Journal of Machine Learning Research 22 (2021) 1-44 Submitted 11/18; Revised 9/21; Published 11/21

Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals

Emilie Kaufmann
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France EMILIE.KAUFMANN@UNIV-LILLE.FR

Wouter M. Koolen
Centrum Wiskunde & Informatica, Science Park 123, Amsterdam, Netherlands WMKOOLEN@CWI.NL

Editor: Csaba Szepesvári

Abstract

This paper presents new deviation inequalities that are valid uniformly in time under adaptive sampling in a multi-armed bandit model. The deviations are measured using the Kullback-Leibler divergence in a given one-dimensional exponential family, and take into account *multiple arms* at a time. They are obtained by constructing for each arm a mixture martingale based on a hierarchical prior, and by multiplying those martingales. Our deviation inequalities allow us to analyze stopping rules based on generalized likelihood ratios for a large class of sequential identification problems. We establish asymptotic optimality of sequential tests generalising the track-and-stop method to problems beyond best arm identification. We further derive sharper stopping thresholds, where the number of arms is replaced by the newly introduced pure exploration problem rank. We construct tight confidence intervals for linear functions and minima/maxima of the vector of arm means.

Keywords: mixture methods, test martingales, multi-armed bandits, best arm identification, adaptive sequential testing

1. Introduction

We are interested in making decisions under uncertainty in its myriad forms, including sequential allocation and hypothesis testing problems. In this paper our goal is the design of tight confidence regions that are valid uniformly in time, as well as the design of efficient stopping rules for a large class of sequential tests.

We will develop our results in the standard multi-armed bandit model with K independent one-dimensional exponential family *arms* that are parameterised by their means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. In this setup, samples $X_1, X_2 \dots$ are sequentially gathered from the different arms: X_t is drawn from the distribution that has mean μ_{A_t} where $A_t \in \{1, \dots, K\}$ is the arm selected at round t. Our techniques all make use of *self-normalised sums*, which are defined after t rounds by

$$\sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a). \tag{1}$$

Here $\mathcal{S}$ is a subset of the arms $\{1, \dots, K\}$, $N_a(t)$ is the *random* number of observations from arm a, $\hat{\mu}_a(t)$ is the empirical mean of these observations after t rounds, and $d(\mu, \lambda) \geq 0$ is the relative entropy (Kullback-Leibler divergence) from the exponential family distribution with mean μ to that with mean λ. The more the empirical means of arms in $\mathcal{S}$ deviate from the true means, the larger

the self-normalised sum. We call the summands self-normalised as they are KL-based analogues of the (squared) t-statistic. Namely, a second-order Taylor expansion in μ around $\hat{\mu}(t)$ reveals that $N(t)d(\hat{\mu}(t),\mu) \approx N(t)\frac{(\hat{\mu}(t)-\mu)^2}{2\mathbb{V}(\hat{\mu}(t))}$, where $\mathbb{V}(\mu)$ is the variance of the model with mean μ. One of the reasons why self-normalized sums show up in different sequential learning problems is their relation to (generalized) log likelihood ratio statistics. For example, it can be shown that

$$\ln \frac{\ell(X_1,\dots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\dots,X_t;\boldsymbol{\mu})} = \sum_{a=1}^{K} N_a(t)d(\hat{\mu}_a(t),\mu_a)$$

where $\ell(X_1,\dots,X_t;\boldsymbol{\lambda})$ is the likelihood of the observations under a bandit model whose vector of means is $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\mu}}(t) = (\hat{\mu}_1(t),\dots,\hat{\mu}_K(t))$.

The proposed analyses of the sequential procedures discussed in this paper all rely on a tight control of the deviations of self-normalized sums of the form (1), which inform us about possible values of the means. Our first contribution is the construction of explicit *calibration functions* $\mathcal{C}(x) = x + o(x)$ for which, under any sampling rule (effecting the $N_a(t)$ sampling counts), any bandit model $\boldsymbol{\mu}$ and any confidence $\delta \in (0,1)$, the self-normalised sum associated to *any subset of arms* $\mathcal{S}$ satisfies

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a\in\mathcal{S}}\Big[N_a(t)d(\hat{\mu}_a(t),\mu_a) - O(\ln\ln N_a(t))\Big] \geq |\mathcal{S}|\mathcal{C}\left(\frac{\ln\frac{1}{\delta}}{|\mathcal{S}|}\right)\right) \leq \delta. \tag{2}$$

The salient features of this result are that it is uniform in time, exploits the information geometry (KL) intrinsic to the exponential family (rather than relying on non-parametric relaxations including sub-Gaussianity), and, more importantly, it generalises confidence ellipses by combining in the strong summation sense the evidence from *multiple arms*. Furthermore, as we develop inequalities that hold for any subset $\mathcal{S}$, at the moderate price of a weighted union bound we may apply the bound to any arbitrary (random) subset of the arms, and thereby control the model-selection trade off between the amount of evidence on the left and the magnitude of the threshold on the right.

We may recognise two well-known statistical effects (i.e. fundamental barriers) in the form of the bound (2). First, the Law of the Iterated Logarithm informs us that, in the Gaussian case, $\limsup_{N_a(t)\to\infty} \frac{N_a(t)d(\hat{\mu}_a(t),\mu_a)}{\ln\ln N_a(t)} = \limsup_{N_a(t)\to\infty} \frac{N_a(t)(\hat{\mu}_a(t)-\mu_a)^2}{\ln\ln N_a(t)}$ is a universal constant a.s., whence the correction in the sum. Moreover, it follows from the Wilks phenomenon (Wilks, 1938), which gives the limit distribution of Generalized Likelihood Ratio statistics, that, when $\mathcal{S} = \{1,\dots,K\}$, twice the self-normalised sum (1) converges in distribution to a χ^2_K distribution. The K degrees of freedom are reflected in the perspective scaling of the threshold to which $\sum_{a=1}^{K} N_a(t)d(\hat{\mu}_a(t),\mu_a)$ is compared in (2).

The formal statement of our concentration inequalities is given in Section 3, in which we prove a general result that holds for any exponential family (Theorem 7) and state two improved results for Gaussian and Gamma distributions (Theorems 9 and 10 respectively). We now compare our results to previous work and explain why measuring deviations over multiple arms simultaneously is crucial for applications to sequential learning, which we discuss in Sections 4 to 6.

1.1 Novelty of our Concentration Results

Due to the sequential nature of the data collection process, the analysis of virtually any bandit algorithm relies on deviation inequalities that can take into account the random number of observations from each arm. Several such inequalities have thus been developed in this literature and beyond.

However, most of these results measure deviations for one arm only, which can be rephrased in the form of the following time-uniform deviation inequality

$$\mathbb{P}_\mu \left(\exists t \in \mathbb{N} : td(\hat{\mu}_t, \mu) - O(f(t)) \geq \mathcal{C}\left(\ln \frac{1}{\delta}\right) \right) \leq \delta, \tag{3}$$

where $\hat{\mu}_t$ is the empirical average of t i.i.d. observations with mean μ in a one-parameter exponential family and $f(t) = \ln(t)$ or $\ln\ln(t)$.[1] Further, the majority of existing inequalities were obtained for Gaussian (or sub-Gaussian) distributions with thresholds featuring $f(t) = \ln(t)$ (e.g. de la Peña et al., 2004; Maillard, 2019) or $f(t) = \ln\ln(t)$ (Robbins, 1970; Jamieson et al., 2014; Kaufmann et al., 2016; Zhao et al., 2016; Howard et al., 2018). For other one-dimensional exponential families, time-uniform deviation inequalities with $f(t) = \ln(t)$ have been stated for Bernoulli (Lai, 1976; Jonsson et al., 2020) and Gamma distributions (Lai, 1976). Lai (1976) also provides a generic recipe for general one-parameter exponential families, but that leads to intractable thresholds. On the contrary, our Theorem 7 applied to $|\mathcal{S}| = 1$ leads to an explicit inequality of the form (3) for any exponential family, with a scaling in $f(t) = \ln\ln(t)$. The closest existing result is that of Garivier and Cappé (2011), which controls the deviations uniformly for t in a finite time range $\{1, \dots, n\}$.

To the best of our knowledge, the only prior result that controls deviations over multiple arms simultaneously is Theorem 2 of Magureanu et al. (2014), which also bounds deviations for t in a finite time range $\{1, \dots, n\}$. We provide a detailed comparison with this result in Section 3, showing that our Theorem 7 leads to tighter thresholds, which are furthermore valid for the entire time range $t \in \mathbb{N}$. Given the large number of results that are available for $|\mathcal{S}| = 1$, a natural question is whether inequalities like (3) for different arms can be combined to obtain an inequality like (2). There is no straightforward way to do so and obtain the right scaling in δ: using a naive union bound leads to an inequality of the form (2) in which the right-hand side is $|\mathcal{S}|\mathcal{C}(\ln(1/\delta)) \simeq |\mathcal{S}| \ln(1/\delta)$ instead of $|\mathcal{S}|\mathcal{C}(\ln(1/\delta)/|\mathcal{S}|) \simeq \ln(1/\delta)$. Hence, specific techniques are needed to propose deviation inequalities that sum evidence across arms, which we provide.

In this work we obtain essentially tight calibration functions by building suitable martingales. We show that a calibration function $\mathcal{C}$ satisfying (2) can be obtained by exhibiting a martingale that multiplicatively dominates $\exp\left(\lambda\left[N_a(t)d(\hat{\mu}_a(t), \mu_a) - O(\ln\ln N_a(t))\right]\right)$ for a suitable $\lambda \in (0, 1)$. This central assumption to derive deviation inequalities that sum evidence across arms is formalized in Section 2. Our results are then obtained by leveraging some particular martingales called *mixture martingales* that have this property, which are defined in Section 2.3.

Using martingales to obtain time-uniform inequalities is an old idea that can be traced back to Ville (1939) and all the concentration results quoted above also rely on martingales. We refer the reader to the recent survey of Howard et al. (2020) who study in great detail the power of elementary martingales for deriving time-uniform inequalities, yet without the particular focus on exponential families or multiple arms that we adopt here. Two important techniques based on martingales are the use of a peeling trick (see, e.g. Cappé et al. 2013) or the "method of mixtures" that has been popularized by de la Peña et al. (2004, 2009), and is sometimes also referred to as the Laplace method (Maillard, 2019). We refer the reader to the discussion in Section 2.3 for examples of use of mixture martingales. Our results rely on new constructions of mixture martingales that are tailored

1. Some existing results rather upper bound the probability that $|\hat{\mu}_t - \mu|$ exceeds some threshold. For one-parameter exponential families, we think that it is more natural to measure deviations with the KL-divergence function as the Cramér-Chernoff inequality for such distributions can be expressed as $\mathbb{P}(td(\hat{\mu}_t, \mu) > \ln(1/\delta), \hat{\mu}_t > \mu) \leq \delta$. This form is also more convenient for measuring deviations for multiple arms, which is supported by our new inequalities.

for exponential families. Interestingly, we note that the result of Magureanu et al. (2014) is not based on mixture martingales: its proof relies on a peeling technique which requires the knowledge of n, and a stochastic dominance argument. Our proof technique based on mixture martingales is more flexible as it allows to easily bound deviations uniformly over the entire domain $t \in \mathbb{N}$, which is crucial for the analysis of sequential tests that involve random stopping.

1.2 Applications to Sequential Learning

In this section we give more context, review our contribution, and illustrate its advantage on a simple example.

1.2.1 Related Work on Bandits

Stochastic multi-armed bandit models can be traced back to the work of Thompson (1933) motivated by clinical trials. They were later studied by Robbins (1952); Lai and Robbins (1985) who introduced the regret minimization objective: the samples $X_1, \dots, X_t$ are seen as rewards and the goal is to find a sequential strategy to maximize the (expected) cumulated reward, which is equivalent to minimizing some notion of regret (see e.g. Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvari, 2019, for surveys).

In the meantime, pure-exploration problems in bandit models have also received increased attention (Even-Dar et al., 2006; Bubeck et al., 2011). In this context, a common objective is to identify as quickly and accurately as possible the arm with the largest mean, relinquishing the incentive to maximize the sum of rewards. In the fixed-confidence setting, the minimal number of samples needed to identify the best arm with accuracy larger than $1-\delta$ when arms belong to a one-dimensional family has been identified by Garivier and Kaufmann (2016), in a regime of small values of δ. Their Track-and-Stop algorithm is shown to asymptotically match this optimal sample complexity. Extensions of this best arm identification problem in which one should answer quickly and accurately some more general query about the means of the arms have also been studied (Huang et al., 2017; Chen et al., 2017). Prototypical queries beyond Best Arm include Top-M (Kalyanakrishnan and Stone, 2010), Thresholding (Locatelli et al., 2016), Minimum Threshold (Kaufmann et al., 2018), Combinatorial Bandits (Chen et al., 2014), pure-strategy Nash equilibria (Zhou et al., 2017) or Monte-Carlo Tree Search (Teraoka et al., 2014). We note that Track-and-Stop has recently been generalized by Juneja and Krishnasamy (2019) to a generic "partition identification" problem similar to the one that we consider in Section 4, while Degenne and Koolen (2019) have studied its extension to queries with multiple correct answers. Finally, recent research has also focused on developing alternatives to Track-and-Stop that are more efficient numerically, like Degenne et al. (2019) who develop algorithms based on iterative saddle point solving.

1.2.2 Our Contributions

The first impact of our concentration results is that they permit to analyse new stopping rules based on Generalized Likelihood Ratios, which extend the stopping rule originally proposed for Track-and-Stop (Garivier and Kaufmann, 2016) to generic sequential identification problems. Our generic stopping rule is presented in Section 4, in which we further show that under some assumptions on the identification problem itself, such a stopping rule combined with a suitable sampling rule is (asymptotically) optimal in terms of sample complexity. We then provide in Section 5 refined

stopping criteria for some particular tests that replace the number of arms K in the threshold by a new notion of rank.

Next, we explain in Section 6 how our deviation inequalities can be used to build tight confidence regions on (functions of) the unknown parameter $\boldsymbol{\mu}$. Indeed, the sum form of the left-hand quantity in (2) allows us to build confidence regions that exclude the configuration of all (many) empirical estimates $\hat{\mu}_a(t)$ being far from their means μ_a simultaneously. We show how this effect yields improved confidence intervals for functions of the mean $\boldsymbol{\mu}$ in the cases of linear functions and minima. In concrete examples, we can quantify the benefit precisely.

1.2.3 ILLUSTRATION OF THE BENEFIT OF (2) ON A SIMPLE EXAMPLE

A common task in sequential learning is to construct a confidence interval on the difference $\mu_1 - \mu_2$ in mean between two arms, for example to decide whether μ_1 can plausibly be higher than μ_2 in a best arm identification scenario. We now quantify the benefit of using the self-normalized sum (2) compared to the classical approach of combining per-arm intervals using the union bound, with an illustration provided in Figure 1.

For maximum interpretability, we instantiate (2) for Gaussian arms with variance 1 (so that $d(x,y) = (x-y)^2/2$), we ignore the $\ln\ln$ terms, and we approximate $K\mathcal{C}(\ln\frac{1}{\delta}/K) \approx \ln\frac{1}{\delta}$. Then if we follow the classical per-arm approach, we obtain a confidence interval on μ_a for each arm a separately using (2) (which now reduces to the standard Chernoff bound), combine these into a rectangular confidence region on the pair (μ_1, μ_2) using the union bound over arms (called "Box" in Figure 1), and work out what we know about the difference $\mu_1 - \mu_2$ by projecting. Doing so, we obtain a confidence interval on $\mu_1 - \mu_2$ that has diameter $\sqrt{8\ln\frac{2}{\delta}}\left(\sqrt{\frac{1}{N_a(t)}} + \sqrt{\frac{1}{N_b(t)}}\right)$. In contrast, the self-normalised sum of 2 arms directly provides a confidence ellipse on the pair (μ_1, μ_2) (called "Sum" in Figure 1), and projecting that to the difference $\mu_1 - \mu_2$ yields a tighter interval of diameter $\sqrt{8\ln\frac{1}{\delta}\left(\frac{1}{N_a(t)} + \frac{1}{N_b(t)}\right)}$. The advantage of the second approach can be up to a factor $\sqrt{2}$, which occurs for equal sample sizes $N_a(t) = N_b(t)$. In typical adaptive stopping problems, a reduction by $\sqrt{2}$ in confidence width leads to an improvement by a factor 2 of the sample complexity.

