 Kernel regression and the Laplace method

In this section we consider a large class of problems, that we refer as sequential regression, and applies to situations when the observations are real-valued. At each time step $t \in \mathbb{N}$, a learner picks a point $x_t \in \mathcal{X} \subset \mathbb{R}^d$ and gets the observation

$$y_t = f_\star(x_t) + \xi_t \in \mathbb{R}\,,$$

where $f_\star$ is an unknown function assumed to belong to some function space $\mathcal{F}$, and ξ_t is a random noise. In the following, we assume a sub-Gaussian streaming predictable model:

Assumption 4.1 (Predictability) *The process generating the observations is* predictable *in the sense that there is a filtration $\mathcal{H} = (\mathcal{H}_t)_{t\in\mathbb{N}}$ such that x_t is $\mathcal{H}_{t-1}$-measurable and y_t is $\mathcal{H}_t$-measurable. Such an example is given by $\mathcal{H}_t = \sigma(x_1, \dots, x_{t+1}, y_1, \dots, y_t)$.*

Assumption 4.2 (Sub-Gaussian streaming model) *In the sub-Gaussian streaming predictable model, for some non-negative constant σ^2, the following holds*

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E}\Big[\exp(\gamma \xi_t)\Big|\mathcal{H}_{t-1}\Big] \leqslant \frac{\gamma^2\sigma^2}{2}\,.$$

Hence we consider a quadratic error loss function $\ell(y, y') = \frac{(y-y')^2}{2\sigma^2}$, as it is adapted to the sub-Gaussian assumption (see chapter 1). A typical first approach is to build at each time t an estimate of $f_\star$ that minimizes the quadratic error:

$$\min_{f\in\mathcal{F}} \sum_{t'=1}^{t} \ell(y_{t'}, f(x_{t'})) = \min_{f\in\mathcal{F}} \sum_{t'=1}^{t} \frac{(y_{t'} - f(x_{t'}))^2}{2\sigma^2}\,.$$

However there is in general no unique solution to this minimization problem (think of a high-dimensional $\mathcal{F}$), and it is thus classical to resort to regularization. We consider the general setting of Reproducing Kernel Hilbert Spaces (RKHS), as it captures many other settings (e.g. linear regression) as a special case, and we provide a powerful result to handle estimation error for a standard regularized kernel least-squares estimate.

RKHS Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function (that is continuous, symmetric positive definite) on a compact set $\mathcal{X}$ equipped with a positive finite Borel measure μ, and denote $\mathcal{K}$ the corresponding reproducing kernel Hilbert Space: Indeed under these two conditions, there exists an at most countable sequence $(\sigma_i, \psi_i)_{i\in\mathbb{N}^\star}$ where $\sigma_i \geqslant 0$, $\lim_{i\to\infty} \sigma_i = 0$ and $(\psi_i)_i$ form an orthonormal basis of $L_{2,\mu}(\mathcal{X})$, such that

$$k(x,y) = \sum_{j=1}^{\infty} \sigma_j \psi_j(x)\psi_j(y') \quad \text{and} \quad \mathcal{K} = \left\{ f \in L_{2,\mu}(\mathcal{X}) : \|f\|_{\mathcal{K}} < \infty \right\} \text{ where } \|f\|_{\mathcal{K}}^2 = \sum_{j=1}^{\infty} \frac{\langle f, \psi_j\rangle_{L_{2,\mu}}^2}{\sigma_j}.$$

Introducing $\varphi_i = \sqrt{\sigma_i}\psi_i$, we note that $\|\varphi_i\|_{L_2} = \sqrt{\sigma_i}$, $\|\varphi_i\|_{\mathcal{K}} = 1$, and further that if $f = \sum_i \theta_i \varphi_i$, then $\|f\|_{\mathcal{K}}^2 = \sum_i \theta_i^2$ and $\|f\|_{L_2}^2 = \sum_i \theta_i^2 \sigma_i$. In particular f belongs to the RKHS if and only if $\sum_i \theta_i^2 < \infty$. For $\varphi(x) = (\varphi_1(x), \dots)$ and $\theta = (\theta_1, \dots)$, we denote $\theta^\top \varphi(x)$ for $\sum_{i\in\mathbb{N}} \theta_i \varphi_i(x)$, by analogy with the finite dimensional case. Note that with such notations, $k(x,y) = \varphi(x)^\top \varphi(y)$.

Remark 4.1 *An example when $\mathcal{X} \subset \mathbb{R}^d$ is the linear kernel $k(x,x') = x^\top x'$ that corresponds to the finite-dimensional space $\mathcal{K} = \{f_\theta : f_\theta(x) = \theta^\top x, \theta \in \mathbb{R}^d\}$. Now, in general, an RKHS may be infinite dimensional (think of Sobolev or Besov spaces)*

Given a RKHS $\mathcal{K}$, it is natural to consider the function space $\mathcal{F} = \{f \in \mathcal{K} : \|f\|_{\mathcal{K}} \leqslant B\}$ for some positive constant $B \in \mathbb{R}_+$, and our goal is to estimate "the unknown", that is the function $f_\star$, assuming it belongs to $\mathcal{F}$. This can be done by building an appropriate estimate of $f_\star$, called a regularized least-squares kernel estimate, and adapting the Laplace method of mixture to RKHS, as we explain now:

Theorem 4.1 (*Streaming kernel least-squares*) *Assume we are in the sub-Gaussian streaming predictable model. For a fixed regularization parameter $\lambda \in \mathbb{R}_+$, let us define the posterior mean and variances after observing $Y_t = (y_1, \dots, y_t)^\top \in \mathbb{R}^{t\times 1}$ as*

$$\begin{cases} f_{\lambda,t}(x) = & k_t(x)^\top (\mathbf{K}_t + \lambda I_t)^{-1} Y_t \\ s^2_{\lambda,t}(x) = & \frac{\sigma^2}{\lambda} k_{\lambda,t}(x,x) \text{ with } k_{\lambda,t}(x,x) = k(x,x) - k_t(x)^\top (\mathbf{K}_t + \lambda I_t)^{-1} k_t(x)\,. \end{cases}$$

where $k_t(x) = (k(x, x_{t'}))_{t'\leqslant t}$ is a $t \times 1$ (column) vector and $\mathbf{K}_t = (k(x_s, x_{s'}))_{s,s'\leqslant t}$. Then $\forall \delta \in [0,1]$, with probability higher than $1-\delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geqslant 0$,

$$|f_\star(x) - f_{\lambda,t}(x)| \leqslant \sqrt{\frac{k_{\lambda,t}(x,x)}{\lambda}} \left[\sqrt{\lambda}\,\|f_\star\|_{\mathcal{K}} + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_t(\lambda)}\right],$$

where the quantity $\gamma_t(\lambda) = \frac{1}{2}\sum_{t'=1}^{t} \ln\left(1 + \frac{1}{\lambda} k_{\lambda,t'-1}(x_{t'}, x_{t'})\right)$ *is the information gain.*

Before proceeding with the proof, let us provide some remarks about this result.

Remark 4.2 *This result should be considered as an extension of Abbasi-Yadkori et al. (2011, Theorem 2) from finite-dimensional to possibly infinite dimensional function space. More specifically, when considering the linear kernel, the result of Theorem 4.1 recovers exactly Theorem 2 from Abbasi-Yadkori et al. (2011). The generalization is non trivial as the Laplace method must be amended in order to be applied beyond the linear case.*

Remark 4.3 *This result holds uniformly over all $x \in \mathcal{X}$ and most importantly over all $t \geqslant 0$, thanks to a random stopping time construction (related to the occurrence of* bad events*) and a self-normalized inequality handling this stopping time. This is in contrast with results such as Wang and de Freitas (2014), that are only stated separately for each t.*

Information gain This quantity measures the information obtained about function $f_\star$ by sampling at points $(x_1, \dots, x_t)$. It is defined (Cover and Thomas, 1991) as the *mutual information* between $f_\star$ and the observations $(y_1, \dots, y_t)$:

$$I(y_1, \dots, y_t; f_\star) = H(y_1, \dots, y_t) - H(y_1, \dots, y_t | f_\star),$$

that is the difference between the *marginal entropy* and the *conditional entropy* of the distributions of observations. The information gain thus quantifies the reduction of uncertainty about $f_\star$ following these observations. For a multidimensional Gaussian, we have $H(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2} \ln |2\pi e \boldsymbol{\Sigma}|$, such that for $\lambda = \sigma^2$ (Srinivas et al., 2010),

$$\gamma_t(\sigma^2) = I(y_1, \dots, y_t; f_\star) = \frac{1}{2} \ln \det(I_t + \sigma^{-2} \mathbf{K}_t),$$

where $\mathbf{K}_t = (k(s, s'))_{s,s' \leqslant t}$. In the linear case when $k(x, x') = x^\top x'$ for $x \in \mathbb{R}^d$, the information gain typically scales as $\gamma_t(\sigma^2) = \mathcal{O}(d \ln t)$ (Srinivas et al., 2010). The information gain can be shown to scale with the *effective dimensionality* (Valko et al., 2013) instead of the dimension, where effective dimensions correspond to the most informative ones. More effective dimensions require more observations for a good space coverage, which increases the information gain. We extend the information gain to any regularization λ.

Definition 4.3 (*Information gain with unknown variance*) *We define the information gain at time t for a regularization parameter λ to be*

$$\gamma_t(\lambda) = \frac{1}{2} \sum_{t'=1}^{t} \ln \left(1 + \frac{1}{\lambda} k_{\lambda, t'-1}(x_{t'}, x_{t'}) \right).$$

This generalization is natural in view of Theorem 4.1. The information gain is inversely proportional to the regularization λ. By controlling the flexibility of the regression model, the regularization limits the impact of a new observation on the resulting model, therefore limiting the information that can be gained out of it.

The following Martingale control is a key component of the analysis.

Lemma 4.1 (*Hilbert Martingale Control*) *Assume that the noise sequence $\{\xi_t\}_{t=0}^\infty$ is conditionally σ^2-sub-Gaussian*

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E}[\exp(\gamma \xi_t) | \mathcal{H}_{t-1}] \leqslant \frac{\gamma^2 \sigma^2}{2}.$$

Let τ be a stopping time with respect to the filtration $\{\mathcal{H}_t\}_{t=0}^\infty$ generated by the variables $\{x_t, \xi_t\}_{t=0}^\infty$. For any $\mathbf{q} = (q_1, q_2, \dots)$ such that $\mathbf{q}^\top \varphi_i(x) = \sum_{i \in \mathbb{N}} q_i \varphi(x) < \infty$, and deterministic positive λ, let us denote

$$M_{m,\lambda}^{\mathbf{q}} = \exp \left(\sum_{t=1}^{m} \frac{\mathbf{q}^\top \varphi(x_t)}{\sqrt{\lambda}} \xi_t - \frac{\sigma^2}{2} \sum_{t=1}^{m} \frac{(\mathbf{q}^\top \varphi(x_t))^2}{\lambda} \right)$$

Then, for all such $\mathbf{q}$ the quantity $M_{\tau,\lambda}^{\mathbf{q}}$ is well defined and satisfies

$$\ln \mathbb{E}[M_{\tau,\lambda}^{\mathbf{q}}] \leqslant 0.$$

Proof of Lemma 4.1:

The only difficulty in the proof is to handle the stopping time. Indeed, for all $m \in \mathbb{N}$, thanks to the conditional σ-sub-Gaussian property, it is immediate to show that $\{M^{\mathfrak{q}}_{m,\lambda}\}_{m=0}^{\infty}$ is a non-negative super-martingale and actually satisfies $\ln \mathbb{E}[M^{\mathfrak{q}}_{m,\lambda}] \leqslant 0$.

By the convergence theorem for nonnegative super-martingales, $M^{\mathfrak{q}}_{\infty} = \lim_{m\to\infty} M^{\mathfrak{q}}_{m,\lambda}$ is almost surely well-defined, and thus $M^{\mathfrak{q}}_{\tau,\lambda}$ is well-defined (whether $\tau < \infty$ or not) as well. In order to show that $\ln \mathbb{E}[M^{\mathfrak{q}}_{\tau,\lambda}] \leqslant 0$, we introduce a stopped version $Q^{\mathfrak{q}}_m = M^{\mathfrak{q}}_{\min\{\tau,m\},\lambda}$ of $\{M^{\mathfrak{q}}_{m,\lambda}\}_m$. Now $\mathbb{E}[M^{\mathfrak{q}}_{\tau,\lambda}] = \mathbb{E}[\liminf_{m\to\infty} Q^{\mathfrak{q}}_m] \leqslant \liminf_{m\to\infty} \mathbb{E}[Q^{\mathfrak{q}}_m] \leqslant 1$ by Fatou's lemma, which concludes the proof. We refer to (Abbasi-Yadkori et al., 2011) for further details. □

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1:

Decomposition step Let N be a random stopping time for the filtration generated by the observations. We let $\Phi_N = (\varphi(x_t))_{t\leqslant N}$ be an $N \times \infty$ matrix and introduce the bi-infinite matrix $V_N = I + \frac{1}{\lambda}\Phi_N^\top \Phi_N$ as well as the noise vector $E_N = (\xi_1, \dots, \xi_N)$. In order to control the term $|f^\star(x) - f_{k,N}(x)|$, we now use the following easy-to-derive decomposition

$$|f^\star(x) - f_{k,N}(x)| \leqslant \frac{1}{\sqrt{\lambda}} \|\varphi(x)\|_{V_N^{-1}} \Big[\Big\| \frac{1}{\sqrt{\lambda}} \Phi_N^\top E_N \Big\|_{V_N^{-1}} + \sqrt{\lambda}\, \|\theta^\star\|_{V_N^{-1}} \Big] ,$$

which is valid provided that all terms involved are finite[1]: Indeed, using the feature map, it holds

$$\begin{aligned} f_{k,N}(x) &= k_N(x)^\top (\mathbf{K_N} + \lambda I_N)^{-1} Y_N \\ &= \varphi(x)^\top \mathbf{\Phi}_N^\top (\mathbf{\Phi}_N \mathbf{\Phi}_N^\top + \lambda I_N)^{-1} Y_N \\ &= \varphi(x)^\top \mathbf{\Phi}_N^\top \left(\frac{I_N}{\lambda} - \frac{1}{\lambda} \mathbf{\Phi}_N \big(\lambda I + \mathbf{\Phi}_N^\top \mathbf{\Phi}_N\big)^{-1} \mathbf{\Phi}_N^\top \right) Y_N \\ &= \varphi(x)^\top (\mathbf{\Phi}_N^\top \mathbf{\Phi}_N + \lambda I)^{-1} \mathbf{\Phi}_N^\top (\mathbf{\Phi}_N \theta^\star + E_N) \end{aligned}$$

where in the third line, we used the Shermann-Morrison formula. From this, simple algebra yields

$$f_{k,N}(x) - f^\star(x) = \frac{1}{\lambda} \varphi(x)^\top V_N^{-1} \big(\mathbf{\Phi}_N^\top E_N - \lambda \theta^\star \big) .$$

We obtain the claim from a simple Holder inequality using the appropriate matrix norm.

Bounding each term (geometry) Now, we note that a simple application of the Shermann-Morrison formula yields

$$\frac{1}{\lambda} \|\varphi(x)\|^2_{V_N^{-1}} = \frac{\sigma^2_{k,N}(x)}{\sigma^2} = k_N(x,x) .$$

[1] The right way to do so is first to replace all infinite sequences with their d first components, for each $d \in \mathbb{N}$, then check the validity of the bound for a each d, and finally that all the limiting quantities make sense.

On the other hand since $\|\theta^\star\|_2 < \infty$ then

$$\sqrt{\lambda}\,\|\theta^\star\|_{V_N^{-1}} \leqslant \sqrt{\lambda}\,\|\theta^\star\|_2 = \sqrt{\lambda}\,\|f^\star\|_{\mathcal{K}}\,.$$

Bounding the last term (concentration) In order to control the remaining term $\left\| \frac{1}{\sqrt{\lambda}}\Phi_N^\top E_N \right\|_{V_N^{-1}}$, we resort to the Laplace method. To this end, we introduce the quantity $M_{m,\lambda}^{\mathbf{q}}$ from Lemma 4.1, and in order to integrate over $\mathbf{q}$, we introduce $Q \sim \mathcal{N}(0, I)$ to be an infinite standard Gaussian random sequence which is independent of all other random variables. We denote $Q^\top\varphi(x) = \sum_{i\in\mathbb{N}} Q_i\varphi_i(x)$. This is justified since for all x, $k(x,x) = \sum_{i\in\mathbb{N}}\varphi_i^2(x) < \infty$ and thus $\mathbb{V}(Q^\top\varphi(x)) < \infty$. Hence, like for the Laplace method in dimension 1, we define $M_{m,\lambda} = \mathbb{E}[M_{m,\lambda}^Q]$. We still have the key property $\mathbb{E}[M_{N,\lambda}] = \mathbb{E}[\mathbb{E}[M_{m,\lambda}^Q]|Q] \leqslant 1$.

Let $V_N = I + \frac{1}{\lambda}\Phi_N^\top\Phi_N$. Elementary algebra gives

$$\begin{aligned}\det(V_N) &= \det(V_{N-1} + \frac{1}{\lambda}\varphi(x_t)\varphi(x_t)^\top) = \det(V_{N-1})(1 + \frac{1}{\lambda}\,\|\varphi(x_t)\|_{V_{t-1}^{-1}}^2) \\ &= \det(V_0)\prod_{t=1}^N\left(1 + \frac{1}{\lambda}\,\|\varphi(x_t)\|_{V_{t-1}^{-1}}^2\right),\end{aligned}$$

where we used the fact that the eigenvalues of a matrix of the form $I + xx^\top$ are all ones except for the eigenvalue $1 + \|x\|^2$ corresponding to x. Then, note that $\det(V_0) = 1$ and thus

$$\begin{aligned}\ln(\det(V_N)) &= \sum_{t=1}^N \ln\left(1 + \frac{1}{\lambda}\,\|\varphi(x_t)\|_{V_{t-1}^{-1}}^2\right) \\ &= \frac{1}{2}\sum_{t=1}^N \ln\Big(1 + k_{t-1}(x_t,x_t)\Big)\,.\end{aligned}$$

In particular, $\ln(\det(V_N))$ is finite. The only difficulty in the proof is now to handle the possibly infinite dimension. To this end, it is enough to take a look at the approximations using the d first element of the sequence for each d. We note $Q_d, M_{N,\lambda}, \Phi_{N,d}$ and $V_{N,d}$ the restriction of the corresponding quantities to the components $\{1,\dots,d\}$. Note that Q_d is Gaussian $\mathcal{N}(0, I_d)$. Following similar steps from Abbasi-Yadkori et al. (2011), we obtain that

$$M_{m,d,\lambda} = \frac{1}{\det(V_{m,d})^{1/2}}\exp\left(\frac{1}{2\lambda}\,\left\|\,\Phi_{m,d}^\top E_m\,\right\|_{V_{m,d}^{-1}}^2\right).$$

Note also that $\mathbb{E}[M_{N,d,\lambda}] \leqslant 1$ for all $d \in \mathbb{N}$. Thus, by an application of Fatou's lemma, it holds that

$$\begin{aligned}\mathbb{P}\left(\lim_{d\to\infty}\frac{\left\|\,\Phi_{N,d}^\top E_N\,\right\|_{V_{N,d}^{-1}}^2}{2\log\Big(\det(V_{N,d})^{1/2}/\delta\Big)} > 1\right) &\leqslant \mathbb{E}\left[\lim_{d\to\infty}\frac{\delta\exp\left(\frac{1}{2\lambda}\,\left\|\,\Phi_{N,d}^\top E_N\,\right\|_{V_{N,d}^{-1}}^2\right)}{\det(V_{N,d})^{1/2}}\right] \\ &\leqslant \delta\lim_{d\to\infty}\mathbb{E}[M_{N,d,\lambda}] \leqslant \delta\,.\end{aligned}$$

Finally, using the above application of Laplace method to the control of the self-normalized term $\left\| \frac{1}{\sqrt{\lambda}} \Phi_N^\top E_N \right\|_{V_N^{-1}}$, and combining it with the previous remarks we obtain that

$$\mathbb{P}\left(\exists x \in \mathcal{X}, |f^\star(x) - f_{k,N}(x)| \geqslant k_N(x,x)^{1/2} \left[\sqrt{2\sigma^2 \ln\left(\frac{\det(V_N)^{1/2}}{\delta} \right)} + \sqrt{\lambda} \left\| f^\star \right\|_{\mathcal{K}} \right] \right) \leqslant \delta \,.$$

In order to get the result uniformly for all t, we simply pick the random stopping time N to be the first time t such that the threshold is crossed:

$$N = \min \left\{ t \in \mathbb{N} : \exists x \in \mathcal{X}, |f^\star(x) - f_{k,t}(x)| \geqslant k_t(x,x)^{1/2} \left[\sqrt{2\sigma^2 \ln\left(\frac{\det(V_t)^{1/2}}{\delta} \right)} + \sqrt{\lambda} \left\| f^\star \right\|_{\mathcal{K}} \right] \right\} .$$

□

Extensions We have seen how to build a high-probability confidence estimate for the function $f_\star$. It is given in the form of the set

$$\widehat{\Theta}_{t,\delta} = \left\{ f \in \mathcal{F} : \forall x \in \mathcal{X}, \forall t' \leqslant t, |f_\star(x) - f_{\lambda,t'}(x)| \leqslant \sqrt{\frac{k_{\lambda,t'}(x,x)}{\lambda}} \left[\sqrt{\lambda} \left\| f_\star \right\|_{\mathcal{K}} + \sigma \sqrt{2\ln(1/\delta) + 2\gamma_{t'}(\lambda)} \right] \right\},$$

and satisfies by construction that for all $\delta \in [0,1]$, $\quad \mathbb{P}\Big(\exists t \in \mathbb{N}, f_\star \notin \widehat{\Theta}_{t,\delta} \Big) \leqslant \delta$.

As we see, this set depends on quantities such as the sub-Gaussian constant σ and the regularization parameter λ. We study in Durand et al. (2017) how to extend these results when σ is unknown and one wants to adapt the parameter λ sequentially, that is, depending on past observations. We also derive complementary results in Maillard (2016) for an ordinary least-squares estimates, and for variance estimation.

Some numerical illustration We conclude this section by providing an illustrative numerical experiment that enables to visualize the empirical confidence intervals that is built from this analysis, and how fast it shrinks. More precisely, we consider here a time series when $x_t = t$ for all t. We plot at each time t the confidence interval built from all observations gathered before time t, and instantiated on the observation point $x_t = t$. Hence by doing so, we see a collection of confidence intervals that progressively shrinks as the number of observations increases.

We consider for the purpose of illustration a low-dimension function space built from a combinations of a few sinus functions (of the form $f_\theta(t) = \theta^\top \varphi(t)$, where $\varphi_i(t) = \sin(2^i t)$ for each $i = 0, \dots, 3$, however the actual function space does not really matter).

In Figure 4.1, we consider the confidence intervals computed from Theorem 4.1, and study the influence of the estimation of the noise parameter σ on the resulting bounds. We plot in Orange the confidence interval when a bound on σ is given (Here $\sigma = 3$, and the actual noise level has been chosen uniformly in $[0,3]$). We plot in Red (respectively Yellow) the intervals using for σ the upper bound (respectively lower bound) provided in Maillard (2016), when no bound on σ is known. Despite the fact the noise level is completely unknown, the confidence interval shrinks reasonably fast. Both the yellow and green intervals correspond to other variants that can be qualified as being "optimistic". Similar figures can be obtained when studying the estimation of σ in the context of *ordinary* rather than regularized least-squares estimates.

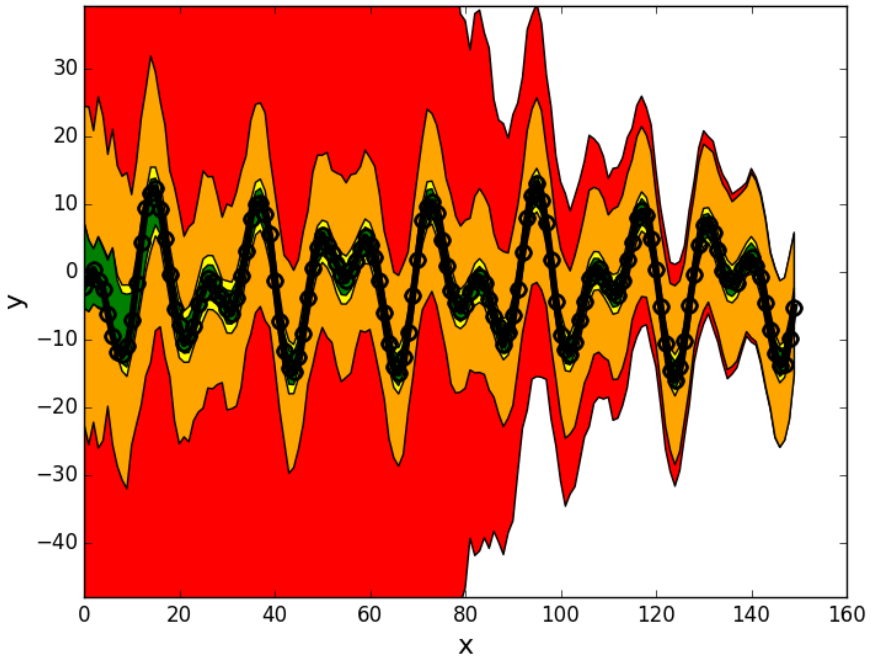


Figure 4.1: Confidence intervals built for the regularized least-squares estimate with $\lambda = 1$ and various methods to handle the noise. Orange: Bound built from Theorem 4.1 when σ is given to the learner. Red: Upper-bound built without any knowledge of the noise (see Maillard (2016)). Green and Yellow: other "optimistic" variants introduced in Maillard (2016).

2 MARKOV CONCENTRATION

In this section, we now turn to a second wide class of setups, when observations are no longer real-valued, but instead belong to a discrete set Σ of symbols. In this context, the traditional approach is to consider processes that are m-order Markov (where m is the size of the memory). Controlling estimation is not trivial in this setup due to the dependencies induced by the memory. We show below a possible way to leverage this structure.

More precisely, we consider a Markov chain of order m, such that the law of Y^t only depends on the m most recent observations before t, which we denote $p(\cdot|Y^{t-1}, \dots, Y^{(\max\{t-m,1\})})$. For a word $w = w_1 \dots w_\ell \in \Sigma^t$ of length $t \in \mathbb{N}$, with symbols $w_i \in \Sigma$ we denote the probability of emission of w by

$$\overline{p}(w) = \prod_{t'=m+1}^{t} p(Y^{(t')} = w_{t'}|w_{t'-1}, \dots, w_{t'-m}) \prod_{i=1}^{m} p(Y^{(i)} = w_i|w_{i-1}, \dots, w_1)\,.$$

For a set $U \subset \Sigma^\star$, where $\Sigma^\star$ denotes the set of all finite words on Σ, we also denote $\overline{p}(Uw) = \sum_{u\in\mathcal{U}} \overline{p}(uw)$.

2.1 Entry-wise concentration

In this section, we first consider the terms of the decomposition separately, starting from the simplest case when $m = 0$. When dealing with probability distributions over a finite alphabet, the following adaptation of

the Laplace method is useful. We first provide below an easy extension of the Laplace method to the control of a discrete distribution generating i.i.d. observations ($m = 0$), obtained by a combination with the method of types (see e.g. Weissman et al. (2003) for the original bound without the Laplace method).

Corollary 4.1 (*Weissman-Laplace concentration*) *For any random stopping time τ with respect to the filtration of the past observations, and any discrete distribution p on $\mathcal{S}$ with support of size $K \leqslant |\mathcal{S}|$,*

$$\mathbb{P}\left(\|p_\tau - p\|_1 \geqslant \sqrt{\frac{2(1+\frac{1}{\tau}) \log\left(\sqrt{\tau+1}\frac{2^K-2}{\delta}\right)}{\tau}} \right) \leqslant \delta \,.$$

Proof of Corollary 4.1:

For discrete measures, it holds $\mathbb{P}\left(\|p_\tau - p\|_1 \geqslant \varepsilon \right) \leqslant \sum_{B \subset \mathcal{S}} \mathbb{P}\left(p_\tau(B) - p(B) \geqslant \frac{1}{2}\varepsilon \right)$. We then remark that there are only $2^K - 2$ sets B (all the sets but the support of p and $\emptyset$) for which the contribution to left hand side term is non 0. This is called the method of types. We conclude by applying the Laplace method to the control of each random variable $\mathbb{I}\{X \in B\}$, as it is in $[0, 1]$. □

This result naturally extends to Markov m models. Let $w \in \Sigma^\star$ be a word of length $\ell \geqslant m$. The previous result then applies to the conditional distribution $p(\cdot|w)$ with support of size $K(w) \leqslant |\Sigma|$ and its estimate $p_t(\cdot|w)$ based on $N_t(w)$ observations at time t, leading to a time-uniform concentration

$$\mathbb{P}\left(\exists t \in \mathbb{N}, \quad \|p_t(\cdot|w) - p(\cdot|w)\|_1 \geqslant \sqrt{\frac{2(1+\frac{1}{N_t(w)}) \log\left(\sqrt{N_t(w)+1}\frac{2^{K(w)}-2}{\delta}\right)}{N_t(w)}} \right) \leqslant \delta \,.$$

Remark 4.4 *The method of type is generally a bit crude, and one may prefer to substitute it by individual Bernstein concentration bounds for each $p_t(s|w)$, seen as an average of $N_t(w)$ i.i.d Bernoulli random variables.*

2.2 Trajectory-wise concentration

In this section, we turn to a different control that targets $\overline{p}$ instead of $p(\cdot|w)$. In general, such a control requires some mixing properties (which we revisit in chapter 8). However, a martingale difference approach can be provided without resorting to any kind of mixing, as we explain now. The following result provides a simple concentration inequality for the frequency estimate of a word w in a sequence of observations that is Markov of order m.

Lemma 4.2 (*Hoeffding concentration for Markov processes*) *Let $x_1, \dots, x_\ell$ be a sequence of ℓ symbols generated by a Markov chain of order m, where ℓ is deterministic. Let $w \in \Sigma^\star$ be a given word and define the following Bernoulli random variables,*

$$\forall s \in [1, \ell - |w| + 1], b_{s,w} = \mathbb{I}\{x_s \dots x_{s+|w|-1} = w\} \text{ and } \forall s > \ell - |w| + 1, b_s = 0\,.$$

Let $\mathbb{E}_{<s}$ denotes the conditional expectation on the variables $x_1, \dots, x_{s-1}$. Then for all $\delta \in [0, 1]$, it holds

$$\mathbb{P}\left[\frac{1}{t}\sum_{s=1}^{t}(b_{s,w} - \mathbb{E}_{<s}[b_{s,w}]) \geqslant \sqrt{\frac{(m + |w|)\ln((m + |w|)/\delta)}{2t}}\right] \leqslant \delta\,.$$

Note that $\mathbb{E}_{<s}[b_{s,w}] = \overline{p}(x_1 \dots x_{s-1}w)$ is the probability of observing w following the sequence of observations. When $m = 0$ and $|w| = 1$, we recover precisely the Hoeffding inequality for an i.i.d. sequence of bounded random variables (here $\mathbb{I}\{x_s = w\}$). There is no difficulty in extending this result to other concentration bounds and peeling or Laplace techniques.

The previous result shows how cumulative errors of occurrence estimate can be controlled. However, it gives no clue at the rate of estimation of $\overline{p}(w)$ for a given word w. This in general requires resorting to mixing properties of the chain. We make use of a powerful result from Kontorovich and Weiss (2014) later in Chapter 8 and Balle and Maillard (2017), in order to address this problem, in the context of stochastic languages.

Proof of Lemma 4.2:

We note that the random variables $(b_s)_s$ are dependent. However, if the sequence of observations is Markov m, then $x_{s+|w|+m}$ is independent from $x_{s+|w|-1}$ conditionally on $x_{s+|w|}, \dots, x_{s+|w|+m-1}$, and thus b_s is independent from $b_{s+|w|+m}$ conditionally on the same sequence. Thus, if we let $k = |w| + m$, then b_{jk+1} and $b_{(j+1)k+1}$ are independent conditionally on $x_{jk+1+|w|}, \dots, x_{(j+1)k}$. Note also that by the Markov assumption, $b_{(j+1)k+1}$ does not depend on the observations before $(j + 1)k - m = jk + 1 + |w|$, thus $\mathbb{E}[b_{(j+1)k+1}|x_{(j+1)k-m}, \dots x_1] = \mathbb{E}[b_{(j+1)k+1}] = \overline{p}(\Sigma^{s-1}w)$.

More generally, we consider the shifted decompositions $B^k_{j,i_0} = b_{jk+1+i_0}$, $j \in [0 : J_{k,i_0} - 1]$ for each $i_0 \in [0, k-1]$, where $J_{k,i_0} = \lfloor (t - i_0 - 1)/k \rfloor + 1$. By construction, it holds $(J_{k,k-1} - 1)k + 1 + k - 1 = \lfloor (t-k)/k \rfloor k + k \leqslant t$ and $\sum_{i_0=0}^{k-1} J_{k,i_0} = t$, as well as

$$\sum_{s=1}^{t} b_s = \sum_{i_0=0}^{k-1}\sum_{j=0}^{J_{k,i_0}-1} B^k_{j,i_0}\,.$$

Also by construction, B^q_{j+1,i_0} is independent from B^q_{j,i_0} conditionally on the random variables on which B^q_{j,i_0} depends, and thus forms a weakly dependent sequence of observations. Thus, we can apply a concentration result.

Applying a standard concentration inequality k times, each with J_{k,i_0} summands we can recover a concentration inequality for $\sum_{s=1}^{t} b_s$. Since the concentration of a sum of ℓ Bernoulli random variables to its means typically scales with $O(\sqrt{\ell})$, it is interesting to note that

$$\frac{1}{t}\sum_{i_0=0}^{k-1}\sqrt{J_{k,i_0}} \leqslant \sqrt{\frac{k}{t}}\,. \tag{4.1}$$

Indeed, by Jensen's inequality, followed by the construction of the J_{k,i_0}, it comes

$$\frac{1}{t}\sum_{i_0=0}^{k-1}\sqrt{J_{k,i_0}} \leqslant \frac{k}{t}\sqrt{\frac{1}{k}\sum_{i_0=0}^{k-1} J_{k,i_0}} = \frac{k}{t}\sqrt{\frac{t}{k}}$$

We will show that for each $i_0 \in [0, k-1]$, for all $\delta \in [0,1]$, then

$$\mathbb{P}\Big[\sum_{j=0}^{J_{k,i_0}-1}(B^k_{j,i_0} - \mathbb{E}_{<jk+i_0+1}[B^k_{j,i_0}]) \geqslant \sqrt{\frac{J_{k,i_0}}{2}\ln(1/\delta)}\Big] \leqslant \delta\,.$$

We then conclude by combining these k inequalities with a union bound and using (4.1).

To this end, we note that $B^k_{j,i_0} \in [0,1]$. In order to handle the dependency between the random variables, let $Z_j = B^k_{j,i_0} - \mathbb{E}_{<jk+i_0+1}[B^k_{j,i_0}]$. We remark that for all $\lambda > 0$,

$$\begin{aligned}
&\mathbb{E}[\exp(\lambda Z_{j+1} + \lambda\sum_{j'=1}^{j} Z_j)|x_1, \dots x_{jk+i_0+|w|}] \\
&= \mathbb{E}[\exp(\lambda Z_{j+1})|x_1, \dots x_{jk+i_0+|w|}]\exp(\lambda\sum_{j'=1}^{j} Z_j) \\
&= \mathbb{E}[\exp(\lambda Z_{j+1})|x_1, \dots x_{(j+1)k+i_0-m}]\exp(\lambda\sum_{j'=1}^{j} Z_j) \\
&= \mathbb{E}[\mathbb{E}[\exp(\lambda Z_{j+1})|x_1, \dots, x_{(j+1)k+i_0}]|x_1, \dots x_{(j+1)k+i_0-m}]\exp(\lambda\sum_{j'=1}^{j} Z_j) \\
&= \mathbb{E}[\mathbb{E}_{<(j+1)k+i_0+1}[\exp(\lambda Z_{j+1})]|x_1, \dots x_{(j+1)k+i_0-m}]\exp(\lambda\sum_{j'=1}^{j} Z_j)\,,
\end{aligned}$$

where the second equality is because $\{Z_{j'}\}_{j'\leqslant j}$ are measurable function of $x_1, \dots x_{jk+i_0+|w|}$ and the third and fourth ones by definition of k. Now, since Z_{j+1} is a shifted Bernoulli variable that is conditionally centered ($\mathbb{E}_{<(j+1)k+i_0+1}[Z_{j+1}] = 0$), then $\mathbb{E}_{<(j+1)k+i_0+1}[\exp(\lambda Z_{j+1})]$ is controlled by a standard argument by $\lambda^2/8$ (it is $1/2$-sub-Gaussian). Hence, we conclude that

$$\mathbb{E}[\exp(\lambda\sum_{j=1}^{J} Z_j)] \leqslant \exp(\lambda^2 J/8)\,.$$

This in turn leads, by a classical Chernoff argument, to

$$\begin{aligned}
\mathbb{P}\Big[\sum_{j=0}^{J_{k,i_0}-1}(B^k_{j,i_0} - \mathbb{E}_{<jp+i_0+1}[B^k_{j,i_0}]) \geqslant \varepsilon\Big] &\leqslant \min_{\lambda>0}\exp(-\lambda\varepsilon + \lambda^2 J_{k,i_0}/8) \\
&= \exp(-2\varepsilon^2/J_{k,i_0})\,.
\end{aligned}$$

□

3 FORECASTERS OF STATIONARY PROCESSES OVER A FINITE ALPHABET

Before concluding this chapter, we describe for completeness a strong family of forecasters for stationary processes over a discrete set. Hence, in this section, we consider that the observation space $\mathcal{Y}$ is a finite alphabet of size S. Although we do not provide confidence sets, we recall their fundamental cumulative regret minimization guarantees.

Definition 4.6 (*Stationary process*) *A stochastic process μ on $\mathcal{Y}^\star$ is stationary if for all $i, m \in \mathbb{N}$ and all $A \in \mathcal{A}_m$, where $\mathcal{A}_m$ is the σ-algebra on $\mathcal{Y}^m$, it holds $\mu(Y_i, \dots, Y_{i+m} \in A) = \mu(Y_1, \dots, Y_m \in A)$.*

Definition 4.9 (*Expert*) *An expert f is a sequence of functions $\mathbf{f}_t : \mathcal{Y}^{t-1} \to \mathcal{P}(\mathcal{Y})$, such that having seen the observations $y^{t-1} \in \mathcal{Y}^{t-1}$, expert f outputs the probability vector $\mathbf{f}_t(\cdot|y^{t-1}) \in \mathcal{P}(\mathcal{Y})$. The probability assigned to observation $y \in \mathcal{Y}$ is the value of the y^{th}-component, that is $f_t(y|y^{t-1}) \in [0,1]$.*

Definition 4.12 (*Self-information loss*) *We measure the loss of a forecaster f sequentially predicting a sequence $y^n = (y_1, \dots, y_n)$ by the cumulative self-information loss of its prediction process:*

$$L(f, y^n) = \sum_{t=1}^{n} \ell(y_t, \mathbf{f}_t(\cdot|y^{t-1})) = \sum_{t=1}^{n} -\ln(f_t(y_t|y^{t-1}))\,,$$

where the self-information loss $\ell : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$ is defined by $\ell(y,p) = -\log(p(y))$. Further, it holds

$$L(f, y^n) = -\ln(f(y^n)) \text{ where } \quad f(y^n) := f_1(y_1) \prod_{t=2}^{n} f_t(y_t|y^{t-1})\,.$$

Let us remark that for each $n \in \mathbb{N}$, the Kullback-Leibler divergence between the learning process f and μ on $\mathcal{Y}^n$ can be interpreted as the expected regret of the forecaster f:

$$\mathtt{KL}_n(\mu, f) \;:=\; \mathbb{E}_\mu \ln \frac{\mu(Y^n)}{f(Y_n)} = \mathbb{E}_\mu[L(f, Y^n) - L(\mu, Y^n)]\,.$$

This is also known as the expected redundancy of f.