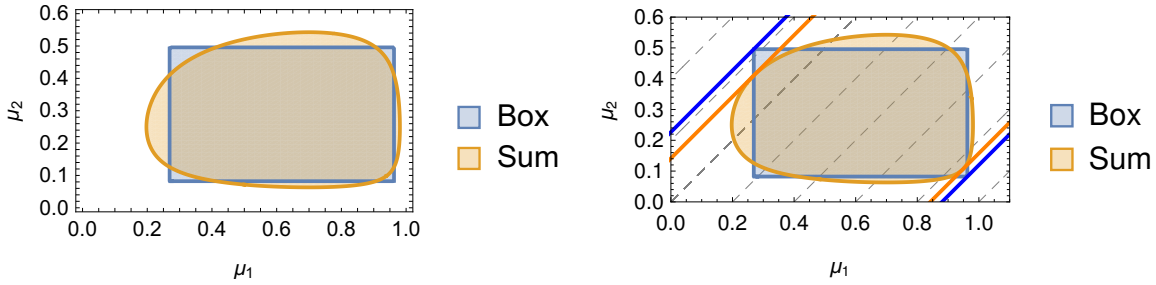
In Section 6.1, we quantify the obtained improvement for the more general task of building a confidence interval on a linear function $\boldsymbol{v}^\intercal\boldsymbol{\mu}$ of the means $\boldsymbol{\mu} \in \mathbb{R}^K$, which can be as large as $\sqrt{K}$.

2. Martingales and Deviation Inequalities for Exponential Family Bandit Models

In this section, we formally introduce the stochastic processes for which we want to obtain deviation inequalities. We then present a general method for obtaining deviation inequalities for any such stochastic process. It relies on the crucial assumption that one can find martingales multiplicatively dominating exponential transforms of the process. We further introduce the general class of martingales that we shall exhibit in order to obtain the particular deviation results of this paper, namely mixture martingales.

2.1 Exponential Family Bandit Models

A one-parameter canonical exponential family is a class $\mathcal{P}$ of probability distributions characterized by a set $\Theta \subset \mathbb{R}$ of natural parameters, a strictly convex and twice-differentiable function $b : \Theta \to \mathbb{R}$



(a) Confidence region for $\boldsymbol{\mu}$

(b) Confidence interval for the difference $\mu_1 - \mu_2$ obtained by projecting confidence regions for $\boldsymbol{\mu}$. The dashed grey help lines connect points of equal difference $\mu_1 - \mu_2$. The largest and smallest values for the difference are obtained by squeezing the confidence interval between diagonal tangents (solid lines). We see that the confidence width, which is the distance between the intercepts, is strictly larger for Box than for Sum: the rounded nature of Sum provides tighter control on the difference.

Figure 1: Visual two-arm comparison of confidence regions for $\boldsymbol{\mu}$ and the implied confidence interval for the difference $\mu_1 - \mu_2$. A union bound over traditional per-arm confidence intervals gives the "Box" region. Our new bound (2) results in a confidence region of the egg-shape marked "Sum".

(called the log-partition function) and a reference measure m. It is defined as

$$\mathcal{P} = \left\{ \nu_\theta, \theta \in \Theta : \nu_\theta \text{ has density } f_\theta(x) = e^{x\theta - b(\theta)} \text{ with respect to } m \right\}.$$

Example of exponential families include the set of Bernoulli distribution, Poisson distributions, Gaussian distribution with known variance or Gamma distributions with known shape parameter. For any exponential family $\mathcal{P}$ it can be shown that the mean $\mu(\theta)$ of the distribution ν_θ satisfies $\mu(\theta) = \dot{b}(\theta)$. Observe that μ is a strictly increasing function of the natural parameter θ, hence distributions in $\mathcal{P}$ can be alternatively parameterized by their means.

We adopt this parameterization in this paper. Letting $\mathcal{I} := \dot{b}(\Theta)$ be the set of possible mean parameters, for all $\mu \in \mathcal{I}$ we define ν^μ to be the distribution in $\mathcal{P}$ that has mean μ. We also define the Kullback-Leibler divergence between two distributions in $\mathcal{P}$ as a function of their means by

$$d(\mu, \mu') := \mathrm{KL}\left(\nu^\mu, \nu^{\mu'}\right) = \int \ln \frac{f_{\dot{b}^{-1}(\mu)}(x)}{f_{\dot{b}^{-1}(\mu')}(x)} f_{\dot{b}^{-1}(\mu)}(x)\, \mathrm{d}m(x).$$

This divergence function has a closed form expression in the classical exponential families mentioned above. For example for Gaussian distribution with variance σ^2 one has $d(\mu, \mu') = (\mu - \mu')^2/(2\sigma^2)$ and for Bernoulli distributions $d(\mu, \mu') = \mu \ln(\mu/\mu') + (1-\mu)\ln((1-\mu)/(1-\mu'))$. Further examples can be found in Cappé et al. (2013).

An exponential family bandit model is a sequence of K probability distributions $\nu^{\mu_1}, \ldots, \nu^{\mu_K}$ that belong to some one-dimensional canonical exponential family $\mathcal{P}$: it can be fully parametrized by the vector of means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathcal{I}^K$. In a bandit model, data is collected sequentially: an arm A_t is selected at round t and a sample X_t from the distribution $\nu^{\mu_{A_t}}$ is observed. We denote by $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s = a)}$ the number of selections of arm a in the first t rounds and $S_a(t) =$

$\sum_{s=1}^{t} X_s \mathbb{1}_{(A_s=a)}$ the sum of these observations. The empirical mean of the observations obtained from arm a up to round t is therefore defined as $\hat{\mu}_a(t) = S_a(t)/N_a(t)$ once $N_a(t) \neq 0$. We let $\mathcal{F}_t = \sigma(A_1, X_1, \ldots, A_t, X_t)$ be the filtration generated by the observations gathered within the first t rounds and assume the sampling rule is such that A_t is mesurable with respect to $\sigma(\mathcal{F}_{t-1}, U_t)$ where U_t is a uniform random variable that is independent from $\mathcal{F}_{t-1}$ (allowing randomized algorithms).

In this paper, our objective is to prove *time-uniform* deviation inequalities for sums involving the terms $N_a(t)d(\hat{\mu}_a(t), \mu_a)$ (or some one-sided versions of these). The price for uniformity in time will be some $\ln \ln(N_a(t))$ term and we shall for example obtain deviation inequalities for sums of the entries of a stochastic process $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ of the form

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t), \mu_a) - c \ln(d + \ln N_a(t)) \tag{4}$$

for some constants c and d. We now describe a general method to obtain time-uniform deviation inequalities for *any* arm-dependent stochastic process $\boldsymbol{X}(t)$.

2.2 A General Method for Obtaining Deviation Inequalities

Let $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ be a stochastic process indexed by arms. Here we introduce a central assumption under which it is easy to obtain deviation inequalities for sums of the entries of $\boldsymbol{X}(t)$ by combining Ville's inequality for martingales with the Cramér-Chernoff method. For this reason, we call such processes g-VCC (in reference to the Ville-Cramér-Chernoff trio). We will also follow Shafer et al. (2011) in calling any non-negative martingale $M(t) \geq 0$ of unit initial value $M(0) = 1$ a *test martingale*.

Definition 1 *Let $g : \Lambda \to \mathbb{R}$ be a function defined on a non-empty interval $\Lambda \subseteq \mathbb{R}$. A stochastic process $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ is called* g-VCC *if it satisfies the following properties.*

1. *For any arm a and $\lambda \in \Lambda$ there exists a test martingale $M_a^\lambda(t)$ such that*

$$\forall t \in \mathbb{N}, \quad M_a^\lambda(t) \geq e^{\lambda X_a(t) - g(\lambda)}. \tag{$*$}$$

2. *For any subset $\mathcal{S} \subseteq \{1, \ldots, K\}$ and for any $\lambda \in \Lambda$, the product $\prod_{a \in \mathcal{S}} M_a^\lambda(t)$ is a martingale.*

We note that the independent work of Howard et al. (2020) also presents a general method based on the Cramér-Chernoff method to derive time-uniform concentration inequalities. The authors propose deviation inequalities for a two-dimensional stochastic processes (S_t, V_t) under an assumption that bears similarities with $(*)$: $\exp(\lambda S_t - \phi(\lambda)V_t)$ has to be upper bounded by a martingale, for a known function ϕ and for all λ in a certain range. Yet the proposed applications of these two general methods differ, in particular there is no emphasis on measuring deviations for multiple arms in the work of Howard et al. (2020).

Remark 2 *To calibrate what to expect for g, we can use knowledge of the asymptotic distribution of the $X_a(t)$ given in (4). In our applications, Wilks' phenomenon (see de la Peña et al., 2009, Chapter 17) tells us that $2X_a(t)$ is asymptotically (for $N_a(t) \to \infty$) χ^2 distributed when $c = 0$ in (4). For $2Y \sim \chi^2$, we have $\mathbb{E}[e^{\lambda Y}] = (1-\lambda)^{-1/2}$. This strongly suggests (and this is what we will find) that $g(\lambda)$ should be at least $\frac{1}{2}\ln(1-\lambda)$, plus a mild additional cost for uniformity in time. For this reason we will refer to $g_{\chi^2}(\lambda) = \frac{1}{2}\ln(1-\lambda)$ as the "ideal function".*

For a g-VCC stochastic process $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$, we provide a general deviation inequality for the sum of the entries $X_a(t)$ over any subset of arms. The threshold is related to the function g through the following quantities.

Definition 3 *For $g : \Lambda \to \mathbb{R}^+$, we define for all $x > 0$,*

$$C^g(x) \quad := \quad \min_{\lambda \in \Lambda} \frac{g(\lambda) + x}{\lambda}.$$

We also define the convex conjugate of g, $g^*(x) := \max_{\lambda \in \Lambda} \ (\lambda x - g(\lambda))$.

With these functions in hand, we can now state our g-VCC deviation inequality.

Lemma 4 *Fix $\mathcal{S} \subseteq \{1, \dots, K\}$. Let $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ be a g-VCC stochastic process. Then*

$$\forall x > 0, \qquad \mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| C^g\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x},$$

$$\forall u > 0, \qquad \mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > u\right) \leq \exp\left(-|\mathcal{S}| g^*\left(\frac{u}{|\mathcal{S}|}\right)\right).$$

Proof Fix $\lambda \in \Lambda$. As $\boldsymbol{X}(t)$ is g-VCC (see Definition 1), we find

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > u\right) &= \mathbb{P}\left(\exists t \in \mathbb{N} : e^{\lambda\left[\sum_{a \in \mathcal{S}} X_a(t)\right]} > e^{\lambda u}\right) \\
&\leq \mathbb{P}\left(\exists t \in \mathbb{N} : \prod_{a \in \mathcal{S}} M_a^\lambda(t) > e^{\lambda u - |\mathcal{S}| g(\lambda)}\right).
\end{aligned}$$

As $\prod_{a \in \mathcal{S}} M_a^\lambda(t)$ is a test martingale, it follows from Ville's inequality ($\mathbb{P}(\exists t \in \mathbb{N}^* : M(t) \geq 1/x) \leq x$ for any non-negative super-martingale starting from $\mathbb{E}[M(0)] = 1$ and any $x \in (0, 1]$, Ville 1939) that

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > u\right) \leq e^{-[\lambda u - |\mathcal{S}| g(\lambda)]} \tag{5}$$

Equivalently, one can also establish that for all $x > 0$, for all $\lambda \in \Lambda$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) > \frac{|\mathcal{S}| g(\lambda) + x}{\lambda}\right) \leq e^{-x} \tag{6}$$

Picking the best possible λ in (6) yields the first inequality in Lemma 4 while picking the best possible λ in (5) yields the second inequality. ■

The deviation inequalities given in Lemma 4 are either expressed in terms of the threshold function C^g or in terms of the convex conjugate g^*. Depending on g, one of these two quantities might be easier to compute that the other one. Note that if g^* is well-behaved, the threshold function can be obtained by inverting g^*, as stated below.

Proposition 5 *Assume g^* is increasing. For all $u \in g^*(\mathbb{R}^+)$, $C^g(u) = (g^*)^{-1}(u)$.*

Proof As g^* is increasing on $\mathbb{R}^+$, the inverse function $(g^*)^{-1}$ is well defined on $\mathcal{I} := g^*(\mathbb{R}^+)$. From the definitions of C^g and g^*, it is easy to check that

$$\forall x > 0, \;\; g^*(C^g(x)) \geq x \;\text{ and }\; C^g(g^*(x)) \leq x.$$

These two inequalities respectively yield that for all $u \in \mathcal{I}$, $(g^*)^{-1}(u) \leq C^g(u)$ and $C^g(u) \leq (g^*)^{-1}(u)$, which concludes the proof. ■

2.3 Mixture Martingales

Introducing the cumulant generating function $\phi_\mu(\eta) := \ln \mathbb{E}_{X\sim\nu_\mu}\left[e^{\eta X}\right]$ for all $\mu \in \mathcal{I}$, it holds for all $\eta \in \mathbb{R}$ that

$$Z_a^\eta(t) \;\coloneqq\; \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) \tag{7}$$

is a test martingale with respect to the filtration $\mathcal{F}_t$, for any sampling rule. Indeed, when $A_t = a$ we have $\mathbb{E}\left[Z_a^\eta(t) | A_t, \mathcal{F}_{t-1}\right] = Z_a^\eta(t-1)\mathbb{E}\left[e^{\eta X_t - \phi_{\mu_a}(\eta)} \middle| A_t, \mathcal{F}_{t-1}\right] = Z_a^\eta(t-1)$, and the same trivially holds when $A_t \neq a$. So by the tower rule $\mathbb{E}\left[Z_a^\eta(t)|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\mathbb{E}\left[Z_a^\eta(t)|A_t, \mathcal{F}_{t-1}\right]|\mathcal{F}_{t-1}\right] = Z_a^\eta(t-1)$. More generally, for any probability distribution π on η, the *mixture martingale*

$$Z_a^\pi(t) \;\coloneqq\; \int Z_a^\eta(t)\, \mathrm{d}\pi(\eta) \tag{8}$$

is also a test martingale, as can be seen by applying Tonelli's theorem

$$\mathbb{E}\left[Z_a^\pi(t) | A_t, \mathcal{F}_{t-1}\right] \;=\; \int \underbrace{\mathbb{E}\left[Z_a^\eta(t) | A_t, \mathcal{F}_{t-1}\right]}_{= Z_a^\eta(t-1)} d\pi(\eta) \;=\; Z_a^\pi(t-1).$$

Finally, given a family of priors $\boldsymbol{\pi} = \{\pi_a\}_{a=1}^K$, the *product martingale* $Z_{\mathcal{S}}^{\boldsymbol{\pi}}(t) \coloneqq \prod_{a\in\mathcal{S}} Z_a^{\pi_a}(t)$ is also a test martingale with respect to $\mathcal{F}_t$, for any subset $\mathcal{S}$. Namely, when $A_t \in \mathcal{S}$ we have

$$\mathbb{E}\left[Z_{\mathcal{S}}^{\boldsymbol{\pi}}(t) | A_t, \mathcal{F}_{t-1}\right] \;=\; Z_{\mathcal{S}\setminus\{A_t\}}^{\boldsymbol{\pi}}(t-1) \underbrace{\mathbb{E}\left[Z_{A_t}^{\pi_{A_t}}(t) \middle| A_t, \mathcal{F}_{t-1}\right]}_{= Z_{A_t}^{\pi_{A_t}}(t-1)} \;=\; Z_{\mathcal{S}}^{\boldsymbol{\pi}}(t-1),$$

and the same result holds trivially when $A_t \notin \mathcal{S}$. The martingale property follows by the tower rule. Hence, a sufficient condition to establish that a stochastic process $\boldsymbol{X}(t)$ is g-VCC is to exhibit for all $\lambda \in \Lambda$ a family of priors $\pi_{a,\lambda}$ such that $M_a^\lambda(t) := Z_a^{\pi_{a,\lambda}}(t)$ satisfies (∗). This is how we proceed in the next sections.

2.3.1 EXAMPLE OF MIXTURE MARTINGALES

Among the first occurrence of such mixture martingales, one can mention the works of Darling and Robbins (1968); Robbins (1970) which consider the martingale $\int \exp\left(\eta S_t - \frac{\eta^2\sigma^2}{2} t\right) \mathrm{d}\pi(\eta)$ where S_t is a sum of t i.i.d. standard Gaussian random variables and π is a Gaussian prior. This martingale coincides with our $Z_a^\pi(t)$ for a single standard Gaussian arm a. The choice of Gaussian

prior π results in a threshold growing like $\sqrt{t \ln t}$. We use different priors, which asymptote at $\eta = 0$, to obtain a deviation inequality for S_t that is uniform in time and compatible with the Law of the Iterated Logarithm: S_t is compared to a threshold that grows like $\sqrt{2t \ln \ln(t)}$.