Definition 4.15 (*Consistent forecaster*) *We say that a forecaster f is consistent if $\lim_{n\to\infty} \frac{1}{n} KL_n(\rho, \mu_f) = 0$.*

We are now ready to introduce popular forecasters for stationary processes. For convenience, for a sequence of observations $(y_1, \dots, y_n) \in \mathcal{Y}^n$ of length n, and $t_1 \leqslant t_2 \leqslant n$, we denote $y_{t_1..t_2} \in \mathcal{Y}^{t_2-t_1+1}$ the sequence $(y_{t_1}, \dots, y_{t_2})$. For two sequences $\mathbf{u} = u_{1..m} \in \mathcal{Y}^m$ and $\mathbf{v} = v_{1..l} \in \mathcal{Y}^l$ we denote their concatenation by $\mathbf{uv} \in \mathcal{Y}^{m+l}$, and by $\nu_{\mathbf{u}}(\mathbf{v})$ the rate at which $\mathbf{v}$ occurs in $\mathbf{u}$, that is

$$\nu_{\mathbf{u}}(\mathbf{v}) := \frac{N_{\mathbf{u}}(\mathbf{v})}{m-l+1} \text{ where } N_{\mathbf{u}}(\mathbf{v}) = \sum_{i=1}^{m-l+1} \mathbb{I}\{u_{i..i+l-1} = \mathbf{v}\}.$$

For i.i.d. observations, the Krichevsky-Trofimov estimates aggregates all possible points in the simplex $q \in \mathcal{P}_S$, seen as constant experts. For a sequence y^n, the forecaster is obtained, by

$$\begin{aligned} \texttt{kt}_0(y^n) &= \int_{\mathcal{P}_S} \prod_{s=1}^{S} p(s)^{N_{y^n}(s)} \varphi(\mathbf{p}) d\mathbf{p}, \text{ where } \mathbf{p} = (p(1), \dots, p(S)) \in \mathcal{P}_S \\ \text{for the prior} \quad \varphi(\mathbf{p}) &= \frac{\Gamma(S/2)}{\Gamma(1/2)} \prod_{s=1}^{S} \frac{1}{\sqrt{p(s)}}, \quad \text{where } \Gamma \text{ denotes the } \Gamma \text{ function}. \end{aligned}$$

Luckily this complicated expression can be computed sequentially, by only requires maintaining S counters in memory, $kt_0(Y_t = y|y^{t-1})$ for each $y \in \mathcal{Y}$. Further, the estimate can be extended to handle Markov models of any order $m \in \mathbb{N}$ over $\mathcal{Y}$, by considering the last m-observations as a side information, considering one predictor for each of the S^m-many side information values. We recall now the standard definition of Krichesky-Trofimov forecasters of all Markov models:

Definition 4.18 (*Markov KT forecasters*) *We denote by $\texttt{kt}_0$ the Krichesky-Trofimov estimate designed for i.i.d. distributions over the finite set $\mathcal{Y}$ of size S. It predicts the following distribution at time t*

$$\forall y \in \mathcal{Y}, kt_0(Y_t = y|Y_{1..t-1}) = \frac{(t-1)\nu_{Y_{1..t-1}}(y) + 1/2}{t - 1 + S/2}.$$

We denote by $\texttt{kt}_m$ the extension of $\texttt{kt}_0$ to Markov distributions of order m on $\mathcal{Y}$ obtained by considering the last m observations as being a side information, see e.g. (Cesa-Bianchi and Lugosi, 2006, chapter 9)

$$\forall y \in \mathcal{Y}, \quad kt_m(Y_t = y|Y_{1..t-1}) = \begin{cases} \frac{1}{S}, & \text{if } t-1 \leqslant m \\ \frac{(t-m-1)\nu_{Y_{1..t-1}}(Y_{t-m..t-1}y)+1/2}{(t-m-1)\nu_{Y_{1..t-2}}(Y_{t-m..t-1})+S/2}, & \text{if } t-1 > m. \end{cases}$$

We can further aggregate predictions of all Markov forecasters of all order simultaneously and thus compete with the best Markov model (of any order), using an aggregation of this countable set of models. The idea is to introduce a prior mass π_m to the Markov model of order m. In that case, the resulting aggregate predictor $\texttt{kt}_\infty$ only looses over the best Markov (see Proposition 1.1) the amount

$$L(\texttt{kt}_\infty, y^n) \leqslant \min_{m \in \mathbb{N}} \left(L(\texttt{kt}_m, y^n) + \ln(1/\pi_m) \right).$$

Definition 4.21 (*Universal KT forecaster*) *We denote by $\texttt{kt}_\infty$ the universal forecaster that aggregates the predictions of all the forecasters $\texttt{kt}_m, m \in \mathbb{N}$ as described in Lysyak and Ryabko (2016) (see also Ryabko (1988)). It is defined, for weights $w_m = \ln(2)/\ln(m+1) - \ln(2)/\ln(m+2)$, by*

$$kt_\infty(Y_t = y|Y_{1..t-1}) = \sum_{m=0}^{\infty} w_{m+1} kt_m(Y_t = y|Y_{1..t-1}).$$

Note: Any weights such that $\sum_{m=0}^{\infty} w_{m+1} = 1$ are valid alternative weights.

We are now ready to restate some useful known properties of the $\texttt{kt}_m$ and $\texttt{kt}_\infty$ forecasters. The $\texttt{kt}_m$ forecaster competes with the family $\mathcal{C}_m$ of all the constant forecasters defined for each of the S^m-many side information values (see Cesa-Bianchi and Lugosi (2006)). For instance $\mathcal{C}_0 = \{f : \forall t, y^{t-1} \in \mathcal{Y}^{t-1}\, f(\cdot|y^{t-1}) = q \text{ for some } q \in \mathcal{P}(\mathcal{Y})\}$. Let $N(\mathbf{v})$ be the number of occurrences of word $\mathbf{v} \in \mathcal{Y}^m$ in the side information sequence. It is known Cesa-Bianchi and Lugosi (2006) that

Lemma 4.3 (*Performance of KT forecasters*) *The regret of the $\texttt{kt}_m$ predictor using side information against the best Markov model of order m is controlled, for any sequence $y(n) \in \mathcal{Y}^n$ by*

$$L(\texttt{kt}_m, y^n) - \inf_{f\in\mathcal{C}_m} L(f, y^n) \leqslant \frac{S-1}{2}\sum_{\mathbf{v}\in\mathcal{Y}^m} \ln(N(\mathbf{v})) + S^m\left(\ln\frac{\Gamma(1/2)^S}{\Gamma(S/2)} + \frac{S-1}{2}\ln(2)+o(1)\right).$$

where $\sum_{\mathbf{v}\in\mathcal{Y}^m} \ln(N(\mathbf{v})) \leqslant S^m \ln(n)$ by Jensen's inequality, and where Γ is the gamma function. Further,

$$L(\texttt{kt}_\infty, y^n) \leqslant \min_{m\in\mathbb{N}}\left(L(\texttt{kt}_m, y^n) - \ln(w_{m+1})\right).$$

Corollary 4.2 (*Consistent KT estimates*) *Let $\mathcal{Y}$ be a finite set with S symbols and $\rho \in \mathcal{P}(\mathcal{Y}^\star)$ be a process that is Markov of order m. Then, the $\texttt{kt}_m$ predictor satisfies for all $n \in \mathbb{N}$,*

$$\frac{1}{n}KL_n(\rho, \texttt{kt}_m) \leqslant \frac{S-1}{2}S^m\frac{\ln(n)}{n} + \frac{S^m}{n}\left(\ln\frac{\Gamma(1/2)^S}{\Gamma(S/2)} + \frac{S-1}{2}\ln(2)+o(1)\right).$$

Further, let ρ be any stationary process such that for all $n \in \mathbb{N}$, there exists a process ρ_n that is Markov of order m_n, such that $\rho(Y^n) = \rho_n(Y^n)$, and such that $S^{m_n} = o(\frac{n}{\ln(n)})$ (this holds for any Markov process). Then

$$\frac{1}{n}KL_n(\rho, \texttt{kt}_\infty) \to 0\,.$$

Proof of corollary 4.2:

The first result for $\texttt{kt}_m$ is immediate by definition of the loss and Lemma 4.3. For the $\texttt{kt}_\infty$ predictor, we first note that by definition, it holds

$$\begin{aligned} -\ln(w_{m+1}) &= -\ln(\ln(2)/\ln(m_n+2) - \ln(2)/\ln(m_n+3)) \\ &= \ln\left(\frac{\ln(m_n+2)\ln(m_n+3)}{\ln(2)\ln(\frac{m_n+3}{m_n+2})}\right). \end{aligned}$$

Then, the $\texttt{kt}_\infty$ forecaster satisfies

$$\begin{aligned}\texttt{KL}_n(\rho, r) &= \texttt{KL}_n(\rho_n, r) \leqslant \frac{S-1}{2} S^{m_n} \ln(n) + S^{m_n}\left(\ln \frac{\Gamma(1/2)^S}{\Gamma(S/2)} + \frac{S-1}{2}\ln(2) + o(1)\right) \\ &\quad + \ln\left(\frac{\ln(m_n+2)\ln(m_n+3)}{\ln(2)\ln(\frac{m_n+3}{m_n+2})}\right).\end{aligned}$$

It is thus consistent when $S^{m_n} = o(\frac{n}{\ln(n)})$. □

CHAPTER 5
$\sup_q \langle q, f \rangle - \mathcal{B}^\psi(q, p)$

Contents

1 Risk-aversion in multi-armed bandits 81

1.1 Regrets for Risk-averse Multi-armed Bandits 82

1.2 The Price for Risk-aversion 83

1.3 A Generic Decomposition of the Empirical Regret 84

1.4 The Risk-Averse Upper Confidence Bound algorithm 85

2 Aggregation of experts: insights from duality. 86

2.1 Bregman duality 86

2.2 Aggregation of growing experts 88

2.3 An application to prediction of changing processes 91

Take-home message

Measure of risk-aversion and bandits

(Entropic **risk measure**) $\kappa_{-\lambda,\nu} = \inf_{\boldsymbol{q}\in\mathcal{P}(\mathbb{R})} \mathbb{E}_{\boldsymbol{q}}(X) + \frac{1}{\lambda}\mathtt{KL}(\boldsymbol{q},\nu)\,.$

(Risk-averse regret) $\overline{\mathfrak{R}}_T(\lambda) = \sum_{a\in\mathcal{A}} \Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}\Big)\mathbb{E}\Big[N_T(a)\Big]\,.$

Strategy: Use the optimistic principle to derive a `KL-ucb` style strategy for risk-averse regret minimization.

Bregman duality aggregation of experts

(η,ψ)**-mixable loss** $\ell : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$:

$$\forall \boldsymbol{x}\in\mathcal{X}^M, \mathbf{p}\in\mathcal{P}_M, \exists x_{\mathbf{x},\mathbf{p}}\in\mathcal{X}, \forall y\in\mathcal{Y},\quad \ell(x_{\mathbf{x},\mathbf{p}},y) \leqslant \inf_{\mathbf{q}\in\mathcal{P}_M} \langle \mathbf{q}, \ell(\boldsymbol{x}_\cdot,y)\rangle + \frac{1}{\eta}\mathcal{B}^\psi(\mathbf{q},\mathbf{p})\,.$$

Strategy: Based on expert proposals $\boldsymbol{x}_t\in\mathcal{X}^M$, choose $x_t = x_{\boldsymbol{x}_t,\mathbf{p}_t}$ where $\mathbf{p}_1 = \boldsymbol{\pi}$, and after receiving observation $y_t\in\mathcal{Y}$, $\mathbf{p}_{t+1} = \arg\min_{\mathbf{q}\in\mathcal{P}_M} \langle \mathbf{q}, \ell(\boldsymbol{x}_{t,\cdot},y_t)\rangle + \frac{1}{\eta}\mathcal{B}^\psi(\mathbf{q},\mathbf{p}_t)\,.$ This achieves,

$$\forall \mathbf{q}\in\mathcal{P}_M,\quad \sum_{t=1}^T \ell(x_t,y_t) - \langle \mathbf{q}, \sum_{t=1}^T \ell(\boldsymbol{x}_{t,\cdot},y_t)\rangle \leqslant \frac{1}{\eta}\Big(\mathcal{B}^\psi(\mathbf{q},\boldsymbol{\pi}) - \mathcal{B}^\psi(\mathbf{q},\boldsymbol{p}_{T+1})\Big)\,.$$

In this chapter, we focus on Bregman duality, as a tool to deviate from the mean estimation goal considered in the previous chapters. Indeed, we first show that Bregman duality naturally defines a notion of risk that can be used for instance to build risk-averse strategies in multi-armed bandits. Then, we briefly recall how Bregman duality naturally yields aggregation of expert techniques. This enables, given a set of learners, each corresponding to a different family of processes, to build strategies that are near uniformly optimal over the union of all these families. This is especially useful when we do not know in advance to which family the observed signal corresponds.

1 RISK-AVERSION IN MULTI-ARMED BANDITS

Let us recall that for arbitrary random variable X admitting a finite cumulant generative function around 0, then the two following properties hold (this is by a simple application of Markov's inequality)

$$\mathbb{P}\Big[X \geqslant \inf\Big\{\frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X) + \frac{\log(1/\delta)}{\lambda} \;:\; \lambda > 0\Big\}\Big] \leqslant \delta\,, \tag{5.1}$$

$$\mathbb{P}\Big[X \leqslant \sup\Big\{-\frac{1}{\lambda}\log\mathbb{E}\exp(-\lambda X) - \frac{\log(1/\delta)}{\lambda} \;:\; \lambda > 0\Big\}\Big] \leqslant \delta\,. \tag{5.2}$$

Note that (5.1) measures the probability that X is big, while (5.2) measures the probability that X is small, which is what we want to be protected against. Now, for the sake of clarity, it makes sense to introduce the value of the cumulant generative function of the variable X at point λ, rescaled by λ, that we denote

$$\kappa_{\lambda,\nu} \stackrel{\text{def}}{=} \frac{1}{\lambda}\log\mathbb{E}_\nu\exp\Big(\lambda X\Big), \tag{5.3}$$

and similarly we denote $\kappa_{-\lambda,\nu}$ the value of $\kappa_{\lambda',\nu}$ for $\lambda' = -\lambda$. We already saw that this quantity is at the heart of many key-results and tools of concentration of measure (e.g. the Cramer-Chernoff method, the Chernoff transform, the log-Laplace transform). More importantly here, $\kappa_{-\lambda,\nu}$ is a key quantity to control the probability that X is small.

Example: To understand (5.1) and (5.2), let us consider t Gaussian random variables $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from a distribution ν with mean μ and variance σ^2, then $X = \sum_{k=1}^t Z_k$ is Gaussian with mean μt and variance $\sigma^2 t$, and simple computations show that $\kappa_{\lambda,\nu} = \mu t + \frac{\lambda\sigma^2 t}{2}$, which yields, after optimizing the previous bounds in λ, to the optimal value $\lambda = \sqrt{\frac{2\log(1/\delta)}{\sigma^2 t}}$ and the familiar concentration bounds for Gaussian random variables

$$\mathbb{P}\Big(\frac{1}{t}\sum_{k=1}^t Z_k - \mu \geqslant \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\Big) \leqslant \delta \;\text{ and }\; \mathbb{P}\Big(\mu - \frac{1}{t}\sum_{k=1}^t Z_k \geqslant \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\Big) \leqslant \delta\,.$$

Let us comment on this example. First, the quantity $\kappa_{-\lambda,\nu} = \mu t - \frac{\lambda\sigma^2 t}{2}$ (sometimes called the mean-variance risk) takes the form of an operator that measures the mean of a random variable, penalized by some higher moment (the variance in that case). This is actually a general property, since by the variational formula for the Kullback-Leibler divergence, we have for a random variable X distributed according to $\nu \in \mathcal{P}(\mathbb{R})$ that

$$\kappa_{-\lambda,\nu} = \inf\Big\{\mathbb{E}_{\nu'}(X) + \frac{1}{\lambda}\mathtt{KL}(\nu'||\nu) \;:\; \nu' \in \mathcal{P}(\mathbb{R})\Big\} \leqslant \mathbb{E}_\nu\big[X\big]\,. \tag{5.4}$$

where $\mathtt{KL}(\nu'||\nu)$ denotes the Kullback-Leibler divergence between two distributions ν and ν'. Using $\kappa_{-\lambda,\nu}$ as a measure of risk-aversion is natural for several reasons: Additionally to the formulation (5.4) and the control (5.2) that are important for interpretability it is also a standard *coherent* risk-measure (see Rockafellar (2007)). Also, due to its deep link for concentration of measure, it is especially natural for analysis. (We however do not pretend this is the "best" choice of risk-measure.)

Mixability gaps Finally, for completeness, we also introduce the two fundamental quantities $m^+_{\lambda,\nu}\big[X\big]$ and $m^-_{\lambda,\nu}\big[X\big]$ that we call here the upper (and respectively lower) mixability gap and that are defined by

$$m^+_{\lambda,\nu} = \kappa_{\lambda,\nu} - \mathbb{E}_\nu\Big[X\Big] \text{ and } m^-_{\lambda,\nu} = \mathbb{E}_\nu\Big[X\Big] - \kappa_{-\lambda,\nu}\,.$$

Note that the mixability gaps are always non-negative by Jensen's inequality, and that an upper bound on them immediately provides a high probability confidence interval. Indeed, with these notations, the previous equations (5.1) and (5.2), can thus be rewritten more compactly as

$$\mathbb{P}\bigg[X - \mathbb{E}_\nu\Big[X\Big] \geqslant \inf_{\lambda>0}\Big\{m^+_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\bigg] \leqslant \delta\,, \tag{5.5}$$

$$\mathbb{P}\bigg[\mathbb{E}_\nu\Big[X\Big] - X \geqslant \inf_{\lambda>0}\Big\{m^-_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\bigg] \leqslant \delta\,. \tag{5.6}$$

1.1 Regrets for Risk-averse Multi-armed Bandits

Optimal arm We now naturally define the optimal arm $a^\star$ as the one maximizing the risk aversion at some fixed level λ, that is we define

$$a^\star \in \operatorname*{arg\,max}_{a=1,\dots,A} \kappa_{-\lambda,\nu_{a^\star}}.$$

Note again that in the case of Gaussian distributions with mean μ_a and variance σ_a^2, we simply have $\kappa_{-\lambda,\nu_{a^\star}} = \mu_a - \frac{\lambda\sigma_a^2}{2}$, and that in general we always have $\kappa_{-\lambda,\nu_{a^\star}} \leqslant \mathbb{E}_{\nu_a}\big[X\big]$. In the sequel, we assume for simplicity that $a^\star$ is unique.

Regret Now we define the *empirical regret* $\mathfrak{R}_T(\lambda)$ of the strategy $\mathfrak{A}$ with respect to the strategy $\star$ that constantly pulls the same arm $a^\star \in \{1,\dots,A\}$ by the difference between the cumulated reward received by algorithm $\mathfrak{A}$ and the cumulated reward that the strategy $\star$ would have received during the same game, that is, by introducing the fictitious plays $\{X_{i,a^\star}\}_{N^{\mathfrak{A}}_{T,a^\star}<i\leqslant T}$,

$$\mathfrak{R}_T(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{T} X_{i,a^\star} - \sum_{a=1}^{A}\sum_{i=1}^{N^{\mathfrak{A}}_{T,a}} X_{i,a} = \sum_{i=N^{\mathfrak{A}}_{T,a^\star}+1}^{T} X_{i,a^\star} - \sum_{a\neq a^\star}\sum_{i=1}^{N^{\mathfrak{A}}_{T,a}} X_{i,a}\,. \tag{5.7}$$

Note that we are not interested here in controlling the *expected regret* $\overline{\mathfrak{R}}_T$ as it gives no information on the risk of the strategy $\mathfrak{A}$ and of pulling one arm. Indeed, we have the following standard decomposition

$$\overline{\mathfrak{R}}_T = T\mathbb{E}_{\nu_{a^\star}}\big[X\big] - \mathbb{E}\bigg[\sum_{s=1}^{T} Y_s\bigg] = \sum_{a\in\mathcal{A}}\Big(\mathbb{E}_{\nu_{a^\star}}[X] - \mathbb{E}_{\nu_a}[X]\Big)\mathbb{E}\Big[N_{T,a}\Big]\,, \tag{5.8}$$

while one would prefer to have a more informative measure, taking into account for instance the variance of the arms or some control of the tails. For this purpose, another natural notion of regret is the *risk-averse regret* $\overline{\mathfrak{R}}_T(\lambda)$ defined by

$$\overline{\mathfrak{R}}_T(\lambda) = \sum_{a\in\mathcal{A}}\Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}\Big)\mathbb{E}\Big[N_{T,a}\Big]\,. \tag{5.9}$$

In the sequel, we control both (5.7) and (5.9) as they both offer interesting interpretations.

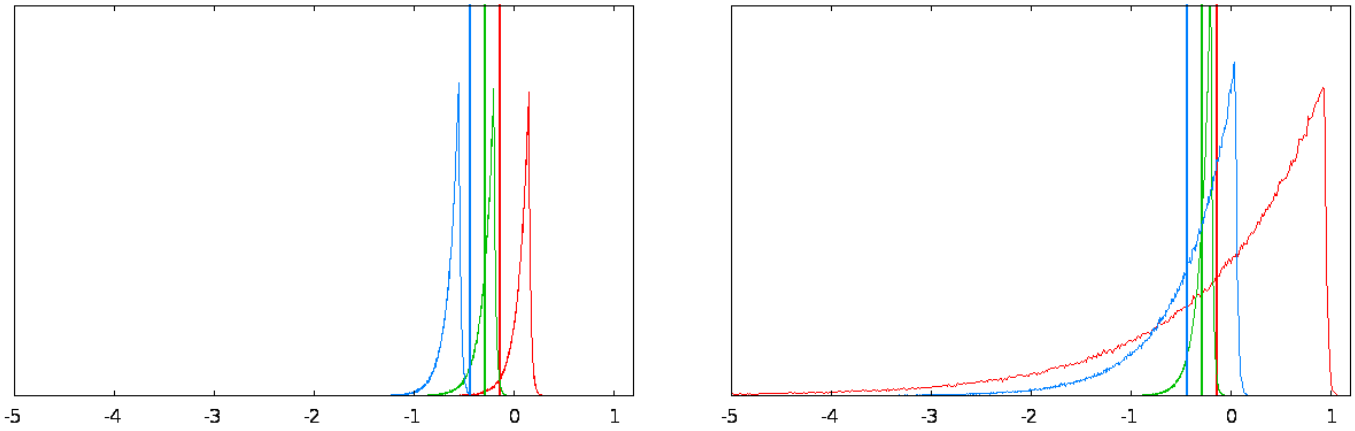


Figure 5.1: Plot of arms' densities and their mean: left) an environment where no arm has fat lower tail. right) an environment where for some λ, the best arm (green) does not have best mean, and sub-optimal arms (red, blue) have fat lower tails.

1.2 The Price for Risk-aversion

At a high-level, there is obviously a trade-off between trying to get maximal rewards and being risk-averse. Being too cautious (such as, arguably, `Exp3` see Auer et al. (2003)) avoids getting linear regret, but prevents from getting high rewards as well. On the other hand, simply targeting the maximal mean (such as `UCB` see Auer et al. (2002)) enables to get close to optimal rewards on average, but possibly very bad rewards in difficult environments (e.g. when sub-optimal arms have fat lower tails). In connection to this remark, see also Audibert et al. (2009) where it is shown that with relatively high probability, `UCB` may indeed incurs bad rewards.

A similar situation appears in the standard expected regret setting for the class of `UCB-`ρ algorithms as shown by Salomon and Audibert (2012): for, $\rho > \rho'$, `UCB-`ρ can compete with a larger class of environments than `UCB-`ρ'. However `UCB-`ρ' will beat `UCB-`ρ on simpler environments.

Simple and complex environments The risk-averse regret (5.9) captures the sub-optimality of an algorithm in terms of risk-aversion at some fixed level λ. As such, it is the direct equivalent of the expected regret in multi-armed bandits, and we control this regret for our `RA-UCB` procedure in Theorem 1, Maillard (2013b). If such control may seem satisfactory for many reasons, it also has some drawbacks. Namely, the level of risk-aversion is not related in any way to the actual distribution of rewards, since it is some parameter chosen a priori by the practitioner who wants to be protected against sampling possibly very low rewards. As a result, in easy situations when the rewards distributions have very light tails, a high risk-averse algorithm will be too cautious, and will get lower cumulative rewards than a less risk-averse algorithm, such as `UCB`. Similarly, if the actual distributions have very fat lower tails, a low risk-averse algorithm may not be cautious enough and thus get bad rewards compared to a more risk-averse algorithm, such as `Exp3`. See also figure 1.2.

Since such situations, that are of immediate practical interest, are not captured by the risk-averse regret (5.9) defined for some level λ, this motivates the study of the empirical risk-aversion regret (5.7) as this one is able to capture such behaviors (this is because it makes the empirical rewards coming from the actual distribution appear explicitly).

Note that this also raises the question of automatically adapting the level of risk-aversion to some bandit problem, or equivalently getting the best of all `RA-UCB-`λ algorithms (in terms of cumulated reward), which is very hard, (or even impossible, see Salomon and Audibert (2012) for impossibility results regarding `UCB-`ρ in the related problem of adaptivity in bandit problems). Since this involves orthogonal ideas that would worsen

readability and interpretation, add a difficult layer of complexity, and is little justified in practice (where the level of risk-aversion is often simply fixed), we do not study this question in the present work.

The difficult situation for risk-aversion appears when the sub-optimal arms produce rewards much lower than their mean (heavy lower tail) while the best arm produces rewards much higher than its mean (heavy upper tail): this creates maximal regret. We introduce in section 1.4 the `RA-UCB` algorithm that guarantees a low regret in such difficult environments (contrary to e.g. `UCB`). We refer to Maillard (2013b) for the technical proofs.

1.3 A Generic Decomposition of the Empirical Regret

We now introduce a generic decomposition of the regret, valid for any strategy $\mathfrak{A}$, that is the direct equivalent of (5.9) for the empirical regret.

Theorem 5.1 (*Risk-averse regret decomposition*) *Let us define, for some non-negative constants* $\{u_a\}_{a=1,\dots,A}$ *the event that sub-optimal arms are pulled too often*

$$\Omega \stackrel{\text{def}}{=} \left\{ \exists a \neq a^\star \,:\, N_{T,a}^{\mathfrak{A}} > u_a \right\},$$

and let us fix some value of λ *such that* $\kappa_{-\lambda,\nu_a}$ *exists for all* $a = 1,\dots,A$*. Then, for all* $\delta \in (0,1)$*, with probability higher than* $1-\delta-\mathbb{P}(\Omega)$*, the regret of the strategy* $\mathfrak{A}$ *is upper bounded by*

$$\begin{aligned} \mathfrak{R}_T(\lambda) \;\leqslant\; & \sum_{a\neq a^\star} u_a \left(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \right) + \left(m^-_{\lambda,\nu_{a^\star}} \sum_{a\neq a^\star} u_a + \frac{(A-1)\log(2A/\delta)}{\lambda} \right) \\ & + \inf_{\lambda'>0} \left\{ m^+_{\lambda',\nu_{a^\star}} \sum_{a\neq a^\star} u_a + \frac{\log(2A/\delta)}{\lambda'} \right\}. \end{aligned} \tag{5.10}$$

The first term of (5.10) makes appear a quantity very similar to that of the optimal regret bounds for the expected regret in the stochastic setting, where the standard optimality gaps $\mathbb{E}_{\nu_{a^\star}}\big[X\big] - \mathbb{E}_{\nu_a}\big[X\big]$ are replaced by $\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}$, as expected. Now the second and third terms involve the mixability gaps of the optimal arm. The third term is intuitive: indeed, a regret minimizing algorithm will try to understand $\kappa_{-\lambda,\nu_a}$ for each arm, and prevent from large deviations below the mean (bad rewards). However, this does not prevent the optimal arm to have large deviations above the mean (that is, unexpected good rewards), which is precisely captured by the third term. Now the presence of the second term comes from another phenomenon: λ is a parameter of the algorithm that tries to pull the arm with highest risk-aversion at level λ. As such, this goal may be successful or not depending on intrinsic properties of the environment. We say that λ is well-adapted to the environment if it is such that the second term in (5.10) is negligible before the first term.

So as to provide some intuition, let us now specialize Proposition 5.1 to the case of Example 1 for illustration purpose. In this case, the mixability gaps of the optimal arm $a^\star$ equal $\frac{\lambda}{2}\sigma^2_{a^\star}$ and $\frac{\lambda'}{2}\sigma^2_{a^\star}$, so that if we introduce for convenience the quantity $u \stackrel{\text{def}}{=} \sum_{a\neq a^\star} u_a$, one can rewrite (5.10) as

$$\mathfrak{R}_T(\lambda) \;\leqslant\; \sum_{a\neq a^\star} u_a \left(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \right) + \left(\frac{u\lambda}{2}\sigma^2_{a^\star} + \frac{(A-1)\log(A/\delta)}{\lambda} \right) + \sqrt{2u\log(A/\delta)}\sigma_{a^\star}\,. \tag{5.11}$$

Thus λ is well-adapted to the environment for instance when $\lambda = \Omega(u^{-1/2})$. Since any reasonable algorithm will pull sub-optimal arms only $u_a = O(\log(T))$ times with high probability, this indicates that a well-adapted

level of risk aversion for a Gaussian game of length T is of order[1] $\lambda = \Omega(\log(T)^{-1/2})$. A similar reasoning holds for the sub-Gaussian and thus the bounded case as well, since we only need an upper-bound on the mixability gaps rather than an equality here. In the sequel, we consider such a case, disregarding the extremely challenging question of defining and estimating a distribution-dependent optimally-adapted value of λ (it also conveys difficult interpretation since the optimal arm depends on λ). Note finally that contrary to the empirical regret, the risk-averse regret (5.9) is completely blind to such situations, as it basically corresponds to the first term in (5.10).

1.4 The Risk-Averse Upper Confidence Bound algorithm

We introduce in this section a strategy $\mathfrak{A}$ that we call the `RA-UCB` algorithm. From now on, we restrict to the case when all distributions belong to $\mathcal{P}(\mathbb{R}_B)$, where $\mathbb{R}_B = (-\infty, B]$ for some known value of B. Thus, let us introduce for all $a \in \mathcal{A}$, the empirical distribution $\widehat{\nu}_t(a) \in \mathcal{P}(\mathbb{R}_B)$ associated to ν_a, built using the past observations $Y_1, \dots, Y_t$; let $\delta_y \in \mathcal{P}(\mathbb{R}_B)$ denotes the Dirac mass at point y. We define

$$\widehat{\nu}_t(a) \stackrel{\text{def}}{=} \frac{1}{N_{t,a}^{\mathfrak{A}}} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}\{A_s = a\} \text{ where } N_{t,a}^{\mathfrak{A}} \stackrel{\text{def}}{=} \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}.$$

Further, for clarity purpose, we now use the notation $\widehat{\nu}_{n,a}$ (with a in subscript) in order to denote the empirical distribution built from the n first samples drawn from ν_a, while we reserve the functional notation $\widehat{\nu}_t(a)$ for the empirical distribution built from the samples received from arm a up to time t. Naturally, we have that $\widehat{\nu}_t(a) = \widehat{\nu}_{N_t^{\mathfrak{A}}(a),a}$. More generally, for some distribution ν, we also write $\widehat{\nu}_n$ for its empirical distribution built from n samples.

The `RA-UCB` algorithm is inspired from the strategies introduced by Lai and Robbins (1985a), Burnetas and Katehakis (1996), Maillard et al. (2011), Garivier and Cappé (2011), Cappé et al. (2013) as it selects at time $t+1$ the arm $A_{t+1} = \arg\max_{a \in \mathcal{A}} U_t(a)$, where $U_t(a)$ is an upper confidence bound on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\text{def}}{=} \sup \left\{ \kappa_{-\lambda,\nu} \,:\, \nu \in \mathcal{P}(\mathbb{R}_B),\, \mathbf{K}(\widehat{\nu}_t(a), \kappa_{-\lambda,\nu}) \leqslant \frac{f(t)}{N_{t,a}} \right\}, \tag{5.12}$$

and where we introduced the following quantity

$$\mathbf{K}(\widehat{\nu}_t(a), r) \stackrel{\text{def}}{=} \inf \left\{ \mathtt{KL}(\widehat{\nu}_t(a), \nu) \,:\, \nu \in \mathcal{P}(\mathbb{R}_B),\, \kappa_{-\lambda,\nu} \geqslant r \right\}. \tag{5.13}$$

Note that `UCB`-like algorithms are unnatural in this setting: they are based on empirical *means* only, while we really need to control the tail distributions here. KL-based algorithm are more suitable, and produce much stronger results. Note also that the parameter λ is here the same that defines the level of risk aversion used in the definition of the regret. The algorithm requires another parameter, that is a non-decreasing function of the time f. A typical choice is such that $f(t) = O(\log(t))$.

A Useful Formulation with Dual Optimality Conditions The definition of the bound (5.12) may seem quite abstract. In order to make it more computable and explicit, we now provide the following result, that is a dual formulation of the optimization problem given by $\mathbf{K}(\widehat{\nu}_t(a), r)$ (see the proof in Maillard (2013a)).

[1] Such (weak) dependency with T is intuitive: if we only have 10 trials do to something, we would be much more risk-averse (big λ) than with 1000 trials.

Lemma 5.1 (*Risk-averse dual formulation*) *Let $\widehat{\nu}_n$ denotes an empirical distribution built with a finite number n of atoms $\{x_i\}_{1\leqslant i\leqslant n}$. Then the following dual formulation holds*

$$\mathbf{K}(\widehat{\nu}_n, r) = \max\left\{ \frac{1}{n}\sum_{i=1}^n \log\left(1 - \frac{\gamma^\star}{\lambda}\Big(1 - e^{-\lambda(x_i - r)}\Big)\right) : 0 \leqslant \gamma^\star \leqslant \frac{\lambda}{1-e^{-\lambda(B-r)}}\right\}.$$

This result shows that the optimization problem (5.12) can actually be solved numerically and is deeply linked to the numerically efficient dual formulation considered for instance in Borwein and Lewis (1991), Harari-Kermadec (2006), or re-derived more recently in Honda and Takemura (2010) for the related problem of optimal regret bounds in the stochastic multi-armed bandit with expected regret criterion. For completeness, it makes sense to introduce the following quantity for general distributions $\nu \in \mathcal{P}(\mathbb{R}_B)$

$$\tilde{\mathbf{K}}(\nu, r) = \sup\left\{ \mathbb{E}\left[\log\left(1 - \frac{\gamma^\star}{\lambda}\big(1 - e^{-\lambda(X-r)}\big)\right)\right] \,:\, 0 \leqslant \gamma^\star \leqslant \frac{\lambda}{1-e^{-\lambda(B-r)}}\right\}.$$

2 AGGREGATION OF EXPERTS: INSIGHTS FROM DUALITY.

We now turn to revisit aggregation of M experts, where experts can be built from different sources, such as kernel estimates for regression, or KT estimates for prediction in Markov models, and provide a few additional results over Chapter 1. Let us recall that a loss function $\ell : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+$ is η-mixable for some $\eta > 0$ if

$$\forall \boldsymbol{x}\in\mathcal{X}^M, \mathbf{p}\in\mathcal{P}_M, \exists x_{\mathbf{x},\mathbf{p}}\in\mathcal{X}, \forall y\in\mathcal{Y}, \quad \ell(x_{\mathbf{x},\mathbf{p}}, y) \leqslant -\frac{1}{\eta}\log \mathbb{E}_{m\sim\mathbf{p}} \exp(-\eta\ell(\boldsymbol{x}_m, y))\,.$$

In particular an η-mixable loss satisfies

$$\forall \boldsymbol{x}\in\mathcal{X}^M, \mathbf{p}\in\mathcal{P}_M, \exists x_{\mathbf{x},\mathbf{p}}\in\mathcal{X}, \forall y\in\mathcal{Y}, \quad \ell(x_{\mathbf{x},\mathbf{p}}, y) \leqslant \inf_{\mathbf{q}\in\mathcal{P}_M} \langle \mathbf{q}, \ell(\boldsymbol{x}_\cdot, y)\rangle + \frac{1}{\eta}\mathtt{KL}(\mathbf{q},\mathbf{p})\,.$$

We say the loss is exactly η-mixable if the property holds with an equality. Written in this form, the connection between the previous section becomes clearer.

We have seen that choosing $x_t = x_{\boldsymbol{x}_t,\mathbf{p}_t}$ with $\mathbf{p}_1 = \boldsymbol{\pi}$, then $\mathbf{p}_{t+1} = \arg\min_{\mathbf{q}\in\mathcal{P}_M} \langle \mathbf{q}, \ell(\boldsymbol{x}_{t,\cdot}, y_t)\rangle + \frac{1}{\eta}\mathtt{KL}(\mathbf{q},\mathbf{p}_t)\,,$ updated using expert proposals $\boldsymbol{x}_t$ and loss y_t, leads to a cumulative loss controlled for each $\mathbf{q}\in\mathcal{P}_M$ by

$$\sum_{t=1}^T \ell(x_t, y_t) - \langle \mathbf{q}, \sum_{t=1}^T \ell(\boldsymbol{x}_{t,\cdot}, y_t)\rangle \leqslant \frac{1}{\eta}\Big(\mathtt{KL}(\mathbf{q},\boldsymbol{\pi}) - \mathtt{KL}(\mathbf{q},\boldsymbol{p}_{T+1})\Big)\,.$$

2.1 Bregman duality

This easy result can be extended beyond the use of Kullback-Leibler divergence, by considering instead a generic Bregman divergence $\mathcal{B}^\psi$.