More broadly, the term mixture martingale can refer to any martingale of the form $\int M^\eta(t) d\pi(\eta)$ where $M^\eta(t)$ is some martingale (not necessarily $Z_a(t)$) and π is some probability distribution (that we call the prior). For example the likelihood ratio martingales introduces by Lai (1976) are of this form. Mixture martingale constructions are also at the heart of the game-theoretic approach to probability (Dawid and Vovk, 1999). The "method of mixtures" has then been popularized by de la Peña et al. (2004, 2009) who use it to prove self-normalized deviation inequalities for general stochastic processes. Examples of its use include the work of Abbasi-Yadkori et al. (2011) who propose a self-normalized deviation inequality for a vector-valued martingale applied to the linear bandit problem or that of Balsubramani (2015) who derive time-uniform Hoeffding or Bernstein deviation inequalities. Most of these works present mixture martingales with specific choices of continuous priors for which the corresponding mixture can be either computed in closed form or well-approximated. In this paper, we will rely on priors constructed in hierarchical fashion from discrete and continuous ingredients with the goal of obtaining explicit near-optimal thresholds.

3. New Deviation Inequalities for Exponential Families

In this section, we first provide a general deviation result that holds for any one-dimensional exponential family and can also accommodate *one-sided deviations* (Theorem 7). Next, we present in Section 3.2 tighter deviation inequalities that measure two-sided deviations for the special cases of Gaussian and Gamma distributions. The two sets of results rely on proving that a stochastic process is g-VCC for certain functions g, which we do by constructing appropriate mixture martingales based on hierarchical priors in Section 3.4.

3.1 Main Result

To state our result, we introduce one-sided versions of the Kullback-Leibler divergence, namely $d^+(u,v) = d(u,v)\mathbb{1}_{(u \leq v)}$ and $d^-(u,v) = d(u,v)\mathbb{1}_{(u \geq v)}$. We further introduce the notation

$$\begin{aligned} Y_a(t) &:= [N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\ln(1 + \ln(N_a(t))]^+ \\ Y_a^-(t) &:= [N_a(t) d^-(\hat{\mu}_a(t), \mu_a) - 3\ln(1 + \ln(N_a(t))]^+ \\ Y_a^+(t) &:= [N_a(t) d^+(\hat{\mu}_a(t), \mu_a) - 3\ln(1 + \ln(N_a(t))]^+ \end{aligned}$$

and let $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ be a stochastic process such that, for all a, either $\forall t, X_a(t) = Y_a(t)$ (for two-sided deviations) or $\forall t, X_a(t) = Y_a^+(t)$ or $\forall t, X_a(t) = Y_a^-(t)$ (for one-sided deviations).

Remark 6 *We use as our correction function* $3\ln(1 + \ln N_a(t))$*, which is vacuous when* $N_a(t) = 0$ *because* $\ln N_a(t) = -\infty$*. Most algorithms for bandits avoid considering this situation, and start by pulling all arms once. In some scenarios, especially with many arms, it may be desirable to include the case* $N_a(t) = 0$*. There is no essential bottleneck, and one could adjust the analysis to, for example, replace it by* $3\ln(1 + \ln(1 + N_a(t)))$*.*

We provide in Theorem 7 below a new self-normalized deviation inequality featuring a calibration function $\mathcal{C}_{\text{exp}}$. To give the expression of $\mathcal{C}_{\text{exp}}$, we need to introduce two functions. First for $u \geq 1$ the

function $h(u) = u - \ln u$ and its inverse $h^{-1}(u)$. Secondly, the function defined for any $z \in [1, e]$ and $x \geq 0$ by

$$\tilde{h}_z(x) \;=\; \begin{cases} e^{1/h^{-1}(x)} h^{-1}(x) & \text{if } x \geq h(1/\ln z), \\ z(x - \ln \ln z) & \text{otherwise.} \end{cases} \tag{9}$$

Theorem 7 *Let $\mathcal{C}_{\text{exp}} : \mathbb{R}^+ \to \mathbb{R}^+$ be the function defined by*

$$\mathcal{C}_{\text{exp}}(x) \;=\; 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x) + \ln(2\zeta(2))}{2}\right) \tag{10}$$

where $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$. For $\mathcal{S}$ a subset of arms and $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| \mathcal{C}_{\text{exp}}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

Moreover, if $X_a(t)$ measures only one-sided deviation (that is for all a, $X_a = Y_a^+$ or $X_a = Y_a^-$), the calibration function can be replaced by the smaller $\tilde{\mathcal{C}}_{\text{exp}}(x) \;=\; 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x)+\ln(\zeta(2))}{2}\right)$.

As can be seen in the proof given in Section 3.4, this result follows by exhibiting *a family of functions* g_ξ such that $\boldsymbol{X}(t)$ is g_ξ-VCC, applying Lemma 4 and then optimizing the parameters to obtain the best possible calibration function. Proposition 8 below (proved in Appendix A) gives a tight bound on the inverse function h^{-1}, which yields an upper bound on the calibration function $\mathcal{C}_{\text{exp}}$. On can easily see that $\mathcal{C}_{\text{exp}}(x) \sim x$ when x tends to infinity. For $x \geq 5$, a good approximation of the threshold is $\mathcal{C}_{\text{exp}}(x) \simeq x + 4\ln(1 + x + \sqrt{2x})$, which is slightly larger than the approximation $\simeq x + \ln(x)$ that is added for comparison to Figure 2.

Proposition 8 *The function h is increasing on $[1, +\infty[$ and its inverse function, defined on $[1, +\infty[$, satisfies $h^{-1}(x) = -W_{-1}(-e^{-x})$ with W_{-1} the negative branch of the Lambert function. Moreover,*

$$\forall x \geq 1, \;\; h^{-1}(x) \leq x + \ln(x + \sqrt{2(x-1)}).$$

3.2 Refined Results for Gaussian and Gamma Distribution

For Gaussian and Gamma distributions, a different martingale construction, explained in detail in Appendix C, permits to establish the following results.

In a bandit model with Gaussian arms with means μ_a and known variance σ^2, the associated divergence is $d(\mu, \mu') = \frac{(\mu - \mu')^2}{2\sigma^2}$ and one can prove the following theorem.

Theorem 9 *In a Gaussian bandit model, introducing for all a the process $X_a(t) = N_a(t) d(\hat{\mu}_a(t), \mu_a) - 2\ln(4 + \ln N_a(t))$, the stochastic process $\boldsymbol{X}(t)$ is g_G-VCC where*

$$\begin{array}{rccl} g_G :]1/2, 1] & \longrightarrow & \mathbb{R} \\ \lambda & \mapsto & 2\lambda - 2\lambda \ln(4\lambda) + \ln \zeta(2\lambda) - \frac{1}{2}\ln(1-\lambda). \end{array}$$

Hence, letting $\mathcal{C}_{\text{G}} := C^{g_G}$, it follows from Lemma 4 that for every subset $\mathcal{S}$ and $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} \left[N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) - 2\ln(4 + \ln N_a(t))\right] \geq |\mathcal{S}| \mathcal{C}_{\text{G}}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

In a bandit model with arms that are Gamma distributed with means μ_a and known shape parameter α, the associated divergence is $d(\mu,\mu') = \alpha\left(\frac{\mu}{\mu'} - 1 - \ln\frac{\mu}{\mu'}\right)$ and one can prove the following theorem.

Theorem 10 *In a Gamma bandit model, introducing for all a the process $X_a(t) = N_a(t)d(\hat{\mu}_a(t),\mu_a) - 2\ln(4+\ln N_a(t))$, the stochastic process $\boldsymbol{X}(t)$ is g_Γ-VCC where*

$$\begin{array}{rcl} g_\Gamma :]1/2,1] & \longrightarrow & \mathbb{R} \\ \lambda & \mapsto & 2\lambda - 2\lambda\ln(4\lambda) + \ln\zeta(2\lambda) - \ln(1-\lambda)\,. \end{array}$$

Hence, letting $\mathcal{C}_\Gamma := C^{g_\Gamma}$, it follows from Lemma 4 that for every subset $\mathcal{S}$ and $x>0$,

$$\mathbb{P}\left(\exists t\in\mathbb{N} : \sum_{a\in\mathcal{S}}\left[N_a(t)d\left(\hat{\mu}_a(t),\mu_a\right) - 2\ln(4+\ln N_a(t))\right] \geq |\mathcal{S}|\mathcal{C}_\Gamma\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

3.3 Discussion

The three deviation inequalities given Theorems 7, 9 and 10 all provide a control of the two-sided deviations of the empirical means from the true means, of the form

$$\mathbb{P}\left(\exists t\in\mathbb{N} : \sum_{a\in\mathcal{S}} N_a(t)d(\hat{\mu}_a(t),\mu_a) > \sum_{a\in\mathcal{S}} c\ln(d+\ln(N_a(t))) + |\mathcal{S}|\mathcal{C}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}$$

where c and d are two constants and $\mathcal{C}(x)$ is a calibration function. For Gaussian or Gamma distributions one can use $c=2, d=4$ while $c=3, d=1$ apply for other one-dimensional exponential families. A more crucial difference is the calibration function $\mathcal{C}$, which can be set to $\mathcal{C}_{\mathrm{G}}$ for Gaussian distributions, to $\mathcal{C}_\Gamma$ for Gamma distributions and to $\mathcal{C}_{\exp}$ in general.

Those three calibration functions are hard to compare at first as they have no closed-form expressions. Equation (10) provides an explicit expression for $\mathcal{C}_{\exp}$ but that still requires to numerically invert the function h, while $\mathcal{C}_{\mathrm{G}}$ and $\mathcal{C}_\Gamma$ can be numerically approximated using Definition 3 which requires to minimize a convex function. In Figure 2 we compare those three thresholds to the "ideal" calibration function $C^{g_{\chi^2}}$ where $g_{\chi^2}(\lambda) = -\frac{1}{2}\ln(1-\lambda)$ (see Remark 2). We see that that this idealized calibration satisfies $C^{g_{\chi^2}}(x) \simeq x + \ln(x)$ and that the calibration functions obtained for Gaussian and Gamma distributions are very close to it. The function $\mathcal{C}_{\exp}$ seems to be off by an additive term of order 10.

Despite this slightly larger calibration function, the general result of Theorem 7 is interesting for the following reasons. First, obviously it covers more distributions like Bernoulli and Poisson distributions that are often relevant for applications of multi-armed bandits. Then, we noted that Theorem 7 can be made tighter in case only one-sided deviations are measured (when $N_a(t)d^+(\hat{\mu}_a(t),\mu_a)$ or $N_a(t)d^-(\hat{\mu}_a(t),\mu_a)$ are used): $\mathcal{C}_{\exp}$ can be replaced by the slightly smaller threshold $\tilde{\mathcal{C}}_{\exp}$. In contrast, the construction presented in Appendix C cannot be easily adapted to obtain better results for one-sided deviations for Gaussian or Gamma distributions. Finally, the presence of the positive part in the definition of $Y_a(t)^\pm$ leads to the following improved result holding uniformly over subsets:

$$\mathbb{P}\left(\exists t\in\mathbb{N} : \exists\mathcal{S}'\subseteq\mathcal{S}, \sum_{a\in\mathcal{S}'} N_a(t)d^\pm(\hat{\mu}_a(t),\mu_a) > \sum_{a\in\mathcal{S}'} 3\ln(1+\ln(N_a(t))) + |\mathcal{S}|\mathcal{C}_{\exp}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

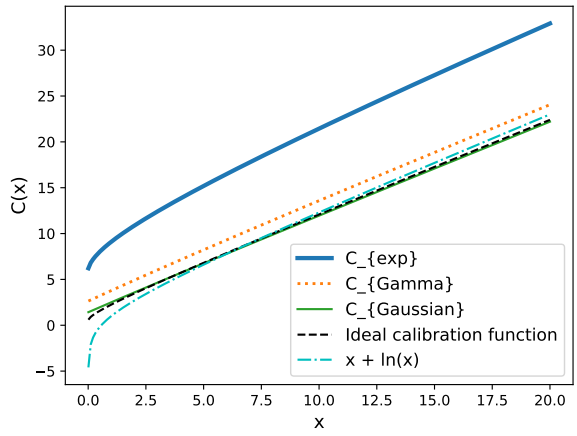


Figure 2: Several calibration functions $\mathcal{C}(x)$ as a function of x.

3.3.1 COMPARISON TO THE STATE-OF-THE-ART

To the best of our knowledge, the only available result that controls deviations over multiple arms simultaneously is the one of Magureanu et al. (2014). More precisely, Theorem 2 in Magureanu et al. (2014) can be rephrased as follows[2], introducing the function $\tilde{f}(u) = u - 2\ln(u)$ for $u \geq 2$:

$$\mathbb{P}\left(\exists t \leq n : \sum_{a \in \mathcal{S}} N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq |\mathcal{S}| \tilde{f}^{-1}\left(1 + \ln\ln(n) + \frac{x+1}{|\mathcal{S}|}\right)\right) \leq e^{-x}. \tag{11}$$

We note that here the deviations are uniform over a finite time range $\{1, \dots, n\}$. This style of deviation inequalities can also be deduced from Theorem 7 for general exponential families:

$$\mathbb{P}\left(\exists t \leq n : \sum_{a \in \mathcal{S}} N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq 3|\mathcal{S}| \ln(1 + \ln(n)) + |\mathcal{S}| \mathcal{C}_{\text{exp}}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}, \tag{12}$$

or from Theorem 9 and Theorem 10:

$$\mathbb{P}\left(\exists t \leq n : \sum_{a \in \mathcal{S}} N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq 2|\mathcal{S}| \ln(4 + \ln(n)) + |\mathcal{S}| \mathcal{C}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}, \tag{13}$$

where $\mathcal{C}(x) = \mathcal{C}_{\text{G}}(x)$ for Gaussian distributions and $\mathcal{C}(x) = \mathcal{C}_{\Gamma}(x)$ Gamma distributions. In Figure 3, we plot the thresholds (right-hand side of the deviation inequalities) featured in (11) and (12) for different values of n, $|\mathcal{S}|$ and x, revealing that the threshold in (12) can be much smaller.

3.3.2 IMPROVED RESULT WHEN $|\mathcal{S}| = 1$

Theorem 7 can be made slightly tighter for a subset of size 1 (see Appendix B.3) and we obtain, with $\hat{\mu}_t$ the empirical mean of t observations from a distribution with mean μ in an exponential family,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : t d\left(\hat{\mu}_t, \mu\right) \geq 3\ln(1 + \ln(t)) + 2\tilde{h}_{3/2}\left(\frac{x + \ln(2\zeta(2))}{2}\right)\right) \leq e^{-x}. \tag{14}$$

2. The result only considers Bernoulli arms and $\mathcal{S} = [K]$, but their analysis can be easily extended to cover the more general case of exponential families and any subset $\mathcal{S}$

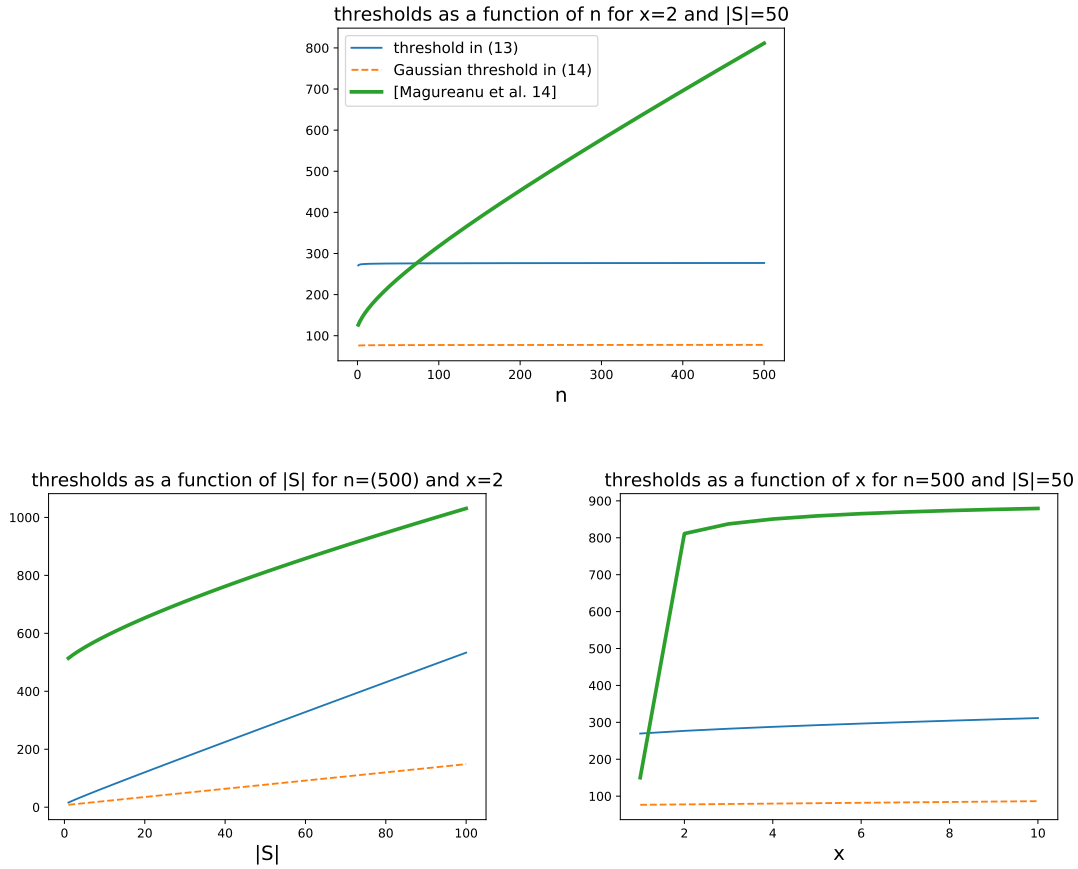


Figure 3: Thresholds that follow from Theorem 7 and Theorem 9 compared to the threshold in (11) obtained by Magureanu et al. (2014).