Definition 5.3 ((η,ψ)-*mixable loss*) *We say the loss ℓ is (η,ψ)-mixable if*

$$\forall \boldsymbol{x}\in\mathcal{X}^M, \mathbf{p}\in\mathcal{P}_M, \exists x_{\mathbf{x},\mathbf{p}}\in\mathcal{X}, \forall y\in\mathcal{Y}, \quad \ell(x_{\mathbf{x},\mathbf{p}}, y) \leqslant \inf_{\mathbf{q}\in\mathcal{P}_M} \langle \mathbf{q}, \ell(\boldsymbol{x}_\cdot, y)\rangle + \frac{1}{\eta}\mathcal{B}^\psi(\mathbf{q},\mathbf{p})\,.$$

Lemma 5.2 (*Bregman duality aggregation*) *Let ℓ be an (η,ψ)-mixable loss and $\boldsymbol{\pi}$ be such that $\boldsymbol{\pi}(m) > 0$ for all m. Assume ψ to ensure that $\forall m, p(m)=0 \implies (\nabla\psi(p))_m = +\infty$. Let us consider the strategy defined by $x_t = x_{\boldsymbol{x}_t,\mathbf{p}_t}$ with $\mathbf{p}_1 = \boldsymbol{\pi}$, then $\mathbf{p}_{t+1} = \underset{\mathbf{q}\in\mathcal{P}_M}{\arg\min}\langle \mathbf{q}, \ell(\boldsymbol{x}_{t,.}, y_t)\rangle + \frac{1}{\eta}\mathcal{B}^\psi(\mathbf{q},\mathbf{p}_t)$, updated using expert proposals $\boldsymbol{x}_t$ and loss y_t. Then, the following holds for each $\mathbf{q}\in\mathcal{P}_M$*

$$\sum_{t=1}^T \ell(x_t,y_t) - \langle \mathbf{q}, \sum_{t=1}^T \ell(\boldsymbol{x}_{t,.},y_t)\rangle \leqslant \frac{1}{\eta}\Big(\mathcal{B}^\psi(\mathbf{q},\boldsymbol{\pi}) - \mathcal{B}^\psi(\mathbf{q},\boldsymbol{p}_{T+1})\Big)\,.$$

Proof :

Step 1 First, let us introduce $\mathcal{B}_p^\psi : q \to \mathcal{B}^\psi(q,p)$, so that

$$\begin{aligned}\mathcal{B}_p^\psi(q) &= \psi(q) - \psi(p) - \langle \nabla\psi(p), q-p\rangle \\ \nabla\mathcal{B}_p^\psi(q) &= \nabla\psi(q) - \nabla\psi(p)\,.\end{aligned}$$

Now, assuming that $\boldsymbol{p}_t(m) > 0$ for each m, an optimality condition for $\boldsymbol{p}_{t+1}$ is that for some constant C independent on m, it holds

$$\forall m\in\mathcal{M},\quad \ell(\boldsymbol{x}_{t,m},y_t) + \frac{1}{\eta}(\nabla\mathcal{B}^\psi_{\boldsymbol{p}_t}(\boldsymbol{p}_{t+1}))_m = C \quad \text{that is} \quad \ell(\boldsymbol{x}_{t,m},y_t) = \frac{1}{\eta}\langle \nabla\psi(\boldsymbol{p}_t) - \nabla\psi(\boldsymbol{p}_{t+1}), e_m\rangle + C\,,$$

where we introduced the hot vector e_m of m. The condition $\nabla\psi(p)_m = +\infty$ whenever $p(m)=0$ then ensures $\boldsymbol{p}_{t+1}(m) > 0$ must hold for all m. Since $\boldsymbol{p}_1(m) > 0$ by assumption, this ensures that the previous decomposition indeed holds for all t.

Step 2 Now, since on the other hand, it holds

$$\begin{aligned}\ell(x_t,y_t) &\leqslant \langle \boldsymbol{p}_{t+1}, \ell(\boldsymbol{x}_{t,.},y_t)\rangle + \frac{1}{\eta}\mathcal{B}^\psi_{\boldsymbol{p}_t}(\boldsymbol{p}_{t+1}) \\ &= \frac{1}{\eta}\Big[\langle \boldsymbol{p}_{t+1}, \nabla\psi(\boldsymbol{p}_t) - \nabla\psi(\boldsymbol{p}_{t+1})\rangle + \mathcal{B}^\psi_{\boldsymbol{p}_t}(\boldsymbol{p}_{t+1})\Big] + C \\ &= \frac{1}{\eta}\Big[\psi(\boldsymbol{p}_{t+1}) - \langle \boldsymbol{p}_{t+1}, \nabla\psi(\boldsymbol{p}_{t+1})\rangle - \psi(\boldsymbol{p}_t) + \langle \nabla\psi(\boldsymbol{p}_t), \boldsymbol{p}_t\rangle\Big] + C\,,\end{aligned}$$

we deduce that for each m, introducing the function $\varphi(\boldsymbol{p}) = \psi(\boldsymbol{p}) - \langle \boldsymbol{p}, \nabla\psi(\boldsymbol{p})\rangle$, it holds

$$\ell(x_t,y_t) - \ell(\boldsymbol{x}_{t,m},y_t) \leqslant \frac{1}{\eta}\Big[\varphi(\boldsymbol{p}_{t+1}) - \varphi(\boldsymbol{p}_t) + \langle \nabla\psi(\boldsymbol{p}_{t+1}) - \nabla\psi(\boldsymbol{p}_t), e_m\rangle\Big]\,.$$

Step 3 Summing over t and considering some $\boldsymbol{q}\in\mathcal{P}_M$, it finally holds

$$\begin{aligned}\sum_{t=1}^T \ell(x_t,y_t) - \langle \boldsymbol{q}, \ell(\boldsymbol{x}_{t,.},y_t)\rangle &\leqslant \frac{1}{\eta}\Big[\varphi(\boldsymbol{p}_{T+1}) + \langle \nabla\psi(\boldsymbol{p}_{T+1}), \boldsymbol{q}\rangle - \varphi(\boldsymbol{p}_1) - \langle \nabla\psi(\boldsymbol{p}_1), \boldsymbol{q}\rangle\Big] \\ &= \frac{1}{\eta}\Big[\psi(\boldsymbol{p}_{T+1}) + \langle \nabla\psi(\boldsymbol{p}_{T+1}), \boldsymbol{q} - \boldsymbol{p}_{t+1}\rangle - \psi(\boldsymbol{p}_1) - \langle \nabla\psi(\boldsymbol{p}_1), \boldsymbol{q} - \boldsymbol{p}_1\rangle)\Big] \\ &= \frac{1}{\eta}\Big[\mathcal{B}^\psi(\boldsymbol{q},\boldsymbol{p}_1) - \mathcal{B}^\psi(\boldsymbol{q},\boldsymbol{p}_{T+1})\Big]\,.\end{aligned}$$

□

2.2 Aggregation of growing experts

While aggregation is generally applied to the situation when the number of experts M is fixed, in practice it is natural to consider situations when $M = M_t$ may change with t. A natural setup in which this happens is when considering prediction of a signal whose underlying process does not correspond to a fixed process from a given set, but instead changes with time. In that case, tracking the process can be handled by aggregating different estimates of the process started at different time steps. We now propose to quickly revisit the aggregation result in this case. For concreteness, we consider using the KL regularization instead of a general Bregman divergence. We discuss further extensions in chapter 9.

Let $\boldsymbol{\tau} = (\tau_c)_{c\in\mathbb{N}}$ be a sequence partitioning $\mathbb{N} = \cup_{c\in\mathbb{N}}\mathbb{T}_c$ into disjoint intervals $\mathbb{T}_c = [\tau_c + 1, \tau_{c+1}]$, and let $C_T = \min\{c : T \in \mathbb{T}_c\}$, and assume without loss of generality that $\tau_{C_T+1} = T$. Let $\mathcal{M}_t$ be the set of experts available at time t, of cardinality M_t, and $\mathcal{M}_{(c)} = \bigcap_{t\in\mathbb{T}_c} \mathcal{M}_t \setminus \mathcal{M}_{\tau_c}$ be the set of experts available at all times in $\mathbb{T}_c$ but not before τ_c. Let $\mathcal{M}^\star(\boldsymbol{\tau}) \subset \mathcal{M}^\star$ be the set of all sequences of experts $(m_t)_{t\in\mathbb{N}}$ that are constant on each interval $\mathbb{T}_c$ with common value that belongs to $\mathcal{M}_{(c)}$ for $c \in \mathbb{N}$. We naturally call this a sequence of "fresh" experts. For any $\boldsymbol{U}$ distributed on $\mathcal{M}^\star(\boldsymbol{\tau})$, let us define the regret after T steps by

$$\begin{aligned}\mathfrak{R}_T(\boldsymbol{U}) &= \sum_{t=1}^T \ell(x_t, y_t) - \sum_{\boldsymbol{m}\in\mathcal{M}^\star(\boldsymbol{\tau})} \boldsymbol{U}(\boldsymbol{m}) \sum_{t=1}^T \ell(x_{t,\boldsymbol{m}_t}, y_t) \\ &= \sum_{c=1}^{C_T} \sum_{t\in\mathbb{T}_c} \sum_{m\in\mathcal{M}_{(c)}} \boldsymbol{u}_c(m)(\ell(x_t, y_t) - \ell(x_{t,m}, y_t))\,,\end{aligned}$$

where we introduced the marginal distribution $\boldsymbol{u}_c$ of $\boldsymbol{U}$ on each $\mathcal{M}_{(c)}$, defined for any $m_{(c)} \in \mathcal{M}_{(c)}$ by

$$\boldsymbol{u}_c(m_{(c)}) = \sum_{c'\in\mathbb{N}\setminus\{c\}} \sum_{m_{(c')}\in\mathcal{M}_{(c')}} \boldsymbol{U}(m_{(1)}, \dots, m_{(2)}, \dots)\,.$$

We finally introduce the following simple strategy, parameterized by a sequence of non negative input weights $(\boldsymbol{q}_t)_{t\in\mathbb{N}}$, where $\boldsymbol{q}_t \in \mathbb{R}_+$:

Algorithm 3 Aggregation of fresh experts

1: **for** $t = 1, \dots$ **do**

2: Define

$$\forall m \in \mathcal{M}_t, \quad w_t(m) = \begin{cases} w_{t-1}(m)\exp(-\eta\ell(x_{t-1,m}, y_{t-1})) & \text{if } m \in \mathcal{M}_{t-1} \\ \boldsymbol{q}_t & \text{else}\,. \end{cases}$$

3: Given $\boldsymbol{x}_t$, predict $x_t = x_{\boldsymbol{x}_t, \boldsymbol{p}_t}$ where $\boldsymbol{p}_t(m) = \dfrac{w_t(m)}{\sum_{m\in\mathcal{M}_t} w_t(m)}\,.$

Lemma 5.3 (*Aggregation of fresh experts*) *The regret after T time steps of Algorithm 3 using input weights $(\boldsymbol{q}_t)_{t\in\mathbb{N}}$ against any $\boldsymbol{U} \in \mathcal{P}(\mathcal{M}^\star(\boldsymbol{\tau}))$ satisfies*

$$\begin{aligned}\mathfrak{R}_T(\boldsymbol{U}) &\leqslant \frac{1}{\eta}\sum_{c=1}^{C_T}\left[\sum_{m\in\mathcal{M}_{(c)}} \boldsymbol{u}_c(m)\ln\left(\frac{\boldsymbol{p}_{\tau_{c+1}+1}(m)}{\boldsymbol{p}_{\tau_c+1}(m)}\right) + \sum_{t\in\mathbb{T}_c}\ln\left(\frac{W_{t+1}}{\overline{W}_t}\right)\right],\\ &\leqslant \frac{1}{\eta}\sum_{c=1}^{C_T}\left[KL(\boldsymbol{u}_c, \boldsymbol{p}_{\tau_c+1}) + \sum_{t\in\mathbb{T}_c}\ln\left(\frac{W_{t+1}}{\overline{W}_t}\right)\right],\end{aligned}$$

where $W_t = \sum_{m\in\mathcal{M}_t} w_t(m)$ and $\overline{W}_t = \sum_{m\in\mathcal{M}_t} w_{t+1}(m)$. In the special case when the loss is exactly η-mixable, the inequality becomes an equality. Further it holds $\boldsymbol{p}_{\tau_c+1}(m) = \frac{\boldsymbol{q}_{\tau_c+1}}{W_t}$ for each $m \in \mathcal{M}_{(c)}$.

Since by construction $W_{t+1} = \overline{W}_t + |\mathcal{M}_{t+1}\backslash\mathcal{M}_t|\boldsymbol{q}_{t+1}$, this suggests to choose q_{t+1} proportional to $\frac{\overline{W}_t}{|\mathcal{M}_{t+1}\backslash\mathcal{M}_t|}$. With such a tuning, we immediately obtain the following result

Corollary 5.1 (*Aggregation of fresh experts*) *Let $(a_t)_{t\in\mathbb{N}}$ be a sequence of non-negative real values, and define the input weights by*

$$\boldsymbol{q}_1 = 1 \quad \textit{and} \quad \boldsymbol{q}_{t+1} = \frac{\overline{W}_t}{a_t m_{+,t+1}} \textit{ where } m_{+,t+1} = |\mathcal{M}_{t+1}\backslash\mathcal{M}_t|\,.$$

The regret after T time steps of Algorithm 3 using these input weights is bounded as

$$\mathfrak{R}_T(\boldsymbol{U}) \leqslant \frac{1}{\eta}\sum_{c=1}^{C_T}\left[\sum_{m\in\mathcal{M}_{(c)}} \boldsymbol{u}_c(m)\ln\left(\boldsymbol{u}_c(m)m_{+,\tau_c+1}(1+a_{\tau_c})\right) + \sum_{t\in\mathbb{T}_c}\ln\left(1+\frac{\mathbb{I}\{m_{+,t+1}>0\}}{a_t}\right)\right].$$

In particular, let $\mathcal{M}^\star_{(c)} = \mathrm{Argmin}_{m\in\mathcal{M}_{(c)}}\sum_{t\in\mathbb{T}_c}\ell(\boldsymbol{x}_{t,m}, y_t)$ be the set of locally optimal experts on $\mathbb{T}_c$. Then, if $m_{+,t} = M > 0$ for all t and $|\mathcal{M}^\star_{(c)}| = K$ for all c, performing growing aggregation using the choice $a_t = t$ ensures that for any $m^\star \in \mathcal{M}^\star(\boldsymbol{\tau})$,

$$\mathfrak{R}_T(\boldsymbol{m}^\star) \leqslant \frac{1}{\eta}\left[C_T\ln\left(\frac{M}{K}\right) + \left(\sum_{c=1}^{C_T}\ln(1+\tau_c)\right) + \ln(T+1)\right].$$

Proof of Lemma 5.3:

Step 1. For convenience, let us denote $\ell_t = \ell(x_t, y_t)$ and $\ell_{m,t} = \ell(\boldsymbol{x}_{t,m}, y_t)$. With these notations, let us first remark that by the η-mixable property, it holds

$$\ell_t \leqslant -\frac{1}{\eta}\log\mathbb{E}_{m\sim\mathbf{p}_t}\exp(-\eta\ell_{m,t})\,.$$

Thus, for all $m \in \mathcal{M}_t$, we get

$$\begin{aligned}
\ell_t - \ell_{m,t} &\leqslant -\frac{1}{\eta}\log \mathbb{E}_{m\sim \mathbf{p}_t} \exp(-\eta \ell_{m,t}) + \frac{1}{\eta}\log\exp(-\eta\ell_{m,t}) \\
&= -\frac{1}{\eta}\log \sum_{m\in\mathcal{M}_t} \frac{w_t(m)}{W_t}\exp(-\eta\ell_{m,t}) + \frac{1}{\eta}\log\frac{w_t(m)}{W_t}\exp(-\eta\ell_{m,t}) - \frac{1}{\eta}\log\left(\frac{w_t(m)}{W_t}\right) \\
&= \frac{1}{\eta}\log\left(\frac{w_t(m)\exp(-\eta\ell_{m,t})}{\sum_{m\in\mathcal{M}_t} w_t(m)\exp(-\eta\ell_{m,t})}\right) - \frac{1}{\eta}\log\left(\frac{w_t(m)}{W_t}\right) \\
&= \frac{1}{\eta}\log\left(\frac{w_{t+1}(m)}{\overline{W}_t}\right) - \frac{1}{\eta}\log\left(\frac{w_t(m)}{W_t}\right).
\end{aligned}$$

where $\overline{W}_t = \sum_{m\in\mathcal{M}_t} w_t(m)\exp(-\ell_{m,t})$. Let us remark that

$$W_{t+1} = \overline{W}_t + m_{+,t+1}\boldsymbol{q}_{t+1} = \overline{W}_t\Big(1 + \frac{m^+_{+,t+1}\boldsymbol{q}_{t+1}}{\overline{W}_t}\Big).$$

From the definition of $\boldsymbol{p}_t$, we deduce that

$$\ell_t - \ell_{m,t} \leqslant \frac{1}{\eta}\left[\ln\Big(\boldsymbol{p}_{t+1}(m)\Big) - \ln\Big(\boldsymbol{p}_t(m)\Big) + \ln\Big(\frac{W_{t+1}}{\overline{W}_t}\Big)\right]$$

Step 2. Now, by definition, the regret writes as

$$\mathfrak{R}_T(\boldsymbol{U}) = \sum_{c=1}^{C_T}\sum_{t\in\mathbb{T}_c}\sum_{m\in\mathcal{M}_{(c)}} \boldsymbol{u}_c(m)(\ell_t - \ell_{m,t}).$$

where we used the marginal distribution $\boldsymbol{u}_c$. Note that by construction $\sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m) = 1$. Plugging-in the result of step 1, we obtain

$$\begin{aligned}
\mathfrak{R}_T(\boldsymbol{U}) &\leqslant \frac{1}{\eta}\sum_{c=1}^{C_T}\left[\sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\sum_{t\in\mathbb{T}_c}\ln\left(\frac{\boldsymbol{p}_{t+1}(m)}{\boldsymbol{p}_t(m)}\right) + \left(\sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\right)\sum_{t\in\mathbb{T}_c}\ln\Big(\frac{W_{t+1}}{\overline{W}_t}\Big)\right] \\
&= \frac{1}{\eta}\sum_{c=1}^{C_T}\left[\sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\ln\left(\frac{\boldsymbol{p}_{\tau_{c+1}+1}(m)}{\boldsymbol{p}_{\tau_c+1}(m)}\right) + \sum_{t\in\mathbb{T}_c}\ln\Big(\frac{W_{t+1}}{\overline{W}_t}\Big)\right], \qquad (5.14)
\end{aligned}$$

which concludes the first regret bound.

Step 3. Now, note that $\boldsymbol{u}_c$ is a probability distributions with support in $\mathcal{M}_{(c)}$, and that $\boldsymbol{p}_{\tau_c+1}, \boldsymbol{p}_{\tau_{c+1}+1}$ are probability distributions with support containing $\mathcal{M}_{(c)}$. This enables to write

$$\begin{aligned}
\sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\ln\left(\frac{\boldsymbol{p}_{\tau_{c+1}+1}(m)}{\boldsymbol{p}_{\tau_c+1}(m)}\right) &= \sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\ln\left(\frac{\boldsymbol{u}_c(m)}{\boldsymbol{p}_{\tau_c+1}(m)}\right) - \sum_{m\in\mathcal{M}_{(c)}}\boldsymbol{u}_c(m)\ln\left(\frac{\boldsymbol{u}_c(m)}{\boldsymbol{p}_{\tau_{c+1}+1}(m)}\right) \\
&= \mathtt{KL}(\boldsymbol{u}_c, \boldsymbol{p}_{\tau_c+1}) - \mathtt{KL}(\boldsymbol{u}_c, \boldsymbol{p}_{\tau_{c+1}+1}) \leqslant \mathtt{KL}(\boldsymbol{u}_c, \boldsymbol{p}_{\tau_c+1}).
\end{aligned}$$

Further, when $m \in \mathcal{M}_{(c)}$, it satisfies that $m \in \mathcal{M}_t$ for each $t \in \mathbb{T}_c$ and $m \notin \mathcal{M}_{\tau_c}$, which means that $\boldsymbol{p}_{\tau_c+1}(m) = \frac{\boldsymbol{q}_{\tau_c+1}}{W_{\tau_c+1}}$. This concludes the proof. □

Remark 5.1 *This result actually holds for any sequence and any segmentation. Thus the bounds are valid from a minimax perspective, as is usual for aggregation results.*

2.3 An application to prediction of changing processes

In order to better understand how to use the growing aggregation procedure, let us consider an application to the task of predicting a sequence of observations whose characteristics are "changing". What we mean here is that we have access to a set of processes $\mathcal{R}$ (e.g. i.i.d. σ-sub-Gaussian, or Markov 1 Bernoulli), but the process generating the observations does not belong to $\mathcal{R}$. Instead, there is a sequence of times $(\tau_c)_{c\in\mathbb{N}}$ such that for each c, all observations in $[\tau_c+1, \tau_{c+1}]$ are generated from a process in $\mathcal{R}$, but different processes are used for any two consecutive time segments. We call this an $\mathcal{R}$-Piecewise process:

Definition 5.6 ($\mathcal{R}$-*Piecewise process*) *An $\mathcal{R}$-piecewise process ρ on $\mathcal{Y}^\infty$ is a couple $(\mathcal{R}, \chi_\rho)$, where $\mathcal{R} \subset \mathcal{P}(\mathcal{Y}^\infty)$ is called the set of root processes, and $\chi_\rho \in \mathcal{P}((\mathcal{R}\times\mathbb{N}_\star)^\infty)$ is a choice process. A sample $Y_{1..\infty} \sim \rho$ writes*

$$Y_{1..\infty} = Y^1_{1..\ell_1} Y^2_{1..\ell_2} \dots \text{ where } Y^c_{1..\ell_c} = (Y^c_1, \dots, Y^c_{\ell_c}) \sim \rho_c \in \mathcal{R} \text{ for each } c \text{ and } (\rho_1,\ell_1),(\rho_2,\ell_2),\dots \sim \chi_\rho\,.$$

We assume this decomposition is minimal (hence $\rho_c \neq \rho_{c+1}$ for each c), denote the change times $\tau_c = \sum_{c'<c}\ell_{c'}$ and the corresponding time intervals $\mathbb{T}_c = [\tau_c+1, \tau_{c+1}]$, for each $c\in\mathbb{N}$.

We want to predict the observations when only the set $\mathcal{R}$ is known but the change times $(\tau_c)_{c\in\mathbb{N}}$ and root processes $(\rho_c)_{c\in\mathbb{N}}$ are unknown. In the general case when no two pieces are generated by the same root process ($\forall c\neq c', \rho_c \neq \rho_{c'}$), one cannot hope to achieve better performance than training a learner specialized for $\mathcal{R}$ on each separate time interval $\mathbb{T}_c$. Hence, if $f_\mathcal{R}$ denotes such a base learner, and $\mathbf{f}_{\mathcal{R},t}(\cdot|(y_{t'})_{t'\in[\tau_c+1,t-1]})$ denotes its prediction at time t having seen past observations in$\mathbb{T}_c$, then a natural strategy is to run a novel instance of $f_\mathcal{R}$ on each $\mathbb{T}_c$. Denoting this forecaster $f^\star_\mathcal{R}$, its cumulative loss writes

$$L(f^\star_\mathcal{R}, y^T) \;=\; \sum_{c\in\mathbb{N}} L_{\mathbb{T}_c}(f_\mathcal{R}, (y_t)_{t\in\mathbb{T}_c}) \text{ where } L_{\mathbb{T}_c}(f_\mathcal{R}, (y_t)_{t\in\mathbb{T}_c}) = \sum_{t\in\mathbb{T}_c} L(y_t, \mathbf{f}_{\mathcal{R},t}(\cdot|(y_{t'})_{t'\in[\tau_c+1,t-1]}))\,.$$

Unfortunately, the change times are unknown, hence such a strategy is not applicable. Instead, we can apply an aggregation of growing experts (Algorithm 3), where at each time step, we consider the set of experts $\mathcal{M}_t = \{f^{(t')}_\mathcal{R} : t' \leqslant t\}$, where $f^{(t')}_\mathcal{R}$ denotes the base learner that outputs $\mathbf{f}_{\mathcal{R},t}(\cdot|(y_{t'})_{t'\in[t'+1,t-1]})$ at each time $t > t'$ (equivalently, the base learner $f_\mathcal{R}$ that only receives observations after time t'). In this case, $\mathcal{M}_t$ increases by one element at each time step, hence $m_{+,t} = 1$ for all t, and we maintain a growing set of t experts at time t.

More generally, when one has access to a set $\mathcal{B}$ of base forecasters instead of a single expert $f_\mathcal{R}$ (A typical situation is when $\mathcal{R} = \bigcup_{b=1}^B \mathcal{R}_b$, and $\mathcal{B} = \{f_{\mathcal{R}_b}, b = 1,\dots,B\}$), we can define an oracle strategy $\mathcal{B}_\star$ that chooses an optimal forecaster in $\operatorname{Argmin}_{f\in\mathcal{B}} L_{\mathbb{T}_c}(f, (y_t)_{t\in\mathbb{T}_c})$ in each piece. Its cumulative loss is $L(\mathcal{B}_\star, y^T) = \sum_{c\in\mathbb{N}} \min_{f\in\mathcal{B}} L_{\mathbb{T}_c}(f, (y_t)_{t\in\mathbb{T}_c})$. In that case, it is natural to apply Algorithm 3 with the set of experts $\mathcal{M}_t = \{f^{(t')} : t' \leqslant t, f\in\mathcal{B}\}$, hence in this case, $m_{+,t} = |\mathcal{B}| = B$ for each t. We denote by $\texttt{Agg}$–$\mathcal{B}$ this strategy. Applying Corollary 5.1, we obtain immediately a sharp bound on the loss of such a strategy:

Corollary 5.2 ($\mathcal{R}$-*piecewise process prediction loss*) *Let ρ be any $\mathcal{R}$-piecewise process, and apply $\texttt{Agg}$–$\mathcal{B}$ with a set of base forecasters $\mathcal{B}$ of size B. Using the same notations as for Corollary 5.1, consider that $|\mathcal{M}^\star_{(c)}| = K$ for each c. Then the cumulative loss of $\texttt{Agg}$–$\mathcal{B}$, assuming the loss is η-mixable, is*

$$L(\texttt{Agg}\text{–}\mathcal{B}, y^T) \leqslant \frac{1}{\eta}\left[C_T \ln(B/K) + \sum_{c=1}^{C_T} \ln(1+\tau_c) + \ln(T+1)\right] + \sum_{c=1}^{C_T} \min_{b\in\mathcal{B}} L_{\mathbb{T}_c}(b, (y_t)_{t\in\mathbb{T}_c})\,.$$

To give a visual illustration, Figure 5.2 depicts, for a small set base of forecasters $\mathcal{B}$ designed for regression on low-degree polynomials (with adaptive variance estimation), and observations coming from a piecewise low-degree polynomial with abrupt changes, the confidence intervals around the next observation built by several strategies. As expected, the Bayes aggregation algorithm on $\mathcal{B}$ is naturally unable to handle the changes of the signal, while `Agg`–$\mathcal{B}$ successfully adapts to the change of stationarity, and competes with the oracle. The confidence intervals appear a little tighter because of the combination of many experts (the oracle only considers a single best expert in each piece). Note that the confidence intervals computed by the method fail at the abrupt change points and recover only one step later, contrary to the oracle (blue) that perfectly knows when a change point will occur: indeed they do not try to predict when a change point will occur, nor to which piece. Rather, they predict the most plausible confidence interval for the current piece.

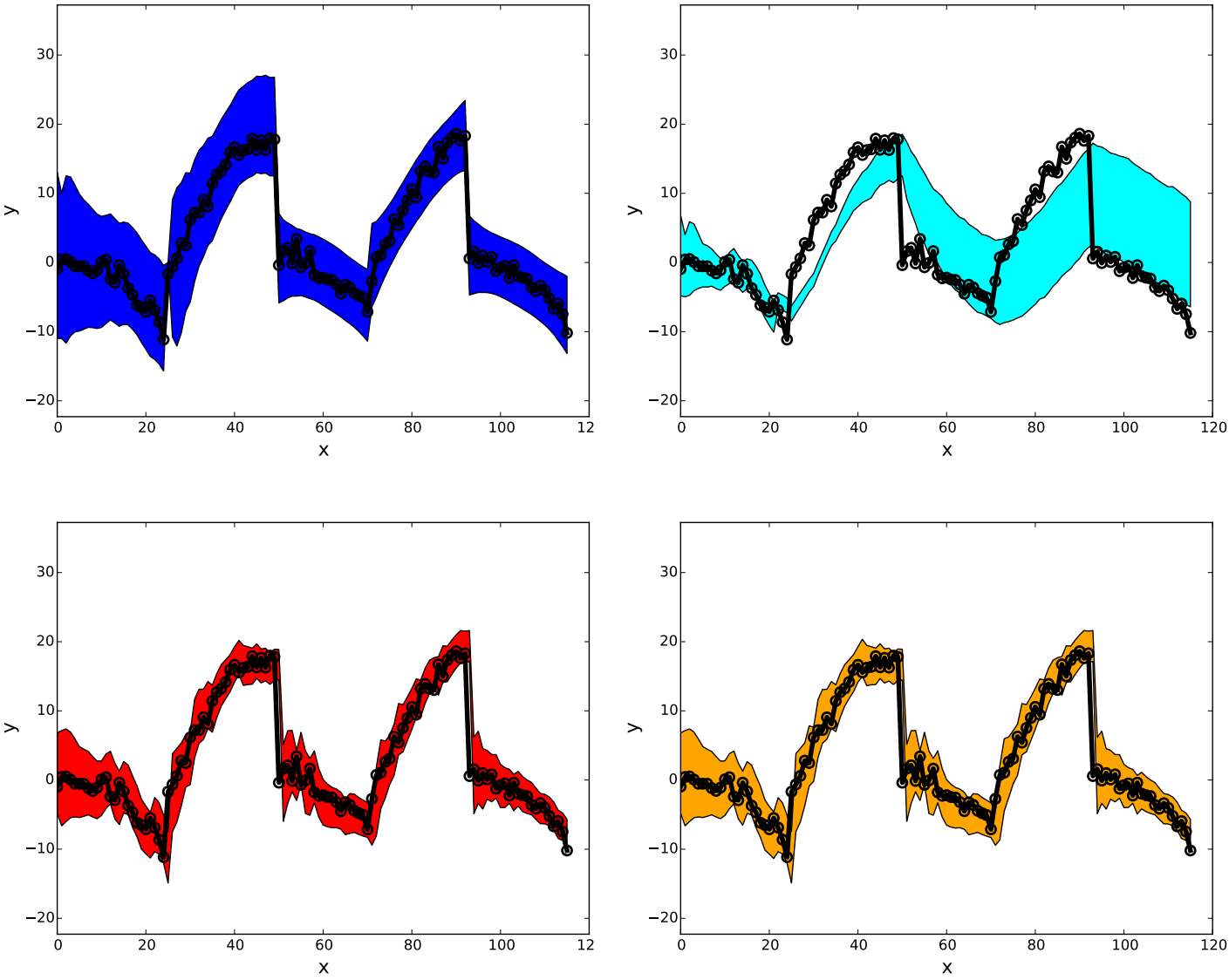


Figure 5.2: Predictive confidence intervals for the oracle $\mathcal{B}^\star$ (top-left), Bayes strategy (top-right), $\texttt{Agg}\text{-}\mathcal{B}$ (bottom-left) and a variant that deletes inadequate experts (bottom-right). At each time t, the confidence interval is built from all past data before time t. the observed signal is in black.

CHAPTER 6
$f = T[f]$ and $\|.\|_?$

Contents

1 MDPs and average reward criterion 97

1.1 Value iteration and the span semi-norm 100

2 Reinforcement learning in the average-reward criterion 105

2.1 A distribution-norm and its dual 106

2.2 A transportation lemma for MDPs 108

2.3 Structured MDPs 110

Take-home message

$$\text{(Gain)} \quad g_\pi(s_1) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} (P_\pi^{t-1}\mu_\pi)(s_1) = (\overline{P}_\pi \mu_\pi)(s_1)\,,$$

$$\text{(Bias)} \quad b_\pi = \sum_{t=1}^{\infty} (P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi\,,$$

Pseudo regret of near-gain optimal policies Let $\star$ be gain-optimal and π be such that $g_\star - g_\pi \leqslant \varepsilon$. Then its cumulative regret when run for T steps is controlled by

$$\mathfrak{R}_{\pi,T} \leqslant \sum_{t=1}^{T} (P_\star^{t-1} - \overline{P}_\star)\mu_\star + (P_\pi^T - I)b_\star + \varepsilon T + \underbrace{\sum_{s,a} \mathbb{E}[N_T^\pi(s,a)]\varphi_a(s)}_{Pseudo-regret}\,,$$

with gap function $\varphi_a(s) = \mu_\star(s) + (P_\star b_\star)(s) - \mu_a(s) + (P_a b_\star)(s)$.

Intrinsic contraction of Bellman operator For any policy π and function f, it holds

$$\mathbb{S}(P_\pi f) \leqslant \frac{1}{2}\|P(\cdot|\overline{s},\pi(\overline{s})) - P(\cdot|\underline{s},\pi(\underline{s}))\|_1 \mathbb{S}(f) \leqslant (1-\gamma^\pi)\mathbb{S}(f)$$

where $\mathbb{S}(f) = \max_s f(s) - \min_s f(s)$, $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_\pi f)(s)$, $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_\pi f)(s)$ and

$$\gamma^\pi = \min_{s_1,s_2} \sum_{s'\in\mathcal{S}} \min\{P(s'|s_1,\pi(s_1)), P(s'|s_2,\pi(s_2))\}$$

Norms should be used with care. Replacing a distribution-independent norm $\|\cdot\|_1$ with a distribution-dependent norm may yield significant improvement.

$$(\nu - \tilde{\nu}, b_\pi) \leqslant \|\nu - \tilde{\nu}\|_{\star,\nu} \|b_\pi - \mathbb{E}_{P_\pi} b_\pi\|_\nu \text{ with } \nu = P_\pi(\cdot|s)$$

In this chapter, we now turn to the setup of reinforcement learning in Markov Decision Processes (MDP). We first revisit some key concepts in the setup of average-regret minimization, before questioning the use of the span semi-norm and its corresponding $\|\cdot\|_1$ dual norm for MDP learning.

1 MDPs AND AVERAGE REWARD CRITERION

We consider an MDP $\mathcal{M}$ with transition function P and reward function R. Let $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ denote a stochastic stationary policy. Let $P(s'|s,\pi(s)) = \mathbb{E}_{A\sim\pi(s)}[P(s'|s,A)]$. Let $P_\pi f$ denote the function such that for all $s \in \mathcal{S}$, $(P_\pi f)(s) = \sum_{s'\in\mathcal{S}} P(s'|s,\pi(s))f(s')$. Likewise, let $\mu_\pi(s) = \mathbb{E}_{A\sim\pi(s)}[\mu(s,A)]$ define the mean reward after choosing action $\pi(s)$ in step s, where $\mu(s,a)$ is the mean of distribution $R(s,a)$.

Definition 6.3 (*Expected cumulative reward*) *The expected cumulative reward of policy π when run for T steps from initial state s_1 is defined as*

$$R_{\pi,T}(s_1) = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t,a_t)\right] = \mu_\pi(s_1) + (P_\pi\mu_\pi)(s_1) + \cdots = \sum_{t=1}^{T}(P_\pi^{t-1}\mu_\pi)(s_1)\,,$$

where $a_t \sim \pi(s_t)$, $s_{t+1} \sim P(\cdot|s_t,a_t)$, and finally $r(s,a) \sim R(s,a)$ with mean $\mu(s,a)$.

Definition 6.6 (*Average gain and bias of proper policies*) *A policy is proper if the limit $\overline{P}_\pi = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} P_\pi^{t-1}$ exists. We call in this case $\overline{P}_\pi$ the average transition operator of policy π. The average gain g_π and the bias function b_π are then defined by*

$$\begin{aligned} g_\pi(s_1) &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(P_\pi^{t-1}\mu_\pi)(s_1) = (\overline{P}_\pi\mu_\pi)(s_1)\,, \\ b_\pi &= \sum_{t=1}^{\infty}(P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi\,. \end{aligned}$$

Definition 6.9 (*Optimal gain and policy*) *The optimal gain is defined by $g_\star(s) = \max_\pi g_\pi(s)$ for each $s \in \mathcal{S}$. Any policy that achieves the optimal gain simultaneously for all s is called optimal, and will be denoted by $\star$ (whenever it exists).*

We recall the following immediate lemma:

Lemma 6.1 (*Key property*)

$$\overline{P}_\pi P_\pi = P_\pi\overline{P}_\pi = \overline{P}_\pi\overline{P}_\pi = \overline{P}_\pi\,.$$

This enables to show several key relations:

Corollary 6.1 (*Bias and Gain*) *The bias function satisfies the following relations*

$$\begin{aligned} \textit{(Fundamental equality)} \quad b_\pi &= [I - P_\pi + \overline{P}_\pi]^{-1}[I - \overline{P}_\pi]\mu_\pi \\ \textit{(Poisson equation)} \quad b_\pi + g_\pi &= \mu_\pi + P_\pi b_\pi \end{aligned}$$

Likewise, the average gain satisfies $P_\pi g_\pi = g_\pi$.

Proof :

For the first relation, we note that by direct application of Lemma 6.1,

$$\begin{aligned} (I - P_\pi + \overline{P}_\pi)b_\pi &= \sum_{t=1}^{\infty}(I - P_\pi)(P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi + \underbrace{\overline{P}_\pi(P_\pi^{t-1} - \overline{P}_\pi)}_{0}\mu_\pi \\ &= \sum_{t=1}^{\infty}(I - P_\pi)P_\pi^{t-1}\mu_\pi - \underbrace{(I - P_\pi)\overline{P}_\pi}_{0}\mu_\pi = \sum_{t=1}^{\infty}\left(P_\pi^{t-1} - P_\pi^t\right)\mu_\pi \,. \end{aligned}$$

Thus, it remains to show that the latter sum equals $I - \overline{P}_\pi$. When P_π is aperiodic, then the limit $\lim_t P_\pi^t$ exists and is equal to $\overline{P}_\pi$. Thus, we easily get

$$\sum_{t=1}^{\infty} P_\pi^{t-1} - P_\pi^t = \lim_{T\to\infty}(I - P_\pi^T) = I - \lim_{T\to\infty} P_\pi^T = I - \overline{P}_\pi \,.$$

The general case is more intricate. See Puterman (1994).

For the second relation, we note that

$$\begin{aligned} P_\pi b_\pi &= \sum_{t=1}^{\infty}(P_\pi^t - \overline{P}_\pi)\mu_\pi = \sum_{t=2}^{\infty}(P_\pi^{t-1} - \overline{P}_\pi)\mu_\pi \\ &= b_\pi - (I - \overline{P}_\pi)\mu_\pi = b_\pi - \mu_\pi + g_\pi \,. \end{aligned}$$

□

Regret to pseudo-regret For any policy π, its cumulative regret $\mathfrak{R}_{\pi,T} = \sum_{t=1}^{T}\left(P_\star^{t-1}\mu_\star - P_\pi^{t-1}\mu_\pi\right)$ with respect to an optimal policy can be decomposed in order to make appear quantities similar to multi-armed bandits. Let $\varphi_a(s) = \mu_\star(s) + (P_\star b_\star)(s) - \mu_a(s) - (P_a b_\star)(s)$ denote the sub-optimality gap of action a in state s. The following regret decomposition result justifies to introduce the notion of pseudo-regret of a policy π, $\sum_{s,a} N_T^\pi(s,a)\varphi_a(s)$. See Puterman (1994) or Burnetas and Katehakis (1997) for similar results. For convenience, the following result is stated in the case of communicating MDPs for which g_π is a constant function. In this case we denote by $g_\pi \in \mathbb{R}$ this constant value.