Albeit not our main focus, this result is interesting as it provides a deviation result in which the threshold features a $\ln(\ln(t))$ term, which is known to be unavoidable in the Gaussian case due to the Law of Iterated Logarithm. In the Gaussian (or sub-Gaussian) case, several results of this type already exist (Robbins, 1970; Jamieson et al., 2014; Kaufmann et al., 2016; Zhao et al., 2016; Howard et al., 2018) and we do not claim ours to be the tightest in general. Beyond Gaussian distribution, (14) is the first result that controls deviations uniformly in $t \in \mathbb{N}$ for general exponential families and with a threshold scaling in $\ln(\ln(t))$.

3.4 Sketch of Proofs

In this section, we provide a detailed proof of Theorem 7, leaving the proof of some intermediate lemmas to Appendix B. The proofs of Theorem 9 and 10 are given in Appendix C, but we provide below a high-level description of the martingale that is used for these results.

Proof of Theorem 7 Fix $\xi > 0$ and define for all $\lambda \in [0, 1/(1+\xi))$,

$$g_\xi(\lambda) = \lambda(1+\xi)\ln\left(C(\xi)\right) - \ln(1-\lambda(1+\xi)) \quad \text{with} \quad C(\xi) = \frac{2\zeta(2)}{(\ln(1+\xi))^2} \ .$$

The proof hinges on the fact that for the stochastic process $\boldsymbol{X}$, there exists a martingale satisfying (∗).

Lemma 11 *For $\xi \in [0, 1/2]$, $\boldsymbol{X}$ is g_ξ-VCC (see Definition 1).*

As will be seen in the proof of Lemma 11, given shortly, in case the stochastic process $\boldsymbol{X}$ only measures one-sided deviations, that is for all a either $X_a(t) = Y_a^-(t)$ or $X_a(t) = Y_a^+(t)$, then $C(\xi)$ can be replaced by the smaller $C(\xi) = \zeta(2)/(\ln(1+\xi))^2$: the factor 2 that is removed corresponds to picking a one-sided versus a two-sided prior and leads to the slightly smaller threshold $\tilde{\mathcal{C}}_{\text{exp}}$ given in the second statement of Theorem 7.

Using Lemma 4, we obtain the following deviation inequality expressed with the function C^{g_ξ} associated to g_ξ (see Definition 3): for all $\xi > 0$, for all $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} X_a(t) \geq |\mathcal{S}| C^{g_\xi}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

Recalling that $C^{g_\xi}(x) = \min_{\lambda \in [0, 1/(1+\xi))} \frac{x + g_\xi(\lambda)}{\lambda}$, the proof is completed by applying Lemma 12 below, proved in Appendix B.2, to compute the optimal tuning of $\xi \in [0, 1/2]$.

Lemma 12 *Let $C(\xi) = \frac{2\zeta(2)}{(\ln(1+\xi))^2}$. Fix $z \in [0, e-1]$ and $x \geq 0$. Then*

$$\inf_{\substack{\xi \in [0,z] \\ \lambda \in [0, 1/(1+\xi)]}} \frac{x - \ln(1 - \lambda(1+\xi))}{\lambda} + (1+\xi)\ln C(\xi) \;=\; 2\tilde{h}_{1+z}\left(\frac{h^{-1}(1+x) + \ln(2\zeta(2))}{2}\right).$$

□

We now provide the proof of the crucial Lemma 11.

Proof of Lemma 11: building the martingale Lemma 13 below shows that the deviations of $X_a(t)$ can be related to the deviations of a well-chosen mixture martingale $Z_a^\pi(t)$, where π has a discrete support. The proof of Lemma 13 is given in Appendix B.

Lemma 13 (mixture martingales) *Fix $\xi \in]0, 1/2[$ and $x > 0$. There exists a (discrete) prior $\pi(x) = \pi(x, \xi)$ such that the corresponding mixture martingale (8), denoted by $Z_a^{\pi(x)}(t)$, satisfies, for all $t \in \mathbb{N}$,*

$$\left\{X_a(t) - (1+\xi)\ln\left(\frac{2\zeta(2)}{(\ln(1+\xi))^2}\right) \geq x\right\} \subseteq \left\{Z_a^{\pi(x)}(t) \geq e^{\frac{x}{1+\xi}}\right\}.$$

If $X_a(t) = Y_a^+(t)$ or $X_a(t) = Y_a^-(t)$, there exists a prior $\pi(x)$ such that

$$\left\{X_a(t) - (1+\xi)\ln\left(\frac{\zeta(2)}{(\ln(1+\xi))^2}\right) \geq x\right\} \subseteq \left\{Z_a^{\pi(x)}(t) \geq e^{\frac{x}{1+\xi}}\right\}. \tag{15}$$

A consequence of Lemma 13 is that, for every $z > 1$, and every $\lambda > 0$

$$\begin{aligned}\left\{e^{\lambda(X_a(t) - (1+\xi)\ln C(\xi))} \geq z\right\} \;&\subseteq\; \left\{Z_a^{\pi(\ln(z)/\lambda)}(t) \geq e^{\frac{\ln(z)}{\lambda(1+\xi)}}\right\} \\ &\subseteq\; \Big\{\underbrace{Z_a^{\pi(\ln(z)/\lambda)}(t) e^{-\frac{\ln(z)}{\lambda(1+\xi)}}}_{:= W_a^{z,\lambda}(t)} \geq 1\Big\},\end{aligned}$$

where $W_a^{z,\lambda}(t)$ is a martingale that satisfies $\mathbb{E}[W_a^{z,\lambda}(0)] = e^{-\frac{\ln(z)}{\lambda(1+\xi)}}$ and, due to the above inclusion,

$$W_a^{z,\lambda}(t) \geq \mathbb{1}_{\left(e^{\lambda(X_a(t)-(1+\xi)\ln C(\xi))} \geq z\right)}. \tag{16}$$

We now define another mixture martingale, for $\lambda \in \left]0, \frac{1}{1+\xi}\right[$:

$$W_a^{\lambda}(t) = 1 + \int_1^{\infty} W_a^{z,\lambda}(t) dz.$$

Using inequality (16) yields

$$W_a^{\lambda}(t) \geq e^{\lambda(X_a(t)-(1+\xi)\ln C(\xi))}.$$

Moreover, a direct computation shows that $W_a^{\lambda}(0) = \frac{1}{1-\lambda(1+\xi)}$. Finally defining

$$M_a^{\lambda}(t) = (1-\lambda(1+\xi))W_a^{\lambda}(t),$$

one has that $M_a^{\lambda}(t)$ is a test martingale, i.e. $\mathbb{E}[M_a^{\lambda}(t)] = 1$, that satisfies

$$\begin{aligned} M_a^{\lambda}(t) &\geq \exp\left(\lambda X_a(t) - \lambda(1+\xi)\ln(C(\xi)) + \ln(1-\lambda(1+\xi))\right) \\ &= \exp\left(\lambda X_a(t) - g_{\xi}(\lambda)\right), \end{aligned}$$

which concludes the proof. Note that if for all a, $X_a(t) = Y_a^{\pm}(t)$, using the tighter statement (15) allows to replace the constant $C(\xi)$ by the smaller value $\frac{\zeta(2)}{(\ln(1+\xi))^2}$.

Above, we are in essence building a test martingale of value $M_t \geq e^{\lambda X_t}$ from test martingales guaranteeing $Z_t \geq e^x \mathbb{1}\{X_t \geq x\}$. The possibilities and limits of doing this are exactly characterised by Dawid et al. (2011) in the process of characterising the so-called *admissible capital calibrators*. By changing the mixture on thresholds x from exponential (as we do here) to polynomial, it is technically possible to guarantee $M_t \geq e^{X_t - O(\ln X_t)}$. We do not pursue this direction, as the additional $\ln X_t$ is not convenient for combining evidence of arms, and moreover it is not at all clear that the high cost in terms of multiplicative constants (i.e. the $g(\lambda)$) is worth it.

□

3.4.1 ALTERNATIVE MIXTURE MARTINGALES

The martingale $M_a^{\lambda}(t)$ built in the proof of Lemma 11 can be viewed as a mixture martingale with a hierarchical prior, which is a continuous mixture of some discrete priors $\pi(\cdot)$ defined in Lemma 13. Indeed, one can write

$$M_a^{\lambda}(t) = (1-\lambda(1+\xi))Z_a^0(t) + \int_1^{\infty} \left(\int Z_a^{\eta}(t)\pi(\ln(z)/\lambda)(d\eta) \right) (1-\lambda(1+\xi))e^{-\frac{\ln(z)}{\lambda(1+\xi)}} dz.$$

To prove Theorem 9 and Theorem 10, we build different mixture martingales in Appendix C. Interestingly, they also rely on a hierarchical prior but this time the prior is a discrete average of continuous priors. More precisely, the martingale used in each case can be written of the form

$$M_a^{\lambda}(t) = \sum_{i=1}^{\infty} \gamma_i \int Z_a^{\eta}(t) p_{T_i}^{\lambda,\mu_a}(\eta) d\eta$$

for a well chosen family of $(\gamma_i, T_i) \in (\mathbb{R}^+, \mathbb{N}^*)$, where $p_t^{\lambda,\mu_a}(\eta)$ is a continuous function satisfying

$$\forall x, \int e^{\eta(tx)-\phi_\mu(\eta)t} p_n^{\lambda,\mu_a}(\eta) d\eta = e^{\lambda t d(x,\mu_a)}.$$

We see that in both cases, elementary non-negative martingales of the form $Z_a^\eta(t)$ from (7) are mixed under a (hierarchically constructed) prior distribution on η. Both approaches are similar in spirit, both implementing the Laplace technique of achieving a value close to that of the maximiser $\hat{\eta}$ (which is a function of the random data), and the peeling technique (to adapts to the random sample size $N_a(t)$). The combined density has an asymptote at $\eta = 0$, with density nearby proportional to $\frac{1}{\eta(\ln \eta)^2}$. The log factor is necessary for making the prior proper, and it is also the reason for the $\ln \ln N_a(t)$ terms in our deviation inequality.

An interesting alternative approach, which goes slightly outside the mixture martingale framework, is taken by Koolen and van Erven (2015). There, a mixture martingale is constructed by mixing the increment $Z_a^\eta(t) - Z_a^\eta(0)$ under the *improper* prior with density $1/\eta$. Subtracting $Z_a^\eta(0)$ solves the problem that, without it, the improper mixture would be infinite. However, it has the effect that the mixture can become negative, interfering with Ville's inequality. Yet the mixture value can be shown (for bounded outcomes) to be at least $-\ln N_a(t)$. Taking that into account properly yields, again, the $\ln \ln N_a(t)$ term in the resulting deviation inequality. We believe exploring these ideas further to be a worthwhile direction for future research.

4. Asymptotically Optimal Adaptive Sequential Testing

In this section, we explain how our new deviation inequalities can be useful to prove the correctness of a stopping strategy for generic sequential adaptive hypothesis testing problems, that we refer to as *sequential identification problems*.

Given a bandit model, we consider M hypotheses $\mathcal{H}_1 = (\boldsymbol{\mu} \in \mathcal{O}_1), \dots, \mathcal{H}_M = (\boldsymbol{\mu} \in \mathcal{O}_M)$ where $\mathcal{O}_1, \dots, \mathcal{O}_M$ are open sets forming a partition of the set of possible means $\mathcal{O}$. Our goal is to adaptively sample the arms until a decision is made that one of the hypotheses $\hat{\imath}$ is correct. Our goal is to identify the correct hypothesis for all possible means $\boldsymbol{\mu} \in \mathcal{O}$. More precisely, we aim for *δ-correct strategies*, for which $\forall \boldsymbol{\mu} \in \mathcal{O}$, $\mathbb{P}_{\boldsymbol{\mu}}(\boldsymbol{\mu} \in \mathcal{O}_{\hat{\imath}}) \geq 1 - \delta$. This problem falls into the framework of Sequential Adaptive Hypothesis Testing as introduced by Chernoff (1959) –who studied only discrete hypotheses and considered a different performance metric– and is called General-Samp by Chen et al. (2017), who study Gaussian arms with unit variance.

For general exponential family bandits, we analyse below a natural stopping rule based on Generalized Likelihood Ratio (GLR) tests. We prove that this stopping rule is δ-correct for any sequential identification problem and that in some cases it attains the minimal sample complexity (in a regime of small risk δ) when coupled with an appropriate sampling rule. We note that the independent work of Juneja and Krishnasamy (2019) studies the same problem as ours under the name "partition identification", also for exponential families. However, that work puts less emphasis on stopping rules, and uses the deviation inequality of Magureanu et al. (2014) for its analysis.

4.1 A General Stopping Rule

For every $\boldsymbol{\mu}$, we define

$$\mathrm{Alt}(\boldsymbol{\mu}) = \bigcup_{i:\boldsymbol{\mu} \notin \mathcal{O}_i} \mathcal{O}_i.$$

If $\boldsymbol{\mu} \in \mathcal{O}$, we let $i^*(\boldsymbol{\mu})$ be the index of the unique element in the partition to which $\boldsymbol{\mu}$ belongs; in particular $\boldsymbol{\mu} \in \mathcal{O}_{i^*(\boldsymbol{\mu})}$ and $\text{Alt}(\boldsymbol{\mu}) = \mathcal{O} \backslash \mathcal{O}_{i^*(\boldsymbol{\mu})}$. We let $\hat{\boldsymbol{\mu}}(t)$ be the vector of empirical means of the arms based on the observations available up to round t. If $\hat{\boldsymbol{\mu}}(t) \in \mathcal{O}$, we let $\hat{\imath}(t) = i^*(\hat{\boldsymbol{\mu}}(t))$ so that $\hat{\boldsymbol{\mu}}(t) \in \mathcal{O}_{\hat{\imath}(t)}$.