Lemma 6.2 (*Pseudo-regret*) *Consider a communicating MDP. Let π be a policy such that $g_\star - g_\pi \leqslant \varepsilon$. Then,*

$$\begin{aligned}\mathfrak{R}_{\pi,T} &= \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + (P_\pi^T - I)b_\star + \sum_{t=1}^{T}(1-P_\pi^{t-1})\big(g_\star - g_\pi\big) + \sum_{s,a}\mathbb{E}[N_T^\pi(s,a)]\varphi_a(s)\\ &\leqslant \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + (P_\pi^T - I)b_\star + \varepsilon T + \sum_{s,a}\mathbb{E}[N_T^\pi(s,a)]\varphi_a(s)\,.\end{aligned}$$

Further, as $T\to\infty$, $\sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + (P_\pi^T - I)b_\star \to \overline{P}_\pi b^\star$ is a finite constant.

Proof of Lemma 6.2:

The regret accumulated during T steps by policy π is given by

$$\begin{aligned}\mathfrak{R}_{\pi,T} &= \sum_{t=1}^{T}\Big(P_\star^{t-1}\mu_\star - P_\pi^{t-1}\mu_\pi\Big)\\ &= \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + T\underbrace{(g_\star - g_\pi)}_{\leqslant\varepsilon} + \sum_{t=1}^{T}\Big(g_\pi - P_\pi^{t-1}\mu_\pi\Big)\,.\end{aligned}$$

Then, we use the fact that g_π is a constant function, so that $P_\pi^{t-1}g_\pi = g_\pi$. Thus, it comes

$$\begin{aligned}\sum_{t=1}^{T}\Big(g_\pi - P_\pi^{t-1}\mu_\pi\Big) &= \sum_{t=1}^{T}P_\pi^{t-1}\Big(g_\pi - \mu_\pi\Big) = \sum_{t=1}^{T}P_\pi^{t-1}(P_\pi - I)b_\pi\\ &= \sum_{t=1}^{T}P_\pi^{t-1}\Big[(P_\pi - I)b_\star + (P_\pi - I)(b_\pi - b_\star)\Big]\,,\end{aligned}$$

where in the first line we introduced the bias function b_π. Now, let us introduce the sub-optimality gap function $\varphi_\pi(s) = \mu_\star(s) + (P_\star b_\star)(s) - \mu_\pi(s) - (P_\pi b_\star)(s)$. Since $\mu_\pi = g_\pi + (I - P_\pi)b_\pi$, it comes that

$$\begin{aligned}\varphi_\pi &= g_\star - g_\pi + (I-P_\star)b_\star + P_\star b_\star - (I-P_\pi)b_\pi - P_\pi b_\star\\ &= g_\star - g_\pi + (I-P_\pi)(b_\star - b_\pi)\,.\end{aligned}$$

Hence we deduce that

$$\mathfrak{R}_{\pi,T} = \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + (P_\pi^T - I)b_\star + \underbrace{\sum_{t=1}^{T}(1-P_\pi^{t-1})\big(g_\star - g_\pi\big)}_{\leqslant\varepsilon T} + \sum_{t=1}^{T}P_\pi^{t-1}\varphi_\pi\,.$$

This result shows that in order to control the regret, it is enough to control the following term

$$\sum_{t=1}^{T}P_\pi^{t-1}\varphi_\pi = \sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi_a(s)\,,$$

that we naturally call the pseudo-regret, by analogy with the bandit setting. Indeed, it comes

$$\begin{aligned}\sum_{t=1}^{T} P_\pi^{t-1}\varphi_\pi &= \sum_{t=1}^{T}\mathbb{E}_{s_{t-1}}[\varphi_\pi(s_{t-1})] \\ &= \sum_{s,a}\varphi_a(s)\sum_{t=1}^{T}\mathbb{E}_{s_{t-1}}[\mathbb{I}\{s_{t-1}=s,\pi(s)=a\}] = \sum_{s,a}\varphi_a(s)\mathbb{E}[N_T^\pi(s,a)]\,,\end{aligned}$$

□

1.1 Value iteration and the span semi-norm

A natural strategy in order to find an optimal policy in a (perfectly known) MDP is the value iteration strategy, defined as follows from P and μ:

Definition 6.12 (*Value iteration*) *The value iteration procedure defines a sequence of functions $(u_n)_{n\in\mathbb{N}}$ and policies $(\pi_n)_{n\in\mathbb{N}}$ according to the following equations*

$$\forall n\in\mathbb{N}\begin{cases} u_{n+1}(s) = \max_{a\in\mathcal{A}}\mu(s,a) + (P_a u_n)(s)\,, & \text{where } u_0 = 0 \\ \pi_{n+1}(s) = \mathcal{U}\Big(\mathrm{Argmax}_{a\in\mathcal{A}}\mu(s,a)+(P_a u_n)(s)\Big) & \text{where } \mathcal{U}(\mathcal{B}) \text{ denotes the uniform distribution over } \mathcal{B}\end{cases}$$

Corollary 6.2 (*Value and gain*) *It holds that*

$$\forall n\in\mathbb{N},\quad \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \leqslant g_{\pi_{n+1}} \leqslant g_\star \leqslant \overline{P}_\star[u_{n+1}-u_n]\,.$$

Further, for any n such that $\mathbb{S}(u_{n+1}-u_n)\leqslant\varepsilon$, then

$$g_\star - g_{\pi_{n+1}} \leqslant \varepsilon,\quad |u_{n+1}-u_n-g_\star|\leqslant\varepsilon \quad \text{and} \quad |u_{n+1}-u_n-g_{\pi_n+1}|\leqslant\varepsilon\,.$$

Proof :

Let us first note that since $\overline{P}_\star P_\star = \overline{P}_\star$, then for any function f we have $g_\star = \overline{P}_\star[\mu_\star + P_\star f - f]$. Applying this to the function u_n yields

$$\begin{aligned} g_\star &= \overline{P}_\star[\mu_\star + P_\star u_n - u_n] \\ &\leqslant \overline{P}_\star[\mu_{\pi_{n+1}} + P_{\pi_{n+1}}u_n - u_n] \\ &= \overline{P}_\star(u_{n+1}-u_n)\,,\end{aligned}$$

where in the second line we used the maximal property of π_{n+1}. On the other hand, we use the equality

$$g_{\pi_{n+1}} = \overline{P}_{\pi_{n+1}}[\mu_{\pi_{n+1}} + P_{\pi_{n+1}}u_n - u_n] = \overline{P}_{\pi_{n+1}}(u_{n+1}-u_n)\,,$$

together with the fact that by optimality of $\star$, $g_\star \geqslant g_{\pi_{n+1}}$

Indeed, it holds on the one hand

$$\begin{aligned} g_\star - g_{\pi_{n+1}} &\leqslant \overline{P}_\star[u_{n+1}-u_n] - \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \\ &\leqslant \max_{s\in\mathcal{S}}(u_{n+1}-u_n)(s) - \min_{s\in\mathcal{S}}[u_{n+1}-u_n] = \mathbb{S}(u_{n+1}-u_n)\,. \end{aligned}$$

On the other hand, using similar steps,

$$\begin{aligned} 0 &\leqslant \overline{P}_\star[u_{n+1}-u_n] - g_\star \leqslant \max_{s\in\mathcal{S}}[u_{n+1}-u_n] - g_\star \\ &\leqslant \max_{s\in\mathcal{S}}[u_{n+1}-u_n] - \overline{P}_{\pi_{n+1}}[u_{n+1}-u_n] \leqslant \mathbb{S}(u_{n+1}-u_n)\,. \end{aligned}$$

Thus, for all $s \in \mathcal{S}$, $(u_{n+1}-u_n)(s) - g_\star \leqslant \varepsilon$. Likewise, we get the reverse inequality $0 \leqslant g_\star - \min_{s\in\mathcal{S}}(u_{n+1}-u_n)(s) \leqslant \mathbb{S}(u_{n+1}-u_n) \leqslant \varepsilon$. The last bound is immediate from the relation $g_{\pi_{n+1}} = \overline{P}_{\pi_{n+1}}(u_{n+1}-u_n)$. □

Corollary 6.3 (*Value and bias*) *It holds for all $n \in \mathbb{N}$,*

$$|(I - P_{\pi_{n+1}})(u_n - b_{\pi_{n+1}})| \leqslant \mathbb{S}(u_{n+1}-u_n)$$

Proof :

Indeed, we have by Corollary 6.2 $|u_{n+1}-u_n-g_{\pi_{n+1}}| \leqslant \mathbb{S}(u_{n+1}-u_n)$, which rewrites $|(I-P_{\pi_{n+1}})u_n - \mu_{\pi_{n+1}} - g_{\pi_{n+1}}| \leqslant \mathbb{S}(u_{n+1}-u_n)$. We conclude by remarking that $\mu_{\pi_{n+1}} - g_{\pi_{n+1}} = (I - P_{\pi_{n+1}})b_{\pi_{n+1}}$. □

It is interesting to take a look at the regret of running the policy π_{n+1} for T steps. We slightly revisit the regret decomposition lemma below.

Proposition 6.1 (*Regret of value iteration policy*) *Let ε be a non-negative constant and let n be such that $\mathbb{S}(u_{n+1}-u_n) \leqslant \varepsilon$. Assume that for all policy π, $\overline{P}_\pi$ is well defined. Then*

$$\begin{aligned} \mathfrak{R}_{\pi_{n+1},T}(s_1) &= R_{\star,T}(s_1) - R_{\pi_{n+1},T}(s_1) \\ &\leqslant \sum_{t=1}^{T}(P_\star^{t-1} - \overline{P}_\star)\mu_\star + T\varepsilon + (P_{\pi_{n+1}}^T - I)b_{\pi_{n+1}}\,. \end{aligned}$$

The first term is MDP dependent and depends on the mixing time of the chain induced by the policy $\star$. The remaining term satisfies

$$(P_{\pi_{n+1}}^T - I)b_{\pi_{n+1}} \leqslant \frac{1}{2}\|P_{\pi_{n+1}}^T - I\|_1 \mathbb{S}(b_{\pi_{n+1}}) \leqslant \mathbb{S}(b_{\pi_{n+1}})\,.$$

Proof :

$$\begin{aligned}\mathfrak{R}_{\pi_{n+1},T} &= \sum_{t=1}^{T}\Big(P_\star^{t-1}\mu_\star - P_{\pi_{n+1}}^{t-1}\mu_{\pi_{n+1}}\Big)\\ &= \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + T(g_\star - g_{\pi_{n+1}}) + \sum_{t=1}^{T}(\overline{P}_{\pi_{n+1}} - P_{\pi_{n+1}}^{t-1})\mu_{\pi_{n+1}}\,.\end{aligned}$$

In order to take care of the last sum, let us note that $\mu_{\pi_{n+1}} = g_{\pi_{n+1}} + (I - P_{\pi_{n+1}})b_{\pi_{n+1}}$. Thus, plugging-in this equality in the regret expression, and using the fact that $\overline{P}_{\pi_{n+1}} g_{\pi_{n+1}} = g_{\pi_{n+1}}$, it comes

$$\begin{aligned}(\overline{P}_{\pi_{n+1}} - P_{\pi_{n+1}}^{t-1})\mu_{\pi_{n+1}} &= (\overline{P}_{\pi_{n+1}} - P_{\pi_{n+1}}^{t-1})\Big(g_{\pi_{n+1}} + (I - P_{\pi_{n+1}})b_{\pi_{n+1}}\Big)\\ &= (\overline{P}_{\pi_{n+1}} - P_{\pi_{n+1}}^{t-1})(I - P_{\pi_{n+1}})b_{\pi_{n+1}}\\ &= (P_{\pi_{n+1}}^{t} - P_{\pi_{n+1}}^{t-1})b_{\pi_{n+1}}\,.\end{aligned}$$

Combining this equality together with the relation $g_\star - g_{\pi_{n+1}} \leqslant \mathbb{S}(u_{n+1} - u_n)$, we get

$$\begin{aligned}\mathfrak{R}_{\pi_{n+1},T} &= \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + T(g_\star - g_{\pi_{n+1}}) + \sum_{t=1}^{T}(P_{\pi_{n+1}}^{t} - P_{\pi_{n+1}}^{t-1})b_{\pi_{n+1}}\\ &\leqslant \sum_{t=1}^{T}(P_\star^{t-1}-\overline{P}_\star)\mu_\star + T\varepsilon + (P_{\pi_{n+1}}^{T} - I)b_{\pi_{n+1}}\,.\end{aligned}$$

□

A more precise look at contraction coefficients In order to better understand the fundamental reason why value iteration makes sense, we study the contraction properties of the operator associated to the Poisson fixed-point equation. Indeed, despite being in an average-reward setup, there is a contraction property, as explained for instance in Puterman (1994).

Lemma 6.3 (*Sandwich bounds*)

$$\forall k \in \mathbb{N},\quad P_{\pi_{k+2}}(u_{k+1} - u_k) \geqslant u_{k+2} - u_{k+1} \geqslant P_{\pi_{k+1}}(u_{k+1} - u_k)\,.$$

Proof of Lemma 6.3:

The difference of values after one step is given by:

$$\forall s \in \mathcal{S},\quad u_{k+2}(s) - u_{k+1}(s) = \mu(s,\pi_{k+2}) - \mu(s,\pi_{k+1}) + (P_{\pi_{k+2}}u_{k+1})(s) - (P_{\pi_{k+1}}u_k)(s)$$

Using this equality, together with the maximal property of π_{k+2} and π_{k+1}, we deduce the two following inequalities

$$\begin{array}{rcl} \forall s \in \mathcal{S}, \quad u_{k+2}(s) - u_{k+1}(s) & \geqslant & \mu(s,\pi_{k+1}) - \mu(s,\pi_{k+1}) \\ & & +(P_{\pi_{k+1}}(u_{k+1}-u_k))(s) = (P_{\pi_{k+1}}(u_{k+1}-u_k))(s) \\ u_{k+2}(s) - u_{k+1}(s) & \leqslant & (P_{\pi_{k+2}}(u_{k+1}-u_k))(s)\,. \end{array}$$

That is, with function notations,

$$\forall k \in \mathbb{N}, \quad P_{\pi_{k+2}}(u_{k+1}-u_k) \geqslant u_{k+2}-u_{k+1} \geqslant P_{\pi_{k+1}}(u_{k+1}-u_k)\,.$$

□

Likewise, the following generalization holds:

Lemma 6.4 (*Multi-steps sandwich bounds*) *Let us introduce the notation $P_\star^{k':k} = P_{\pi_k} \circ \cdots \circ P_{\pi_{k'}}$ for all $k' < k$. Then, it holds*

$$\forall k \in \mathbb{N}, \forall k' < k \quad P_\star^{k'+2:k+1}(u_{k'+1}-u_{k'}) \geqslant u_{k+1}-u_k \geqslant P_\star^{k'+1:k}(u_{k'+1}-u'_k)\,.$$

We now make use of Lemma 6.4 in order to recall a fundamental contraction inequality (see Puterman (1994)). Let us introduce for convenience the notation $\Delta_k(s) = u_{k+1}(s) - u_k(s)$. Also, for a function f defined on $\mathcal{S}$, let us introduce the span defined by $\mathbb{S}(f) = \max_{s\in\mathcal{S}} f(s) - \min_{s\in\mathcal{S}} f(s)$, which is a semi-norm.

Lemma 6.5 (*Value contraction of the Span*) *Let $\Delta_k(s) = u_{k+1}(s) - u_k(s)$ be the difference of value functions between iterations k and $k+1$. Then*

$$\mathbb{S}(\Delta_{k+1}) \leqslant \underbrace{\Big(1 - \sum_{s'\in\mathcal{S}} \min\{P(s'|\overline{s},\pi_{k+2}(\overline{s})), P(s'|\underline{s},\pi_{k+1}(\underline{s}))\}\Big)}_{\frac{1}{2}\|P(\cdot|\overline{s},\pi_{k+2}(\overline{s}))-P(\cdot|\underline{s},\pi_{k+1}(\underline{s}))\|_1} \mathbb{S}(\Delta_k) \leqslant \big(1-\gamma\big)\mathbb{S}(\Delta_k)\,,$$

where we introduced $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_{\pi_{k+2}}\Delta_k)(s)$, $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_{\pi_{k+1}}\Delta_k)(s)$ and finally

$$\gamma \;=\; \min_{s_1,s_2\in\mathcal{S}} \min_{\pi,\pi'} \sum_{s'\in\mathcal{S}} \min\{P(s'|s_1,\pi(s_1)), P(s'|s_2,\pi'(s_2))\}\,.$$

Proof :

By Lemma 6.3, it holds

$$\mathbb{S}(\Delta_{k+1}) \leqslant \max_{s\in\mathcal{S}}(P_{\pi_{k+2}}\Delta_k)(s) - \min_{s\in\mathcal{S}}(P_{\pi_{k+1}}\Delta_k)(s)\,.$$

At this point it is convenient to introduce $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_{\pi_{k+2}}\Delta_k)(s)$ and $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_{\pi_{k+1}}\Delta_k)(s)$. Developing each operator term, and introducing the quantity $\gamma(s') = \min\{P(s'|\overline{s},\pi_{k+2}(\overline{s})), P(s'|\underline{s},\pi_{k+1}(\underline{s}))\}$, it comes

$$\begin{aligned}
\mathbb{S}(\Delta_{k+1}) &\leqslant \sum_{s'\in\mathcal{S}}\Big(P(s'|\overline{s},\pi_{k+2}(\overline{s})) - P(s'|\underline{s},\pi_{k+1}(\underline{s}))\Big)\Delta_k(s') \\
&= \sum_{s'\in\mathcal{S}}\underbrace{\Big(P(s'|\overline{s},\pi_{k+2}(\overline{s})) - \gamma(s')\Big)}_{\geqslant 0}\Delta_k(s') - \sum_{s'\in\mathcal{S}}\underbrace{\Big(P(s'|\underline{s},\pi_{k+1}(\underline{s})) - \gamma(s')\Big)}_{\geqslant 0}\Delta_k(s') \\
&\leqslant \sum_{s'\in\mathcal{S}}\underbrace{\Big(P(s'|\overline{s},\pi_{k+2}(\overline{s})) - \gamma(s')\Big)}_{\geqslant 0}\max_{s\in\mathcal{S}}\Delta_k(s) - \sum_{s'\in\mathcal{S}}\underbrace{\Big(P(s'|\underline{s},\pi_{k+1}(\underline{s})) - \gamma(s')\Big)}_{\geqslant 0}\min_{s\in\mathcal{S}}\Delta_k(s) \\
&= \Big(1 - \sum_{s'\in\mathcal{S}}\gamma(s')\Big)\mathbb{S}(\Delta_k)\,.
\end{aligned}$$

Further, note that

$$\begin{aligned}
\sum_{s'\in\mathcal{S}}\gamma(s') &\geqslant \min_{\pi,\pi'}\sum_{s'\in\mathcal{S}}\min\{P(s'|\overline{s},\pi(\overline{s})), P(s'|\underline{s},\pi'(\underline{s}))\} \\
&\geqslant \min_{s_1,s_2\in\mathcal{S}}\min_{\pi,\pi'}\sum_{s'\in\mathcal{S}}\min\{P(s'|s_1,\pi(s_1)), P(s'|s_2,\pi'(s_2))\}
\end{aligned}$$

□

Likewise, this lemma generalizes to multi-steps. To this end, we introduce the matrix $P_\star^{k':k}(s'|s)$ such that $(P_\star^{k':k}f)(s) = \sum_{s'\in\mathcal{S}}P_\star^{k':k}(s'|s)f(s')$, and for an arbitrary stationary policy π, we introduce P_π^k its k-step transition matrix.

Lemma 6.6 (*Multi-steps value contraction of the Span*)

$$\forall k\in\mathbb{N}, \forall k' < k, \quad \mathbb{S}(\Delta_k) \leqslant \Big(1 - \sum_{s'\in\mathcal{S}}\min\{P_\star^{k'+2:k+1}(s'|\overline{s}), P_\star^{k'+1:k}(s'|\underline{s})\}\Big)\mathbb{S}(\Delta_{k'}) \leqslant \big(1-\gamma_{k'-k}\big)\mathbb{S}(\Delta_{k'})\,,$$

where we introduced $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_\star^{k'+2:k+1}\Delta_{k'})(s)$, $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_\star^{k'+1:k}\Delta_{k'})(s)$ *and finally*

$$\gamma_{k'-k} = \min_{s_1,s_2\in\mathcal{S}}\min_{\pi,\pi'}\sum_{s'\in\mathcal{S}}\min\{P_\pi^{k-k'}(s'|s_1), P_\pi^{k-k'}(s'|s_2)\}\,.$$

Another closely related property is the following one:

Lemma 6.7 (*Policy contraction of the Span*) *For any policy π and function f, it holds*

$$\mathbb{S}(P_\pi f) \leqslant \frac{1}{2}\|P(\cdot|\overline{s},\pi(\overline{s})) - P(\cdot|\underline{s},\pi(\underline{s}))\|_1 \mathbb{S}(f) \leqslant (1-\gamma^\pi)\mathbb{S}(f)\,,$$

where we introduced $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_\pi f)(s)$, $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_\pi f)(s)$ and finally

$$\gamma^\pi \;=\; \min_{s_1,s_2}\sum_{s'\in\mathcal{S}} \min\{P(s'|s_1,\pi(s_1)), P(s'|s_2,\pi(s_2))\}\,.$$

Likewise, $$\mathbb{S}(P_\pi^k f) \leqslant \frac{1}{2}\|P_\pi^k(\cdot|\overline{s}) - P_\pi^k(\cdot|\underline{s})\|_1 \mathbb{S}(f) \leqslant (1-\gamma_k^\pi)\mathbb{S}(f)\,,$$

where we introduced $\overline{s} = \arg\max_{s\in\mathcal{S}}(P_\pi^k f)(s)$, $\underline{s} = \arg\min_{s\in\mathcal{S}}(P_\pi^k f)(s)$ and finally

$$\gamma_k^\pi \;=\; \min_{s_1,s_2}\sum_{s'\in\mathcal{S}} \min\{P_\pi^k(s'|s_1), P_\pi^k(s'|s_2)\}\,.$$

2 Reinforcement learning in the average-reward criterion

The previous section recalls some fundamental principles of MDP, mostly focusing on the span semi-norm and contraction properties. When the transition and reward functions are unknown, learning in a single-stream of interactions in order to minimize the regret is challenging. Similarly to the multi-armed bandit setup, lower-bounds have been derived using a change of measure argument for specific sets of MDPs such as ergodic MDPs, see Burnetas and Katehakis (1997) or Graves and Lai (1997). This is achieved first thanks to the form of the pseudo-regret $\sum_{s,a}\mathbb{E}[N_T(s,a)]\varphi_a(s)$ that resembles that of multi-armed bandit problems but with a different gap function $\varphi_a(s)$, second that in an ergodic MDP, $\liminf_{T\to\infty}\mathbb{E}[N_T(s)]/T > 0$ for every strategy, which enables to asymptotically replace the possibly complicated random variables $N_T(s)$ with a constant T. However, informative lower bounds in full generality are still missing. Further, one could argue that the existing lower bounds are weak in some sense.

Despite these holes in the RL literature, strategies have been derived to address regret minimization in average-reward MDPs. One of the most promising one is the `UCRL2` strategy introduced in Auer et al. (2009), following intuitions from the multi-armed bandit literature. However, the initial version and analysis of the strategy is a little crude. We provide in Maillard et al. (2014), and Talebi and Maillard (2018) a refinement of one of the key steps of the proofs.

`UCRL2` We now briefly present the `UCRL2` algorithm from Auer et al. (2009). At a high level, `UCRL2` follows the optimistic principle by trying to compute $\overline{\pi}_t^+ = \arg\max_{\pi:\mathcal{S}\to\mathcal{A}}\max\{g_\pi^{\mathcal{M}} : \mathcal{M}\in\mathscr{M}_t\}$ where $g_\pi^{\mathcal{M}}$ is the average-gain for policy π in MDP $\mathcal{M}$, and

$$\mathscr{M}_t \;=\; \Big\{(\mathcal{S},\mathcal{A},\tilde{p},\tilde{\nu}) : \forall (s,a)\in\mathcal{S}\times\mathcal{A},\quad |\mu_{N_t(s,a)}(s,a) - \tilde{\mu}(\cdot|s,a)| \leqslant \tilde{b}_t^H(s,a,\frac{\delta}{2SA})$$
$$\text{and}\quad \|p_{N_t(s,a)}(\cdot|s,a) - \tilde{p}(\cdot|s,a)\|_1 \leqslant \tilde{b}_t^W(s,a,\frac{\delta}{2SA})\Big\}\,.$$

Here $\mu_n(s,a)$ denotes the empirical mean built using n i.i.d. rewards from $\nu(s,a)$, $p_n(\cdot|s,a)$ is the empirical distribution built using n i.i.d. observations from $p(\cdot|s,a)$, $N_t(s,a)$ is the total number of observations of state action pair (s,a) up to time t, and finally $\tilde{b}_t^H$ and $\tilde{b}_t^W$ are the two functions

$$\tilde{b}_t^H(s,a,\delta) = \sqrt{\frac{3.5\log(t/\delta)}{N_t(s,a)\vee 1}}\,, \quad \tilde{b}_t^W(s,a,\delta) = \sqrt{\frac{7S\log(t/S\delta)}{N_t(s,a)\vee 1}}\,,$$

respectively based on a Hoeffding and Weissman inequality where $\vee$ denotes the max operator.

The computation of $\overline{\pi}_t^+$ is achieved approximately by an Extended Value Iteration (EVI) algorithm that builds a near-optimistic policy π_t^+ and MDP $\mathcal{M}_t^+$ such that $g_{\pi_t^+}^{\mathcal{M}_t^+} \geqslant \max_{\pi, M\in\mathscr{M}_t} g_\pi^{\mathcal{M}} - \frac{1}{\sqrt{t}}$.

Finally, `UCRL2` does not recompute π_t^+ at each time step. Instead, it proceeds in internal episodes $k = 0, \dots$ and computes π_t^+ only at the starting time t_k of each episode, defined as $t_1 = 1$ and for all $k > 1$,

$$t_k = \min\left\{t > t_{k-1}; \exists s,a\colon n_{t_k:t}(s,a) \geqslant \max\{N_{t_k}(s,a),1\}\right\},$$

where $n_{t_1:t_2}(s,a)$ denotes the number of observations of state-action pair (s,a) between time t_1 and t_2.

Remark 6.1 *The bounds $\tilde{b}^H$ and $\tilde{b}^W$ were obtained from simple union bounds with a slightly loose analysis. However, we know from the Laplace method that $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$ and any $[0,1]$-bounded distribution with mean $\mu(s,a)$:*

$$\mathbb{P}\left(\exists t\in\mathbb{N} \quad |\mu_{N_t(s,a)}(s,a) - \mu(s,a)| \geqslant b^H_{N_t(s,a)}(\delta)\right) \leqslant \delta\,, \text{ with } b_n^H(\delta) = \sqrt{\frac{(1+\frac{1}{n})\log(2\sqrt{n+1}/\delta)}{2n}}\,.$$

Further, for any discrete distribution $p(\cdot|s,a)$ on $\mathcal{S}$ with support of size $K \leqslant |\mathcal{S}|$

$$\mathbb{P}\left(\exists t\in\mathbb{N}\, \|p_{N_t(s,a)}(\cdot|s,a) - p(\cdot|s,a)\|_1 \geqslant b^W_{N_t(s,a)}(\delta)\right) \leqslant \delta\,, \text{ with } b_n^W(\delta) = \sqrt{\frac{2(1+\frac{1}{n})\log\left(\sqrt{n+1}\frac{2^K-2}{\delta}\right)}{n}}\,.$$

This suggests to replace $b_t^H(s,a,\frac{\delta}{2SA})$ with $b^H_{N_t(s,a)}(\frac{\delta}{2SA})$, and $b_t^W(s,a,\frac{\delta}{2SA})$ with $b^W_{N_t(s,a)}(\frac{\delta}{2SA})$ in the definition of $\mathscr{M}_t$, to ensure that the true MDP belongs to $\mathscr{M}_t$ for all time steps with probability higher than $1-\delta$.

There is also no reason to use the $\|\cdot\|_1$ distance to control the transition probabilities. Other norms, or simply using individual Bernstein bounds for each element of the transition vector, or a `KL` contrast between the distributions makes more sense. We have explored such questions in Maillard et al. (2014) and Talebi and Maillard (2018).

2.1 A distribution-norm and its dual

In this section, we further question the use of the $\|\cdot\|_1$ norm. In Machine Learning (ML), norms often play a crucial role in obtaining performance bounds. One typical example is the following. Let $\mathcal{X}$ be a measurable space equipped with an unknown probability measure $\nu \in \mathcal{M}_1(\mathcal{X})$ with density p. Based on some procedure, an algorithm produces a candidate measure $\tilde{\nu} \in \mathcal{M}_1(\mathcal{X})$ with density $\tilde{p}$. One is then interested in the loss with respect to a continuous function f. It is natural to look at the mismatch between ν and $\tilde{\nu}$ on f. That is

$$(\nu - \tilde{\nu}, f) = \int_{\mathcal{X}} f(x)(\nu-\tilde{\nu})(dx) = \int_{\mathcal{X}} f(x)(p(x)-\tilde{p}(x))dx\,.$$

Such a situation appears in the context of Markov decision processes, due to the Bellman and Poisson equations. Indeed,

$$b_\pi(s) + g_\pi(s) = \mu_\pi(s) + \int_{\mathcal{S}} P_\pi(s'|s) b_\pi(s') ds'\,.$$

Hence, we are naturally led towards controlling terms such as $(\nu - \tilde{\nu}, b_\pi)$, for $\nu = P_\pi(\cdot|s)$ and its estimate $\tilde{\nu}$.

A typical bound on this quantity is obtained by applying a Hölder inequality to f and $p - \tilde{p}$, which gives $(\nu - \tilde{\nu}, f) \leqslant \|p - \tilde{p}\|_1 \|f\|_\infty$. Assuming a bound is known for $\|f\|_\infty$, this inequality can be controlled with a bound on $\|p - \tilde{p}\|_1$. When $\mathcal{X}$ is finite and $\tilde{p}$ is the empirical distribution $\widehat{p}_n$ estimated from n i.i.d. samples of p, results such as Weissman et al. (2003) can be applied to bound this term with high probability.

However, in this learning problem, what matters is not f but the way f behaves with respect to ν. Thus, trying to capture the properties of f via the distribution-free $\|f\|_\infty$ bound is not satisfactory. So we propose, instead, a norm $\|\cdot\|_\nu$ driven by ν. A well-behaving f will have a small norm $\|f\|_\nu$, whereas a badly-behaving f will have a large norm $\|f\|_\nu$. Every distribution has a natural norm associated with it that measures the quadratic variations of f with respect to ν. This quantity is at the heart of many key results in mathematical statistics, and is formally defined by

$$\|f\|_\nu = \sqrt{\int_{\mathcal{X}} \Big(f(x) - \mathbb{E}_\nu f\Big)^2 \nu(dx)}\,. \tag{6.1}$$

To get a norm, we restrict $\mathcal{C}(\mathcal{X})$ to the space of continuous functions $\mathcal{E}_\nu = \{f \in \mathcal{C}(\mathcal{X}) : \|f\|_\nu < \infty, \mathrm{supp}(\nu) \subset \mathrm{supp}(f), \mathbb{E}_\nu f = 0\}$. We then define the corresponding dual space in a standard way by $\mathcal{E}_\nu^\star = \{\mu : \|\mu\|_{\star,\nu} < \infty\}$ where

$$\|\mu\|_{\star,\nu} = \sup_{f \in \mathcal{E}_\nu} \frac{\int_x f(x)\mu(dx)}{\|f\|_\nu}\,.$$

Interestingly, this optimization problem has a fully closed form expression. Also, note that for $f \in \mathcal{E}_\nu$, using the fact the $\nu(\mathcal{X}) = \tilde{\nu}(\mathcal{X}) = 1$ and that $x \to f(x) - \mathbb{E}_\nu f$ is a zero mean function, we immediately have

$$\begin{aligned}(\nu - \tilde{\nu}, f) &= (\nu - \tilde{\nu}, f - \mathbb{E}_\nu f) \\ &\leqslant \|p - \tilde{p}\|_{\star,\nu} \|f - \mathbb{E}_\nu f\|_\nu\,.\end{aligned} \tag{6.2}$$

The key difference with the generic Hölder inequality is that $\|\cdot\|_\nu$ is now capturing the behavior of f with respect to ν, as opposed to $\|\cdot\|_\infty$. Conceptually, using a quadratic norm instead of an L1 norm, as we do here, is analogous to moving from Hoeffding's inequality to Bernstein's inequality in the framework of concentration inequalities.

We are interested in situations where $\|f\|_\nu$ is much smaller than $\|f\|_\infty$. That is, f is well-behaving with respect to ν. In such cases, we can get an improved bound $\|p - \tilde{p}\|_{\star,\nu} \|f - \mathbb{E}_\nu f\|_\nu$ instead of the best possible generic bound $\inf_{c \in \mathbb{R}} \|p - \tilde{p}\|_1 \|f - c\|_\infty$.

Simply controlling either $\|p - \tilde{p}\|_{\star,\nu}$ (respectively $\|p - \tilde{p}\|_1$) or $\|f\|_\nu$ (respectively $\|f\|_\infty$) is not enough. What matters is the product of these quantities. For our choice of norm, we show that $\|p - \tilde{p}\|_{\star,\nu}$ concentrates at essentially the same speed as $\|p - \tilde{p}\|_1$, but $\|f\|_\infty$ is typically much larger than $\|f\|_\nu$ for the typical functions met in the analysis of MDPs. We do not claim that the norm defined in equation (6.1) is the best norm that leads to a minimal $\|p - \tilde{p}\|_{\star,\nu} \|f - \mathbb{E}_\nu f\|_\nu$, but we show that it is an interesting candidate.

We proceed in two steps. First, we design a concentration bound for $\|p - \widehat{p}_n\|_{\star,\nu}$ that is not much larger than the Weissman et al. (2003) bound on $\|p - \widehat{p}_n\|_1$. (Note that $\|p - \widehat{p}_n\|_{\star,\nu}$ must be larger than $\|p - \widehat{p}_n\|_1$ as it captures a refined property). Second, we consider RL in an MDP where p represents the transition kernel of a station-action pair and f represents the value (bias) function of the MDP for a policy. The value function and p are strongly linked by construction, and the distribution-norm helps us capture their interplay. We show in Maillard et al. (2014) that common benchmark MDPs have optimal value functions with small $\|\cdot\|_\nu$ norm. This

naturally introduces a new way to capture the hardness of MDPs, besides the diameter (Jaksch et al., 2010) or the span (Bartlett and Tewari, 2009). Our formal notion of MDP hardness is summarized in Definitions 6.15 and 6.18, for discounted and undiscounted MDPs, respectively:

Definition 6.15 (*Hardness of discounted MDP*) *Let $M = (\mathcal{S}, \mathcal{A}, r, p, \gamma)$ be a γ-discounted MDP, with reward function r and transition kernel p. We denote V^π the value function corresponding to a policy π (Puterman, 1994). We define the hardness of policy π in MDP M by*

$$C_M^\pi = \max_{s,a \in \mathcal{S}\times\mathcal{A}} \|V^\pi\|_{p(\cdot|s,a)} .$$

Definition 6.18 (*Hardness of undiscounted MDP*) *Let $M = (\mathcal{S}, \mathcal{A}, r, p)$ be an undiscounted MDP, with reward function r and transition kernel p. We denote by h^π the bias function for policy π (Puterman, 1994, Jaksch et al., 2010). We define the hardness of policy π in MDP M by the quantity*

$$C_M^\pi = \max_{s,a \in \mathcal{S}\times\mathcal{A}} \|h^\pi\|_{p(\cdot|s,a)} .$$

In the discounted setting with bounded rewards in $[0,1]$, $V^\pi \leqslant \frac{1}{1-\gamma}$ and thus $C_M^\pi \leqslant \frac{1}{1-\gamma}$ as well. In the undiscounted setting, then $\|h^\pi\|_{p(\cdot|s,a)} \leqslant \mathbb{S}(h^\pi)$, and thus $C_M^\pi \leqslant \mathbb{S}(h^\pi)$. We define the class of C-"hard" MDPs by $\mathfrak{M}_C = \left\{ M : C_M^{\pi^*} \leqslant C \right\}$. That is, the class of MDPs with optimal policy having a low hardness, or for short, *MDPs with low hardness*.

Important note It may be tempting to think that, since the above definition captures a notion of variance, an MDP that is very noisy will have a high hardness. However this reasoning is incorrect. The hardness of an MDP is not the variance of a roll-out trajectory, but rather captures the variations of the value (or the bias value) function with respect to the transition kernel. For example, consider a fully connected MDP with transition kernel that transits to every state uniformly at random, but with a constant reward function. In this trivial MDP, $C_M^\pi = 0$ for all policies π, even though the MDP is extremely noisy because the value function is constant. In general MDPs, the hardness depends on how varying the value function is at the possible next states and on the distribution over next states. Note also that we use the term hardness rather than complexity to avoid confusion with such concepts as Rademacher or VC complexity.

2.2 A transportation lemma for MDPs

In Filippi (2010), the authors have introduced another variant based on a KL contrast instead of a $\|\cdot\|_1$ norm, thus leading to an algorithm called KL-UCRL2. Unfortunately, the original analysis of KL-UCRL2 was however not showing any improvement over that of UCRL2. In Talebi and Maillard (2018), we revisit this analysis in order to make appear more explicitly the benefit of using the KL divergence instead of the $\|\cdot\|_1$ distance. The fundamental result on which this improvement is based is a transportation lemma, that we recall now. We refer to the full paper for more details and the results one can obtain for KL-UCRL2.

We now recall a powerful result, directly related to duality formulas for the KL divergence, known as the transportation lemma. This gives an alternative to the bound $(Q - P, f)$, for two probability measures Q, P.

Lemma 6.8 (*Transportation Lemma*) *For any function f, let us introduce $\varphi_f : \lambda \mapsto \log \mathbb{E}_P \exp(\lambda(f(X) - \mathbb{E}_P[f]))$. Whenever φ_f is defined on some possibly unbounded interval $0 \in I$, define its dual $\varphi_{\star,f}(x) = \sup_{\lambda \in I} \lambda x - \varphi_f(\lambda)$. Then it holds*

$$\begin{aligned} \forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \varphi_{+,f}^{-1}(KL(Q,P)) \quad \text{where } \varphi_{+,f}^{-1}(t) = \inf\{x \geqslant 0 : \varphi_{\star,f}(x) > t\} \\ \forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\geqslant \varphi_{-,f}^{-1}(KL(Q,P)) \quad \text{where } \varphi_{-,f}^{-1}(t) = \sup\{x \leqslant 0 : \varphi_{\star,f}(x) > t\}\,. \end{aligned}$$

We apply it to f being the value function, P a local transition $p(\cdot|s,a)$ and Q a plausible candidate $\tilde{p}(\cdot|s,a)$.

Proof :

Let us recall the fundamental equality

$$\forall \lambda \in \mathbb{R}, \log \mathbb{E}_P \exp(\lambda(X - \mathbb{E}_P[X]) = \sup_{Q \ll P} \left[\lambda\Big(\mathbb{E}_Q[X] - \mathbb{E}_P[X]\Big) - \mathtt{KL}(Q,P) \right].$$

In particular, we obtain on the one hand that

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \leqslant \min_{\lambda \in \mathbb{R}^+} \frac{\varphi_f(\lambda) + \mathtt{KL}(Q,P)}{\lambda}\,.$$

Since $\varphi_f(0) = 0$, then the right hand side quantity is non-negative. Let us call it u. Then, we note that for any t such that $u \geqslant t \geqslant 0$, then by construction of u, it holds $\mathtt{KL}(Q,P) \geqslant \varphi_{\star,f}(t)$. Thus, $\{t \geqslant 0 : \varphi_{f,\star}(t) \geqslant \mathtt{KL}(Q,P)\} = (u, \infty)$ and thus $u = \varphi_{+,f}^{-1}(\mathtt{KL}(Q,P))$.

On the other hand, it holds

$$\forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] \geqslant \max_{\lambda \in \mathbb{R}^-} \frac{\varphi_f(\lambda) + \mathtt{KL}(Q,P)}{\lambda}\,.$$

Since $\varphi(0) = 0$, then the right hand side quantity is non-positive. Let us call it v. Then, we note that for any t such that $v \leqslant t \leqslant 0$, then by construction of v, it holds $\mathtt{KL}(Q,P) \geqslant \varphi_{\star,f}(t)$. Thus, $\{t \leqslant 0 : \varphi_{\star,f}(t) \geqslant \mathtt{KL}(Q,P)\} = (-\infty, v)$ and thus $v = \varphi_{-,f}^{-1}(\mathtt{KL}(Q,P))$.

□

Corollary 6.4 (*Transportation and KL*) *Assume that f is such that $\mathbb{V}_P[f]$ and $\mathbb{S}(f) = \max_x f(x) - \min_x f(x)$ are finite. Then it holds*

$$\begin{aligned} \forall Q \ll P, \quad \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f] KL(Q,P)} + \frac{2\mathbb{S}(f)}{3} KL(Q,P)\,, \\ \forall Q \ll P, \quad \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f] KL(Q,P)}\,. \end{aligned}$$

In particular, this shows that it is enough to control the Kullback-Leibler divergence between a distribution and its empirical counter part in order to derive immediately a concentration result for the empirical mean of virtually any function (with finite variance and span).

Proof :

Indeed, by a standard Bernstein argument, it holds

$$\begin{aligned} \forall \lambda \in [0, \frac{3}{\mathbb{S}(f)}), \quad \varphi_f(\lambda) &\leqslant \frac{\mathbb{V}_P[f]}{2} \frac{\lambda^2}{1 - \frac{\mathbb{S}(f)\lambda}{3}}, \\ \forall x \geqslant 0, \quad \varphi_{\star,f}(x) &\geqslant \frac{x^2}{2(\mathbb{V}_P[f] + \frac{\mathbb{S}(f)}{3}x)}. \end{aligned}$$

Then, a direct computation shows that

$$\varphi_{+,f}^{-1}(t) \leqslant \frac{\mathbb{S}(f)}{3}t + \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2},$$

$$\varphi_{-,f}^{-1}(t) \geqslant \frac{\mathbb{S}(f)}{3}t - \sqrt{2t\mathbb{V}_P[f] + \Big(\frac{\mathbb{S}(f)}{3}t\Big)^2}.$$

Combining these two bounds, we obtain that

$$\begin{aligned} \mathbb{E}_Q[f] - \mathbb{E}_P[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P) + \Big(\frac{\mathbb{S}(f)}{3}\Big)^2 \mathtt{KL}(Q,P)^2} + \frac{\mathbb{S}(f)}{3}\mathtt{KL}(Q,P), \\ \mathbb{E}_P[f] - \mathbb{E}_Q[f] &\leqslant \sqrt{2\mathbb{V}_P[f]\mathtt{KL}(Q,P) + \Big(\frac{\mathbb{S}(f)}{3}\Big)^2 \mathtt{KL}(Q,P)^2} - \frac{\mathbb{S}(f)}{3}\mathtt{KL}(Q,P). \end{aligned}$$

We conclude by using that $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for non-negative a, b. □

2.3 Structured MDPs

We close this chapter with an advocacy for studying a specific notion of structure in MDPs, that is especially efficient at reducing the cost of learning as it appears in most typical MDPs. More precisely, structure is helpful when observations gathered on a transition can be transfered to another one. We introduce the following notion that enables to capture such a phenomenon in great generality.

Definition 6.21 (*Similar state-action pairs*) *The pair (s', a') is ε-similar to the pair (s, a), for $\varepsilon = (\varepsilon_p, \varepsilon_\mu) \in \mathbb{R}_+^2$, if*

$$\begin{aligned} \|p(\sigma_{s,a}(\cdot)|s,a) - p(\sigma_{s',a'}(\cdot)|s',a')\|_1 &\leqslant \varepsilon_p, \quad \textit{(similar profile)} \\ \textit{and} \qquad |\mu(s,a) - \mu(s',a')| &\leqslant \varepsilon_\mu, \quad \textit{(similar rewards)} \end{aligned}$$

where $\sigma_{s,a} : \{1, \dots, S\} \to \mathcal{S}$ indexes a permutation of states such that $p(\sigma_{s,a}(1)|s,a) \geqslant p(\sigma_{s,a}(2)|s,a) \geqslant \dots \geqslant p(\sigma_{s,a}(S)|s,a)$. We call it a profile *mapping (it may not be unique).*

Remark 6.2 *$(0,0)$-similarity is an equivalence relation over $\mathcal{S} \times \mathcal{A}$. It thus induces a canonical partition of $\mathcal{S} \times \mathcal{A}$, which we denote by $\mathcal{C}$.*

Grid-world	Figure 6.1	Figure 6.2	Figure 6.3	Figure 6.4
SA	84	800	736	$\sim 10^4$
$\lvert\mathcal{C}\rvert$	6	6	7	7

We now show that in typical grid-world MDPs, the number of classes of state-action pairs using Definition 6.21 stays small even for large SA. We consider to this end a grid-world MDP with four actions $a \in \{u, d, l, r\}$. Playing action $a = u$ moves the current state up with probability 0.8, does not change the current state with probability 0.1, and moves left or right with same probability 0.05 (it never goes down). When the resulting state is a wall, the distribution is modified: the probability mass is reported on the current state. Other actions have similar effects. Finally, the goal-state with reward 1 is put in the bottom-right corner of the MDP. The table summarizes the size of the state-action space as well as the number of classes.

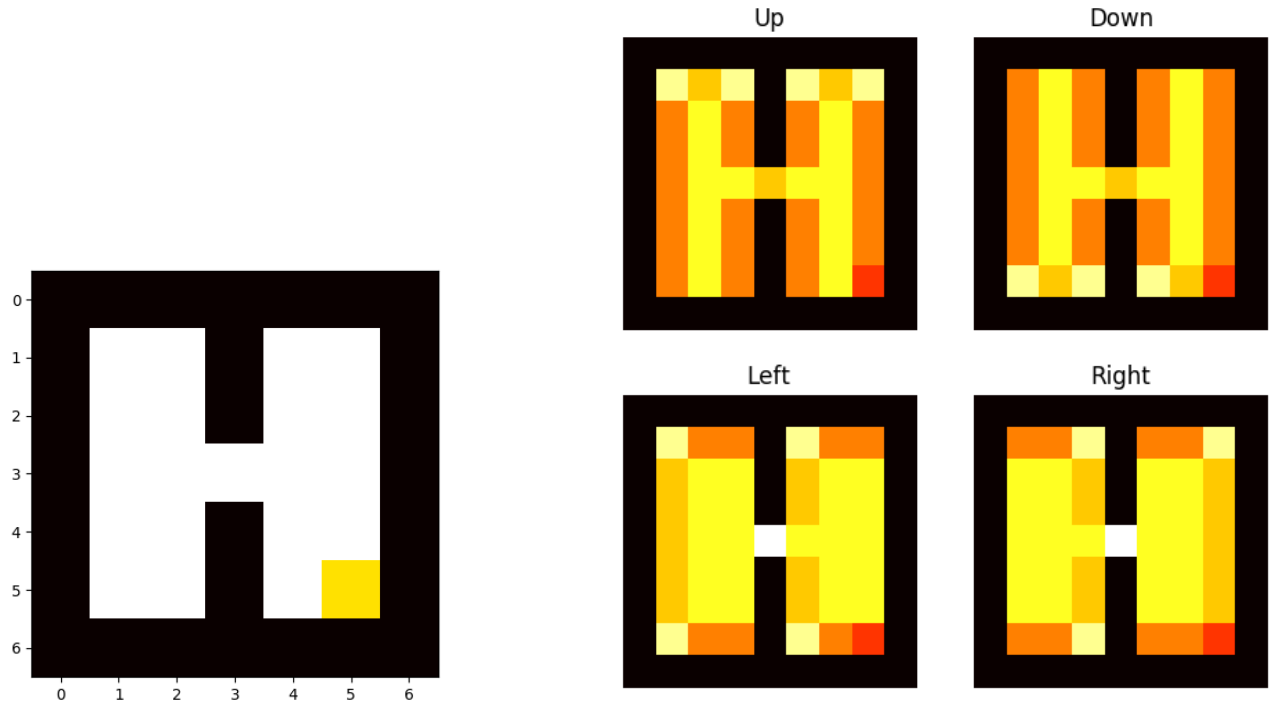


Figure 6.1: Left: Two-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

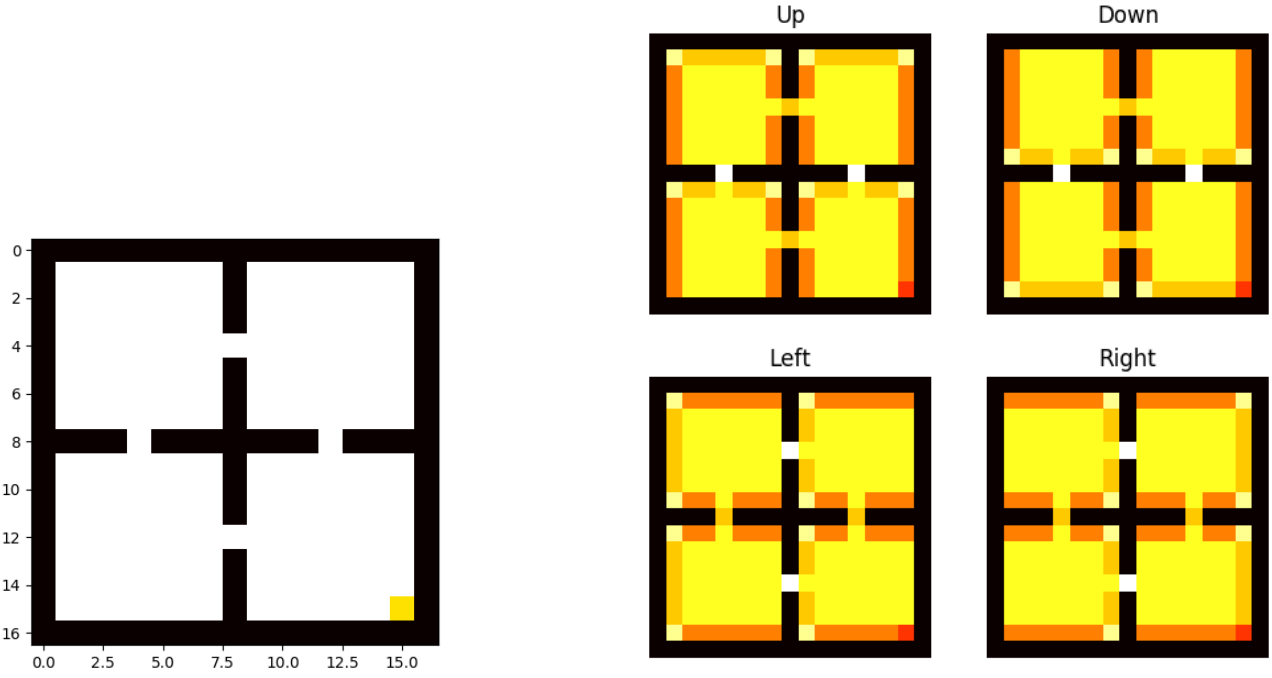


Figure 6.2: Left: Four-room grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

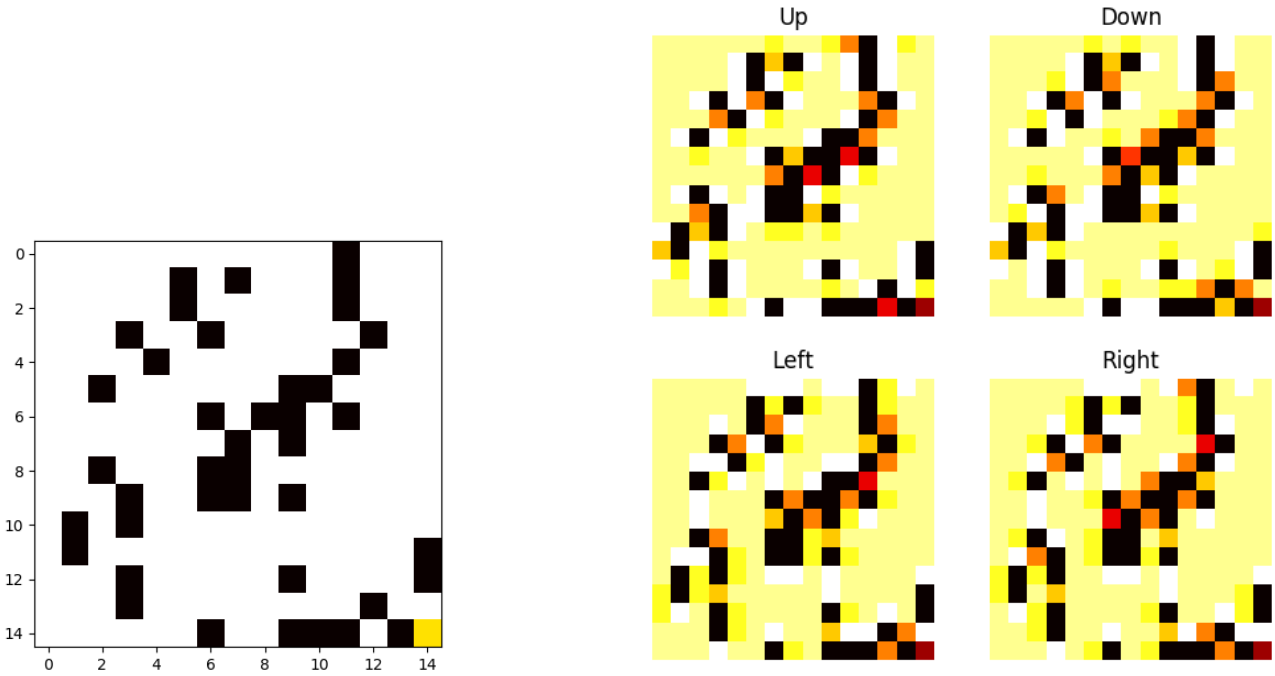


Figure 6.3: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

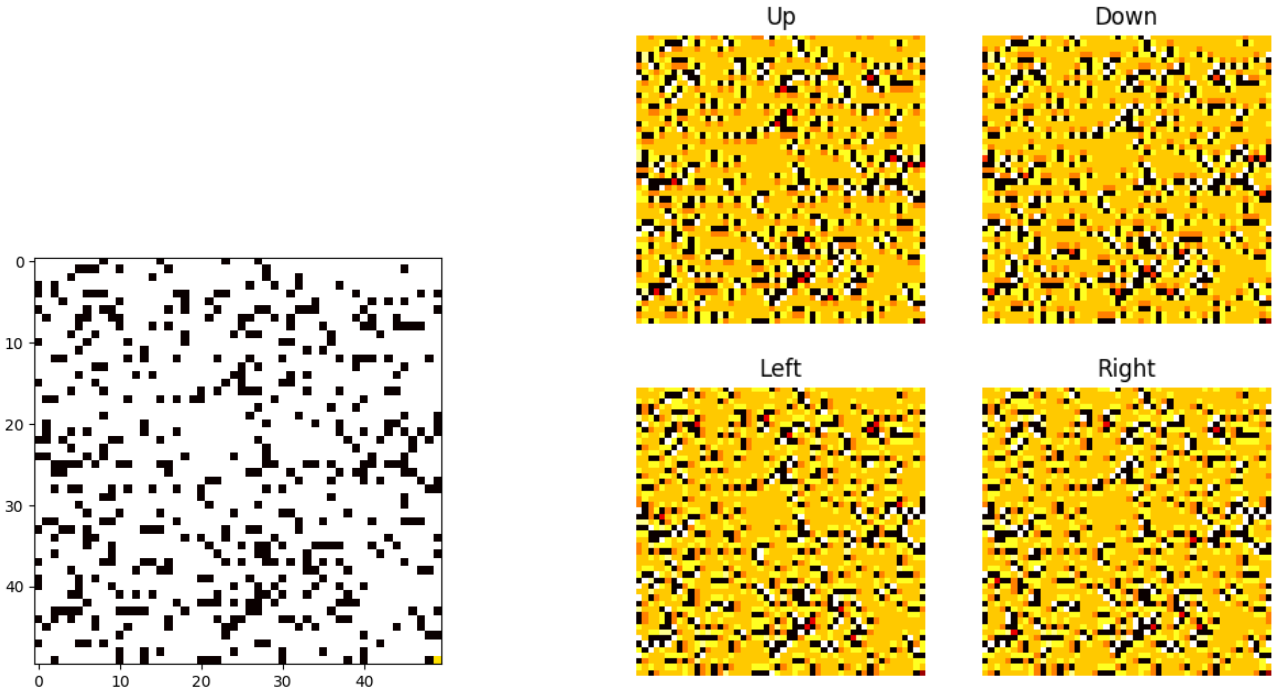


Figure 6.4: Left: A more complex grid-world (left) with walls in black, and goal state in yellow. Right: equivalence classes for state-action pairs (one color per class).

Other notions Other notions of similarity have been introduced in the RL literature. However, they do not scale well compared to the concept we consider here (that is, they produce a number of classes that is large when SA increases). For instance, in Ortner (2013), a partition $\mathcal{S}_1, \dots \mathcal{S}_n$ of the state space $\mathcal{S}$ is considered to define an aggregated MDP, in case it satisfies

$$\forall s, s' \in \mathcal{S}_i, \forall a \in \mathcal{A}, \mu(s,a) = \mu(s',a) \text{ and } \forall j, \sum_{s'' \in \mathcal{S}_j} p(s''|s,a) = \sum_{s'' \in \mathcal{S}_j} p(s''|s',a)\,.$$

This readily prevents any two states s, s' such that $p(\cdot|s,a)$ and $p(\cdot|s',a)$ have disjoint support from being in the same set $\mathcal{S}_i$. Thus, since in grid-world MDP where transitions are local, the number of pairs with disjoint support is (about linearly) increasing with S, this implies a potentially large number of classes for grid-worlds with many states. A similar criticism can be formulated for Anand et al. (2015), even though it considers sets of state-action instead of states only, thus slightly reducing the total number of classes.

Usage Interestingly, the notion of similarity can be used directly to benefit a reinforcement learning strategy such as `UCRL2`, by combining observations from similar state-actions pairs in order to produce more accurate estimates. Indeed the typical scaling of the regret of `UCRL2` with S, A, T and K (K bounding the size of the support of transitions) is $O(\sqrt{SAKT})$, up to logarithmic factors. This can be reduced to $O(\sqrt{CKT})$ when the structure is known (up to logarithmic factors). We have investigated such an approach in a recent work that shows the regret can be greatly reduced when considering there is such a structure, and that, perhaps more surprisingly, it is possible to design a strategy that approximately maintains such performances even without the knowledge of the structure beforehand.

Part II

Focus on three contributions

CHAPTER 7

Boundary Crossing Probabilities

Contents

1 Multi-armed bandit setup and notations . . . 117

2 Boundary crossing probabilities for the generic `KL-ucb` strategy. . . . 118

3 Boundary crossing for K-dimensional exponential families . . . 122

3.1 Previous work on boundary-crossing probabilities . . . 122

4 Main analysis . . . 126

4.1 Peeling and cone covering . . . 126

4.2 Change of measure . . . 127

4.3 Localized change of measure . . . 128

4.4 Concentration of measure . . . 128

4.5 Combining the different steps . . . 129

This chapter corresponds to the article Maillard (2018).

1 MULTI-ARMED BANDIT SETUP AND NOTATIONS

Let us consider a stochastic multi-armed bandit problem $(\mathcal{A}, \nu)$, where $\mathcal{A}$ is a finite set of cardinality $A \in \mathbb{N}$ and $\nu = (\nu_a)_{a\in\mathcal{A}}$ is a set of probability distribution over $\mathbb{R}$ indexed by $\mathcal{A}$. The game is sequential and goes as follows:

At each round $t \in \mathbb{N}$, the player picks an arm a_t (based on her past observations) and receives a stochastic payoff Y_t drawn independently at random according to the distribution ν_{a_t}. She only observes the payoff Y_t, and her goal is to maximize her expected cumulated payoff, $\sum_{t=1} Y_{a_t}$, over a possibly unknown number of steps.

Although the term multi-armed bandit problem was probably coined during the 60's in reference to the casino slot machines of the 19th century, the formulation of this problem is due to Herbert Robbins – one of the most brilliant mind of his time, see Robbins (1952) and takes its origin in earlier questions about optimal stopping policies for clinical trials, see Thompson (1933, 1935), Wald (1945). We refer the interested reader to Robbins (2012) regarding the legacy of the immense work of H. Robbins in mathematical statistics for the sequential design of experiments, compiling his most outstanding research for his 70's birthday. Since then, the field of multi-armed bandits has grown large and bold, and we humbly refer to the introduction of Cappé et al. (2013) for key historical aspects about the development of the field. Most notably, they include first the introduction of dynamic allocation indices (a.k.a. Gittins indices, Gittins (1979)) suggesting that an

optimal strategy can be found in the form of an index strategy (that at each round selects an arm with highest "index"); second, the seminal work of Lai and Robbins (1985a) that shows indexes can be chosen as "upper confidence bounds" on the mean reward of each arm, and provided the first asymptotic lower-bound on the achievable performance for specific distributions; third, the generalization of this lower bound in the 90's to generic distributions by Burnetas and Katehakis (1997) (see also the recent work from Garivier et al. (2016)) as well as the asymptotic analysis by Agrawal (1995) of generic classes of upper-confidence-bound based index policies and finally Auer et al. (2002) that popularized a simple sub-optimal index strategy termed `UCB` and most importantly opened the quest for finite-time, as opposed to asymptotic, performance guarantees. For the purpose of this chapter, we now remind the formal definitions and notations for the stochastic multi-armed bandit problem, following Cappé et al. (2013).

Quality of a strategy For each arm $a \in \mathcal{A}$, let μ_a be the expectation of the distribution ν_a, and let $a^\star$ be any optimal arm in the sense that

$$a^\star \in \operatorname*{Argmax}_{a \in \mathcal{A}} \mu_a \,.$$

We write $\mu^\star$ as a short-hand notation for the largest expectation $\mu_{a^\star}$ and denote the *gap* of the expected payoff μ_a of an arm a to $\mu^\star$ as $\Delta_a = \mu^\star - \mu_a$. In addition, we denote the number of times each arm a is pulled between the rounds 1 and T by $N_a(T)$,

$$N_a(T) \overset{\text{def}}{=} \sum_{t=1}^{T} \mathbb{I}_{\{a_t = a\}} \,.$$

Definition 7.3 (*Expected regret*) *The quality of a strategy is evaluated using the notion of expected regret (or simply, regret) at round $T \geqslant 1$, defined as*

$$\mathfrak{R}_T \overset{\text{def}}{=} \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} Y_t\right] = \mathbb{E}\left[T\mu^\star - \sum_{t=1}^{T} \mu_{a_t}\right] = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}\big[N_a(T)\big] \,, \tag{7.3}$$

where we used the tower rule for the first equality. The expectation is with respect to the random draws of the Y_t according to the ν_{a_t} and to the possible auxiliary randomization introduced by the decision-making strategy.

2 Boundary crossing probabilities for the generic KL-UCB strategy.

The first appearance of the `KL-ucb` strategy can be traced at least to Lai (1987) although it was not given an explicit name at that time. It seems the strategy was forgotten after the work of Auer et al. (2002) that opened a decade of intensive research on finite-time analysis of bandit strategies and extensions to variants of the problem (Audibert et al. (2009), Audibert and Bubeck (2010), see also Bubeck et al. (2012) for a survey of relevant variants of bandit problems), until the work of Honda and Takemura (2010) shed a novel light on the asymptotically optimal strategies. Thanks to their illuminating work, the first finite-time regret analysis of `KL-ucb` was obtained by Maillard et al. (2011) for discrete distributions, soon extended to handle exponential families of dimension 1 as well, in the unifying work of Cappé et al. (2013). However, as we will see in this paper, we should all be much in debt of the outstanding work of T.L. Lai. regarding the analysis of this index strategy, both asymptotically and in finite-time, as a second look at his papers shows how to bypass the

limitations of the state-of-the-art regret bounds for the control of *boundary crossing probabilities* in this context (see Theorem 3, Maillard (2018)). Actually, the first focus of the present chapter is not on stochastic bandits but boundary crossing probabilities, and the bandit setting that we provide here should be considered only as giving a solid motivation for the contribution of this paper.

Let us now introduce formally the `KL-ucb` strategy. We assume that the learner is given a family $\mathcal{D} \subset \mathcal{P}(\mathbb{R})$ of probability distributions that satisfies $\nu_a \in \mathcal{D}$ for each arm $a \in \mathcal{A}$, where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions over the set $\mathcal{X}$. For two distributions $\nu, \nu' \in \mathcal{P}(\mathbb{R})$, we denote by $\mathtt{KL}(\nu, \nu')$ their Kullback-Leibler divergence and by $E(\nu)$ and $E(\nu')$ their respective expectations (this operator is denoted by E while expectations of a function f with respect to underlying randomizations are referred to as $\mathbb{E}[f]$, or $\mathbb{E}_{X\sim\nu}[f(X)]$ to make explicit the law of the random variable X).

The generic form of the algorithm of interest in this paper is described as Algorithm 4. It relies on two parameters: an operator $\Pi_{\mathcal{D}}$ (in spirit, a projection operator) that associates with each empirical distribution $\widehat{\nu}_a(t)$ an element of the model $\mathcal{D}$; and a non-decreasing function f, which is typically such that $f(t) \approx \log(t)$.

At each round $t \geqslant K+1$, a upper confidence bound $U_a(t)$ is associated with the expectation μ_a of the distribution ν_a of each arm; an arm a_{t+1} with highest upper confidence bound is then played.

Algorithm 4 The `KL-ucb` algorithm (generic form).

Parameters: An operator $\Pi_{\mathcal{D}} : \mathcal{P}(\mathbb{R}) \to \mathcal{D}$; a non-decreasing function $f : \mathbb{N} \to \mathbb{R}$

Initialization: Pull each arm of $\{1, \dots, K\}$ once

for each round $t+1$, where $t \geqslant K$, **do**

compute for each arm a the quantity

$$U_a(t) = \sup\left\{ E(\nu) : \quad \nu \in \mathcal{D} \quad \text{and} \quad \mathtt{KL}\Big(\Pi_{\mathcal{D}}\big(\widehat{\nu}_a(t)\big),\, \nu\Big) \leqslant \frac{f(t)}{N_a(t)} \right\};$$

pick an arm $\quad a_{t+1} \in \underset{a\in\mathcal{A}}{\arg\max}\, U_a(t).$

In the literature, another variant of `KL-ucb` is introduced where the term $f(t)$ is replaced with $f(t/N_a(t))$. We refer to this algorithm as `KL-ucb+`. While `KL-ucb` has been analyzed and shown to be provably near-optimal, the variant `KL-ucb+` has not been analyzed yet.

Alternative formulation of `KL-ucb` We wrote the `KL-ucb` algorithm so that the optimization problem resulting from the computation of $U_a(t)$ is easy to handle. Now, under some assumption, one can rewrite this term, in an equivalent form more suited for the analysis. We refer to Cappé et al. (2013):

Assumption 7.1 *There is a known interval $\mathcal{I} \subset \mathbb{R}$ with boundary $\mu^- \leqslant \mu^+$, for which each model $\mathcal{D} = \mathcal{D}_a$ of probability measures is included in $\mathcal{P}(\mathcal{I})$ and such that $\forall \nu \in \mathcal{D}_a, \forall \mu \in \mathcal{I} \setminus \{\mu^+\}$,*

$$\inf\Big\{ KL(\nu,\nu') :\ \ \nu' \in \mathcal{D}_a \ \text{ s.t. } \ E(\nu') > \mu \Big\} = \min\Big\{ KL(\nu,\nu') :\ \ \nu' \in \mathcal{D}_a \ \text{ s.t. } \ E(\nu') \geqslant \mu \Big\}.$$

Lemma 7.1 (*Rewriting*) *Under Assumption 7.1, the upper bound used by the* `KL-ucb` *algorithm satisfies the following equality*

$$U_a(t) \;=\; \max\left\{\mu\in\mathcal{I}\setminus\{\mu^+\}:\; \mathcal{K}_a\Big(\Pi_a\left(\widehat{\nu}_a(t)\right),\mu\Big)\leqslant\frac{f(t)}{N_a(t)}\right\}$$
$$\text{where}\quad \mathcal{K}_a(\nu_a,\mu^\star)\overset{\text{def}}{=}\inf_{\nu\in\mathcal{D}_a:\,E(\nu)>\mu^\star} KL(\nu_a,\nu)\quad\text{and}\quad \Pi_a\overset{\text{def}}{=}\Pi_{\mathcal{D}_a}.$$

Likewise, a similar result holds for `KL-ucb+` *but where $f(t)$ is replaced with $f(t/N_a(t))$.*

Remark 7.1 *For instance, this assumption is valid when $\mathcal{D}_a=\mathcal{P}([0,1])$ and $\mathcal{I}=[0,1]$. Indeed we can replace the strict inequality with an inequality provided that $\mu<1$ by Honda and Takemura (2010), and the infimum is reached by lower semi-continuity of the* `KL` *divergence and convexity and closure of the set $\{\nu'\in\mathcal{P}([0,1])\ \text{s.t.}\ E(\nu')\geqslant\mu\}$.*

Using boundary-crossing probabilities for regret analysis We continue this warming-up by restating a convenient way to decompose the regret and make appear the *boundary crossing probabilities* that are at the heart of this chapter. The following lemma is a direct adaptation from Cappé et al. (2013):

Lemma 7.2 (*From Regret to Boundary Crossing Probabilities*) *Let $\varepsilon\in\mathbb{R}^+$ be a small constant such that $\varepsilon\in(0,\min\{\,\mu^\star-\mu_a\,,\,a\in\mathcal{A}\,\})$. For $\mu,\gamma\in\mathbb{R}$, let us introduce the following set*

$$\mathcal{C}_{\mu,\gamma}\;=\;\Big\{\nu'\in\mathfrak{M}_1(\mathbb{R}):\;\mathcal{K}_a(\Pi_a(\nu'),\mu)<\gamma\Big\}.$$

Then, the number of pulls of a sub-optimal arm $a\in\mathcal{A}$ by Algorithm `KL-ucb` *satisfies*

$$\mathbb{E}\big[N_T(a)\big]\leqslant 2+\inf_{n_0\leqslant T}\left\{n_0+\sum_{n\geqslant n_0+1}^{T}\mathbb{P}\Big\{\widehat{\nu}_{a,n}\in\mathcal{C}_{\mu^\star-\varepsilon,f(T)/n}\Big\}\right\}$$
$$+\sum_{t=|\mathcal{A}|}^{T-1}\underbrace{\mathbb{P}\Big\{N_{a^\star}(t)\;\mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\,\mu^\star-\varepsilon\big)>f(t)\Big\}}_{\textit{Boundary Crossing Probability}}.$$

Likewise, the number of pulls of a sub-optimal arm $a\in\mathcal{A}$ by Algorithm `KL-ucb+` *satisfies*

$$\mathbb{E}\big[N_T(a)\big]\leqslant 2+\inf_{n_0\leqslant T}\left\{n_0+\sum_{n\geqslant n_0+1}^{T}\mathbb{P}\Big\{\widehat{\nu}_{a,n}\in\mathcal{C}_{\mu^\star-\varepsilon,f(T/n)/n}\Big\}\right\}$$
$$+\sum_{t=|\mathcal{A}|}^{T-1}\underbrace{\mathbb{P}\Big\{N_{a^\star}(t)\;\mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\,\mu^\star-\varepsilon\big)>f(t/N_{a^\star}(t))\Big\}}_{\textit{Boundary Crossing Probability}}.$$

Proof of Lemma 7.2:

The first part of this lemma for `KL-ucb` is proved in Cappé et al. (2013). The second part that is about `KL-ucb+`can be proved straightforwardly following the very same lines. We thus only provide the main steps here for clarity: We start by introducing a small $\varepsilon > 0$ that satisfies $\varepsilon < \min\{\, \mu^\star - \mu_a \,,\, a \in \mathcal{A} \,\}$, and then consider the following inclusion of events:

$$\Big\{a_{t+1} = a\Big\} \subseteq \Big\{\mu^\star - \varepsilon < U_a(t) \text{ and } a_{t+1} = a\Big\} \cup \Big\{\mu^\star - \varepsilon \geqslant U_{a^\star}(t)\Big\} ;$$

indeed, on the event $\Big\{a_{t+1} = a\Big\} \cap \Big\{\mu^\star - \varepsilon < U_{a^\star}(t)\Big\}$, we have, $\mu^\star - \varepsilon < U_{a^\star}(t) \leqslant U_a(t)$ (where the last inequality is by definition of the strategy). Moreover, let us note that

$$\Big\{\mu^\star - \varepsilon < U_a(t)\Big\} \subseteq \Big\{\exists \nu' \in \mathcal{D} : E(\nu') > \mu^\star - \varepsilon \text{ and } N_a(t)\ \mathcal{K}_a\big(\Pi_a(\widehat{\nu}_{a,N_a(t)}),\, \mu^\star - \varepsilon\big) \leqslant f(t/N_a(t))\Big\},$$
$$\text{and} \quad \Big\{\mu^\star - \varepsilon \geqslant U_{a^\star}(t)\Big\} \subseteq \Big\{\exists \nu' \in \mathcal{D} : N_{a^\star}(t)\ \mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\, \mu^\star - \varepsilon\big) > f(t/N_{a^\star}(t))\Big\},$$

since $\mathcal{K}_a$ is a non-decreasing function in its second argument and $\mathcal{K}_a\big(\nu, E(\nu)\big) = 0$ for all distributions ν. Therefore, this simple remark leads us to the following decomposition

$$\mathbb{E}\big[N_T(a)\big] \leqslant 1 + \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{ N_{a^\star}(t)\ \mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\, \mu^\star - \varepsilon\big) > f(t/N_{a^\star}(t))\Big\}$$
$$+ \sum_{t=|\mathcal{A}|}^{T-1} \mathbb{P}\Big\{ N_a(t)\ \mathcal{K}_a\big(\Pi_a(\widehat{\nu}_{a,N_a(t)}),\, \mu^\star - \varepsilon\big) \leqslant f(t/N_a(t)) \ \text{ and } \ A_{t+1} = a\Big\}.$$

The remaining steps of the proof of the result from Cappé et al. (2013), equation (10) can now be straightforwardly modified to work with $f(t/N_a(t))$ instead of $f(t)$, thus concluding this proof. □

Lemma 7.2 shows that two terms need to be controlled in order to derive regret bounds for the considered strategy. The *boundary crossing probability* term is arguably the most difficult to handle and is the focus of the next sections. The other term involves the probability that an empirical distribution belongs to a convex set, which can be handled either directly as in Cappé et al. (2013) or by resorting to finite-time Sanov-type results such as that of (Dinwoodie, 1992, Theorem 2.1 and comments on page 372), or its variant from (Maillard et al., 2011, Lemma 1). For completeness, the exact result from Dinwoodie (1992) writes

Lemma 7.3 (*Non-asymptotic Sanov's lemma*) *Let $\mathcal{C}$ be an open convex subset of $\mathcal{P}(\mathcal{X})$ such that $\Lambda_\nu(\mathcal{C}) = \inf_{\kappa\in\mathcal{C}} KL(\kappa,\nu)$ is finite. Then, for all $t \geqslant 1$, $\mathbb{P}_\nu\{\widehat{\nu}_t \in \mathcal{C}\} \leqslant \exp\big(-t\Lambda_\nu(\overline{\mathcal{C}})\big)$ where $\overline{\mathcal{C}}$ is the closure of $\mathcal{C}$.*

Scope and focus of this work We focus on the setting of stochastic multi-armed bandits because this gives a strong and natural motivation for studying boundary crossing probabilities. However, one should understand that the primary goal of this work is to give credit to the work of T.L. Lai regarding the neat understanding of boundary crossing probabilities and not necessarily to provide a regret bound for such bandit algorithms as `KL-ucb` or `KL-ucb+`. Also, we believe that results on boundary crossing probabilities are useful beyond the bandit problem in hypothesis testing. Thus, and in order to avoid obscuring the main result regarding boundary crossing probabilities, we choose not to provide regret bounds here and to leave them as an exercise for the interested reader; controlling the remaining term appearing in the decomposition of Lemma 7.2 is indeed mostly technical and does not seem to require especially illuminating or fancy idea. We refer to Cappé et al. (2013) for an example of bound in the case of exponential families of dimension 1.

High-level overview of the contribution We are now ready to explain the main results of this paper. For the purpose of clarity, we provide them as an informal statement before proceeding with the technical material.

Our contribution is about the behavior the of the *boundary crossing probability* term for exponential families of dimension K when choosing the threshold function $f(x) = \log(x) + \xi \log\log(x)$. Our result reads as follows.
Theorem (Informal statement) *Assuming that the observations are generated from a distribution that belongs to an exponential family of dimension K that satisfies some mild conditions, then for any non-negative ε and some class-dependent but fully explicit constants c, C (also depending on ε) it holds*

$$\mathbb{P}\Big\{ N_{a^\star}(t)\ \mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\ \mu^\star - \varepsilon\big) > f(t) \Big\} \ \leqslant\ \frac{C}{t}\log(t)^{K/2-\xi} e^{-c\sqrt{f(t)}}$$
$$\mathbb{P}\Big\{ N_{a^\star}(t)\ \mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,N_{a^\star}(t)}),\ \mu^\star - \varepsilon\big) > f(t/N_{a^\star}(t)) \Big\} \ \leqslant\ \frac{C}{t}\log(tc)^{K/2-\xi-1},$$

where the first inequality holds for all t and the second one for large enough $t \geqslant t_c$ where t_c is class dependent but explicit and "reasonably" small.

The rigorous statement is provided in Theorem 3 and Corollaries 1,2 in Maillard (2018). The main interest of this result is that it shows how to tune ξ with respect to the dimension K of the family. Indeed, in order to ensure that the probability term is summable in t, the bound suggests that ξ should be at least larger than $K/2 - 1$. The case of exponential families of dimension 1 ($K = 1$) is especially interesting, as it supports the fact that both `KL-ucb` and `KL-ucb+` can be tuned using $\xi = 0$ (and even negative ξ for `KL-ucb`). This was observed in numerical experiments in Cappé et al. (2013) although not theoretically supported until now.

3 Boundary crossing for K-dimensional exponential families

In this section, we now study the boundary crossing probability term appearing in Lemma 7.2 for a K-dimensional exponential family $\mathcal{E}(F;\nu_0)$. We first provide an overview of the existing results before detailing our main contribution. As explained in the introduction, the key technical tools that enable to obtain the novel results were already known three decades ago, and thus even though the novel result is impressive due to its generality and tightness, it should be regarded as a modernized version of an existing, but almost forgotten result, that enables to solve a few long-lasting open questions as a by-product.