Definition 14 *The GLR statistic is defined as*

$$\hat{\Lambda}_t = \inf_{\boldsymbol{\lambda} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t) d\left(\hat{\mu}_a(t), \lambda_a\right). \tag{17}$$

Given a sequence of thresholds $(\hat{c}_t(\delta))_{t\in\mathbb{N}}$, the GLR stopping rule with thresholds $\hat{c}_t(\delta)$ is defined by

$$\tau_\delta := \inf\left\{ t \in \mathbb{N} : \hat{\Lambda}_t > \hat{c}_t(\delta) \right\}. \tag{18}$$

A Generalized Likelihood Ratio statistic is usually defined for testing a possibly composite hypothesis $\mathcal{H}_0 : (\mu \in \Omega_0)$ against a possibly composite alternative $\mathcal{H}_1 : (\mu \in \Omega_1)$ by

$$R_t = \frac{\sup_{\lambda \in \Omega_0 \cup \Omega_1} \ell(X_1, \dots, X_t; \lambda)}{\sup_{\lambda \in \Omega_0} \ell(X_1, \dots, X_t; \lambda)},$$

where $X_1, \dots, X_t$ are some observations whose likelihood $\ell(X_1, \dots, X_t; \mu)$ depends on some unknown parameter μ. Large values of R_t tend to reject the hypothesis $\mathcal{H}_0$. When the observations are obtained under a sampling rule (A_t) in an exponential family bandit model and $\hat{\boldsymbol{\mu}}(t) \in \Omega_0 \cup \Omega_1$ it can be shown that

$$\ln(R_t) = \inf_{\boldsymbol{\lambda} \in \Omega_0} \sum_{a=1}^{K} d(\hat{\mu}_a(t), \lambda_a).$$

The GLR statistic $\hat{\Lambda}_t$ can thus be interpreted as a statistic for testing $\mathcal{H}_0 : \{\boldsymbol{\mu} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t))\}$ against $\mathcal{H}_1 : \left\{\boldsymbol{\mu} \in \mathcal{O}_{\hat{\imath}(t)}\right\}$ (if $\hat{\boldsymbol{\mu}}(t) \in \mathcal{O}$, otherwise note that $\hat{\Lambda}_t = 0$ which prevent from stopping). However the two hypotheses that are "tested" at time t are data-dependent. Still, large values $\hat{\Lambda}_t$ tend to reject $(\boldsymbol{\mu} \in \text{Alt}(\hat{\boldsymbol{\mu}}(t)))$: hypothesis $\hat{\imath}(t)$ must be true. Another possible interpretation of the GLR stopping rule is that it is running in parallel M GLR tests of $\mathcal{H}_0 : (\boldsymbol{\mu} \in \mathcal{O} \backslash \mathcal{O}_i)$ against $\mathcal{H}_1 : (\boldsymbol{\mu} \in \mathcal{O}_i)$ for $i = 1, \dots, M$ and stops the first time one of these tests $\hat{\imath}$ rejects $\mathcal{H}_0$. This "Parallel GLRT" view is the one discussed for example by Garivier and Kaufmann (2021).

It can be observed that $\left\{\hat{\Lambda}_t > \hat{c}_t(\delta)\right\} = \left\{\mathcal{C}_t(\delta) \subseteq \mathcal{O}_{\hat{\imath}(t)}\right\}$ where $\mathcal{C}_t(\delta)$ is the *confidence region*

$$\mathcal{C}_t(\delta) := \left\{ \boldsymbol{\lambda} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) \leq \hat{c}_t(\delta) \right\}. \tag{19}$$

The GLR stopping rule (18) can thus be rephrased in the following way: stop when the set of statistically plausible parameters $\mathcal{C}_t(\delta)$ is included in one fold of the partition. Building on Theorem 7, Proposition 15 below provides a choice of thresholds for which the GLR stopping rule yields a δ-correct algorithm. We provide a choice of thresholds for which the GLR rule is δ-correct when the hypothesis $\mathcal{H}_{\hat{\imath}(\tau)}$ is recommended and the corresponding confidence intervals $\mathcal{C}_t(\delta)$ always contain the true parameter with probability larger than $1 - \delta$.

Proposition 15 *Let $\mathcal{C}_{\exp}$ be the threshold function defined in Theorem 7. The sequence of thresholds*

$$\hat{c}_t(\delta) = 3\sum_{a=1}^{K} \ln(1 + \ln N_a(t)) + K\mathcal{C}_{\exp}\left(\frac{\ln(1/\delta)}{K}\right) \tag{20}$$

is such that, for every sampling rule*,*

$$\mathbb{P}_{\boldsymbol{\mu}}(\forall t \in \mathbb{N}, \boldsymbol{\mu} \in \mathcal{C}_t(\delta)) \geq 1 - \delta \quad \textit{and} \quad \mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty, \hat{\imath}(\tau_\delta) \neq i^*(\boldsymbol{\mu})\right) \leq \delta.$$

Proof Using Theorem 7 in the last inequality, one can write

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu}}\left(\tau < \infty, \hat{\imath}(\tau) \neq i^*\right) &\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \hat{\imath}(t) \neq i^*, \hat{\Lambda}_t > \hat{c}_t(\delta)\right) \\
&= \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \exists i \neq i^*, \mathcal{C}_t \subseteq \mathcal{O}_i\right) \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \boldsymbol{\mu} \notin \mathcal{C}_t\right) \\
&= \mathbb{P}_{\boldsymbol{\mu}}\left(\exists t \in \mathbb{N} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \hat{c}_t(\delta)\right) \\
&\leq \delta.
\end{aligned}$$

This proves both claims of Proposition 15. ■

4.2 An Asymptotically Optimal Adaptive Testing Procedure

Proposition 15 provides a threshold for which the GLR stopping rule (18) is δ-correct *for any sampling rule*. We now show that used in conjunction with an appropriate "Tracking" stopping rule, it can even attain the optimal sample complexity. The following lower bound generalizes the sample complexity lower bound obtained by Garivier and Kaufmann (2016) for the particular Best Arm Identification problem and is obtained with the exact same change-of-measure technique.

Proposition 16 *Define the complexity term $T^*(\boldsymbol{\mu})$ as*

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{w} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a),$$

where $\Sigma_K = \left\{\boldsymbol{w} \in [0,1]^K : \sum_{i=1}^K w_i = 1\right\}$. Then any δ-correct strategy satisfies

$$\forall \boldsymbol{\mu} \in \mathcal{O},\ \mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right).$$

We define, when they exist (that is, when the argmax below is unique) the *optimal weights*

$$\boldsymbol{w}^*(\boldsymbol{\mu}) := \operatorname*{argmax}_{\boldsymbol{w} \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a) \tag{21}$$

for $\boldsymbol{\mu} \in \mathcal{O}$. For well-behaved sequential testing problems, those weights indicate the fraction of samples that should be allocated to each arm by an optimal strategy. This motivates the Tracking

rule, originally proposed by Garivier and Kaufmann (2016) as the D-Tracking rule for Best Arm Identification and that we recall here. Letting $\mathcal{U}_t = \{a \in \{1, \dots, K\} : N_a(t) \leq \max(\sqrt{t} - K/2, 0)\}$ be the set of under-sampled arms, at time $t+1$ the selected arm is

$$A_{t+1} \in \begin{cases} \operatorname*{argmin}_{a \in \mathcal{U}_t} N_a(t) & \text{if } \mathcal{U}_t \neq \emptyset \quad (\textit{forced exploration}) \\ \operatorname*{argmax}_{a \in [K]} t\, w_a^*(\hat{\boldsymbol{\mu}}(t)) - N_a(t) & \text{o.w.} \qquad (\textit{tracking the plug-in estimate}) \end{cases} \tag{22}$$

It can be noted that $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$ is defined only if $\hat{\boldsymbol{\mu}}(t) \in \mathcal{O}$. In practice if $\hat{\boldsymbol{\mu}}(t) \notin \mathcal{O}$ the tracking step of the algorithm can be replaced by uniform exploration. Due to the forced exploration, as $\boldsymbol{\mu} \in \mathcal{O}$ (which is an open set by assumption) the law of large numbers ensures that at some point $\hat{\boldsymbol{\mu}}(t) \in \mathcal{O}$, and the tracking step can be applied.

Theorem 17 *Assume that the following assumptions are satisfied:*

1. *For every $\boldsymbol{\mu} \in \mathcal{O}$, there is a unique vector of optimal weights $\boldsymbol{w}^*(\boldsymbol{\mu})$*
2. *For all $i \in \{1, \dots, M\}$, the mapping $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ is continuous on $\mathcal{O}_i$.*

For $\delta \in (0, 1]$ let $\hat{c}_t(\delta)$ be a deterministic sequence of thresholds that is increasing in t and for which there exists constants $C, D > 0$ such that

$$\forall t \geq C, \forall \delta \in (0, 1],\ \hat{c}_t(\delta) \leq \ln\left(\frac{Dt}{\delta}\right).$$

Let τ_δ be the GLR stopping rule (18) with thresholds $\hat{c}_t(\delta)$. The Tracking rule (22) ensures

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} = T^*(\boldsymbol{\mu}).$$

The proof of Theorem 17 is given in Appendix D. Combining this result with Proposition 15 yields that an adaptive sequential test using the Tracking rule and the GLR stopping rule with thresholds (20) is δ-correct for every $\delta \in (0, 1]$ and its sample complexity is asymptotically matching the lower bound of Proposition 16, provided that the optimal weights $\boldsymbol{w}^*(\boldsymbol{\mu})$ are well defined and continuous in $\boldsymbol{\mu}$.

Efficient ways to compute those weights are also needed for the actual implementation of the Tracking rule. In the next section, we will discuss particular examples of adaptive sequential tests in which those requirements are fulfilled and optimal (and efficient) adaptive testing is thus possible. We will see that smaller thresholds than the universal threshold (20) can be used in some cases.

The assumptions of Theorem 17 are frequently satisfied for practical problems (see also Combes et al. 2017, Lemma 1 proving continuity of the highly related oracle regret problem for the structured multi-armed bandit problem in the fixed-budget setting, under a unique optimiser assumption similar to 1). Uniqueness may however fail for other practical problems, including e.g. the Minimum Threshold problem studied by Kaufmann et al. (2018). A solution for such cases was proposed in recent follow-up work by Degenne and Koolen (2019), who propose regarding the oracle weight map $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ from (21) as set-valued, and prove that it is upper hemi-continuous and convex-valued for every sequential identification problem of the form we consider here (in particular with a unique correct answer for each instance). Leveraging these two properties, they analyse a variation of the Tracking rule (22) for which the overall approach is asymptotically optimal in general (Degenne and Koolen, 2019, Theorem 7).

5. Smaller Thresholds for Better Sequential Tests

A stylized form of (two-sided) deviation inequalities obtained in this paper (in Corollaries 9 and 10 and Theorem 7) is the following. For any subset of arms $\mathcal{S} \subseteq \{1, \dots, K\}$, for all $x > 0$,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \sum_{a \in \mathcal{S}} c \ln(d + \ln(N_a(t))) + |\mathcal{S}| \mathcal{C}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x} \tag{23}$$

where c and d are two positive constants and $\mathcal{C}(x)$ is a threshold function. This result holds *for any subset of arms* $\mathcal{S}$. Combining (23) with a weighted union bound, one obtains in Lemma 18 below a deviation inequality that is uniform over subsets belonging to the support of a weight vector $\tilde{\pi}$.

Lemma 18 (weighted union bound) *Assume* (23) *holds. Let* $\tilde{\pi}$ *be a probability distribution over subsets:* $\sum_{\mathcal{S} \subseteq \{1,\dots,K\}} \tilde{\pi}(\mathcal{S}) = 1$. *Then for all* $x > 0$

$$\mathbb{P}\left(\exists t, \exists \mathcal{S} : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \sum_{a \in \mathcal{S}} c \ln(d + \ln(N_a(t))) + |\mathcal{S}| \mathcal{C}\left(\frac{x - \ln(\tilde{\pi}(\mathcal{S}))}{|\mathcal{S}|}\right)\right) \leq e^{-x}.$$

Proof A union bound followed by inequality (23) gives

$$\begin{aligned}
&\mathbb{P}\left(\exists t, \exists \mathcal{S} : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \sum_{a \in \mathcal{S}} c \ln(d + \ln(N_a(t))) + |\mathcal{S}| \mathcal{C}\left(\frac{x - \ln(\tilde{\pi}(\mathcal{S}))}{|\mathcal{S}|}\right)\right) \\
&\leq \sum_{\mathcal{S} \subseteq \{1,\dots,K\}} \mathbb{P}\left(\exists t : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \sum_{a \in \mathcal{S}} c \ln(d + \ln(N_a(t))) + |\mathcal{S}| \mathcal{C}\left(\frac{x - \ln(\tilde{\pi}(\mathcal{S}))}{|\mathcal{S}|}\right)\right) \\
&\leq \sum_{\mathcal{S} \subseteq \{1,\dots,K\}} e^{-(x - \ln(\tilde{\pi}(\mathcal{S})))} = e^{-x} \sum_{\mathcal{S} \subseteq \{1,\dots,K\}} \tilde{\pi}(\mathcal{S}) = e^{-x}.
\end{aligned}$$

■

We now explain how this result can serve to tighten the analysis of the GLR stopping rule for some particular sequential testing problems, to allow for the use of smaller threshold functions. We later discuss in Section 6 the impact of this result on the design of confidence regions.

5.1 Improved Stopping Rules for Best Arm Identification

The (fixed-confidence) Best Arm Identification problem is a particular sequential identification problem as defined in Section 4 with $\mathcal{O}_k = \{\boldsymbol{\mu} : \mu_k > \max_{j \neq k} \mu_j\}$: the goal is to identify the arm with largest mean. For this particular problem, the GLR statistic (17) rewrites to

$$\hat{\Lambda}_t = \min_{b \neq \hat{\imath}(t)} \min_{\lambda_b > \lambda_{\hat{\imath}(t)}} \left[N_{\hat{\imath}(t)}(t) d(\hat{\mu}_{\hat{\imath}(t)}(t), \lambda_{\hat{\imath}(t)}) + N_b(t) d(\hat{\mu}_b(t), \lambda_b)\right] \tag{24}$$

and the associated stopping rule, $\hat{\Lambda}_t > \hat{c}_t(\delta)$, is referred to as the Chernoff stopping rule by Garivier and Kaufmann (2016). In this particular case, it is possible to propose a smaller threshold than the universal threshold (20) that still ensures a δ-correct rule. Indeed, the probability of error of the

strategy that stops when $\hat{\Lambda}_t > \hat{c}_t(\delta)$ and outputs $\hat{\imath}(\tau)$ is upper bounded as follows, assuming arm 1 is the arm with largest mean:

$$\begin{aligned}\mathbb{P}(\text{error}) &\leq \mathbb{P}\left(\exists t \in \mathbb{N}, \exists a \neq 1 : \min_{\lambda_a > \lambda_1} [N_a(t) d(\hat{\mu}_a(t), \lambda_a) + N_1(t) d(\hat{\mu}_1(t), \lambda_1)] > \hat{c}_t(\delta)\right) \\ &\leq \mathbb{P}\left(\exists t \in \mathbb{N}, \exists a \neq 1 : N_a(t) d(\hat{\mu}_a(t), \mu_a) + N_1(t) d(\hat{\mu}_1(t), \mu_1) > \hat{c}_t(\delta)\right) \\ &= \mathbb{P}\left(\exists t, \exists a \neq 1 : \sum_{j \in \{1,a\}} N_j(t) d(\hat{\mu}_j(t), \mu_j) > \hat{c}_t(\delta)\right).\end{aligned}$$

From Theorem 7 and a union bound over the $K-1$ subsets $\{1,2\}, \ldots, \{1,K\}$ (Lemma 18 with a weight vector such that $\tilde{\pi}(\{1,a\}) = 1/(K-1)$ for $a \neq 1$) it holds that

$$\mathbb{P}\left(\exists t, \exists a \neq 1 : \sum_{j \in \{1,a\}} N_j(t) d(\hat{\mu}_j(t), \mu_j) > 3 \sum_{j \in \{1,a\}} \ln(1 + \ln(N_j(t))) + 2\mathcal{C}_{\text{exp}}\left(\frac{\ln \frac{K-1}{\delta}}{2}\right)\right) \leq \delta.$$

This implies that the GLR rule is δ-correct with the threshold

$$\hat{c}_t(\delta) = 6 \ln\left(\ln\left(\frac{t}{2}\right) + 1\right) + 2\mathcal{C}_{\text{exp}}\left(\frac{\ln \frac{K-1}{\delta}}{2}\right). \tag{25}$$

For large t, this will be smaller than the original threshold $\hat{c}_t(\delta) = \ln \frac{2t(K-1)}{\delta}$ proposed by Garivier and Kaufmann (2016) in the Bernoulli case. It can hence lead to earlier stopping while preserving the optimal sample complexity guarantees, as this threshold still satisfies the assumptions of Theorem 17. Note also that this new threshold provides a better motivation for the stylized $\ln((\ln(t)+1)/\delta)$ threshold that is sometimes used in best arm identification experiments, and for which the empirical error probability is reported to remain below δ.