3.1 Previous work on boundary-crossing probabilities

The existing results used in the bandit literature about boundary-crossing probabilities are restricted to a few specific cases. For instance in Cappé et al. (2013), the authors provide the following control

Theorem 7.1 (`KL-ucb`) *In the case of canonical (that is $F(x) = x$) exponential families of dimension $K = 1$, for a function f such that $f(x) = \log(x) + \xi \log\log(x)$, then it holds for all $t > A$*

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=1}^{t-A+1} n\ \mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}),\, \mu^\star\big) > f\big(t\big) \cap \mu_{a^\star} > \widehat{\mu}_{a^\star,n} \Big\} \leqslant e\lceil f(t)\log(t)\rceil e^{-f(t)}\,.$$

Further, in the special case of distributions with finitely many K atoms, it holds for all $t > A, \varepsilon > 0$

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=1}^{t-A+1} n\mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}),\, \mu^\star - \varepsilon\big) > f\big(t\big) \Big\} \leqslant e^{-f(t)}\Big(3e + 2 + 4\varepsilon^{-2} + 8e\varepsilon^{-4}\Big)\,.$$

In contrast in Lai (1988), the authors provide an asymptotic control in the more general case of exponential families of dimension K with some basic regularity condition, as we explained earlier. We now restate this beautiful result from Lai (1988) in a way that is suitable for a more direct comparison with other results. The following holds:

Theorem 7.2 (*Lai, 88*) *Let us consider an exponential family of dimension K. Define for $\gamma > 0$ the cone $\mathcal{C}_\gamma(\theta) = \{\theta' \in \mathbb{R}^K : \langle \theta', \theta\rangle \geqslant \gamma|\theta||\theta'|\}$. Then, for a function f such that $f(x) = \alpha\log(x) + \xi\log\log(x)$ it holds for all $\theta^\dagger \in \Theta$ such that $|\theta^\dagger - \theta^\star|^2 \geqslant \delta_t$, where $\delta_t \to 0$, $t\delta_t \to \infty$ as $t \to \infty$,*

$$\begin{aligned}\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=1}^{t} \widehat{\theta}_n \in \Theta_\rho \cap n\mathcal{B}^\psi(\widehat{\theta}_n, \theta^\dagger) \geqslant f\Big(\frac{t}{n}\Big) \cap \nabla\psi(\widehat{\theta}_n) - \nabla\psi(\theta^\dagger) \in \mathcal{C}_\gamma(\theta^\dagger - \theta^\star)\Big\} \\ \overset{t\to\infty}{=} O\Big(t^{-\alpha}|\theta^\dagger - \theta^\star|^{-2\alpha}\log^{-\xi-\alpha+K/2}(t|\theta^\dagger - \theta^\star|^2)\Big) \\ = O\Big(e^{-f(t|\theta^\dagger-\theta^\star|^2)}\log^{-\alpha+K/2}(t|\theta^\dagger - \theta^\star|^2)\Big)\,.\end{aligned}$$

Discussion The quantity $\mathcal{B}^\psi(\widehat{\theta}_n, \theta^\dagger)$ is the direct analog of $\mathcal{K}_{a^\star}\big(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}), \mu^\star - \varepsilon\big)$ in Theorem 7.1. Note however that $f(t/n)$ replaces the larger quantity $f(t)$, which means that Theorem 7.2 controls a larger quantity than Theorem 7.1, and is thus in this sense stronger. It also holds for general exponential families of dimension K. Another important difference is the order of magnitude of the right hand side terms of both theorems. Indeed, since $e\lceil f(t)\log(t)\rceil e^{-f(t)} = O(\frac{\log^{2-\xi}(t)+\xi\log(t)^{1-\xi}\log\log(t)}{t})$, Theorem 7.1 requires that $\xi > 2$ in order that this term is $o(1/t)$, and $\xi > 0$ for the second term of Theorem 7.1. In contrast, Theorem 7.2 shows that it is enough to consider $f(x) = \log(x) + \xi\log\log(x)$ with $\xi > K/2 - 1$ to ensure a $o(1/t)$ bound. For $K = 1$, this means we can even use $\xi > -1/2$ and in particular $\xi = 0$, which corresponds to the value they recommend in the experiments.

Thus, Theorem 7.2 improves in three ways over Theorem 7.1: it is an extension to dimension K, it provides a bound for $f(t/n)$ (and thus for `KL-ucb+`) and not only $f(t)$, and finally allows for smaller values of ξ. These improvements are partly due to the fact Theorem 7.1 controls a concentration with respect to $\theta^\dagger$, not $\theta^\star$, which takes advantage of the fact there is some gap when going from $\mu^\star$ to distributions with mean $\mu^\star - \varepsilon$. The proof

of Theorem 7.2 directly takes advantage of this, contrary to that of the first part of Theorem 7.1.

On the other hand, Theorem 7.2 is only asymptotic whereas Theorem 7.1 holds for finite t. Furthermore, we notice two restrictions on the control event. First, it requires $\widehat{\theta}_n \in \Theta_\rho$, but we showed in the previous section that this is a minor restriction. Second, there is the restriction to a cone $\mathcal{C}_\gamma(\theta^\dagger - \theta^\star)$ which simplifies the analysis, but is a more dramatic restriction. This restriction cannot be removed trivially as it can be seen from the complete statement of (Lai, 1988, Theorem 2) that the right hand-side blows up to ∞ when $\gamma \to 0$. As we will see, it is possible to overcome this restriction by resorting to a smart covering of the space with cones, and sum the resulting terms via a union bound over the covering. We quickly explain the way of proceeding in the proof of Theorem 3 from Maillard (2018) in section 4.

Hint at proving the first part of Theorem 7.1 We believe it is interesting to give some hint about the proof of the first part of Theorem 7.1, as it involves an elegant step, despite relying quite heavily on two specific properties of the canonical exponential family of dimension 1. Indeed in the special case of the canonical one-dimensional family (that is $K = 1$ and $F_1(x) = x \in \mathbb{R}$), $\widehat{F}_n = \frac{1}{n}\sum_{i=1}^n X_i$ coincides with the empirical mean and it can be shown that $\Phi^\star(F)$ is strictly decreasing on $(-\infty, \mu^\star]$. Thus for any $F \leqslant \mu^\star$, it holds

$$\left\{\widehat{F}_n \leqslant \mu^\star \cap \Phi^\star(\widehat{F}_n) \geqslant \Phi^\star(F)\right\} \subset \left\{\widehat{F}_n \leqslant F\right\}. \tag{7.4}$$

Further, using the notations of Section 4, it also holds in that case $\mathcal{K}_{a^\star}\left(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}),\, \mu^\star\right) = \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) = \Phi^\star(\widehat{F}_n)$, where $\widehat{\theta}_n = \dot{\psi}^{-1}(\widehat{F}_n)$ is uniquely defined. A second non-trivial property that is shown in Cappé et al. (2013) is that for all $F \leqslant \mu^\star$, we can localize the supremum as

$$\Phi^\star(F) = \sup\left\{xF - \Phi(x) : x < 0 \text{ and } xF - \Phi(x) > 0\right\}. \tag{7.5}$$

Armed with these two properties, the proof reduces almost trivially to the following elegant lemma:

Lemma 7.4 (*Dimension 1*) *Consider a canonical one-dimensional family (that is $K = 1$ and $F_1(x) = x \in \mathbb{R}$). Then, for all f such that $f(t/n)/n$ is non-increasing in n,*

$$\mathbb{P}_{\theta^\star}\left\{\bigcup_{m\leqslant n<M} \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star\right\} \;\leqslant\; \exp\left(-\frac{m}{M} f(t/M)\right).$$

The proof of this lemma is directly adapted from the proof of Theorem 7.1, and makes use the of Bregman duality lemma 1.8. The first statement of Theorem 7.1 is obtained by a peeling argument, using $m/M = (f(t) - 1)/f(t)$. However this argument does not extend nicely to using $f(t/n)$, which explains why there is no statement regarding the threshold of `KL-ucb+`.

Proof of Lemma 7.4:

The proof goes as follows. First, we observe that:

$$\begin{aligned}\mathbb{P}_{\theta^\star}\left\{\bigcup_{m\leqslant n<M} \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star\right\} &= \mathbb{P}_{\theta^\star}\left\{\bigcup_{m\leqslant n<M} \Phi^\star(\widehat{F}_n) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star\right\} \\ &\leqslant \mathbb{P}_{\theta^\star}\left\{\bigcup_{m\leqslant n<M} \Phi^\star(\widehat{F}_n) \geqslant f(t/M)/M \cap \widehat{F}_n \leqslant \mu^\star\right\}.\end{aligned}$$

At this point note that if for all $F = \nabla\psi(\theta)$ with mean $\mu_\theta \leqslant \mu^\star$, it holds that $\Phi^\star(F) < f(t/M)/M$ then the probability of interest is 0 and we are done. In the other case, there exists an F_M such that $\Phi^\star(F_M) = f(t/M)/M$. We thus proceed with this case as follows

$$\begin{aligned}
&\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star \Big\} \\
&\leqslant \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \Phi^\star(\widehat{F}_n) \geqslant \Phi^\star(F_M) \cap \widehat{F}_n \leqslant \mu^\star \Big\} \\
&\overset{(a)}{\leqslant} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \widehat{F}_n \leqslant F_M \Big\} \\
&\overset{(b)}{\leqslant} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \exp\Big(\lambda \sum_{i=1}^n F(X_i)\Big) \geqslant \exp\Big(n\lambda F_M\Big) \Big\} \\
&\leqslant \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \exp\Big(\sum_{i=1}^n \Big(\lambda F(X_i) - \Phi(\lambda)\Big)\Big) \geqslant \exp\Big(n[\lambda F_M - \Phi(\lambda)]\Big) \Big\} \\
&\overset{(c)}{\leqslant} \mathbb{P}_{\theta^\star}\Big\{ \max_{m\leqslant n<M} \exp\Big(\sum_{i=1}^n \Big(\lambda F(X_i) - \Phi(\lambda)\Big)\Big) \geqslant \exp\Big(m[\lambda F_M - \Phi(\lambda)]\Big) \Big\},
\end{aligned}$$

where (a) holds by (7.4), (b) holds for all $\lambda < 0$, and (c) for all $\lambda < 0$ such that $\lambda F_M - \Phi(\lambda) > 0$. Now, the process defined by $W_{\lambda,0} = 1$ and $W_{\lambda,n} = \exp\Big(\sum_{i=1}^n \Big(\lambda F(X_i) - \Phi(\lambda)\Big)\Big)$ is a non-negative super-martingale, since it holds

$$\begin{aligned}
\mathbb{E}_{\theta^\star}\Big[\exp\Big(\sum_{i=1}^n \Big(\lambda F(X_i) - \Phi(\lambda)\Big)\Big) \Big| \mathcal{H}_{n-1} \Big] &= W_{\lambda,n-1}\mathbb{E}_{\theta^\star}\Big[\exp\Big(\lambda F(X_n) - \Phi(\lambda)\Big) \Big| \mathcal{H}_{n-1} \Big] \\
&\leqslant W_{\lambda,n-1} \exp\Big(\Phi(\lambda) - \Phi(\lambda)\Big) \leqslant 1\,.
\end{aligned}$$

Thus, we deduce that for all $\lambda < 0$ such that $\lambda F_M - \Phi(\lambda) > 0$

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star \Big\} \leqslant \exp\Big(- m[\lambda F_M - \Phi(\lambda)]\Big).$$

Since by (7.5) this is satisfied by the optimal λ for $\Phi^\star(F_M)$, we thus deduce that

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{m\leqslant n<M} \mathcal{B}^\psi(\widehat{\theta}_n, \theta^\star) \geqslant f(t/n)/n \cap \widehat{F}_n \leqslant \mu^\star \Big\} \leqslant \exp\Big(- m\Phi^\star(F_M)\Big) = \exp\Big(- \frac{m}{M} f(t/M)\Big).$$

□

This result has been extended to handle the boundary crossing probabilities simultaneously over all arms in Magureanu et al. (2014), using the simple idea of "stochastic orderings" from Hoeffding (1963).

4 MAIN ANALYSIS

We refer the interested reader to the journal article Maillard (2018) that explains these contributions in great details, and only provide below the main ingredients of the proof.

At a high level, we closely follow the proof technique used in Lai (1988) for the proof of Theorem 7.2, in order to prove the main result in (Maillard, 2018, Theorem 3). We precise further the constants, remove the cone restriction on the parameter and modify the original proof to be fully non-asymptotic which, using the technique of Lai (1988), forces us to make some parts of the proof a little more accurate.

Let us recall that we consider Θ and ρ such that $\theta^\star \in \Theta_\rho \subset \mathring{\Theta}_I$. The proof is divided in four main steps that we briefly present here for clarity:

In Section 4.1, we take care of the random number of pulls of the arm by a peeling argument. Simultaneously, we introduce a covering of the space with cones, which enables to later use arguments from proof of Theorem 7.2.

In Section 4.2, we proceed with the first change of measure argument: taking advantage of the gap between $\mu^\star$ and $\mu^\star - \varepsilon$, we move from a concentration argument around $\theta^\star$ to one around a shifted point $\theta^\star - \Delta_c$.

In Section 4.3, we localize the empirical parameter $\widehat{\theta}_n$ and make use of the second change of measure, this time to a mixture of measures, following Lai (1988). Even though we follow the same high level idea, we modified the original proof in order to better handle the cone covering, and also make all quantities explicit.

In Section 4.4, we apply a concentration of measure argument. This part requires a specific care since this is the core of the finite-time result. An important complication comes from the "boundary" of the parameter set, and was not explicitly controlled in the original proof from Lai (1988). A very careful analysis enables to obtain the finite-time concentration result without further restriction.

We finally combine all these steps in Sections 4.5.

4.1 Peeling and cone covering

In this section, the intuition we follow is that we want to control the random number of pulls $N_{a^\star}(t) \in [1, t]$ and, to this end, use a standard peeling argument, considering maximum concentration inequalities on time intervals $[b^i, b^{i+1}]$ for some $b > 1$. Likewise, since the term $\mathcal{K}_{a^\star}(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}), \mu^\star - \varepsilon)$ can be seen as an infimum of some quantity over the set of parameters Θ, we use a covering of Θ in order to reduce the control of the desired quantity to that of each cell of the cover. Formally, we show that

Lemma 7.5 (*Peeling and cone covering decomposition*) *For all $\beta \in (0,1), b > 1$ and $\eta \in [0,1)$ it holds*

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1 \leqslant n \leqslant t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}_{a^\star}(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}), \mu^\star - \varepsilon) \geqslant f(t/n)/n \Big\}$$
$$\leqslant \sum_{i=0}^{\lceil \log_b(\beta t+\beta)\rceil - 2} \sum_{c=1}^{C_{p,\eta,K}} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{b^i \leqslant n < b^{i+1}} E_{c,p}(n,t) \Big\} + \sum_{c=1}^{C_{p,\eta,K}} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=b^{\lceil \log_b(\beta t+\beta)\rceil - 1}}^{t} E_{c,p}(n,t) \Big\},$$

where the event $E_{c,p}(n,t)$ is defined by

$$E_{c,p}(n,t) \stackrel{\text{def}}{=} \left\{ \widehat{\theta}_n \in \Theta_\rho \cap \widehat{F}_n \in \mathcal{C}_p(\theta_c^\star) \cap \mathcal{B}^\psi(\widehat{\theta}_n, \theta_c^\star) \geqslant \frac{f(t/n)}{n} \right\}. \tag{7.6}$$

In this definition, $(\theta_c^\star)_{c \leqslant C_{p,\eta,K}}$, constrained to satisfy $\theta_c^\star \notin \mathcal{B}_2(\theta^\star, \eta\rho_\varepsilon)$, parameterize a minimal covering of $\nabla\psi(\Theta_\rho \setminus \mathcal{B}_2(\theta^\star, \rho_\varepsilon))$ with cones $\mathcal{C}_p(\theta_c^\star) := \mathcal{C}_p(\nabla\psi(\theta_c^\star); \theta^\star - \theta_c^\star)$ (That is $\nabla\psi(\Theta_\rho \setminus \mathcal{B}_2(\theta^\star, \rho_\varepsilon)) \subset \bigcup_{c=1}^{C_{p,\eta,K}} \mathcal{C}_p(\theta_c^\star)$), where $\mathcal{C}_p(y;\Delta) = \left\{ y' \in \mathbb{R}^K : \langle y' - y, \Delta \rangle \geqslant p \, \|y' - y\| \, \|\Delta\| \right\}$. For all $\eta < 1$, $C_{p,\eta,K}$ is of order $(1-p)^{-K}$ and $C_{p,\eta,1} = 2$, while $C_{p,\eta,K} \to \infty$ when $\eta \to 1$.

4.2 Change of measure

In this section, we focus on one event $E_{c,p}(n,t)$. The idea is to take advantage of the gap between $\mu^\star$ and $\mu^\star - \varepsilon$, that allows to shift from $\theta^\star$ to some of the $\theta_c^\star$ from the cover. The key observation is to control the change of measure from $\theta^\star$ to each $\theta_c^\star$. Note that $\theta_c^\star \in (\Theta_\rho \cap \mathcal{B}_2(\theta_c^\star, \rho_\varepsilon)) \setminus \mathcal{B}_2(\theta_c^\star, \eta\rho_\varepsilon)$ and that $\mu_{\theta_c^\star} \geqslant \mu^\star - \varepsilon$. We show that

Lemma 7.6 (*Change of measure*) *If $n \to nf(t/n)$ is non-decreasing, then for any increasing sequence $\{n_i\}_{i \geqslant 0}$ of non-negative integers it holds*

$$\mathbb{P}_{\theta^\star}\Big\{ \bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t) \Big\} \leqslant \exp\Big(- n_i\alpha^2 - \chi\sqrt{n_i f(t/n_i)} \Big) \mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t) \Big\}$$

where $\alpha = \alpha(p,\eta,\varepsilon) = \eta\rho_\varepsilon\sqrt{v_\rho/2}$ and $\chi = p\eta\rho_\varepsilon\sqrt{2v_\rho^2/V_\rho}$.

4.3 Localized change of measure

In this section, we decompose further the event of interest in $\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\Big\}$ in order to apply some concentration of measure argument. In particular, since by construction

$$\widehat{F}_n \in \mathcal{C}_p(\theta_c^\star) \Leftrightarrow \langle \Delta_c, \nabla\psi(\theta_c^\star) - \widehat{F}_n\rangle \geqslant p\,\|\Delta_c\|\,\Big\|\nabla\psi(\theta_c^\star) - \widehat{F}_n\Big\|\ ,$$

it is then natural to control $\Big\|\nabla\psi(\theta_c^\star) - \widehat{F}_n\Big\|$. This is what we call localization. More precisely, we introduce for any sequence $\{\varepsilon_{t,i,c}\}_{t,i}$ of positive values, the following decomposition

$$\begin{aligned}\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\Big\} \ \leqslant\ & \mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap ||\nabla\psi(\theta_c^\star) - \widehat{F}_n|| < \varepsilon_{t,i,c}\Big\}\\ &+\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap ||\nabla\psi(\theta_c^\star) - \widehat{F}_n|| \geqslant \varepsilon_{t,i,c}\Big\}. \qquad (7.7)\end{aligned}$$

We handle the first term in (7.7) by another change of measure argument that we detail below, and the second term thanks to a concentration of measure argument that we detail in section 4.4. We will show more precisely that

Lemma 7.7 (*Change of measure*) *For any sequence of positive values $\{\varepsilon_{t,i,c}\}_{i\geqslant 0}$, it holds*

$$\begin{aligned}&\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap \Big\|\nabla\psi(\widehat{\theta}_n) - \nabla\psi(\theta_c^\star)\Big\| < \varepsilon_{t,i,c}\Big\}\\ &\leqslant\ \alpha_{\rho,p}\exp\Big(-f\Big(\frac{t}{n_{i+1}-1}\Big)\Big)\min\Big\{\rho^2 v_\rho^2, \tilde{\varepsilon}_{t,i,c}^2, \frac{(K+2)v_\rho^2}{K(n_{i+1}-1)V_\rho}\Big\}^{-K/2}\tilde{\varepsilon}_{t,i,c}^K.\end{aligned}$$

where $\tilde{\varepsilon}_{t,i,c} = \min\{\varepsilon_{t,i,c}, \mathit{Diam}\big(\nabla\psi(\Theta_\rho)\cap\mathcal{C}_p(\theta_c^\star)\big)\}$ and $\alpha_{\rho,p} = 2\frac{\omega_{p,K-2}}{\omega_{p',K-2}}\Big(\frac{V_\rho}{v_\rho^2}\Big)^{K/2}\Big(\frac{V_\rho}{v_\rho}\Big)^K$ where $p' > \max\{p, \frac{2}{\sqrt{5}}\}$, with $\omega_{p,K} = \int_p^1\sqrt{1-z^2}^K dz$ for $K\geqslant 0$ and $w_{p,-1} = 1$.

4.4 Concentration of measure

In this section, we focus on the second term in (7.7), that is we want to control $\mathbb{P}_{\theta_c^\star}\Big\{\bigcup_{n_i\leqslant n<n_{i+1}} E_{c,p}(n,t)\cap ||\nabla\psi(\theta_c^\star) - \widehat{F}_n|| \geqslant \varepsilon_{t,i,c}\Big\}$. In this term, $\varepsilon_{t,i,c}$ should be considered as decreasing fast to 0 with i, and slowly increasing with t. Note that by definition $\nabla\psi(\widehat{\theta}_n) = \widehat{F}_{a^\star,n} = \frac{1}{n}\sum_{i=1}^n F(X_{a^\star,i}) \in \mathbb{R}^K$ is an empirical mean with mean given by $\nabla\psi(\theta_c^\star)\in\mathbb{R}^K$ and covariance matrix $\frac{1}{n}\nabla^2\psi(\theta_c^\star)$. We thus resort to a concentration of measure argument.

Lemma 7.8 (*Concentration of measure*) *Let $\varepsilon_c^{\max} = Diam(\nabla\psi(\Theta_\rho \cap \mathcal{C}_{c,p}))$ where we introduced the projected cone $\mathcal{C}_{c,p} = \{\theta \in \Theta : \langle \frac{\Delta_c}{\|\Delta_c\|}, \frac{\nabla\psi(\theta_c^\star)-\nabla\psi(\theta)}{\|\nabla\psi(\theta_c^\star)-\nabla\psi(\theta)\|}\rangle \geqslant p\}$. Then, for all $\varepsilon_{t,i,c}$, it holds*

$$\mathbb{P}_{\theta_c^\star}\Big\{ \bigcup_{n=n_i}^{n_{i+1}-1} E_{c,p}(n,t) \cap ||\nabla\psi(\widehat{\theta}_n) - \nabla\psi(\theta_c^\star)|| \geqslant \varepsilon_{t,i,c} \Big\} \leqslant \exp\Big(-\frac{n_i^2 p \varepsilon_{t,i,c}^2}{2V_\rho(n_{i+1}-1)} \Big) \mathbb{I}\{\varepsilon_{t,i,c} \leqslant \overline{\varepsilon}_c\}.$$

4.5 Combining the different steps

In this part, we recap what we have shown so far. Combining the peeling, change of measure, localization and concentration of measure steps of the four previous sections, we have shown that for all $\{\varepsilon_{t,i,c}\}_{t,i}$, then

$$\begin{aligned}
[1] &\overset{\text{def}}{=} \mathbb{P}_{\theta^\star}\Big\{ \bigcup_{1\leqslant n\leqslant t} \widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}_{a^\star}(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}), \mu^\star - \varepsilon) \geqslant f(t/n)/n \Big\} \\
&\leqslant \sum_{c=1}^{C_{p,\eta,K}} \sum_{i=0}^{I_t-1} \underbrace{\exp\Big(-n_i\alpha^2 - \chi\sqrt{n_i f(t/n_i)} \Big)}_{\text{change of measure}} \Big[\underbrace{\exp\Big(-\frac{n_i^2 p \varepsilon_{t,i,c}^2}{2V_\rho(n_{i+1}-1)} \Big) \mathbb{I}\{\varepsilon_{t,i,c} \leqslant \overline{\varepsilon}_c\}}_{\text{concentration}} \\
&\quad + \underbrace{\alpha_{p,K} \exp\Big(-f\Big(\frac{t}{n_{i+1}-1}\Big)\Big) \min\Big\{\rho^2 v_\rho^2, \varepsilon_{t,i,c}^2, \frac{(K+2)v_\rho^2}{K(n_{i+1}-1)V_\rho}\Big\}^{-K/2} \varepsilon_{t,i,c}^K}_{\text{localization + change of measure}} \Big],
\end{aligned}$$

where we recall that $\alpha = \alpha(p,\eta,\varepsilon) = \eta\rho_\varepsilon\sqrt{v_\rho/2}$ and that the definition of n_i is

$$n_i = \begin{cases} b^i \text{ if } i < I_t \overset{\text{def}}{=} \lceil \log_b(\beta t + \beta) \rceil \\ t+1 \text{ if } i = I_t\,. \end{cases}$$

A simple rewriting leads to the form

$$\begin{aligned}
[1] \;\leqslant\; & \sum_{c=1}^{C_{p,\eta,K}} \sum_{i=0}^{I_t-1} \exp\Big(-n_i\alpha^2 - \chi\sqrt{n_i f(t/n_i)} \Big) \Big[\alpha_{p,K} \exp\Big(-f\Big(\frac{t}{n_{i+1}-1}\Big)\Big) \times \\
& \max\Big\{ \frac{\varepsilon_{t,i,c}}{\rho v_\rho}, 1, \sqrt{\frac{(n_{i+1}-1)V_\rho}{1+2/K}} \frac{\varepsilon_{t,i,c}}{v_\rho} \Big\}^K + \exp\Big(-\frac{n_i^2 p \varepsilon_{t,i,c}^2}{2V_\rho(n_{i+1}-1)} \Big) \mathbb{I}\{\varepsilon_{t,i,c} \leqslant \overline{\varepsilon}_c\} \Big],
\end{aligned}$$

which suggests we use $\varepsilon_{t,i,c} = \sqrt{\frac{2V_\rho(n_{i+1}-1)f(t/(n_{i+1}-1))}{pn_i^2}}$. Replacing this term in the above expression, we obtain

$$\begin{aligned}
[1] \;\leqslant\; & \sum_{i=0}^{I_t-1} \exp\Big(-n_i\alpha^2 - \chi\sqrt{n_i f(t/n_i)} - f(t/(n_{i+1}-1)) \Big) f(t/(n_{i+1}-1))^{K/2} \times \\
& C_{p,\eta,K}\Big(\alpha_{p,K} \max\Big\{ \frac{2V_\rho}{p\rho^2 v_\rho^2 b^{i-1}}, 1, \frac{b^2 V_\rho^2}{p v_\rho^2 (\frac{1}{2}+\frac{1}{K})} \Big\}^{K/2} + 1 \Big).
\end{aligned}$$

At this point, using the somewhat crude lower bound $b^i \geqslant 1$, it is convenient to introduce the constant

$$C(K,\rho,p,b,\eta) = C_{p,\eta,K}\Big(\alpha_{p,K}\max\Big\{\frac{2bV_\rho}{p\rho^2 v_\rho^2}, 1, \frac{b^2V_\rho^2}{pv_\rho^2(\frac{1}{2}+\frac{1}{K})}\Big\}^{K/2}+1\Big)\,,$$

which leads to the final bound

$$\begin{aligned}
&\mathbb{P}_{\theta^\star}\Big\{\bigcup_{1\leqslant n\leqslant t}\widehat{\theta}_n \in \Theta_\rho \cap \mathcal{K}_{a^\star}(\Pi_{a^\star}(\widehat{\nu}_{a^\star,n}), \mu^\star-\varepsilon) \geqslant f(t/n)/n\Big\}\\
&\quad\leqslant\quad C(K,\rho,p,b,\eta)\sum_{i=0}^{I_t-1}\exp\Big(-n_i\alpha^2-\chi\sqrt{n_i f(t/n_i)}-f(t/(n_{i+1}-1))\Big)f(t/(n_{i+1}-1))^{K/2}\,.
\end{aligned}$$

CHAPTER 8

Hankel matrices

Contents

1 Weighted automata 131

1.1 Geometry and mixing properties of PFA and SWFA 133

2 Hankel matrices and spectral learning 134

2.1 Learning with Césaro Averages is Consistent 135

2.2 Spectral Learning Algorithm 136

2.3 Concentration Results 136

2.4 Concentration of Hankel Matrices for SWFA 137

This chapter corresponds to the article Balle and Maillard (2017), written in collaboration with Borja Balle (now at Amazon Research). We present below the main results only, and refer to the research article for further details.

1 WEIGHTED AUTOMATA

Let Σ be a finite alphabet, $\Sigma^\star$ denote the set of words of finite length on Σ, Σ^ω the set of all infinite words on Σ, and ε be the empty word. Given two sets of words $\mathcal{U}, \mathcal{V} \subset \Sigma^\star$ we write $\mathcal{U} \cdot \mathcal{V}$ to denote the set of words $\{uv | u \in \mathcal{U}, v \in \mathcal{V}\}$ obtained by concatenating all words in $\mathcal{U}$ with all words in $\mathcal{V}$. Let $\mathcal{P}(\Sigma^\omega)$ be the set of probability distributions over Σ^ω. A member $\rho \in \mathcal{P}(\Sigma^\omega)$ is called a *stochastic process* and a random infinite word $\xi \sim \rho$ is called a *trajectory*.

Definition 8.3 (*Weighted Finite Automaton*) *A* weighted finite automaton *(WFA) with n states is a tuple $\mathbb{A} = (\alpha, \beta, \{A_\sigma\}_{\sigma \in \Sigma})$ where $\alpha, \beta \in \mathbb{R}^n$ are vectors of initial and final weights, respectively, and $A_\sigma \in \mathbb{R}^{n \times n}$ are matrices of transition weights. A weighted automaton $\mathbb{A}$ computes a function $f_\mathbb{A} : \Sigma^\star \to \mathbb{R}$ given by $f_\mathbb{A}(w) = \alpha^\top A_w \beta$ where $A_w = A_{w_1} \cdots A_{w_t}$ for $w = w_1 \cdots w_t$.*

A WFA is *minimal* if there does not exist another WFA with less states computing the same function.

A WFA is *irreducible* if the labelled directed graph with n vertices obtained by adding a transition from i to j with label σ whenever $A_\sigma(i, j) \neq 0$ is strongly connected. It can be shown that irreducibility implies minimality, and that the set of irreducible WFAs is dense in the set of all WFA Balle et al. (2017).

Definition 8.6 (*Stochastic WFA*) *A WFA $\mathbb{A} = \langle \alpha, \beta, \{A_\sigma\}\rangle$ is* stochastic *(is a SWFA) if there exists a stochastic process $\rho_{\mathbb{A}}$ such that for every $w \in \Sigma^\star$, $f_{\mathbb{A}}(w) = \mathbb{P}[\xi \in w\Sigma^\omega]$ where $\xi \sim \rho_{\mathbb{A}}$; that is, $\mathbb{A}$ provides a representation for the probabilities of prefixes under the distribution of ρ. It is immediate to check that this implies that the weights of $\mathbb{A}$ satisfy the properties:*

(i) $\alpha^\top A_x \beta \geqslant 0$ *for all* $x \in \Sigma^\star$, *and*

(ii) $\alpha^\top A^t \beta = \sum_{|w|=t} \alpha^\top A_w \beta = 1$ *for all* $t \geqslant 0$, *where* $A = \sum_{\sigma\in\Sigma} A_\sigma$.

Without loss of generality we assume that $\mathbb{A}$ is a minimal SWFA of dimension n, meaning that any SWFA computing the same probability distribution than $\mathbb{A}$ must have dimension at least n. Importantly, the weights in α, β, and A_σ are *not* required to be non-negative in this definition. Nonetheless, it follows from these properties that β is an eigenvector of A of eigenvalue 1.

Definition 8.9 (*Probabilistic WFA*) *A* probabilistic finite automaton *(PFA) is a stochastic WFA $\mathbb{A} = (\alpha, \mathbf{1}, \{A_\sigma\})$ where the weights have a probabilistic interpretation. Namely, α is a probability distribution over $[n]$, $A_\sigma(i,j)$ is the probability of emitting symbol σ and transitioning to state j starting from state i, and $\mathbf{1}(i) = 1$ for all $i \in [n]$.*

It is immediate to check that a PFA satisfying these conditions induces a stochastic process. However not all stochastic WFA admit an equivalent PFA Jaeger (2000), Denis and Esposito (2008). If $\mathbb{A}$ is a PFA, then the matrix $A = \sum_{\sigma\in\Sigma} A_\sigma$ yields the Markov kernel $A(i,j) = \mathbb{P}[j \mid i]$ on the state space $[n]$ after marginalizing over the observations. It is easily checked that A is row-stochastic, and thus $A\beta = \beta$. Furthermore, for every distribution $\alpha_0 \in \mathbb{R}^n$ over $[n]$ we have $\alpha_0^\top A = \alpha_1$ for some other probability distribution α_1 over $[n]$. In the case of PFA, irreducibility coincides with the usual concept of irreducibility of the Markov chain induced by A.

Mixing and concentration Let $\rho \in \mathcal{P}(\Sigma^\omega)$ be a stochastic process and $\xi = x_1 x_2 \cdots$ a random word drawn from ρ. For $1 \leqslant s < t \leqslant T$ and $u \in \Sigma^s$ we let $\rho_{t:T}(\cdot|u)$ denote the distribution of $x_t \cdots x_T$ conditioned on $x_1 \cdots x_s = u$. With this notation we define the quantity

$$\eta_t(u, \sigma, \sigma') = \|\rho_{t:T}(\cdot|u\sigma) - \rho_{t:T}(\cdot|u\sigma')\|_{TV}$$

for any $u \in \Sigma^{s-1}$, and $\sigma, \sigma' \in \Sigma$. Then the *$\eta$-mixing coefficients* of ρ at horizon T are given by

$$\eta_{s,t} = \sup_{u\in\Sigma^{s-1},\sigma,\sigma'\in\Sigma} \eta_t(u, \sigma, \sigma') \ .$$

Mixing coefficients are useful in establishing concentration properties of functions of dependent random variables. The Lipschitz constant of a function $g : \Sigma^T \to \mathbb{R}$ with respect to the Hamming distance is defined as

$$\|g\|_{Lip} = \sup |g(w) - g(w')| \ ,$$

where the supremum is taken over all pairs of words $w, w' \in \Sigma^T$ differing in exactly one symbol. The following theorem proved in Chazottes et al. (2007), Kontorovich and Ramanan (2008) provides a concentration inequality for Lipschitz functions of weakly dependent random variables.

Theorem 8.1 *Let $\rho \in \mathcal{P}(\Sigma^\omega)$ and $\xi = x_1 x_2 \cdots \sim \rho$. Suppose $g : \Sigma^T \to \mathbb{R}$ satisfies $\|g\|_{Lip} \leqslant 1$ and let $Z = g(x_1, \ldots, x_T)$. Let $\eta_\rho = 1 + \max_{1<s<T} \sum_{t=s+1}^{T} \eta_{s,t}$, where $\eta_{s,t}$ are the η-mixing coefficients of ρ at horizon T. Then the following holds for any $\varepsilon > 0$:*

$$\mathbb{P}_\xi \left[Z - \mathbb{E}Z > \varepsilon T\right] \leqslant \exp\left(\frac{-2\varepsilon^2 T}{\eta_\rho^2}\right) \ ,$$

with an identical bound for the other tail.

Theorem 8.1 shows that the mixing coefficient η_ρ is a key quantity in order to control the concentration of a function of dependent variables. In fact, upper-bounding η_ρ in terms of geometric ergodicity coefficients of a latent variable stochastic process enables Kontorovich and Weiss (2014) to analyze the concentration of functions of HMMs and Azizzadenesheli et al. (2016) to provide PAC guarantees for an RL algorithm for POMDP based on spectral tensor decompositions. Our Lemma 8.1 uses a similar but more refined bounding strategy that directly applies when the transition and observation processes are *not* conditionally independent. Lemma 8.3 refines this strategy further to control η_ρ for stochastic WFA (for which there may be no underlying Markov stochastic process in general). To the best of our knowledge, this yields the first concentration results for the challenging setting of stochastic WFA.

1.1 Geometry and mixing properties of PFA and SWFA

We recall below the traditional definition of mixing for PFA.

Lemma 8.1 (***η-mixing for PFA***) *Let $\mathbb{A}$ be PFA and assume that it is (C,θ)-geometrically mixing in the sense that for some constants $C>0, \theta\in(0,1)$ we have*

$$\forall t\in\mathbb{N},\quad \mu_t^{\mathbb{A}} = \sup_{\alpha,\alpha'} \frac{\|\alpha A^t - \alpha' A^t\|_1}{\|\alpha-\alpha'\|_1} \leqslant C\theta^t\,,$$

where the supremum is over all probability vectors. Then we have $\eta_{\rho_{\mathbb{A}}} \leqslant C/(\theta(1-\theta))$.

Remark 8.1 *A sufficient condition for the geometric control of $\mu_t^{\mathbb{A}}$ is that A admits a spectral gap. In this case θ can be chosen to be the modulus of the second eigenvalue $|\lambda_2(A)|<1$ of the transition kernel A.*

This notion however does not apply nicely to SWFA. Before presenting the correct extension of the mixing coefficients for SWFA, let us provide a short tour of the geometry of SWFA.

Cone norms of SWFA A minimal SWFA $\mathbb{A}$ is naturally associated with a proper (i.e. pointed, closed, and solid) cone in $\mathcal{K}\subset\mathbb{R}^n$ called the *backward cone* Jaeger (2000), and characterized by the following properties: 1) $\beta\in\mathcal{K}$, 2) $A_\sigma\mathcal{K}\subseteq\mathcal{K}$ for all $\sigma\in\Sigma$, and 3) $\alpha^\top v\geqslant 0$ for all $v\in\mathcal{K}$. Condition 2) says that every transition matrix A_σ leaves $\mathcal{K}$ invariant, and in particular the backward vector $A_w\beta$ belongs to $\mathcal{K}$ for all $w\in\Sigma^\star$.

The vector of final weights β plays a singular role in the geometry of the state space of a SWFA. This follows from facts about the theory of invariant cones Berman and Plemmons (1994) which provides a generalization of the classical Perron–Frobenius theory of non-negative matrices to arbitrary matrices. We recall from Berman and Plemmons (1994) that a norm on $\mathbb{R}^n$ can be associated with every vector in the interior of $\mathcal{K}$. In particular, we will take the norm associated with the final weights $\beta\in\mathcal{K}$. This norm, denoted by $\|\cdot\|_\beta$, is completely determined by its unit ball $B_\beta = \{v\in\mathbb{R}^n : -\beta\leqslant_{\mathcal{K}} v\leqslant_{\mathcal{K}}\beta\}$, where $u\leqslant_{\mathcal{K}} v$ means $v-u\in\mathcal{K}$. In particular, $\|v\|_\beta = \inf\{r\geqslant 0 : v\in rB_\beta\}$. Induced and dual norms are derived from $\|\cdot\|_\beta$ as usual. When $\mathbb{A}$ is a PFA, one can take $\mathcal{K}$ to be the cone of vectors in $\mathbb{R}^n$ with non-negative entries, in which case $\beta=(1,\dots,1)$ and $\|\cdot\|_\beta$ reduces to $\|\cdot\|_\infty$ Berman and Plemmons (1994). The following result shows that $\|\cdot\|_\beta$ indeed provides the right generalization to SWFA of the norm $\|\cdot\|_\infty$.