Remark 19 *The improved threshold (25) yields a δ correct stopping rule, however the corresponding confidence interval (19) does not satisfy $\mathbb{P}(\forall t \in \mathbb{N} : \boldsymbol{\mu} \in \mathcal{C}_t(\delta)) \geq 1-\delta$. There is no equivalence between the improved δ-correct stopping rule and improved δ-valid confidence regions. We will discuss the implications of Lemma 18 for confidence regions in Section 6.*

5.2 Smaller Thresholds for More General Tests

The reason why we are able to propose a smaller threshold for the BAI problem is that its GLR statistic (24) only features pairs of arms. In more general tests, the structure of the GLR statistic may also be exploited to allow for a smaller threshold that does not depend on the total number of arms K featuring in the universal threshold (20) but on a smaller *effective number* of arms.

Definition 20 *Consider a sequential identification problem specified by a partition $\mathcal{O} = \bigcup_{i=1}^{M} \mathcal{O}_i$. We say this problem has* rank R *if for every $i \in \{1, \ldots, M\}$ we can write*

$$\mathcal{O} \backslash \mathcal{O}_i = \bigcup_{q \in [Q]} \left\{ \boldsymbol{\lambda} \in \mathcal{I}^K \middle| (\lambda_{k_1^{i,q}}, \ldots, \lambda_{k_R^{i,q}}) \in \mathcal{L}_{i,q} \right\},$$

for a family of arm indices $k_r^{i,q} \in [K]$ and open sets $\mathcal{L}_{i,q}$ indexed by $r \in [R]$, $q \in [Q]$ and $i \in [M]$. In words, the rank is R if every set $\mathcal{O} \setminus \mathcal{O}_i$ is a finite union of sets that are each defined in terms of only R arms.

The BAI problem has rank 2. Indeed, for all $i \in \{1, \dots, K\}$,

$$\mathcal{O} \backslash \mathcal{O}_i = \bigcup_{a \neq i} \left\{ \boldsymbol{\lambda} \in \mathcal{I}^K \,|(\lambda_i, \lambda_a) \in \{(x,y) : x < y\} \right\}.$$

In any testing problem that has rank R, the GLR statistic may be rewritten

$$\hat{\Lambda}_t \;=\; \min_{q \in [Q]} \inf_{\substack{\boldsymbol{\lambda} \\ (\lambda_{k_1^{\hat{\imath}(t),q}}, \dots, \lambda_{k_R^{\hat{\imath}(t),q}}) \in \mathcal{L}_{\hat{\imath}(t),q}}} \sum_{r=1}^{R} N_{k_r^{\hat{\imath}(t),q}}(t) d\left(\hat{\mu}_{k_r^{\hat{\imath}(t),q}}(t), \lambda_{k_r^{\hat{\imath}(t),q}}\right),$$

which yields the expression (24) in the BAI case.

Proposition 21 *Fix an identification problem of rank R. Then the GLR stopping rule (17) is δ-correct with threshold*

$$\hat{c}_t(\delta) \;=\; 3R \ln(1 + \ln(t/R)) + R\mathcal{C}_{exp}\left(\frac{\ln \frac{M-1}{\delta}}{R}\right)$$

Proof Fix $\boldsymbol{\mu} \in \mathcal{O}$. For each $i \neq i^*$, $\boldsymbol{\mu} \in \mathcal{O} \backslash \mathcal{O}_i$, thus from Definition 20 there exists q_i such that $(\mu_{k_1^{i,q_i}}, \dots, \mu_{k_R^{i,q_i}}) \in \mathcal{L}_{i,q_i}$. Then

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}\left\{\tau_\delta < \infty \text{ and } \hat{\imath}(\tau_\delta) \neq i^*\right\} \\
&\leq\; \mathbb{P}_{\boldsymbol{\mu}}\left\{\exists t : \hat{\Lambda}_t \geq \hat{c}_t(\delta) \text{ and } \hat{\imath}(t) \neq i^*\right\} \\
&=\; \mathbb{P}_{\boldsymbol{\mu}}\left\{\exists t, i \neq i^* : \hat{\Lambda}_t \geq \hat{c}_t(\delta) \text{ and } \hat{\imath}(t) = i\right\} \\
&=\; \mathbb{P}_{\boldsymbol{\mu}}\left\{\exists t, i \neq i^* : \min_{q \in [Q]} \inf_{\substack{\boldsymbol{\lambda} \\ (\lambda_{k_1^{i,q}}, \dots, \lambda_{k_R^{i,q}}) \in \mathcal{L}_{i,q}}} \sum_{r=1}^{R} N_{k_r^{i,q}}(t) d\left(\hat{\mu}_{k_r^{i,q}}(t), \lambda_{k_r^{i,q}}\right) \geq \hat{c}_t(\delta)\right\} \\
&\leq\; \mathbb{P}_{\boldsymbol{\mu}}\left\{\exists t, i \neq i^* : \sum_{r=1}^{R} N_{k_r^{i,q_i}}(t) d\left(\hat{\mu}_{k_r^{i,q_i}}(t), \mu_{k_r^{i,q_i}}\right) \geq 3R \ln(1 + \ln(t/R)) + \mathcal{C}_{\text{exp}}\left(\frac{\ln \frac{M-1}{\delta}}{R}\right)\right\} \\
&\leq\; \mathbb{P}_{\boldsymbol{\mu}}\left\{\exists t, i \neq i^* : \sum_{r=1}^{R} N_{k_r^{i,q_i}}(t) d\left(\hat{\mu}_{k_r^{i,q_i}}(t), \mu_{k_r^{i,q_i}}\right) \geq 3 \sum_{r=1}^{R} \ln\left(1 + \ln N_{k_r^{i,q_i}}(t)\right) + \mathcal{C}_{\text{exp}}\left(\frac{\ln \frac{M-1}{\delta}}{R}\right)\right\} \\
&\leq\; \delta,
\end{aligned}$$

where the last inequality follows from Theorem 7 and a union bound over $M-1$ subsets (Lemma 18 with a weight vector $\tilde{\pi}(\{k_1^{i,q_i}, \dots, k_R^{i,q_i}\}) = 1/(M-1)$ for $i \neq i^*$) together with the concavity of $s \mapsto \ln(1 + \ln(s))$ that ensures

$$\sum_{r=1}^{R} \ln\left(1 + \ln N_{k_r^{i,q_i}}(t)\right) \;\leq\; R \ln(1 + \ln(t/R)).$$

■

5.2.1 A RANK 4 EXAMPLE

Assume we are given a collection of K pairs of arms and want to find out which pair has the largest difference (which we think of as profit) between first component (which we think of as revenue) and second component (which we think of as cost). More precisely, we consider a $K \times 2$ array of random sources X_{ij} where $i \in [K]$ and $j \in \{1,2\}$. Let $\mu_{ij} = \mathbb{E}[X_{ij}]$ denote the means. A strategy samples one arm $A_t = (I_t, J_t)$ per round and its goal is to identify the largest profit pair

$$i^*(\boldsymbol{\mu}) \;=\; \arg\max_i \; \mu_{i,1} - \mu_{i,2}.$$

It is easy to check that this problem, which we call *Largest Profit Identification*, has rank 4 and the GLR statistic rewrites to

$$\hat{\Lambda}_t \;=\; \min_{b \neq \hat{\imath}} \; \inf_{\substack{\boldsymbol{\lambda} \in \mathbb{R}^{\{b,\hat{\imath}\}\times\{1,2\}} \\ \lambda_{b,1}-\lambda_{b,2} > \lambda_{\hat{\imath},1}-\lambda_{\hat{\imath},2}}} \; \sum_{\substack{a \in \{b,\hat{\imath}\} \\ j \in \{1,2\}}} N_{a,j}(t) d\left(\hat{\mu}_{a,j}(t), \lambda_{a,j}\right).$$

By Proposition 21 the GLR stopping rule (18) is δ-correct with the threshold

$$\hat{c}_t(\delta) \;=\; 12\ln(1+\ln(t/4)) + 4\mathcal{C}_{\text{exp}}\left(\frac{\ln\frac{K-1}{\delta}}{4}\right).$$

Remark 22 *For Largest Profit Identification the oracle weights $\boldsymbol{w}^*(\boldsymbol{\mu})$, which are needed for implementing the asymptotically optimal procedure of Section 4.2, maximise the concave function $T^*(\boldsymbol{\mu})^{-1}$. For both Gaussian and Bernoulli (and possibly more) we can write the objective as a Disciplined Convex Program and solve it efficiently with e.g. CVX (Grant and Boyd, 2017).*

5.2.2 BEST ACTION IDENTIFICATION IN A GAME TREE

In the bandit literature, a particular structured identification problem that offers a simple model for Monte Carlo Tree Search in games has been recently studied by Teraoka et al. (2014); Garivier et al. (2016); Huang et al. (2017); Kaufmann and Koolen (2017). The goal is to quickly identify the action at the root of a (maxmin) game tree whose value is the largest by querying noisy samples of the leaves' values of that tree.

Lemma 8 in Kaufmann and Koolen (2017) provides an expression for the optimal weights in a depth-two tree, that are then computable using disciplined convex optimization tools (e.g. CVX). Furthermore, it can be checked that this identification problem is of rank $L+1$, where L is the maximum number of actions of the second player. This is much smaller than the number of leaves, which is $K \cdot L$ in a game tree where the first player has K moves and the second player has L moves. Assuming the weights (which are only numerically computable) satisfy the continuity assumption of Theorem 17 (or, if not, by Degenne and Koolen 2019, Theorem 7), the GLR rule with a rank $L+1$ threshold is asymptotically optimal in combination with the Tracking rule. We note that the existing literature does not provide asymptotically optimal algorithms for best action identification in a game tree, even for depth-two trees.

6. Projected Confidence Intervals

The deviation inequalities presented in this paper can also be used to build tight confidence regions on (functions of) the parameter $\boldsymbol{\mu} \in \mathcal{I}^K$. We are particularly interested in building δ-*uniformly valid* confidence regions $\mathcal{C}_t(\delta)$, that satisfy $\mathbb{P}\left(\forall t \in \mathbb{N}, \boldsymbol{\mu} \in \mathcal{C}_t(\delta)\right) \geq 1-\delta$ for every sampling rule.

Lemma 18 in combination with our deviation results allows to build such confidence regions. Indeed for any weight vector $\tilde{\pi}$ over subsets, the following confidence interval is δ-uniformly valid (with c and d as given by the lemma):

$$\mathcal{C}_t^{\tilde{\pi}}(\delta) := \left\{ \boldsymbol{\lambda} : \forall \mathcal{S}, \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \lambda_a) \leq c \sum_{a \in \mathcal{S}} \ln(d + \ln N_a(t)) + |\mathcal{S}| \mathcal{C}_{\exp}\left(\frac{\ln(1/(\tilde{\pi}(\mathcal{S})\delta))}{|\mathcal{S}|} \right) \right\}. \tag{26}$$

A natural question is thus which vector $\tilde{\pi}$ yields the most interesting confidence region. Answering this question would require to compare complicated shapes in $\mathbb{R}^K$ (like we do for $K = 2$ in Figure 1(a) in the Introduction) and the answer would still depend on the *purpose* of those confidence regions.

In this section we investigate their use for computing confidence intervals on derived quantities of the form $f(\boldsymbol{\mu})$, where $f : \mathbb{R}^K \to \mathbb{R}$ is some fixed function. Knowing that $\boldsymbol{\mu} \in \mathcal{C}_t$, we can immediately conclude that $f(\boldsymbol{\mu}) \in \mathcal{I}_t(\delta) \coloneqq \{f(\boldsymbol{\lambda}) | \boldsymbol{\lambda} \in \mathcal{C}_t\}$. The interplay of the structure of the function f and the shape of the confidence region $\mathcal{C}_t$ will jointly determine the tightness of the *projected confidence interval* $\mathcal{I}_t(\delta)$. The principal challenge is to find, for each f of interest, a statistically tight $\mathcal{C}_t$ with a computationally tractable way of computing $\mathcal{I}_t$. In this section we study two classes of examples, linear f and minima/maxima.

6.1 Linear Functions

In this section we consider an arbitrary linear function $f(\boldsymbol{\mu}) = \boldsymbol{v}^\intercal \boldsymbol{\mu}$ where $\boldsymbol{v} \in \mathbb{R}^K$. We will derive our results in the Gaussian case because it admits revealing and explicit closed-form expressions. In that case the confidence region (26) is δ-uniformly valid for $c = 2$ and $d = 4$ and $g = g_G$, as licensed by Corollary 9. The following two confidence intervals on $\boldsymbol{v}^\intercal \boldsymbol{\mu}$ follow from two extreme choices of weight vectors: one supported on all the singleton sets and one supported on the full set.

Proposition 23 (Box) *The following is a δ-uniformly valid confidence interval on $\boldsymbol{v}^\intercal \boldsymbol{\mu}$*

$$\mathcal{I}_t(\delta) = \left[\boldsymbol{v}^\intercal \hat{\boldsymbol{\mu}}(t) \pm \sum_{a \in [K]} \sqrt{2 \left(C^g \left(\ln \frac{K}{\delta} \right) + c \ln(d + \ln(N_a(t))) \right) \frac{v_a^2}{N_a(t)}} \right].$$

Proof Simple algebra show that $\mathcal{I}_t(\delta) = \{\boldsymbol{v}^T \boldsymbol{\lambda}, \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi}}(\delta)\}$ where $\tilde{\pi}$ is uniform on singletons. Indeed, as $\mathcal{C}_t^{\tilde{\pi}}(\delta)$ is δ-uniformly valid, it holds that for all $t \in \mathbb{N}$ and $a \in [K]$, $|\hat{\mu}_a(t) - \mu_a| \leq \sqrt{\frac{2}{N_a(t)} \left(C^g \left(\ln \frac{K}{\delta} \right) + c \ln(d + \ln(N_a(t))) \right)}$. ∎

Proposition 24 (Ellipse) *The following is a δ-uniformly valid confidence interval on $\boldsymbol{v}^\intercal \boldsymbol{\mu}$*

$$\mathcal{I}_t(\delta) = \left[\boldsymbol{v}^\intercal \hat{\boldsymbol{\mu}}(t) \pm \sqrt{2 \left(K C^g \left(\frac{\ln \frac{1}{\delta}}{K} \right) + \sum_{a \in [K]} c \ln(d + \ln(N_a(t))) \right) \sum_{a \in [K]} \frac{v_a^2}{N_a(t)}} \right].$$

Proof We show that $\mathcal{I}_t(\delta) = \{\boldsymbol{v}^T \boldsymbol{\lambda}, \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi}}(\delta)\}$ where $\tilde{\pi}$ is a point-mass on the whole set: $\tilde{\pi}(\{1, \dots, K\}) = 1$. Letting $C = \sum_{a=1}^K \ln(1 + \ln N_a(t)) + K \mathcal{C}_{\exp}(\ln(1/\delta)/K)$, computing the

upper bound of this confidence interval requires to compute

$$\max_{\boldsymbol{\lambda}} \ \boldsymbol{v}^\intercal \boldsymbol{\lambda} \qquad \text{subject to} \qquad \sum_{a\in[K]} N_a(t) \frac{(\hat{\mu}_a(t)-\lambda_a)^2}{2} \ \leq \ C.$$

Introducing Lagrange multiplier ρ, we find that this is equivalent to

$$\min_{\rho\geq 0} \max_{\boldsymbol{\lambda}} \ \boldsymbol{v}^\intercal \boldsymbol{\lambda} + \rho \left(C - \sum_{a\in[K]} N_a(t) \frac{(\hat{\mu}_a(t)-\lambda_a)^2}{2} \right).$$

Solving for $\boldsymbol{\lambda}$ by cancelling the derivative results in $\lambda_a \ = \ \hat{\mu}_a(t) + \frac{v_a}{\rho N_a(t)}$, asking us to solve

$$\min_{\rho\geq 0} \ \boldsymbol{v}^\intercal \hat{\boldsymbol{\mu}}(t) + \sum_{a\in[K]} \frac{v_a^2}{2\rho N_a(t)} + \rho C \ = \ \boldsymbol{v}^\intercal \hat{\boldsymbol{\mu}}(t) + \sqrt{2C \sum_{a\in[K]} \frac{v_a^2}{N_a(t)}}$$

where zero ρ derivative is found at $\rho \ = \ \sqrt{C^{-1} \sum_{a\in[K]} \frac{v_a^2}{2N_a(t)}}$. As $\min_{\boldsymbol{\lambda}} \boldsymbol{v}^\intercal \boldsymbol{\lambda} = -\max_{\boldsymbol{\lambda}} (-\boldsymbol{v})^\intercal \boldsymbol{\lambda}$, the lower bound of $\mathcal{I}_t(\delta)$ also follows. ■

6.1.1 COMPARISON

The major difference between the two above bounds is the appearance of the sum outside vs inside of the square root. To get more intuition, let's compare in the special case $N_a(t) = t/K$ and approximate $C^g(x) \approx x$. Then we need to compare

$$\|\boldsymbol{v}\|_1 \sqrt{2\left(\ln\frac{K}{\delta} + c\ln(d+\ln(t/K))\right)\frac{K}{t}} \quad \text{and} \quad \|\boldsymbol{v}\|_2 \sqrt{2\left(\ln\frac{1}{\delta} + Kc\ln(d+\ln(t/K))\right)\frac{K}{t}}.$$

We see that the box bound depends on the one-norm of $\boldsymbol{v}$, whereas the ellipse bound depends on the two-norm of $\boldsymbol{v}$, which can be smaller by a factor $\sqrt{K}$ (at the price of a factor K multiplying the $\ln\ln t$ term). In a regime of small δ, the ellipse bound can thus be much better than the box bound. Another case of interest is $N_a(t) = t\frac{|v_a|}{\sum_a |v_a|}$, which result from following the oracle weights $\boldsymbol{w}^*(\boldsymbol{\mu})$. Also here the advantage of ellipse over box can again be as large as a factor $\sqrt{K}$.