Lemma 8.2 (***Cone-norm properties***) *For any $w\in\Sigma^\star$: (i) $\|A_w\beta\|_\beta\leqslant 1$, and (ii) $\left\|\alpha^\top A_w\right\|_{\beta,*} = \alpha^\top A_w\beta$.*

It is also natural to consider mixing coefficients for stochastic processes generated by SWFA in terms of the dual β-norm. This provides a direct analog to Lemma 8.1 for PFA:

Lemma 8.3 (***η-mixing for SWFA***) *Let $\mathbb{A}$ be SWFA and assume that it is (C,θ)-geometrically mixing in the sense that for some $C \geqslant 0, \theta \in (0,1)$,*

$$\mu_t^{\mathbb{A}} = \sup_{\alpha_0,\alpha_1 : \alpha_0^\top \beta = \alpha_1^\top \beta = 1} \frac{\left\| \alpha_0^\top A^t - \alpha_1^\top A^t \right\|_{\beta,\star}}{\|\alpha_0 - \alpha_1\|_{\beta,\star}} \leqslant C\theta^t \quad .$$

Then the η-mixing coefficient satisfies $\eta_{\rho_{\mathbb{A}}} \leqslant C/(\theta(1-\theta))$.

Remark 8.2 *A sufficient condition for the geometric control of $\mu_t^{\mathbb{A}}$ is that A admits a spectral gap. In this case θ can be chosen to be the modulus of the second eigenvalue $|\lambda_2(A)| < 1$ of A. Another sufficient condition is that $\theta = \gamma_\beta(A) < 1$, where*

$$\gamma_\beta(A) = \sup \left\{ \frac{||A\nu||_{\beta,\star}}{||\nu||_{\beta,\star}} : \nu \text{ s.t. } ||\nu||_{\beta,\star} \neq 0, \nu^\top \beta = 0 \right\}.$$

2 HANKEL MATRICES AND SPECTRAL LEARNING

The Hankel matrix of a function $f : \Sigma^\star \to \mathbb{R}$ is the infinite matrix $H_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with entries $H_f(u,v) = f(uv)$. Given finite sets $\mathcal{U}, \mathcal{V} \subset \Sigma^\star$, $H_f^{\mathcal{U},\mathcal{V}} \in \mathbb{R}^{\mathcal{U} \times \mathcal{V}}$ denotes the restriction of matrix H_f to prefixes in $\mathcal{U}$ and suffixes in $\mathcal{V}$.

Fliess' Theorem Fliess (1974) states that a Hankel matrix H_f has finite rank n if and only if there exists a WFA $\mathbb{A}$ with n states such that $f = f_{\mathbb{A}}$. This implies that a WFA $\mathbb{A}$ with n states is minimal if and only if $n = rank(H_{f_{\mathbb{A}}})$. The spectral learning algorithm for WFA Balle et al. (2014) provides a mechanism for recovering such a WFA from a finite sub-block $H_f^{\mathcal{U},\mathcal{V}}$ of H_f such that: 1) $\varepsilon \in \mathcal{U} \cap \mathcal{V}$, 2) there exists a set $\mathcal{U}'$ such that $\mathcal{U} = \mathcal{U}' \cup \left(\bigcup_{\sigma \in \Sigma} \mathcal{U}'\sigma\right)$, 3) $rank(H_f) = rank(H_f^{\mathcal{U}',\mathcal{V}})$. A pair $(\mathcal{U},\mathcal{V})$ that satisfies these conditions is called a *complete basis* for f. If these conditions are satisfied, we say that the pair $(\mathcal{U},\mathcal{V})$ is a *complete basis* for f. The pseudo-code of this algorithm is given below:

Algorithm 5 Spectral Learning for WFA

Input: number of states n, Hankel matrix $H^{\mathcal{U},\mathcal{V}}$

1: Find $\mathcal{U}'$ such that $\mathcal{U} = \mathcal{U}' \cup (\cup_{\sigma \in \Sigma} \mathcal{U}'\sigma)$
2: Let $H_\varepsilon = H^{\mathcal{U}',\mathcal{V}}$
3: Compute the rank n SVD $H_\varepsilon \approx UDV^\top$
4: Let $h_{\mathcal{V}} = H^{\{\varepsilon\},\mathcal{V}}$ and take $\alpha = V^\top h_{\mathcal{V}}$
5: Let $h_{\mathcal{U}'} = H^{\mathcal{U}',\{\varepsilon\}}$ and take $\beta = D^{-1}U^\top h_{\mathcal{U}'}$
6: **for** $\sigma \in \Sigma$ **do**
7: Let $H_\sigma = H^{\mathcal{U}'\sigma,\mathcal{V}}$ and take $A_\sigma = D^{-1}U^\top H_\sigma V$
8: $\mathbb{A} = (\alpha, \beta, \{A_\sigma\})$

The main strength of Algorithm 5 is its robustness to noise. Specifically, if only an approximation $\widehat{H}^{\mathcal{U},\mathcal{V}}$ of the Hankel matrix is known, then the error between the target automaton $\mathbb{A}$ and the automaton $\widehat{\mathbb{A}}$ learned

from $\widehat{H}^{\mathcal{U},\mathcal{V}}$ can be controlled in terms of the error $\left\| H^{\mathcal{U},\mathcal{V}} - \widehat{H}^{\mathcal{U},\mathcal{V}} \right\|_2$; see Hsu et al. (2012) for a proof in the HMM case and Balle (2013) for a proof in the general WFA case. These tedious but now standard arguments readily reduce the problem of learning WFA via spectral learning to that of estimating the corresponding Hankel matrix.

Classical applications of spectral learning assume one has access to i.i.d. samples from a stochastic process ρ. In this setting one can obtain a sample $S = (\xi^{(1)}, \ldots, \xi^{(N)})$ containing N finite-length trajectories from ρ, and use them to estimate a Hankel matrix $\widehat{H}_S^{\mathcal{U},\mathcal{V}}$ as follows:

$$\widehat{H}_S^{\mathcal{U},\mathcal{V}}(u,v) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\xi^{(i)} \in uv\Sigma^\omega\}\,.$$

If $\rho = \rho_{\mathbb{A}}$ for some stochastic WFA, then obviously $\mathbb{E}_S[\widehat{H}_S^{\mathcal{U},\mathcal{V}}] = H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}}$ and a large sample size N will provide a good approximation $\widehat{H}_S^{\mathcal{U},\mathcal{V}}$ of $H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}}$. Explicit concentration bounds for Hankel matrices bounding the error $\left\| H_{f_{\mathbb{A}}}^{\mathcal{U},\mathcal{V}} - \widehat{H}_S^{\mathcal{U},\mathcal{V}} \right\|_2$ can be found in Denis et al. (2016).

In this chapter we consider the more challenging setup where we only have access to a sample $S = \{\xi\}$ of size $N = 1$ from ρ. In particular, we show it is possible to replace the empirical average above by a Césaro average and still use the spectral learning algorithm to recover the transition matrices of a stochastic WFA. To obtain a finite-sample analysis of this *single-trajectory* learning algorithm, we prove concentration results for Césaro averages of Hankel matrices. Our analysis relies on concentration inequalities for functions of dependent random variables, which depend on mixing properties of the underlying process.

2.1 Learning with Césaro Averages is Consistent

Let $\mathbb{A} = (\alpha, \beta, \{A_\sigma\})$ be a PFA computing a function $f_{\mathbb{A}} : \Sigma^\star \to \mathbb{R}$ and defining a stochastic process $\rho_{\mathbb{A}} \in \mathcal{P}(\Sigma^\omega)$. For convenience we drop the subscript and just write f and ρ. Since we only have access to a single trajectory ξ from ρ we cannot obtain an approximation of the Hankel matrix for f by averaging over multiple i.i.d. trajectories. Instead, we compute Césaro averages over the trajectory ξ to obtain a Hankel matrix whose expectation is related to $\mathbb{A}$ as follows.

For any $t \in \mathbb{N}$ let $\bar{f}_t : \Sigma^\star \to \mathbb{R}$ be the function given by $\bar{f}_t(w) = (1/t)\sum_{s=0}^{t-1} f(\Sigma^s w)$, where $f(\Sigma^s w) = \sum_{u\in\Sigma^s} f(uw)$. We shall sometimes write $f_s(w) = f(\Sigma^s w)$. Using the definition of the function computed by a WFA it is easy to see that

$$\sum_{u\in\Sigma^s} f(uw) = \sum_{u\in\Sigma^s} \alpha^\top A_u A_w \beta = \alpha^\top A^s A_w \beta \ ,$$

where $A = \sum_\sigma A_\sigma$ is the Markov kernel on the state space of A. Thus, introducing $\bar{\alpha}_t^\top = (1/t)\sum_{s=0}^{t-1} \alpha^\top A^s$, we get $\bar{f}_t(w) = \bar{\alpha}_t^\top A_w \beta$. Since α is a probability distribution, A is a Markov kernel, and probability distributions are closed by convex combinations, then $\bar{\alpha}_t$ is also a probability distribution over $[n]$. Thus, we have just proved the following:

Lemma 8.4 (*Cesaro consistency*) *The Césaro average of f over t steps, $\bar{f}_t$, is computed by the probabilistic automaton $\bar{\mathbb{A}}_t = (\bar{\alpha}_t, \beta, \{A_\sigma\})$. In particular, $\mathbb{A}$ and $\bar{\mathbb{A}}_t$ have the same number of states and the same transition probability matrices. Furthermore, if $\mathbb{A}$ is irreducible then $\bar{\mathbb{A}}_t$ is minimal.*

The irreducibility claim follows from Balle et al. (2017). For convenience, in the sequel we write $\bar{H}_t^{\mathcal{U},\mathcal{V}}$ for the $(\mathcal{U},\mathcal{V})$-block of the Hankel matrix $H_{f_{\bar{\mathbb{A}}_t}}$.

Remark 8.3 *The irreducible condition simply ensures there is a unique stationary distribution, and that the Hankel matrix of $\bar{\mathbb{A}}_t$ has the same rank as the Hankel matrix of $\mathbb{A}$ (otherwise it could be smaller).*

2.2 Spectral Learning Algorithm

Algorithm 6 describes the estimation of the empirical Hankel matrix $\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}$ from the first $t+L$ symbols of a single trajectory using the corresponding Césaro averages. To avoid cumbersome notations, in the sequel we may drop super and subscripts when not needed and write $\widehat{H}_t$ or $\widehat{H}$ when $\mathcal{U}$, $\mathcal{V}$, and ξ are clear from the context. Note that by Lemma 8.4, the expectation $\mathbb{E}[\widehat{H}]$ over $\xi \sim \rho$ is equal to the Hankel matrix $\bar{H}_t$ of the function $\bar{f}_t$ computed by the PFA $\bar{\mathbb{A}}_t$.

Algorithm 6 Single Trajectory Spectral Learning (Generative Case)

Input: number of states n, length t, prefixes $\mathcal{U} \subset \Sigma^\star$, suffixes $\mathcal{V} \subset \Sigma^\star$

Let $L = \max_{w \in \mathcal{U}\cdot\mathcal{V}} |w|$
Sample trajectory $\xi = x_1 x_2 \cdots x_{t+L} \cdots \sim \rho$
for $u \in \mathcal{U}$ and $v \in \mathcal{V}$ **do**
    Let $\widehat{H}(u,v) = \frac{1}{t}\sum_{s=0}^{t-1} \mathbb{I}\{x_{s+1:s+|uv|} = uv\}$
Apply the spectral algorithm to $\widehat{H}$ with rank n

2.3 Concentration Results

Now we proceed to analyze the error $\widehat{H}_t - \bar{H}_t$ in the Hankel matrix estimation inside Algorithm 6. In particular, we provide concentration bounds that depend on the length t, the mixing coefficient η_ρ of the process ρ, and the structure of the basis $(\mathcal{U},\mathcal{V})$. The main result of this section is the matrix concentration bound Theorem 8.3 where we control the spectral norm of the error matrix. For comparison we also provide a simpler entry-wise bound and recall the equivalent matrix bound in the i.i.d. setting.

Before stating the main result of this section, we provide a concentration result for each individual entry of the estimated Hankel matrix as a warm-up.

Theorem 8.2 (*Single-trajectory, entry-wise*) *Let $\mathbb{A}$ be a (C,θ)-geometrically mixing PFA and $\xi \sim \rho_{\mathbb{A}}$ a trajectory of observations. Then for any $u \in \mathcal{U}, v \in \mathcal{V}$ and $\delta \in (0,1)$,*

$$\mathbb{P}\left[\widehat{H}_{t,\xi}^{\mathcal{U},\mathcal{V}}(u,v) - \bar{H}_t^{\mathcal{U},\mathcal{V}}(u,v) \geqslant \frac{|uv|C}{\theta(1-\theta)}\sqrt{\left(1+\frac{|uv|-1}{t}\right)\frac{\log(1/\delta)}{2t}}\right] \leqslant \delta \ ,$$

with an identical bound for the other tail.

A naive way to handle the concentration of the whole Hankel matrix is to control the Frobenius norm $\left\|\widehat{H}_t - \bar{H}_t\right\|_F$ by taking a union bound over all entries using Theorem 8.2. However, the resulting concentration bound would scale as $\sqrt{|\mathcal{U}||\mathcal{V}|}$. To have better dependency with the dimension (the matrix has dimension $|\mathcal{U}| \times$

$|\mathcal{V}|$) can split the empirical Hankel matrix $\widehat{H}$ into blocks containing strings of the same length (as suggested by the dependence of the bound above on $|uv|$). We thus introduce the maximal length $L = \max_{w \in \mathcal{U}\cdot\mathcal{V}} |w|$, and the set $\mathcal{U}_\ell = \{u \in \mathcal{U} : |u| = \ell\}$ for any $\ell \in \mathbb{N}$. We use these to define the quantity $n_\mathcal{U} = |\{\ell \in [0, L] : |\mathcal{U}_\ell| > 0\}|$, and introduce likewise $\mathcal{V}_\ell, n_\mathcal{V}$ with obvious definitions. With this notation we can now state the main result of this section.

Theorem 8.3 (*Single-trajectory, matrix-wise*) *Let $\mathbb{A}$ be as in Theorem 8.2. Let $m = \sum_{u \in \mathcal{U}, v \in \mathcal{V}} \bar{f}_t(uv)$ be the probability mass and $d = \min\{|\mathcal{U}||\mathcal{V}|, 2n_\mathcal{U} n_\mathcal{V}\}$ be the effective dimension. Then, for all $\delta \in (0,1)$ we have*

$$\mathbb{P}\left[\left\| \widehat{H}^{\mathcal{U},\mathcal{V}}_{t,\xi} - \bar{H}^{\mathcal{U},\mathcal{V}}_t \right\|_2 \geqslant \left(\sqrt{L} + \sqrt{\frac{2C}{1-\theta}}\right)\sqrt{\frac{2m}{t}} + \frac{2LC}{\theta(1-\theta)}\sqrt{\left(1+\frac{L-1}{t}\right)\frac{d\ln(1/\delta)}{2t}} \right] \leqslant \delta \ .$$

Remark 8.4 *Note that quantity $n_\mathcal{U} n_\mathcal{V}$ in d can be exponentially smaller than $|\mathcal{U}||\mathcal{V}|$. Indeed, for $\mathcal{U} = \mathcal{V} = \Sigma^{\leqslant L/2}$ we have $|\mathcal{U}||\mathcal{V}| = \Theta(|\Sigma|^L)$ while $n_\mathcal{U} n_\mathcal{V} = \Theta(L^2)$.*

For comparison, we recall a state-of-the-art concentration bound for estimating the Hankel matrix of a stochastic language[1] from N i.i.d. trajectories.

Theorem 8.4 (*Theorem 7 in Denis et al. (2014)*) *Let $\mathbb{A}$ be a stochastic WFA with stopping probabilities and $S = (\xi^{(1)}, \ldots, \xi^{(N)})$ be an i.i.d. sample of size N from the distribution $\rho_\mathbb{A} \in \mathcal{P}(\Sigma^\star)$. Let $m = \sum_{u \in \mathcal{U}, v \in \mathcal{V}} f_\mathcal{A}(uv)$. Then, for all $c > 0$ we have*

$$\mathbb{P}\left[\left\| \widehat{H}^{\mathcal{U},\mathcal{V}}_S - H^{\mathcal{U},\mathcal{V}}_{f_\mathbb{A}} \right\|_2 > \sqrt{\frac{2cm}{N}} + \frac{2c}{3N} \right] \leqslant \frac{2c}{e^c - c - 1} \ .$$

2.4 Concentration of Hankel Matrices for SWFA

We are now ready to extend the proof of Theorem 8.3 to SWFA. Using that both PFA and SWFA define probability distributions over prefixes, it follows that any argument in the proof that only appeals to the function computed by the automaton can remain unchanged.

Recalling that Hölder's inequality can be applied with any pair of dual norms, we start by replacing the norms $\|\cdot\|_\infty$ and $\|\cdot\|_1$ with the cone-norms $\|\cdot\|_\beta$ and $\|\cdot\|_{\beta,\star}$ respectively. Next we use Lemma 8.2 to obtain, for any $w \in \Sigma^\star$, the bound $\|A_w \beta\|_\beta \leqslant 1$ and the equation $\left\| \alpha^\top A_w \right\|_{\beta,*} = \alpha^\top A_w \beta$ which are direct analogs of the results used for PFA. Then it only remains to relate the β-norm of $A^{s'-s-l} - \beta\alpha^\top_{s'-1}$ to the mixing coefficients $\mu^\mathbb{A}_t$: we obtain $\left\| A^{s'-s-l} - \beta\alpha^\top_{s'-1} \right\|_\beta \leqslant 2\mu^\mathbb{A}_{s'-s-l}$. Thus we obtain for SWFA exactly the same concentration result that we obtained for empirical Hankel matrices estimated from a single trajectory of observations generated by a PFA.

[1] A stochastic language is a probability distribution over $\Sigma^\star$.

Theorem 8.5 (*Single-trajectory, SWFA*) *Let $\mathbb{A}$ be a (C, θ)-geometrically mixing SWFA with the definition in Lemma 8.3. Then the concentration bound in Theorem 8.3 also holds for trajectories $\xi \sim \rho_{\mathbb{A}}$.*

CHAPTER 9

Aggregation of growing experts

Contents

1 Introduction 139

2 Preliminary: the exponential weights algorithm 140

3 Overview of the results 141

3.1 Regret against arbitrary sequences of experts 144

3.2 Sparse shifting regret for growing experts 145

This chapter corresponds to the article Mourtada and Maillard (2017), written following the supervision of Jaouad Mourtada during his Master internship.

1 INTRODUCTION

Aggregation of experts is a well-established framework in machine learning (Cesa-Bianchi and Lugosi, 2006, Vovk, 1998, Györfi et al., 1999, Haussler et al., 1998), that provides a sound strategy to combine the forecasts of many different sources. This is classically considered in the sequential prediction setting, where at each time step, a learner receives the predictions of experts, uses them to provide his own forecast, and then observes the true value of the signal, which determines his loss and those of the experts. The goal is then to minimize the *regret* of the learner, which is defined as the difference between his cumulated loss and that of the best expert (or combination thereof), no matter what the experts' predictions or the values of the signal are.

A standard assumption in the existing literature is that the set of experts is known before the beginning of the game. In many situations, however, it is desirable to add more and more forecasters over time. For instance, in a non-stationary setting one could add new experts trained on a fraction of the signal, possibly combined with change point detection. Even in a stationary setting, a growing number of increasingly complex models enables to account for increasingly subtle properties of the signal without having to include them from the start, which can be needlessly costly computationally (as complex models, which take more time to fit, are not helpful in the first rounds) or even intractable in the case of an infinite number of models with no closed form expression. Additionally, in many realistic situations some completely novel experts may appear in an unpredicted way (possibly due to innovation, the discovery of better algorithms or the availability of new data), and one would want a way to safely incorporate them into the aggregation procedure.

In this chapter, we study how to amend aggregation of experts strategies in order to incorporate novel experts that may be added on the fly at any time step. Importantly, since we do not know in advance when new experts are made available, we put a strong emphasis on *anytime* strategies, that do not assume the time horizon is finite and known. Likewise, our algorithms should be agnostic to the total number of experts available at a given time. Three notions of regret of increasing complexity will be defined for growing expert sets, that extend existing notions to a growing expert set. Besides comparing against the best expert, it is natural in a growing experts setting to track the best expert; furthermore, when the number of experts grows large, it becomes

profitable to track the best expert in a small pool of good experts. For each notion, we propose corresponding algorithms with tight regret bounds. As is often the case in structured aggregation of experts, the key difficulty is typically not to derive the regret bounds, but to obtain efficient algorithms. All our methods exhibit minimal time and space requirements that are linear in the number of present experts.

2 Preliminary: the exponential weights algorithm

First, we introduce the simple but fundamental *exponential weights* or *Hedge algorithm* (Vovk, 1998, Cesa-Bianchi and Lugosi, 2006), designed to control the regret $L_T - L_{i,T} = \sum_{t=1}^T \ell_t - \sum_{t=1}^T \ell_{i,t}$ for a fixed set of experts $\{1, \dots, M\}$. The algorithm depends on a *prior distribution* $\boldsymbol{\pi} \in \mathscr{P}_M$ on the experts and predicts as

$$x_t = \frac{\sum_{i=1}^M w_{i,t}\, x_{i,t}}{\sum_{i=1}^M w_{i,t}} \quad \text{with} \quad w_{i,t} = \pi_i\, e^{-\eta L_{i,t-1}}\,. \tag{9.1}$$

Equivalently, it forecasts $x_t = \sum_{i=1}^M v_{i,t}\, x_{i,t}$, where the weights $\boldsymbol{v}_t \in \mathscr{P}_M$ are sequentially updated in the following way: $\boldsymbol{v}_1 = \boldsymbol{\pi}$ and, after each round $t \geqslant 1$, $\boldsymbol{v}_{t+1}$ is set to the *posterior* distribution $\boldsymbol{v}_t^m$ of $\boldsymbol{v}_t$ given the losses $(\ell_{i,t})_{1 \leqslant i \leqslant M}$, defined by

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,. \tag{9.2}$$

All subsequent regret bounds will rely on the following standard regret bound, by reducing complex forecasting strategies to the aggregation of experts under a suitable prior.

Proposition 9.1 (***Cesa-Bianchi and Lugosi (2006, Corollary 3.1)***) *Irrespective of the values of the signal and the experts' predictions, the exponential weights algorithm (9.1) with prior $\boldsymbol{\pi}$ achieves*

$$L_T - L_{i,T} \leqslant \frac{1}{\eta} \log \frac{1}{\pi_i} \tag{9.3}$$

for each $i = 1, \dots, M$ and $T \geqslant 1$. More generally, for each probability vector $\boldsymbol{u} \in \mathscr{P}_M$,

$$L_T - \sum_{i=1}^M u_i\, L_{i,T} \leqslant \frac{1}{\eta}\, KL(\boldsymbol{u}, \boldsymbol{\pi})\,. \tag{9.4}$$

Choosing a uniform prior $\boldsymbol{\pi} = \frac{1}{M}\mathbf{1}$ yields an at most $\frac{1}{\eta} \log M$ regret with respect to the best expert.

Related work. This work builds on the setting of prediction with expert advice (Cesa-Bianchi and Lugosi, 2006, Vovk, 1998, Herbster and Warmuth, 1998) that originates from the work on universal prediction (Ryabko, 1984, 1988, Merhav and Feder, 1998, Györfi et al., 1999). We make use of the notion of *specialists* (Freund et al., 1997, Chernov and Vovk, 2009) and its application to *sleeping experts* (Koolen et al., 2012), as well as the corresponding standard extensions (Fixed Share, Mixing Past Posteriors) of basic strategies to the problem of *tracking the best expert* (Herbster and Warmuth, 1998, Koolen and de Rooij, 2013, Bousquet and Warmuth, 2002); see also Willems (1996), Shamir and Merhav (1999) for related work in the context of lossless compression. Note that, due to its versatility, aggregation of experts has been adapted successfully to a number of applications (Monteleoni et al., 2011, McQuade and Monteleoni, 2012, Stoltz, 2010). It should be noted that the literature on prediction with expert advice is split into two categories: the first one focuses on exp-concave loss functions, whereas the second studies convex bounded losses. While our work belongs to the first category,

it should be possible to transport our regret bounds to the convex bounded case by using time-varying learning rates, as done e.g. by Hazan and Seshadhri (2009) and Gyorgy et al. (2012). In this case, the growing body of work on the automatic tuning of the learning rate (de Rooij et al., 2014, Koolen et al., 2014) as well as alternative aggregation schemes (Wintenberger, 2017, Koolen and van Erven, 2015, Luo and Schapire, 2015) might open the path for even further improvements.

The use of a growing expert ensemble was already proposed by Györfi et al. (1999) in the context of sequentially predicting an ergodic stationary time series, where new higher order Markov experts were introduced at exponentially increasing times (and the weights were reset to uniform); since consistency was the core focus of the paper, this simple "doubling trick" could be used, something we cannot afford when new experts arrive more regularly. Closer to our approach, growing expert ensembles have been considered in contexts where the underlying signal may be non-stationary, see e.g. Hazan and Seshadhri (2009), Shalizi et al. (2011). Of special interest to our problem is Shalizi et al. (2011), which considers the particular case when one new expert is introduced every τ time steps, and propose a variant of the Fixed Share (FS) algorithm analogous to our GrowingMarkovHedge algorithm. However, their algorithms depend on parameters which have to be tuned depending on the parameters of the comparison class, whereas our algorithms are parameter-free and do not assume the prior knowledge of the comparison class. Moreover, we introduce several other algorithms tailored to different notions of regret; in particular, we address the problem of comparing to sequences of experts that alternate between a small number of experts, a refinement that is crucial when the total set of experts grows, and has not been obtained previously in this context.

Another related setting is that of "branching experts" considered by Gofer et al. (2013), where each incumbent expert is split into several experts that may diverge later on. Their results include a regret bound in terms of the number of *leading experts* (whose cumulated loss was minimal at some point). Our approach differs in that it does not assume such a tree-like structure: a new entering forecaster is not assumed to be associated to an incumbent expert. More importantly, while Gofer et al. (2013) compare to the leaders in terms of cumulated loss (since the beginning of the game), our methods compete instead with sequences of experts that perform well on some periods, but can predict arbitrarily bad on others; this is harder, since the loss of the optimal sequence of experts can be significantly smaller than that of the best expert.

3 Overview of the results

Our work is framed in the classical setting of *prediction with expert advice* (Vovk, 1998, Cesa-Bianchi and Lugosi, 2006), which we adapt to account for a growing number of experts. The problem is characterized by its *loss function* $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{X}$ is a convex *prediction space*, and $\mathcal{Y}$ is the *signal space*.

Let M_t be the total number of experts at time t, and $m_t = M_t - M_{t-1}$ be the number of experts introduced at time t. We index experts by their entry order, so that expert i is the i^{th} introduced expert and denote $\tau_i = \min\{t \geqslant 1 : i \leqslant M_t\}$ its *entry time* (the time at which it is introduced). We say we are in the *fixed expert set* case when $M_t = M$ for every $t \geqslant 1$ and in the *growing experts setting* otherwise. At each step $t \geqslant 1$, the experts $i = 1, \dots, M_t$ output their predictions $x_{i,t} \in \mathcal{X}$, which the learner uses to build $x_t \in \mathcal{X}$; then, the environment decides the value of the signal $y_t \in \mathcal{Y}$, which sets the losses $\ell_t = \ell(x_t, y_t)$ of the learner and $\ell_{i,t} = \ell(x_{i,t}, y_t)$ of the experts.

Notations. Let $\mathscr{P}_M$ be the *probability simplex*, *i.e.* the set of probability measures over the set of experts $\{1, \dots, M\}$. We denote by $\mathtt{KL}$ the *Kullback-Leibler divergence*, defined for $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{P}_M$ by $\mathtt{KL}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{i=1}^{M} u_i \ln \frac{u_i}{v_i} \geqslant 0$.

Loss function. Throughout this text, we make the following standard assumption[1] on the loss function (Cesa-Bianchi and Lugosi, 2006).

Assumption 9.1 *The loss function ℓ is* η-exp-concave *for some $\eta > 0$, in the sense that $\exp(-\eta\,\ell(\cdot, y))$ is concave on $\mathcal{X}$ for every observation $y \in \mathcal{Y}$. This is equivalent to the inequality*

$$\ell\left(\sum_{i=1}^{M} v_i\, x_i, y\right) \leqslant -\frac{1}{\eta} \ln \sum_{i=1}^{M} v_i\, e^{-\eta\,\ell(x_i, y)} \tag{9.5}$$

for every $y \in \mathcal{Y}$, $\boldsymbol{x} = (x_i)_{1\leqslant i\leqslant M} \in \mathcal{X}^M$ and $\boldsymbol{v} = (v_i)_{1\leqslant i\leqslant M} \in \mathscr{P}_M$.

Remark 9.1 *An important example in the case when $\mathcal{X}$ is the set of probability measures over $\mathcal{Y}$ is the* logarithmic *or* self-information *loss $\ell(x, y) = -\log x(\{y\})$ for which the inequality holds with $\eta = 1$, and is actually an equality. Another example of special interest is the quadratic loss on a bounded interval: indeed, for $\mathcal{X} = \mathcal{Y} = [a, b] \subset \mathbb{R}$, $\ell(x, y) = (x - y)^2$ is $\frac{1}{2(b-a)^2}$-exp-concave.*

Several notions of regret can be considered in the growing expert setting. We review here three of them, each corresponding to a specific comparison class; we show the kind of bounds that our algorithms achieve, to illustrate the more general results stated in the subsequent sections.

Constant experts. Since the experts only output predictions after their entry time, it is natural to consider the *regret* with respect to each expert $i \geqslant 1$ over its time of activity, namely the quantity

$$\sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \tag{9.6}$$

for every $T \geqslant \tau_i$. Note that this is equivalent to controlling (9.6) for every $T \geqslant 1$ and $i \leqslant M_T$. Algorithm GrowingHedge is particularly relevant in this context; with the choice of (unnormalized) prior weights $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$, it achieves the following regret bound: for every $T \geqslant 1$ and $i \leqslant M_T$,

$$\sum_{t=\tau_i}^{T} (\ell_t - \ell_{i,t}) \leqslant \frac{1}{\eta} \log m_{\tau_i} + \frac{1}{\eta} \log \tau_i + \frac{1}{\eta} \log(1 + \log T)\,. \tag{9.7}$$

This bound has the merit of being simple, virtually independent of T and independent of the number of experts $(m_t)_{t>\tau_i}$ added after i. Several other instantiations of the general regret bound of GrowingHedge are provided in the article Mourtada and Maillard (2017).

Sequences of experts. Another way to study growing expert sets is to view them through the lens of sequences of experts. Given a sequence of experts $i^T = (i_1, \ldots, i_T)$, we measure the performance of a learning algorithm against it in terms of the *cumulative regret*:

$$L_T - L_T(i^T) = \sum_{t=1}^{T} \ell_t - \sum_{t=1}^{T} \ell_{i_t,t}\,, \tag{9.8}$$

In order to derive meaningful regret bounds, some constraints have to be imposed on the comparison sequence; hence, we consider in the sequel different types of comparison classes that lead to different notions of regret, from the least to the most challenging one:

(a) Sequences of fresh experts. These are *admissible* sequences of experts i^T, in the sense that $i_t \leqslant M_t$ for $1 \leqslant t \leqslant T$ (so that $\ell_{i_t,t}$ is always well-defined) that only switch to *fresh* (newly entered) experts, *i.e.* if

[1] This could be readily replaced (up to some cosmetic changes in the statements and their proofs) by the more general η-*mixability* condition (Vovk, 1998), that allows to use higher learning rates η for some loss functions (such as the square loss, but not the logarithmic loss) by using more sophisticated combination functions.

$i_t \neq i_{t-1}$, then $M_{t-1}+1 \leqslant i_t \leqslant M_t$. More precisely, for each $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_k)$ with $1 < \sigma_1 < \cdots < \sigma_k \leqslant T$, $\mathscr{S}_T^{(f)}(\boldsymbol{\sigma})$ denotes the set of sequences of fresh experts whose only shifts occur at times $\sigma_1, \dots, \sigma_k$. Both the switch times $\boldsymbol{\sigma}$ and the number of shifts k are assumed to be unknown, although to obtain controlled regret one typically needs $k \ll T$. Comparing to sequences of fresh experts is essentially equivalent to comparing against constant experts; algorithms GrowingHedge and FreshMarkovHedge with $\pi_i = \frac{1}{m_{\tau_i}}$ achieve, for *every* $T \geqslant 1$, $k \leqslant T-1$ and $\boldsymbol{\sigma} = (\sigma_j)_{1 \leqslant j \leqslant k}$ (details in Mourtada and Maillard (2017)):

$$L_T - \inf_{i^T \in \mathscr{S}_T^{(f)}(\boldsymbol{\sigma})} L_T(i^T) \leqslant \frac{1}{\eta} \left\{ \log m_1 + \sum_{j=1}^{k} (\log m_{\sigma_j} + \log \sigma_j) + \log T \right\}. \tag{9.9}$$

In particular, the regret with respect to any sequence of fresh experts with k shifts is bounded by

$$\frac{1}{\eta} \left((k+1) \log \max_{1 \leqslant t \leqslant T} m_t + (k+1) \log T \right).$$

(b) Arbitrary admissible sequences of experts. Like before, these are admissible sequences of experts that are piecewise constant with a typically small number of shifts k, except that shifts to *incumbent* (previously introduced) experts $i_t \leqslant M_{t-1}$ are now authorized. Specifically, given $\boldsymbol{\sigma}^0 = (\sigma_1^0, \dots, \sigma_{k_0}^0)$ and $\boldsymbol{\sigma}^1 = (\sigma_1^1, \dots, \sigma_{k_1}^1)$, we denote by $\mathscr{S}_T^{(a)}(\boldsymbol{\sigma}^0; \boldsymbol{\sigma}^1)$ the class of admissible sequences whose switches to fresh (resp. incumbent) experts occur only at times $\sigma_1^0 < \cdots < \sigma_{k_0}^0$ (resp. $\sigma_1^1 < \cdots < \sigma_{k_1}^1$). By Theorem 9.1, algorithm GrowingMarkovHedge with $\pi_i = \frac{1}{m_{\tau_i}}$ and $\alpha_t = \frac{1}{t}$ satisfies, for *every* $T \geqslant 1$, k_0, k_1 with $k_0 + k_1 \leqslant T-1$ and $\boldsymbol{\sigma}^0, \boldsymbol{\sigma}^1$:

$$L_T - \inf_{i^T \in \mathscr{S}_T^{(a)}(\boldsymbol{\sigma}^0; \boldsymbol{\sigma}^1)} L_T(i^T) \leqslant \frac{1}{\eta} \left\{ \log m_1 + \sum_{j=1}^{k} (\log m_{\sigma_j} + \log \sigma_j) + \sum_{j=1}^{k_1} \log \sigma_j^1 + 2 \log T \right\} \tag{9.10}$$

where $k = k_0 + k_1$ and $\sigma_1 < \cdots < \sigma_k$ denote *all* shifts (either in $\boldsymbol{\sigma}^0$ or in $\boldsymbol{\sigma}^1$). Note that the upper bound (9.10) may be further relaxed as $\frac{1}{\eta}\left((k+1)\log \max_{1 \leqslant t \leqslant T} m_t + (k_0 + 2k_1 + 2)\log T\right)$.

(c) Sparse sequences of experts. These are admissible sequences i^T of experts that are additionally *sparse*, in the sense that they alternate between a small number $n \ll M_T$ of experts; again, n may be unknown in advance. Denoting $\mathscr{S}_T^{(s)}(\boldsymbol{\sigma}, E)$ the class of sequences with shifts in $\boldsymbol{\sigma}$ and taking values in the subset of experts $E = \{e_1, \dots, e_n\}$, algorithm GrowingSleepingMarkovHedge with $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$ and $\alpha_t = \beta_t = \frac{1}{t}$ achieves, for every $T \geqslant 1$, $E \subset \{1, \dots, M_T\}$ and $\boldsymbol{\sigma}$,

$$L_T - \inf_{i^T \in \mathscr{S}_T^{(s)}(\boldsymbol{\sigma}, E)} L_T(i^T) \leqslant \frac{1}{\eta} \sum_{p=1}^{n} \left(\ln \tau_{e_p} + \ln \frac{m_{\tau_{e_p}}}{n} \right) + \frac{1}{\eta} n \log(2T) + \frac{2}{\eta} \sum_{j=1}^{k} \log \sigma_j. \tag{9.11}$$

In particular, the regret with respect to every admissible sequence of T experts with at most k shifts and taking at most n values is bounded by $\frac{1}{\eta}\left(n \log \frac{\max_{1 \leqslant t \leqslant T} m_t}{n} + 2n \log(\sqrt{2}T) + 2k \log T\right)$.

The main results of this text are Theorem 9.1, a powerful parameter-free generalization of (Shalizi et al., 2011, Theorem 2), and Theorem 9.2, which adapts results of Bousquet and Warmuth (2002), Koolen et al. (2012) to sequentially incoming forecasters, and has no precedent in this context.

Markov prior. If $i^T = (i_1, \dots, i_T)$ is a finite sequence of experts, its predictions up to time T are derived from those of the base experts $i \in \{1, \dots, M\}$ in the following way: $x_t(i^T) = x_{i_t,t}$ for $1 \leqslant t \leqslant T$. Given a prior distribution $\pi = (\pi(i^T))_{i^T}$, we could in principle consider the exponentially weighted aggregation of sequences under this prior; however, such an algorithm would be intractable even for moderately low values of T, since it would require to store and update $O(M^T)$ weights. Fortunately, when $\pi(i_1, \dots, i_T) = \theta_1(i_1)\,\theta_2(i_2 \,|\, i_1) \cdots \theta_T(i_T \,|\, i_{T-1})$ is a Markov probability distribution with initial measure $\boldsymbol{\theta}_1$ and transition matrices $\boldsymbol{\theta}_t$, $2 \leqslant t \leqslant T$, the exponentially weighted aggregation under the prior π collapses to the efficient algorithm MarkovHedge.

Algorithm 7 MarkovHedge — Aggregation of sequences of experts under a Markov prior

1: **Parameters:** Learning rate $\eta > 0$, initial weights $\boldsymbol{\theta}_1 = (\theta_1(i))_{1 \leqslant i \leqslant M}$, and transition probabilities $\boldsymbol{\theta}_t = \big(\theta_t(i \,|\, j)\big)_{1 \leqslant i,j \leqslant M}$ for all $t \geqslant 2$.
2: **Initialization:** Set $\boldsymbol{v}_1 = \boldsymbol{\theta}_1$.
3: **for** $t = 1, 2, \dots$ **do**
4: Receive predictions $\boldsymbol{x}_t \in \mathcal{X}^M$ from the experts, and predict $x_t = \boldsymbol{v}_t \cdot \boldsymbol{x}_t$.
5: Observe $y_t \in \mathcal{Y}$, then derive the losses $\ell_t = \ell(x_t, y_t)$ and $\ell_{i,t} = \ell(x_{i,t}, y_t)$ and the posteriors

$$v_{i,t}^m = \frac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^M v_{j,t}\, e^{-\eta\, \ell_{j,t}}}\,. \tag{9.12}$$

6: Update the weights by $\boldsymbol{v}_{t+1} = \boldsymbol{\theta}_{t+1}\, \boldsymbol{v}_t^m$, *i.e.*

$$v_{i,t+1} = \sum_{j=1}^M \theta_{t+1}(i \,|\, j)\, v_{j,t}^m\,. \tag{9.13}$$

Remark 9.2 *Algorithm MarkovHedge only requires to store and update $O(M)$ weights. Due to the matrix product (9.13), the update may take an $O(M^2)$ time; however, all the transition matrices we consider lead to a simple update in $O(M)$ time.*

3.1 Regret against arbitrary sequences of experts

We now consider the more ambitious objective of comparing against *arbitrary* admissible sequences of experts. This can be done by using another choice of transition matrices, which puts all the weight to admissible sequences of experts (and not just sequences of fresh experts).

Algorithm GrowingMarkovHedge instantiates MarkovHedge on the transition matrices

$$\theta_1(i) = \frac{\pi_i}{\Pi_{M_1}} \mathbf{1}_{i \leqslant M_1} \quad ; \quad \theta_{t+1}(i \,|\, j) = \alpha_{t+1} \frac{\pi_i}{\Pi_{M_{t+1}}} + (1 - \alpha_{t+1})\, \theta_{t+1}^{(f)}(i \,|\, j) \tag{9.14}$$

where $\boldsymbol{\theta}_t^{(f)}$ denotes the transition matrices of algorithm FreshMarkovHedge. As before, this leads to a well-defined growing experts algorithm which predicts $x_t = \sum_{i=1}^{M_t} v_{i,t}\, x_{i,t}$, where the weights $(v_{i,t})_{1 \leqslant i \leqslant M_t}$ are recursively defined by $v_{i,1} = \frac{\pi_i}{\Pi_{M_1}}$ $(1 \leqslant i \leqslant M_1)$ and the update

$$v_{i,t+1} = (1 - \alpha_{t+1}) \frac{\Pi_{M_t}}{\Pi_{M_{t+1}}}\, v_{i,t}^m + \alpha_{t+1} \frac{\pi_i}{\Pi_{M_{t+1}}} \; (1 \leqslant i \leqslant M_t)\,; \; v_{i,t+1} = \frac{\pi_i}{\Pi_{M_{t+1}}} \; (M_t + 1 \leqslant i \leqslant M_{t+1})\,, \tag{9.15}$$

where again $v_{i,t}^m = \dfrac{v_{i,t}\, e^{-\eta\, \ell_{i,t}}}{\sum_{j=1}^{M_t} v_{j,t}\, e^{-\eta\, \ell_{j,t}}}$ for $1 \leqslant i \leqslant M_t$. In this case, (Mourtada and Maillard, 2017, Proposition 5) (Markov Hedge) yields:

Theorem 9.1 (*Growing Markov Hedge regret*) *Algorithm GrowingMarkovHedge based on the weights $\boldsymbol{\pi}$ and parameters $(\alpha_t)_{t\geqslant 2}$ achieves the following regret bound: for every $T \geqslant 1$, and every admissible sequence of experts $i^T = (i_1, \dots, i_T)$ with shifts at times $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_k)$,*

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta} \left\{ \sum_{j=0}^{k} \log \frac{\Pi_{M_{\sigma_{j+1}-1}}}{\pi_{i_{\sigma_j}}} + \sum_{j=1}^{k_1} \log \frac{1}{\alpha_{\sigma_j^1}} + \sum_{2\leqslant t\leqslant T\,:\, t\notin\boldsymbol{\sigma}} \log \frac{1}{1-\alpha_t} \right\} . \tag{9.16}$$

where $\boldsymbol{\sigma}^0 = (\sigma_1^0, \dots, \sigma_{k_0}^0)$ (resp. $\boldsymbol{\sigma}^1 = (\sigma_1^1, \dots, \sigma_{k_1}^1)$) denotes the shifts to fresh (resp. incumbent) experts, with $k = k_0 + k_1$. Moreover, it has an $O(M_t)$ time and space complexity at each step $t \geqslant 1$.

Remark 9.3 *Note that by choosing $\alpha_t = \frac{1}{t}$, we have, since $\frac{1}{1-1/t} = \frac{t}{t-1}$,*

$$\sum_{j=1}^{k_1} \log \frac{1}{\alpha_{\sigma_j^1}} + \sum_{2\leqslant t\leqslant T\,:\, t\notin\boldsymbol{\sigma}} \log \frac{1}{1-\alpha_t} \leqslant \sum_{j=1}^{k_1} \log \sigma_j^1 + \sum_{t=2}^{T} \log \frac{t}{t-1} = \sum_{j=1}^{k_1} \log \sigma_j^1 + \log T \,.$$

Additionally, by setting $\pi_i = 1$ the bound (9.16) becomes $\frac{1}{\eta}(\sum_{j=0}^{k} \log M_{\sigma_{j+1}-1} + \sum_{j=1}^{k_1} \log \sigma_j^1 + \log T)$, which is lower than $\frac{1}{\eta}(k+1)\log M_T + \frac{1}{\eta}(k_1+1)\log T$. We can also recover the bound (9.10) by setting $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$, since in this case we have $\Pi_{M_{\sigma_{j+1}-1}} \leqslant \Pi_{M_T} \leqslant \sum_{t=1}^{T} \frac{1}{t} \leqslant 1 + \log T$.

Algorithm 8 SleepingMarkovHedge: sequences of sleeping experts under a Markov chain prior

1: **Parameters:** Learning rate $\eta > 0$, (normalized) prior $\boldsymbol{\pi}$ on the experts, initial wake/sleep probabilities $\theta_{i,1}(a)$, transition probabilities $\boldsymbol{\theta}_{i,t} = \big(\theta_{i,t}(a \,|\, b)\big)_{a,b\in\{0,1\}}$ for $t \geqslant 2$, $1 \leqslant i \leqslant M$.

2: **Initialization:** Set $v_1(i,a) = \pi_i\, \theta_{i,1}(a)$ for $i = 1, \dots, M$ and $a \in \{0,1\}$.

3: **for** $t = 1, 2, \dots$ **do**

4: Receive predictions $\boldsymbol{x}_t \in \mathcal{X}^M$ from the experts, and predict

$$x_t = \frac{\sum_{i=1}^{M} v_t(i,1)\, x_{i,t}}{\sum_{i=1}^{M} v_t(i,1)} \,. \tag{9.17}$$

5: Observe $y_t \in \mathcal{Y}$, then derive the losses $\ell_t(i,0) = \ell_t = \ell(x_t, y_t)$, $\ell_t(i,1) = \ell_{i,t} = \ell(x_{i,t}, y_t)$ and the posteriors

$$v_t^m(i,a) = \frac{v_t(i,a)\, e^{-\eta\, \ell_t(i,a)}}{\sum_{i',a'} v_t(i',a')\, e^{-\eta\, \ell_t(i',a')}} \,. \tag{9.18}$$

6: Update the weights by

$$v_{t+1}(i,a) = \sum_{b\in\{0,1\}} \theta_{i,t+1}(a \,|\, b)\, v_t^m(i,b) \,. \tag{9.19}$$

3.2 Sparse shifting regret for growing experts

We show here how to instantiate algorithm SleepingMarkovHedge in order to adapt it to the growing experts setting. Again, we use a "muting trick" which attributes a zero weight to experts that have not entered.

Let us consider prior weights $\boldsymbol{\pi} = (\pi_i)_{i\geqslant 1}$ on the experts, which may be unnormalized and chosen at entry time. Let $\alpha_t, \beta_t \in (0,1)$ for $t \geqslant 2$. We set $\theta_{i,1}(1) = \frac{1}{2}$ for $i = 1, \dots, M_1$ and 0 otherwise; moreover, for every $t \geqslant 1$, we take $\theta_{i,t+1}(1\,|\,\cdot) = 0$ for $i > M_{t+1}$ (recall that $\boldsymbol{\theta}_{i,t+1}$ can be chosen at step $t+1$), $\theta_{i,t+1}(1\,|\,\cdot) = \frac{1}{2}$ if $M_t + 1 \leqslant i \leqslant M_{t+1}$, and for $i \leqslant M_t$: $\theta_{i,t+1}(0\,|\,1) = \alpha_{t+1}$, $\theta_{i,t+1}(1\,|\,0) = \beta_{t+1}$. The algorithm obtained with these choices, which we call GrowingSleepingMarkovHedge, is well-defined and predicts $x_t = (\sum_{i=1}^{M_t} v_t(i,1)\, x_{i,t})/(\sum_{i=1}^{M_t} v_t(i,1))$, where the weights $(v_t(i,a))_{1\leqslant i\leqslant M_t,\, a\in\{0,1\}}$ are defined by $v_1(i,a) = \frac{1}{2}\,\pi_i$ $(1 \leqslant i \leqslant M_1)$ and by the update

$$v_{t+1}(i,a) = \sum_{b\in\{0,1\}} \theta_{i,t+1}(a\,|\,b)\, v_t^m(i,b)\ (1 \leqslant i \leqslant M_t)\,;\ v_{t+1}(i,a) = \frac{1}{2}\,\pi_i\ (M_t + 1 \leqslant i \leqslant M_{t+1})\,,$$

with $v_t^m(i,a) = v_t(i,a)\, e^{-\eta\,\ell_t(i,a)} / \sum_{i=1}^{M_t} \sum_{a'\in\{0,1\}} v_t(i',a')\, e^{-\eta\,\ell_t(i',a')}$ for $1 \leqslant i \leqslant M_t$.

Theorem 9.2 (*Sparse shifting regret*) *Algorithm GrowingSleepingMarkovHedge guarantees the following: for each $T \geqslant 1$ and any sequence i^T of experts taking values in the pool $\{e_p \mid 1 \leqslant p \leqslant n\}$, denoting $a_{p,t} = \mathbf{1}_{i_t = e_p}$,*

$$L_T - L_T(i^T) \leqslant \frac{1}{\eta}\sum_{p=1}^{n} \ln\frac{\Pi_{M_T}/n}{\pi_{e_p}} + \frac{1}{\eta} n \log 2 + \frac{1}{\eta}\sum_{t=2}^{T}\left[\log\frac{1}{1-\alpha_t} + (n-1)\log\frac{1}{1-\beta_t}\right] + \frac{1}{\eta}\sum_{j=1}^{k}\left(\log\frac{1}{\alpha_{\sigma_j}} + \ln\frac{1}{\beta_{\sigma_j}}\right), \tag{9.20}$$

where $\boldsymbol{\sigma} = \sigma_1 < \cdots < \sigma_k$ denote the shifting times of i^T. Moreover, the algorithm has an $O(M_t)$ time and space complexity at step t, for every $t \geqslant 1$.

In particular, Theorem 9.2 enables to recover the bound (9.11) for $\alpha_t = \beta_t = \frac{1}{t}$ and $\pi_i = \frac{1}{\tau_i m_{\tau_i}}$.

Part III

Thoughts and perspective

Chapter 9

In this part, we now highlight a few thoughts and promising research perspective regarding the many things that remain to be understood in statistical sequential learning.

Confidence sets, Mismatch and Prediction

Let us come back to our observations $Y_{1:n} = Y^{(1)}, Y^{(2)}, \dots, Y^{(n)}$, and consider that they have been generated by a process $\rho \in \mathcal{P}(\mathcal{Y}^\star)$, with $\mathcal{Y} \subset \mathbb{R}$. We may further assume that ρ belongs to some family $\mathcal{P}_m$, and we have seen in Chapter 4 for some specific families how to build confidence sets for $\rho \in \mathcal{P}_m$ that are time-uniform and satisfy

$$\forall \delta \in [0,1], \quad \mathbb{P}_\rho\Big(\exists n \in \mathbb{N}, \rho \notin \widehat{\mathcal{P}}_m(Y^{(1)}, \dots, Y^{(n)}; \delta) \Big) \leqslant \delta \,.$$

It turns out that it is often possible, with the same tools, to derive a confidence set on the next observation $Y^{(n+1)}$. This leads to the construction of two real-valued functions $\overline{y}_m, \underline{y}_m$ such that for each $\rho \in \mathcal{P}_m$, $\delta \in [0,1]$,

$$\mathbb{P}_\rho\Big(\exists n \in \mathbb{N}, Y^{(n+1)} \geqslant \overline{y}_m(Y^{(1)}, \dots, Y^{(n)}; \delta) \Big) \leqslant \delta, \quad \mathbb{P}_\rho\Big(\exists n \in \mathbb{N}, Y^{(n+1)} \leqslant \underline{y}_m(Y^{(1)}, \dots, Y^{(n)}; \delta) \Big) \leqslant \delta \,,$$

where $\overline{y}_m$ (resp. $\underline{y}_m$) is a decreasing (resp. increasing) function of δ, with limit $+\infty$ (resp. $-\infty$) as $\delta \to 0$. Such confidence functions can be used in turn for some tasks such as mismatch detection as well as predictive loss estimation. We detail below for illustration the notion of adequacy.

Definition 9.3 (*Adequacy function*) *The adequacy function corresponding to the confidence functions $\overline{y}_m, \underline{y}_m$, is defined for any sequence $y_{1:n+1} = y^{(1)}, y^{(2)} \dots$ by* $\alpha_m(y_{1:n+1}) = \min_{n' \leqslant n} \min\{\overline{\delta}_m(y_{1:n'+1}), \underline{\delta}_m(y_{1:n'+1})\}$, *where*

$\overline{\delta}_m(y_{1:n+1}) = \inf\{\delta \in [0,1] \colon y^{(n+1)} \geqslant \overline{y}_m(y_{1:n}, \delta)\}$ *and* $\underline{\delta}_m(y_{1:n+1}) = \inf\{\delta \in [0,1] \colon y^{(n+1)} \leqslant \overline{y}_m(y_{1:n}, \delta)\}$.

This quantity is guaranteed not to be too small since it satisfies by construction (and a union bound)

$$\forall \rho \in \mathcal{P}_m, \forall \delta \in [0,1], \forall n \in \mathbb{N}, \quad \mathbb{P}\Big(\alpha_m(Y^{(1)}, \dots, Y^{(n+1)}) \leqslant \delta \Big) \leqslant 2\delta \,.$$

Hence this is a natural candidate to provide a score of adequacy of the sequence of observations with respect to a family of models. For illustration, we quickly compute this adequacy on two examples, first on Gaussian distributions with $\overline{y}_G, \underline{y}_G$ functions, then on bounded observations in $(0,1)$ with functions $\overline{y}_{(0,1)}, \underline{y}_{(0,1)}$.

Example 1: Iid Gaussian observations We first introduce the estimate $\mu_n = \frac{1}{n}\sum_{i=1}^n Y^{(i)}$ and then

$$\begin{aligned}
\mathbb{P}\Big(\exists n \in \mathbb{N}, Y^{(n+1)} > \mu_n + \sigma\sqrt{2\Big(1+\frac{1}{n}\Big)\log(2\sqrt{n+1}/\delta)}\Big(1+\frac{1}{\sqrt{n}}\Big)\Big) &\leqslant \delta \,, \\
\mathbb{P}\Big(\exists n \in \mathbb{N}, Y^{(n+1)} < \mu_n - \sigma\sqrt{2\Big(1+\frac{1}{n}\Big)\log(2\sqrt{n+1}/\delta)}\Big(1+\frac{1}{\sqrt{n}}\Big)\Big) &\leqslant \delta \,.
\end{aligned}$$

Proof :

Part III

Indeed, for all random stopping time τ, it holds on the one hand

$$\mathbb{P}\left(\mu - \mu_\tau \geqslant \sigma\sqrt{\frac{2(1+\frac{1}{\tau})}{\tau}\log(\sqrt{\tau+1}/\delta)}\right) \leqslant \delta$$

and on the other hand

$$\mathbb{P}\left(Y^{(1+\tau)} - \mu \geqslant \sigma\sqrt{2(1+\frac{1}{\tau})\log(\sqrt{\tau+1}/\delta)}\right) \leqslant \delta\,.$$

We then choose τ to be the first time such that both events hold, and use a union bound argument. □

We deduce that

$$\begin{aligned}
\overline{\delta}_G(Y_{1:n+1}) &= \min\left\{2\sqrt{n+1}\exp\left(-\frac{\left(Y^{(n+1)}-\mu_n\right)^2}{2\sigma^2(1+1/n)(1+1/\sqrt{n})^2}\right)\mathbb{I}\{Y^{(n+1)} \geqslant \mu_n\}, 1\right\},\\
\underline{\delta}_G(Y_{1:n+1}) &= \min\left\{2\sqrt{n+1}\exp\left(-\frac{\left(Y^{(n+1)}-\mu_n\right)^2}{2\sigma^2(1+1/n)(1+1/\sqrt{n})^2}\right)\mathbb{I}\{Y^{(n+1)} \leqslant \mu_n\}, 1\right\}.
\end{aligned}$$

The resulting adequacy function has a Gaussian shape. Similar computations can be extended to processes following a kernel regression model, in which case the adequacy has a flattened Gaussian shape with a plateau having value 1. This corresponds to the range of observations whose parameters are indistinguishable due to the finiteness of the number of observations.

Example 2: Adversarial bounded observations Let us consider the following envelopes

$$\mathbb{P}\left(\exists n \in \mathbb{N}, Y^{(n+1)} > \mathbb{I}\{\delta < 1\}\right) \leqslant \delta\,, \quad \mathbb{P}\left(\exists n \in \mathbb{N}, Y^{(n+1)} < \mathbb{I}\{\delta = 1\}\right) \leqslant \delta.$$

We deduce that $\alpha_{(0,1)}(Y^{(1)},\ldots,Y^{(n+1)}) = \mathbb{I}\{\forall n' \leqslant n, Y^{(n'+1)} \in (0,1)\}$, hence the observations are always perfectly adequate provided that they stay in $(0,1)$, which is intuitive for this setup.

Loss A second way to use the confidence bounds is to derive an estimation of the loss of a decision maker. Indeed, for a decision $Z_{n+1} = Z(Y^{(1)},\ldots,Y^{(n)})$, and a loss $\ell(Y^{(n+1)}, Z_{n+1})$, $Y^{(n+1)}$ is generally unknown. However, since the loss is non-negative, we have that

$$\mathbb{E}[\ell(Y^{(n+1)}, Z_{n+1})|Y^{(1)},\ldots,Y^{(n)}] = \int_{\mathbb{R}^+} \mathbb{P}(\ell(Y^{(n+1)}, Z_{n+1}) \geqslant x|Y^{(1)},\ldots,Y^{(n)})dx\,.$$

This means that using the confidence functions in order to compute for each $x \in \mathbb{R}^+$ an upper bound on the conditional probability $\mathbb{P}(\ell(Y^{(n+1)}, Z_{n+1}) \geqslant x|Y^{(1)},\ldots,Y^{(n)})$ yields an upper bound on its expected loss.

We believe such usage of the confidence sets should be better explored, whether for prediction, model selection, model aggregation or other machine learning tasks.

Parameter estimation Yet another way to use the confidence bounds is to perform an "optimistic" parameter tuning. For instance, let us consider the case of iid Gaussian observations but with unknown variance σ^2. One way is to build an estimate together with a confidence bound for σ^2, and combining it with the confidence bounds on the mean. Alternatively, using the confidence bounds, it is possible to define at time n the smallest value of σ that is compatible with the confidence bounds. For instance, for a given $\delta \in (0,1)$ and time $t \in \mathbb{N}$,

$$\sigma_t^2(\delta) = \max_{n<t} \left(\frac{(Y^{(n+1)} - \mu_n)^2}{(1+1/\sqrt{n})^2} 2\big(1+\frac{1}{n}\big) \log(2\sqrt{n+1}/2\delta) \right)$$

is a valid lower bound on σ^2, since by construction $\mathbb{P}(\exists t \in \mathbb{N}, \sigma_t^2(\delta) \geqslant \sigma^2) \leqslant \delta$, that is increasing with t. A partially open question (see Leadbetter et al. (2012)) is to understand how much smaller than σ^2 it can be, which would allow to build a confidence bound on the variance and combine it with the mean confidence estimates with known variance in order to produce bounds without this knowledge.

Doubly uniformly optimal strategies for multi-armed bandits

Despite many decades of research of the stochastic multi-armed bandit setup, a non trivial question remains open. Indeed, for each given set of bandit configurations $\mathcal{D}$, one is able to build a strategy inspired from the lower bounds. In some cases we are able to prove optimality (such as for unstructured configurations coming from an exponential family, see Maillard (2018), or some specific structures, see Magureanu (2018)); in general, a complete answer to this question is still open. However, even beyond this question, if we now have different sets $\mathcal{D}_1, \mathcal{D}_2, \dots \mathcal{D}_M$ and $\mathcal{D} = \bigcup_m \mathcal{D}_m$, a tricky question naturally appears: Is there a hope to build a strategy that is (near) uniformly optimal simultaneously on all these sets (and not only on $\mathcal{D}$)? Of course in general being uniformly optimal on a specific $\mathcal{D}_m$ prevents us from being uniformly optimal on $\mathcal{D}$ and vice-versa. Hence there is a price to pay, but when the price is small enough there is hope to build such powerful strategies. Such questions open the quest for "doubly" uniformly-good strategies (the terminology is used in reference to the "doubly" universal codes). A good candidate for a doubly uniformly-good strategy over the sets of bandit configurations coming from exponential families of dimension 1 (Bernoulli, Exponential, Poisson, standard Gaussians, etc.) seems to be the BESA algorithm that we introduced in Baransi et al. (2014). Indeed, this strategy that is based on sub-sampling techniques (see Bardenet et al. (2015)) has been shown to be uniformly competitive with each `KL-ucb` and Thompson sampling strategies designed for a specific family. Even though we could provide a simple regret bound for this strategy, a full analysis showing it is indeed near uniformly optimal jointly on all such families is still open.

Regret minimization in MDPs is far from being understood

Likewise, despite many decades or research in MDPs, it seems that very little is actually understood on the fundamental learning challenges of MDPs. For instance, no existing informative lower bound currently seems to capture the 'navigation' challenge of an MDP in full generality, that is the price to pay when learning in a single stream of interactions (hence we need to handle the cost of navigating to a desired state from the current state); there exists lower bounds based on a change of measure (Burnetas and Katehakis (1997), Graves and Lai (1997)), but they are unfortunately weak and restricted to specific MDPs such as ergodic MDPs (for which navigation is not an issue, since all policies eventually visit all states infinitely often). Now, some specific MDP construction have been provided for minimax regret bounds (Auer et al. (2009)), but these are unfortunately not informative to build a strategy against a generic MDP. Further, recent analysis of average-reward minimization strategies suffer from flaws (see Fruit et al. (2018)) that seem hard to correct and make us go back to studying

the basic questions first. Hence, one can say a bit bluntly that today, nobody knows how to solve correctly even a two-state MDP in the average-reward setup from a statistical standpoint (that is, not knowing the process generating the observations). In all cases, this opens exciting research questions.

No-evidence learning, and novelty estimation

The concepts of states and transition make us go beyond the current research trend. In particular, while most of the literature focuses on states that are reachable with high probability or visited often and estimation of "likely" transitions, it seems interesting to take a look at what happens for the states that are especially difficult to reach, not observed much as well as for the transitions that have near zero probability. Indeed, we sometimes remark that human beings can perform complex tasks (e.g. walking) even in an environment that has never been observed before simply because they assume most of the dynamics will remain the same. But we want to point out that this is perhaps not as much the result of an assumption as of an estimation: for instance we may have observed in the past that whenever there is a new environment, performing this action (walking) as in a previously known environment matched our predictions very well. Hence and focusing on what happens when visiting or discovering a state or transition *for the first time* gives information on estimating the dynamics of the system when seeing a new task; that is the frequency of novelty. This shift of paradigm could be given a name, such as "Machine Imagination". It consists in considering states that do not exist (say, by interpolating or composing states from existing ones) setting goals to these states, and guessing the would-be dynamics, thanks to a precise monitoring of transitions having zero empirical probability, observations with zero evidence, and what happens when transiting to a new observation for the first time. What if I take this action in this state ? what will be the output (given we have no evidence, or contradictory evidence) ? Answering such questions can help an agent better understand how to handle a novel situation, and thus behave more autonomously, expanding its own state space and dynamics beyond what has been observed, and setting its own reward functions. Note also that the ability to consider situations that do not exist is often considered as one of the reason for the emergence of consciousness.

Bibliography

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011. 32, 66, 68, 69

Rajeev Agrawal. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(04):1054–1078, 1995. 118

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space. *IEEE Transactions on Automatic Control*, 34(3), 1989. 46

Ankit Anand, Aditya Grover, Parag Singla, et al. Asap-uct: Abstraction of state-action pairs in uct. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 114

J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. 83, 118

J.Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2635–2686, 2010. 118

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. 32, 83, 118

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003. ISSN 0097-5397. 83

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 22 (NIPS)*, pages 89–96, 2009. 105, 151

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. In *29th Annual Conference on Learning Theory*, pages 193–256, 2016. 133

Borja Balle. *Learning Finite-State Machines: Algorithmic and Statistical Aspects*. PhD thesis, Universitat Politècnica de Catalunya, 2013. 135

Borja Balle and Odalric-Ambrym Maillard. Spectral learning from a single trajectory under finite-state policies. In *International Conference on Machine Learning*, pages 361–370, 2017. xi, 73, 131

Borja Balle, Xavier Carreras, Franco M Luque, and Ariadna Quattoni. Spectral learning of weighted automata. *Machine learning*, 96(1-2):33–63, 2014. 134

Borja Balle, Pascale Gourdeau, and Prakash Panangaden. Bisimulation metrics for weighted automata. In *44rd International Colloquium on Automata, Languages, and Programming, ICALP*, 2017. 131, 136

Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 115–131. Springer, 2014. 151

Rémi Bardenet, Odalric-Ambrym Maillard, et al. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. 151

Peter L. Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the 25th conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8. 108

Jay Bartroff, Tze Leung Lai, and Mei-Chiung Shih. *Sequential experimentation in clinical trials: design and analysis*, volume 298. Springer Science & Business Media, 2012. xix

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7, 2013. 12

Abraham Berman and Robert J Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994. 133

Rudolf B Blazek, Hongjoong Kim, Boris Rozovskii, and Alexander Tartakovsky. A novel approach to detection of denial-of-service attacks via adaptive sequential and batch-sequential change-point detection methods. In *Proceedings of IEEE systems, man and cybernetics information assurance workshop*, pages 220–226, 2001. 53

J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problem. *SIAM Journal on Computation and Optimization*, 29(2):325–338, 1991. 86

Olivier Bousquet and Manfred K. Warmuth. Tracking a small set of experts by mixing past posteriors. *The Journal of Machine Learning Research*, 3:363–396, 2002. 140, 143

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967. 18

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. 118

A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996. 85

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997. 98, 105, 118, 151

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 2013. 85

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013. 51, 117, 118, 119, 120, 121, 122, 124

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, New York, USA, 2006. 15, 16, 76, 77, 139, 140, 141, 142, 162

J-R Chazottes, Pierre Collet, Christof Külske, and Frank Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137(1-2):201–225, 2007. 132

Alexey Chernov and Vladimir Vovk. Prediction with expert evaluators' advice. In *Proceedings of the 20th international conference on Algorithmic learning theory*, ALT '09, pages 8–22, Berlin, Heidelberg, 2009. Springer-Verlag. 140

YS Chow and H Teicher. Probability theory. 2nd. *Springer-Verlag*, 1:988, 1988. 50

T. M Cover and J. A. Thomas. Elements of information theory. 1991. 67

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008. 9

Z Daróczy. On the measurable solutions of a functional equation. *Acta Mathematica Academiae Scientiarum Hungarica*, 22(1-2):11–14, 1971. 9

Steven de Rooij, Tim van Erven, Peter Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014. 141

François Denis, Mattias Gybels, and Amaury Habrard. Dimension-free concentration bounds on hankel matrices for spectral learning. *Journal of Machine Learning Research*, 17(31):1–32, 2016. 135

François Denis and Yann Esposito. On rational stochastic languages. *Fundamenta Informaticae*, 86(1, 2): 41–77, 2008. 132

François Denis, Mattias Gybels, and Amaury Habrard. Dimension-free concentration bounds on hankel matrices for spectral learning. In *ICML*, pages 449–457, 2014. 137, 162

Ian H Dinwoodie. Mesures dominantes et théoreme de sanov. In *Annales de l'IHP Probabilités et statistiques*, volume 28, pages 365–373, 1992. 121

Allen B Downey. A novel changepoint detection algorithm. *arXiv preprint arXiv:0812.1237*, 2008. 54

Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *arXiv preprint arXiv:1708.00768*, 2017. 70

S. Filippi. *Optimistic strategies in Reinforcement Learning* (in French). PhD thesis, Telecom ParisTech, 2010. URL `http://tel.archives-ouvertes.fr/tel-00551401/`. 108

M. Fliess. Matrices de Hankel. *Journal de Mathématiques Pures et Appliquées*, 53, 1974. 134

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC)*, pages 334–343, 1997. 140

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. *arXiv preprint arXiv:1807.02373*, 2018. 151

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011. 85

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *arXiv preprint arXiv:1602.07182*, 2016. 43, 118

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *arXiv preprint arXiv:1612.04740*, 2016. 60

J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2):148–177, 1979. 117

Eyal Gofer, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for branching experts. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 618–638, 2013. 141

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws incontrolled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997. 46, 105, 151

László Györfi, Gábor Lugosi, and Gustáv Morvai. A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999. 139, 140, 141

András Gyorgy, Tamás Linder, and Gábor Lugosi. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11):6709–6725, 2012. 141

Hugo Harari-Kermadec. *Vraisemblance empirique généralisée et estimation semi-paramétrique*. PhD thesis, Université Paris–Ouest, December 2006. 86

David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998. 139

Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, ICML '09, pages 393–400, 2009. 141

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, August 1998. 140

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. 125

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Haifa, Israel, 2010. 86, 118, 120

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. 135

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6): 1371–1398, 2000. 132, 133

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010. 108

Emilie Kaufmann. *Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources*. PhD thesis, Paris, ENST, 2014. 43

Michael Kearns and Lawrence Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 311–319. Morgan Kaufmann Publishers Inc., 1998. 12

Aryeh Kontorovich and Roi Weiss. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *Journal of Applied Probability*, 51(04):1100–1113, 2014. 73, 133

Leonid Aryeh Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008. 132

Wouter M. Koolen and Steven de Rooij. Universal codes from switching strategies. *IEEE Transactions on Information Theory*, 59(11):7168–7185, November 2013. 140

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1155–75, 2015. 141

Wouter M. Koolen, Dmitry Adamskiy, and Manfred K. Warmuth. Putting bayes to sleep. In *Advances in Neural Information Processing Systems 25*, pages 135–143. Curran Associates, Inc., 2012. 17, 140, 143

Wouter M. Koolen, Tim van Erven, and Peter Grünwald. Learning the learning rate for prediction with expert advice. In *Advances in Neural Information Processing Systems 27*, pages 2294–2302. Curran Associates, Inc., 2014. 141

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985a. 40, 85, 118

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987. 51, 118

Tze Leung Lai. Boundary crossing problems for sample means. *The Annals of Probability*, pages 375–396, 1988. 123, 124, 126

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985b. 39

Tze Leung Lai and Haipeng Xing. Sequential change-point detection when the pre-and post-change parameters are unknown. *Sequential analysis*, 29(2):162–175, 2010. 53, 60

Malcolm R Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012. 151

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adaptive normalhedge. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1286–1304, 2015. 141

Aleksandr Sergeevich Lysyak and B Ya Ryabko. Time series prediction based on data compression methods. *Problems of Information Transmission*, 52(1):92–99, 2016. 76

Stefan Magureanu. *Efficient Online Learning under Bandit Feedback*. PhD thesis, KTH Royal Institute of Technology, 2018. 52, 151

Stefan Magureanu, Richard Combes, and Alexandre Proutière. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *COLT 2014*, 2014. 125

O-A. Maillard. Robust risk-averse stochastic multi-armed bandits. Technical Report http://hal.inria.fr/hal-00821670, 2013a. URL http://hal.inria.fr/hal-00821670. HAL-INRIA open archive. 85

O.-A. Maillard. Self-normalization techniques for streaming confident regression. working paper or preprint, May 2016. URL https://hal.archives-ouvertes.fr/hal-01349727. 64, 70, 71

O-A Maillard. Boundary crossing probabilities for general exponential families. *Mathematical Methods of Statistics*, 27(1):1–31, 2018. xi, 52, 117, 119, 122, 124, 126, 151

O-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Budapest, Hungary, 2011. 85, 118, 121

Odalric-Ambrym Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013b. 83, 84

Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. How hard is my MDP? "the distribution-norm to the rescue". In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 1835–1843, 2014. 105, 106, 107

Scott McQuade and Claire Monteleoni. Global climate model tracking using geospatial neighborhoods. In *AAAI*, 2012. 140

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998. 14, 140

Claire Monteleoni, Gavin A Schmidt, Shailesh Saroha, and Eva Asplund. Tracking climate models. *Statistical Analysis and Data Mining*, 4(4):372–392, 2011. 140

Jaouad Mourtada and Odalric-Ambrym Maillard. Efficient tracking of a growing number of experts. In *International Conference on Algorithmic Learning Theory*, pages 517–539, 2017. 139, 142, 143, 144

Ronald Ortner. Adaptive aggregation for reinforcement learning in average reward markov decision processes. *Annals of Operations Research*, 208(1):321–336, 2013. 114

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. 53

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2008. 30

Martin L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994. xvi, 4, 98, 102, 103, 108

Maxim Raginsky, Igal Sason, et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013. 12

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. xiv, 117

Herbert Robbins. *Herbert Robbins Selected Papers*. Springer, 2012. xiv, 117

R. Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. *Tutorials in operation Research*, pages 38–61, 2007. 81

Boris Y. Ryabko. Twice-universal coding. *Problems of information transmission*, 20(3):173–177, 1984. 140

Boris Y. Ryabko. Prediction of random sequences and universal coding. *Problems of information transmission*, 24(2):87–96, 1988. 76, 140

A. Salomon and J.-Y. Audibert. Robustness of stochastic bandit policies. *Theoretical Computer Science, Special issue*, 2012. 83

Cosma Rohilla Shalizi, Abigail Z Jacobs, Kristina Lisa Klinkner, and Aaron Clauset. Adapting to non-stationarity with growing expert ensembles. *arXiv preprint arXiv:1103.0949*, 2011. 141, 143

Gil Shamir and Neri Merhav. Low-complexity sequential lossless coding for piecewise-stationary memoryless sources. *IEEE transactions on information theory*, 45(5):1498–1519, 1999. 140

John Skilling. Failures of information geometry. In *AIP Conference Proceedings*, volume 1641, pages 27–42. AIP, 2015. ix

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, 2010. 67

Gilles Stoltz. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, 151(2):66–106, 2010. 140

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805, 2018. 105, 106, 108

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. xiv, xix, 117

William R Thompson. On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *The Annals of Mathematical Statistics*, 6(4):214–219, 1935. xiv, 117

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *Conference on Uncertainty In Artificial Intelligence (UAI)*, pages 654–665, 2013. 67

Vladimir Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2): 153–173, 1998. 15, 139, 140, 141, 142

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945. xiv, 41, 117

Z. Wang and N. de Freitas. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. *arXiv preprint arXiv:1406.7758*, 2014. 66

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Technical Report HPL-2003-97*, 11, 2003. URL `www.hpl.hp.com/techreports/2003/HPL-2003-97R1.pdf`. 72, 107

Frans M. J. Willems. Coding for a binary independent piecewise-identically-distributed source. *IEEE transactions on information theory*, 42(6):2210–2217, 1996. 140

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106(1):119–141, 2017. 141

Index

Corollaries
- $\mathcal{R}$-piecewise process prediction loss, 89
- Aggregation of fresh experts, 87
- Bias and Gain, 96
- Consistent KT estimates, 75
- Sub-Gaussian Concentration Inequality, 11
- Transportation and KL, 107
- Value and bias, 99
- Value and gain, 98
- Weissman-Laplace concentration, 70

Definitions
- (η, ψ)-mixable loss, 84
- $(\mathcal{D}_0, \mathcal{D}_1)$-uniformly-$\delta$-correct detection strategy, 48
- $\mathcal{R}$-Piecewise process, 89
- Adequacy function, 147
- Asymptotic price for uniformly-good strategies, 47
- Average gain and bias of proper policies, 95
- Consistent forecaster, 73
- Empirical distributions, 50
- Expected cumulative reward, 95
- Expected regret, 116
- Expert, 73
- Exponential families, 17
- Hardness of discounted MDP, 106
- Hardness of undiscounted MDP, 106
- Information gain with unknown variance, 65
- Loss-adapted noise, 14
- Markov KT forecasters, 74
- Mixable loss, 15
- Optimal gain and policy, 95
- Probabilistic WFA, 130
- Self-information loss, 14, 73
- Similar state-action pairs, 108
- Stationary process, 73
- Stochastic WFA, 130
- Uniformly-good change-point detection strategies, 54
- Uniformly-good strategy for bandits, 40
- Uniformly-good strategy for separation, 48
- Universal KT forecaster, 74
- Value iteration, 98
- Weighted Finite Automaton, 129

Lemmas
- η-mixing for PFA, 131
- η-mixing for SWFA, 132
- $\mathcal{D}$-constrained regret lower bound, 46
- Aggregation of fresh experts, 87
- Alternative Cramer-Chernoff, 10
- Asymptotic Maximal Hoeffding inequality, 22
- Birge-Massart concentration for predictable process, 29
- Bregman duality, 18
- Bregman duality aggregation, 85
- Cesaro consistency, 133
- Change of measure, 39, 125, 126
- Chernoff's rule, 7
- Concentration inequality for predictable processes, 24
- Concentration of measure, 127
- Cone-norm properties, 131
- Cramer-Chernoff, 9
- Detection lower bounds, 49
- Dimension 1, 122
- Doob's maximal inequality for non-negative submartingale, 21
- Doob's maximal inequality for non-negative supermartingale, 22
- Empirical distributions, 50
- From Regret to Boundary Crossing Probabilities, 118
- Fundamental change of measure inequality, 42
- Hilbert Martingale Control, 65
- Hoeffding concentration for Markov processes, 71
- Key property, 95
- Laplace concentration for Exponential families, 33
- Loss-adapted noise, 13

Multi-steps sandwich bounds, 101
Multi-steps value contraction of the Span, 102
Non-asymptotic Sanov's lemma, 119
Non-negative random variables, 7
Peeling and cone covering decomposition, 125
Performance of KT forecasters, 75
Policy contraction of the Span, 103
Price for uniform optimality, 45
Pseudo-regret, 97
Regret lower bound for uniformly good strategies, 40, 43
Rewriting, 118
Risk-averse dual formulation, 84
Sandwich bounds, 100
Structural properties, 15
Tails and log-Laplace, 8
Time-uniform concentration inequalities, 30
Time-uniform joint concentration, 55
Time-uniform sub-Gaussian concentration, 32
Transportation Lemma, 107
Value contraction of the Span, 101

Propositions
Cesa-Bianchi and Lugosi (2006, Corollary 3.1) , 138
Regret of aggregation, 16
Regret of value iteration policy, 99

Theorems
Detection delay, 57
Doubly-time-uniform concentration, 57
Growing Markov Hedge regret, 143
Lai, 88, 121
Risk-averse regret decomposition, 82
Single-trajectory, entry-wise, 134
Single-trajectory, matrix-wise, 135
Single-trajectory, SWFA, 136
Sparse shifting regret, 144
Streaming kernel least-squares, 64
Theorem 7 in Denis et al. (2014), 135

Odalric-Ambrym Maillard, February 2019, Université de Lille, France.