6.2 Minimum

We now turn our attention to $f(\boldsymbol{\mu}) = \min_a \mu_a$. Estimating the minimum (or, symmetrically, maximum) mean is a natural task in the multi-armed bandit setting (see Kaufmann et al. 2018). Unlike in the linear case, here the situation is not symmetric. We will study separately the lower and upper confidence bounds

$$\mathrm{L}_t^{\tilde{\pi}}(\delta) = \min\left\{\min_a \lambda_a : \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi},-}(\delta)\right\} \quad \text{and} \quad \mathrm{U}_t^{\tilde{\pi}}(\delta) = \max\left\{\min_a \lambda_a : \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi},+}(\delta)\right\}$$

for the confidence regions

$$\mathcal{C}_t^{\tilde{\pi},\pm}(\delta) = \left\{\boldsymbol{\lambda} : \forall \mathcal{S}, \sum_{a\in\mathcal{S}} \left[N_a(t) d^{\pm}(\hat{\mu}_a(t),\lambda_a) - 3\ln(1+\ln N_a(t))\right]^+ \leq |\mathcal{S}| \mathcal{C}_{\exp}\left(\frac{\ln(\tilde{\pi}(\mathcal{S})/\delta)}{\delta}\right)\right\}$$

that are both δ-uniformly valid by Lemma 18. It follows that $\mathbb{P}\left\{\forall t \in \mathbb{N} : \min_a \mu_a \leq \mathrm{U}_t^{\tilde{\pi}}(\delta)\right\} \geq 1-\delta$ and $\mathbb{P}\left\{\forall t \in \mathbb{N} : \min_a \mu_a \geq \mathrm{L}_t^{\tilde{\pi}}(\delta)\right\} \geq 1-\delta$. We investigate in each case the tightest possible confidence bound that can be obtained by optimising the choice of the weight vector $\tilde{\pi}$.

6.2.1 LOWER CONFIDENCE BOUND

A minimum is low whenever *one* entry is low. This means that the $\boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi},-}$ of lowest mean will have all entries equal to $\hat{\boldsymbol{\mu}}$ except for one. This in turn means that we do not get any mileage out of combining evidence from multiple arms. Instead, the best $\mathrm{L}_t^{\tilde{\pi}}$ is obtained for the choice $\tilde{\pi}(\{k\}) = 1/K$ (uniform distribution on singletons). We find the following.

Proposition 25 *At time t, for each arm a, let $\theta_a(t) \leq \hat{\mu}_a(t)$ be the solution to*

$$N_a(t) d^-(\hat{\mu}_a(t), \theta_a(t)) \;=\; 3\ln(1+\ln(N_a(t))) + \mathcal{C}_{exp}\left(\ln\frac{K}{\delta}\right)$$

(note the left-hand side increases with decreasing $\theta_a(t)$, so the solution can be found by binary search). Then

$$\mathbb{P}\left\{\forall t \in \mathbb{N}, \min_a \mu_a \geq \min_a \theta_a(t)\right\} \geq 1-\delta.$$

Proof With the choice $\tilde{\pi}(\{k\}) = 1/K$, $\mathcal{C}_t^{\tilde{\pi},-}(\delta)$ is the set of $\boldsymbol{\lambda}$:

$$\forall a \in [K] : N_a(t) d^-(\hat{\mu}_a(t), \lambda_a) \;\leq\; 3\ln(1+\ln(N_a(t))) + \mathcal{C}_{\exp}\left(\ln\frac{K}{\delta}\right).$$

By definition, $\theta_a(t)$ is the lowest possible value for λ_a, and hence $\min_a \theta_a(t)$ is the lowest possible value for $\min_a \lambda_a$. ■

6.2.2 UPPER CONFIDENCE BOUND

Above, we found that we do not learn much about the lower bound in the presence of many arms. For the upper confidence bound the story is different. We explain in Proposition 26 how to compute $\mathrm{U}_t^{\tilde{\pi}}$ for a general weight vector $\tilde{\pi}$. We then show that empirically a weight vector supported on *all subsets* can be helpful.

Proposition 26 *Let $\theta(t)$ be the solution in θ to the equation*

$$\max_{\mathcal{S}\subseteq[K]} \left[\sum_{a\in\mathcal{S}} \left[N_a(t) d^+(\hat{\mu}_a(t), \theta) - 3\ln(1+\ln(N_a(t)))\right]^+ - |\mathcal{S}| \mathcal{C}_{exp}\left(\frac{\ln \frac{1}{\delta\tilde{\pi}(\mathcal{S})}}{|\mathcal{S}|}\right)\right] \;=\; 0.$$

Then $\mathbb{P}\{\forall t \in \mathbb{N}, \min_a \mu_a \leq \theta(t)\} \geq 1-\delta$.

Proof We prove that $\mathrm{U}_t^{\tilde{\pi}}(\delta) = \theta(t)$. Let $\boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi},+}(\delta)$. By definition,

$$\max_{\mathcal{S}\subseteq[K]} \left[\sum_{a\in\mathcal{S}} \left[N_a(t) d^+(\hat{\mu}_a(t), \lambda_a) - 3\ln(1+\ln(N_a(t)))\right]^+ - |\mathcal{S}| \mathcal{C}_{\exp}\left(\frac{\ln \frac{1}{\delta\tilde{\pi}(\mathcal{S})}}{|\mathcal{S}|}\right)\right] \;\leq\; 0.$$

What does this tell us about $\min_{a\in[K]} \lambda_a$? Well, consider a candidate value $\theta \geq \min_a \hat{\mu}_a(t)$ for the minimum. Among bandit models $\boldsymbol{\lambda}$ with $\min_a \lambda_a = \theta$, the left-hand side above is minimised at $\lambda_a = \max\{\hat{\mu}_a(t), \theta\}$ and the maximal value of $\min_{a\in[K]} \lambda_a$ is the maximal value of θ such that

$$\max_{\mathcal{S}\subseteq[K]} \left[\sum_{a\in\mathcal{S}} \left[N_a(t) d^+(\hat{\mu}_a(t), \theta) - 3\ln(1+\ln(N_a(t))) \right]^+ - |\mathcal{S}| \mathcal{C}_{\text{exp}} \left(\frac{\ln \frac{1}{\delta\tilde{\pi}(\mathcal{S})}}{|\mathcal{S}|} \right) \right] \leq 0.$$

We recover the objective in the statement by noting that the left-hand side is a continuous and non-decreasing function of θ. ■

6.2.3 PRACTICAL CHOICE OF WEIGHT VECTOR

The upper bound for a minimum may benefit from considering many subsets $\mathcal{S} \subseteq [K]$ in the weighted union bound. The reason is that a smaller subset will have a smaller evidence term (summing fewer terms), but it may also have a smaller threshold. Here we investigate the use of cardinality-based weight vectors of the form $\tilde{\pi}(\mathcal{S}) = \pi(|\mathcal{S}|)/\binom{K}{|\mathcal{S}|}$ for some probability vector π on $\{1,\dots,K\}$.

First, let's consider the computation of $\theta(t)$ for those weight vectors: we are looking for the zero crossing of an increasing function, which can be found by e.g. binary search. It remains to efficiently evaluate the objective for a fixed $\theta(t)$. Here we propose to express the objective as

$$\max_{k\in[K]} \underbrace{\max_{\mathcal{S}\subseteq[K]:|\mathcal{S}|=k} \sum_{a\in\mathcal{S}} \left[N_a(t) d^+(\hat{\mu}_a(t), \theta(t)) - 3\ln(1+\ln(N_a(t))) \right]^+}_{\text{the best set takes the } k \text{ largest contributors; implement by sorting once.}} - k\mathcal{C}_{\text{exp}} \left(\frac{\ln \frac{\binom{K}{k}}{\delta\pi(k)}}{k} \right).$$

and observe that the best set of size k takes the k arms of largest contribution, which we can look up after sorting the arms by their contribution. Hence each evaluation of the objective can be obtained in $O(K \ln K)$ time.

We expect combining evidence across arms to be particularly useful when there are several arms with means close to the minimum. We illustrate this empirically in Figure 4 for a Bernoulli bandit model with M arms with mean 0.1 and 4 more arms with means $0.2, 0.3, 0.4, 0.5$ (thus $K = M+4$), for different values of M. We consider the use of a "Box" weight vector that is uniform on the singletons ($\pi(1) = 1$), a weight vector supported on the whole set of arms ($\pi(K) = 1$) and a weight vector that is uniform over subset sizes ($\pi(k) = 1/K$). For each value of M, data is collected using uniform sampling and we set $\delta = 10^{-10}$ to focus on the high confidence regime. We see that the uniform weight vector consistently leads to smaller upper confidence bounds when compared to Box, with an increased gap when M increases. We also experimented with a "Zipf" distribution for π ($\pi(k) \propto 1/k$), which performed almost identically as the uniform vector.

This experiment shows that for small values of δ a uniform cardinality-based weight vector is a robust choice: summing evidence across arms never hurts too much. In the particular case of minimums, we would like to mention that one can go even further and *aggregate* samples from different arms, as explained in Kaufmann et al. (2018), which leads to even smaller upper confidence bounds in experiments.

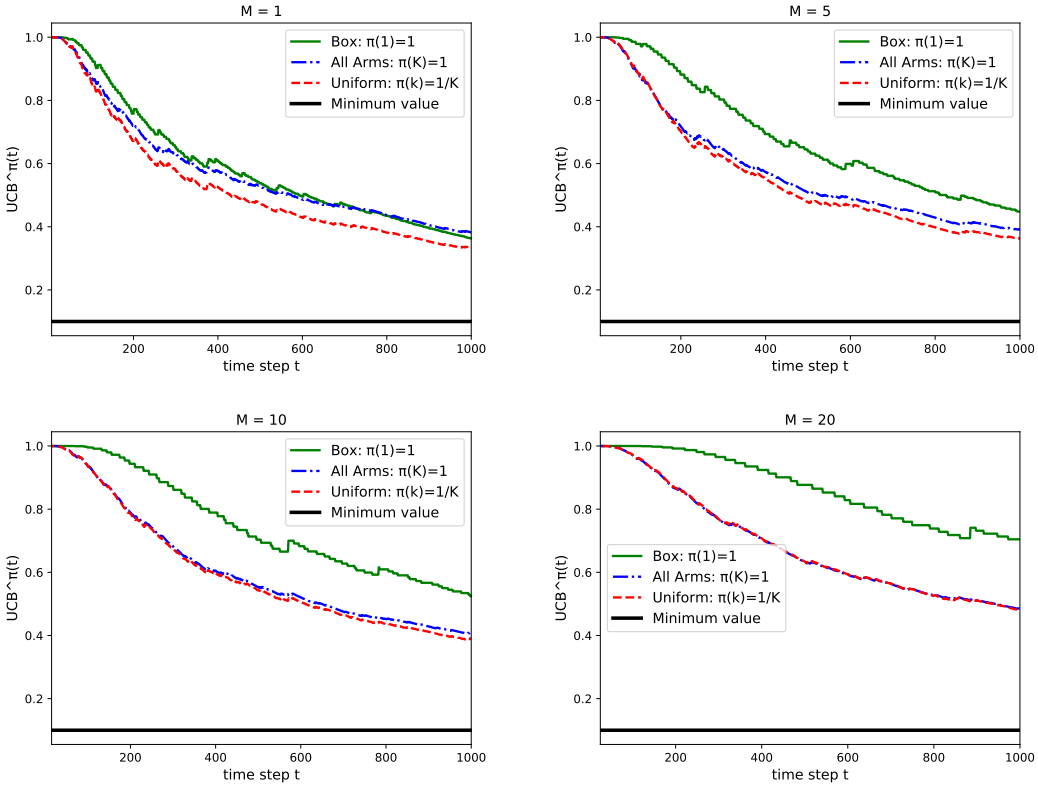


Figure 4: $\mathrm{U}_t^{\tilde{\pi}}(\delta)$ as a function of t for several cardinality-based weight vectors $\tilde{\pi}$ in a presence of M identical arms with the minimal mean, for $M = 1, 5, 10, 20$.

7. Conclusion

Sequential problems are studied in the multi-armed bandit model, where the learner sequentially picks arms to sample. The central question is what the learner infers from the samples that it has seen. This is used for deciding what to do next, when to stop, what to recommend and/or estimate.

We use mixture martingales to design confidence regions, based on self-normalised sums, for exponential family multi-armed bandit models. We argue that these confidence regions are the tightest known, and match, in spirit, established statistical lower bounds.

We then apply the obtained deviation inequalities to the design of confidence intervals by means of explicit projections, stopping rules by means of GLR statistics, and asymptotically optimal sampling rules by a tight analysis of the Track-and-Stop algorithm. The fact that we are pushing the state of the art in each of these areas clearly demonstrates the generic appeal of the mixture martingale approach.

Acknowledgments

Emilie Kaufmann acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grants BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04) and the

European CHIST-ERA project DELTA. Wouter Koolen acknowledges support from the Netherlands Organization for Scientific Research (NWO) under Veni grant 639.021.439. We are grateful to INRIA Associate Team ^{6}PAC.

References

Y. Abbasi-Yadkori, D.Pál, and C.Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, 2011.

A. Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv:1405.2639*, 2015.

C. Berge. *Topological Spaces*. Oliver & Boyd, 1963.

S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Fondations and Trends in Machine Learning*, 5(1):1–122, 2012.

S. Bubeck, R. Munos, and G. Stoltz. Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852, 2011.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

L. Chen, A. Gupta, J. Li, M. Qiao, and R. Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, 2017.

S. Chen, T. Lin, I. King, M. Lyu, and W. Chen. Combinatorial Pure Exploration of Multi-Armed Bandits. In *Advances in Neural Information Processing Systems*, 2014.

H. Chernoff. Sequential design of Experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959.

R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems 30*, pages 1763–1771. Curran Associates, Inc., 2017.

D. Darling and H. Robbins. Some further remarks on inequalities for sample sums. *Proceedings of the National Academy of Sciences*, 60(4):1175–1182, 1968.

A. P. Dawid and V. G. Vovk. Prequential probability: principles and properties. *Bernoulli*, 5(1): 125–162, 02 1999.

A. P. Dawid, S. de Rooij, G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Insuring against loss of evidence in game-theoretic probability. *Statistics & Probability Letters*, 81(1):157 – 162, 2011. ISSN 0167-7152.

V. H. de la Peña, M. Klass, and T. L. Lai. Self-Normalized Processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3A):1902–1933, 2004.

V. H. de la Peña, T. L. Lai, and S. Q. *Self-normalized processes. Limit Theory and Statistical applications*. Springer, 2009.

R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 14591–14600. Curran Associates, Inc., Dec. 2019.

R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, pages 14492–14501. Curran Associates, Inc., Dec. 2019.

E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7: 1079–1105, 2006.

A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory*, 2011.

A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference On Learning Theory (COLT)*, 2016.

A. Garivier and E. Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021.

A. Garivier, E. Kaufmann, and W. M. Koolen. Maximin action identification: A new bandit framework for games. In *Proceedings of the 29th Conference On Learning Theory (COLT)*, 2016.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, Mar. 2017.

S. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *arXiv:1810.08240*, 2018.

S. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

R. Huang, M. M. Ajallooeian, C. Szepesvári, and M. Müller. Structured best arm identification with fixed confidence. In *International Conference on Algorithmic Learning Theory (ALT)*, 2017.

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*, 2014.

A. Jonsson, E. Kaufmann, P. Ménard, O. D. Domingues, E. Leurent, and M. Valko. Planning in Markov decision processes with gap-dependent sample complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

S. Juneja and S. Krishnasamy. Sample complexity of partition identification using multi-armed bandits. In *Conference on Learning Theory (COLT)*, 2019.

S. Kalyanakrishnan and P. Stone. Efficient Selection in Multiple Bandit Arms: Theory and Practice. In *International Conference on Machine Learning (ICML)*, 2010.

E. Kaufmann and W. M. Koolen. Monte-Carlo tree search by best arm identification. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

E. Kaufmann, O. Cappé, and A. Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

E. Kaufmann, W. M. Koolen, and A. Garivier. Sequential test for the lowest mean: From Thompson to Murphy sampling. In *Advances in Neural Information Processing Systems (NeurIPS) 31*, pages 6333–6343, Dec. 2018.

W. M. Koolen and T. van Erven. Second-order quantile methods for experts and combinatorial games. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1155–1175, June 2015.

T. Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.

T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the thresholding bandit problem. In M. Balcan and K. Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1690–1698. JMLR.org, 2016.

S. Magureanu, R. Combes, and A. Proutière. Lipschitz Bandits: Regret lower bounds and optimal algorithms. In *Proceedings on the 27th Conference On Learning Theory*, 2014.

O. Maillard. *Mathematics of Statistical Sequential Decision Making*. 2019.

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

H. Robbins. Statistical Methods Related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

K. Teraoka, K. Hatano, and E. Takimoto. Efficient sampling method for Monte Carlo tree search problem. *IEICE Transactions on Infomation and Systems*, pages 392–398, 2014.

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

J. Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing (NIPS)*, 2016.

Y. Zhou, J. Li, and J. Zhu. Identify the Nash equilibrium in static games with random payoffs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4160–4169, International Convention Centre, Sydney, Australia, Aug. 2017. PMLR.

Appendix A. Proof of Proposition 8

We may write

$$h^{-1}(x) \;=\; \inf_{z \geq 1} z\left(x - 1 + \ln \frac{z}{z-1}\right)$$

Plugging in the sub-optimal feasible choice $z = 1 + \frac{1}{(x-1)+\sqrt{2(x-1)}}$ reveals

$$\begin{aligned} h^{-1}(x) \;&\leq\; \left(1 + \frac{1}{(x-1)+\sqrt{2(x-1)}}\right)\left(x - 1 + \ln\left(x + \sqrt{2(x-1)}\right)\right) \\ &\leq\; 1 + (x-1) + \ln\left(x + \sqrt{2(x-1)}\right). \end{aligned}$$

The last inequality uses $\ln\left(x + \sqrt{2(x-1)}\right) \leq \sqrt{2(x-1)}$ which holds with equality at $x = 1$ and whose gap is increasing (as can be checked by differentiation).

Appendix B. Additional Proofs for Exponential Families

B.1 Proof of Lemma 13

Given any probability distribution π, recall that the associated mixture martingale is defined as

$$Z_a^\pi(t) = \int \exp\left(\lambda S_a(t) - \phi_{\mu_a}(\lambda) N_a(t)\right) \mathrm{d}\pi(\lambda).$$

The first step of the construction is Lemma 27, which relates the deviation of $N_a(t)d^+(\hat{\mu}_a(t), \mu_a)$ and $N_a(t)d^-(\hat{\mu}_a(t), \mu_a)$ to those of $\eta S_a(t) - \phi_{\mu_a}(\eta)N_a(t)$ for a well chosen η, provided that $N_a(t)$ belongs to some "slice" $[(1+\xi)^{i-1}, (1+\xi)^i]$.

Lemma 27 *Fix $i \in \mathbb{N}^*$, $x > 0$ and $\xi > 0$. There exist $\eta_i^+(x,\xi)$ and $\eta_i^-(x,\xi)$ such that, if $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$ it holds that*

$$\begin{aligned} \left\{N_a(t)d^+(\hat{\mu}_a(t), \mu_a) \geq x\right\} \;&\subseteq\; \left\{\eta_i^+ S_a(t) - N_a(t)\phi_{\mu_a}(\eta_i^+) \geq \frac{x}{1+\xi}\right\} \\ \left\{N_a(t)d^-(\hat{\mu}_a(t), \mu_a) \geq x\right\} \;&\subseteq\; \left\{\eta_i^- S_a(t) - N_a(t)\phi_{\mu_a}(\eta_i^-) \geq \frac{x}{1+\xi}\right\}. \end{aligned}$$

The next step is to relate the deviation of $X_a(t)$ to those of a martingale for every $t \in \mathbb{N}$ and not only for $N_a(t)$ is some slice: this will be achieved by a mixture martingale with a well-chosen discrete prior. In the sequel, we consider the (most complicated) case in which $X_a(t) = Y_a(t)$ for all t. Given x, we define the following probability distribution. Let

$$\begin{array}{rclcrcl} \gamma_i & = & \frac{1}{2}\frac{1}{i^2\zeta(2)} & & x_i & = & x + \ln\left(\frac{1}{\gamma_i}\right) \\ \eta_i^+ & = & \eta_i^+(x_i,\xi) & & \eta_i^- & = & \eta_i^-(x_i,\xi), \end{array}$$

where $\eta_i^{\pm}(x,\xi)$ are defined in Lemma 27. We define the discrete prior

$$\pi = \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^+} + \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^-}$$

and the corresponding mixture martingale

$$Z_a^{\pi}(t) = \sum_{i=1}^{\infty} \gamma_i Z_a^{\eta_i^+}(t) + \sum_{i=1}^{\infty} \gamma_i Z_a^{\eta_i^-}(t),$$

where by a slight abuse of notation, $Z_a^{\eta}(t) = Z_a^{\delta_\eta}(t) = \exp(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t))$ for $\eta \in \mathbb{R}$.

In the case $X_a(t) = Y_a^+(t)$, this prior is modified by taking $\gamma_i = \frac{1}{i^2\zeta(2)}$ and $\pi = \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^+}$, while for $X_a(t) = Y_a^-(t)$, one defines $\pi = \sum_{i=1}^{\infty} \gamma_i \delta_{\eta_i^-}$. We continue the proof assuming $X_a(t) = Y_a(t)$ for all t. The proof of the two other cases follow the exact same lines, with the corresponding priors, leading to an improved constant $C(\xi) = \frac{\ln \zeta(2)}{(\ln(1+\xi))^2}$.

$$\begin{aligned} &\{X_a(t) - (1+\xi)\ln C(\xi) \geq x\} \\ &\quad \subseteq \left\{ [N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\ln(1 + \ln(N_a(t)))]^+ \geq x + (1+\xi)\ln C(\xi) \right\} \\ &\quad = \left\{ N_a(t) d(\hat{\mu}_a(t), \mu_a) - 3\ln(1 + \ln(N_a(t))) \geq x + (1+\xi)\ln C(\xi) \right\}, \end{aligned}$$

where we use that $x + (1+\xi)\ln C(\xi) > 0$ as $\xi < 1/2$. Now, as $2(1+\xi) < 3$, one has

$$\begin{aligned} &\{X_a(t) - (1+\xi)\ln C(\xi) \geq x\} \\ &\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) - 2(1+\xi)\ln\left(1 + \ln(N_a(t))\right) \geq x + (1+\xi)\ln\left(\frac{2\zeta(2)}{\ln(1+\xi)^2}\right) \right\} \\ &\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\ln\left(\frac{2\zeta(2)(1 + \ln(N_a(t))^2}{\ln(1+\xi)^2}\right) \right\} \\ &\quad \subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\ln\left(\frac{2\zeta(2)(\ln(1+\xi) + \ln(N_a(t))^2}{\ln(1+\xi)^2}\right) \right\}, \end{aligned}$$

where the last inequality uses $\ln(1+\xi) \leq \ln(3/2) \leq 1$. Now, let $i(t) \geq 1$ be such that $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$. One can observe that $\frac{\ln N_a(t)}{\ln(1+\xi)} \geq i(t) - 1$. Using Lemma 27,

$$
\begin{aligned}
&\{X_a(t) - (1+\xi)\ln C(\xi) \geq x\} \\
&\quad\subseteq \left\{ N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right) \geq x + (1+\xi)\ln\left(\frac{1}{\gamma_{i(t)}}\right) \right\} \\
&\quad\subseteq \left\{ \max_{\eta \in \left\{\eta^+_{i(t)}, \eta^-_{i(t)}\right\}} \left[\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right] \geq \frac{1}{1+\xi}\left[x + (1+\xi)\ln\left(\frac{1}{\gamma_{i(t)}}\right)\right] \right\} \\
&\quad\subseteq \left\{ \max_{\eta \in \left\{\eta^+_{i(t)}, \eta^-_{i(t)}\right\}} \gamma_{i(t)} \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) \geq e^{\frac{x}{1+\xi}} \right\} \\
&\quad\subseteq \left\{ \max_{i \in \mathbb{N}} \max_{\eta \in \left\{\eta_i^+, \eta_i^-\right\}} \gamma_i \exp\left(\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)\right) \geq e^{\frac{x}{1+\xi}} \right\} \\
&\quad\subseteq \left\{ Z_a^\pi(t) \geq e^{\frac{x}{1+\xi}} \right\}.
\end{aligned}
$$

Proof of Lemma 27 We introduce the notation θ for the natural parameter associated to μ_a, defined as $\theta = \dot{b}^{-1}(\mu_a)$. Define $\eta_i^+ < 0$ and $\eta_i^- > 0$ such that

$$
\mathrm{KL}(\theta + \eta_i^+, \theta) = \mathrm{KL}(\theta + \eta_i^-, \theta) = \frac{x}{(1+\xi)^i}.
$$

where $\mathrm{KL}(\theta, \theta')$ is the Kullback-Leibler divergence between the distributions of natural parameter θ and θ'. Moreover, using some properties of the KL-divergence, one can write

$$
\begin{aligned}
\mathrm{KL}(\theta + \eta_i^+, \theta) &= \eta_i^+ \mu_i^+ - \phi_{\mu_a}(\eta_i^+) \text{ with } \mu_i^+ := \dot{b}^{-1}(\theta + \eta_i^+) < \mu_a, \\
\mathrm{KL}(\theta + \eta_i^-, \theta) &= \eta_i^- \mu_i^- - \phi_{\mu_a}(\eta_i^-) \text{ with } \mu_i^- := \dot{b}^{-1}(\theta + \eta_i^-) > \mu_a.
\end{aligned}
$$

For $N_a(t) \in [(1+\xi)^{i-1}, (1+\xi)^i]$, one has

$$
\begin{aligned}
\left\{N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq x\right\} &\subseteq \left\{ d^+(\hat{\mu}_a(t), \mu_a) \geq \frac{x}{(1+\xi)^i} \right\} \\
&\subseteq \left\{\hat{\mu}_a(t) \leq \mu_i^+\right\} \\
&\subseteq \left\{\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+) \geq \mathrm{KL}(\theta + \eta_i^+, \theta)\right\} \\
&\subseteq \left\{ (1+\xi)^{i-1}\left(\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+)\right) \geq \frac{x}{1+\xi} \right\} \\
&\subseteq \left\{ N_a(t)\left(\eta_i^+ \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^+)\right) \geq \frac{x}{1+\xi} \right\},
\end{aligned}
$$

where the third inclusion uses that η_i^+ is negative. Similarly, using this time that $\eta_i^- > 0$ yields

$$
\begin{aligned}
\left\{N_a(t) d^-(\hat{\mu}_a(t), \mu_a) \geq x\right\} &\subseteq \left\{\hat{\mu}_a(t) \geq \mu_i^-\right\} \\
&\subseteq \left\{\eta_i^- \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^-) \geq \mathrm{KL}(\theta + \eta_i^-, \theta)\right\} \\
&\subseteq \left\{ N_a(t)\left(\eta_i^- \hat{\mu}_a(t) - \phi_{\mu_a}(\eta_i^-)\right) \geq \frac{x}{1+\xi} \right\},
\end{aligned}
$$

which concludes the proof.

B.2 Tight Tuning: Proof of Lemma 12

In this section we prove Lemma 12, which gives the tightest possible tuning achievable with our method. We first prove two auxiliary lemmas.

Lemma 28 *Let $x \geq 0$. Then*

$$\inf_{q \in [0,1]} \frac{x - \ln(1-q)}{q} \;=\; h^{-1}(1+x)\,.$$

Proof The objective is convex in $\frac{1}{q}$, and hence minimised at zero derivative. Cancelling the derivative requires

$$1 + x \;=\; \frac{1}{1-q} - \ln \frac{1}{1-q} \;=\; h\left(\frac{1}{1-q}\right) \qquad \text{so that} \qquad q \;=\; 1 - \frac{1}{h^{-1}(1+x)}$$

where the rewrite in terms of h is allowed since $1/(1-q) \geq 1$. Plugging this in, we find the value as stated. ■

Definition 29 *For any $z \in [1, e]$ and $x \geq 0$, we define*

$$\tilde{h}_z(x) \;=\; \min_{y \in [1,z]} y\,(x - \ln \ln y)\,.$$

We can now make the connection to (9).

Lemma 30 *Fix $z \in [1, e]$. Then*

$$\tilde{h}_z(x) \;=\; \begin{cases} \exp\left(\frac{1}{h^{-1}(x)}\right) h^{-1}(x) & \text{if } x \geq h\left(\frac{1}{\ln z}\right), \\ z\,(x - \ln \ln z) & \text{o.w.} \end{cases}$$

Proof The objective in Definition 29 is convex on $y \in [1, e]$, and its derivative is $x - h(1/\ln y)$. When $x \leq h(1/\ln z)$ it is decreasing on the entire domain $y \in [1, z]$, and hence minimised at $y = z$, yielding the second case. If on the other hand $x \geq h(1/\ln z)$, the derivative of the objective is cancelled at $y = e^{\frac{1}{h^{-1}(x)}}$, and substitution reveals that the value equals

$$e^{\frac{1}{h^{-1}(x)}}\left(x + \ln h^{-1}(x)\right) \;=\; e^{\frac{1}{h^{-1}(x)}} h^{-1}(x).$$

■

We are now ready to prove the Lemma.

Proof (of Lemma 12) We reorganise, apply Lemma 28 and then Lemma 30 to find

$$\begin{aligned}
\mathcal{C}_{\exp}(x) &= \inf_{\xi\in[0,z]} (1+\xi)\left(\inf_{q\le 1} \frac{x-\ln(1-q)}{q} + \ln C(\xi)\right) \\
&= \inf_{\xi\in[0,z]} (1+\xi)\left(h^{-1}(1+x) + \ln C(\xi)\right) \\
&= \inf_{\xi\in[0,z]} (1+\xi)\left(h^{-1}(1+x) + \ln(2\zeta(2)) - 2\ln\ln(1+\xi)\right) \\
&= 2\inf_{y\in[1,1+z]} y\left(\frac{h^{-1}(1+x)+\ln(2\zeta(2))}{2} - \ln\ln y\right) \\
&= 2\tilde{h}_{1+z}\left(\frac{h^{-1}(1+x)+\ln(2\zeta(2))}{2}\right).
\end{aligned}$$

■

B.3 A Tighter One-Arm Bound

Lemma 13 allows us to directly derive valid thresholds involving only a single arm. Namely, we have

Corollary 31 *Let $\tilde{h}_z(x)$ be as defined in (9). For every arm a and confidence parameter $x \ge 0$*

$$\mathbb{P}\left\{\exists t\in\mathbb{N} : X_a(t) \ge 2\tilde{h}_{3/2}\left(\frac{x+\ln(2\zeta(2))}{2}\right)\right\} \le e^{-x}.$$

Proof By Lemma 13, for every $\xi \in [0, 1/2]$,

$$\mathbb{P}\left\{X_a(t) - (1+\xi)\ln\left(\frac{2\zeta(2)}{(\ln(1+\xi))^2}\right) \ge (1+\xi)x\right\} \le \mathbb{P}\left\{Z_a^{\pi((1+\xi)x)}(t) \ge e^x\right\} \le e^{-x}$$

Minimising the threshold w.r.t. ξ using Lemma 30 results in

$$\min_{\xi\in[0,1/2]} (1+\xi)\left(x + \ln\left(\frac{2\zeta(2)}{(\ln(1+\xi))^2}\right)\right) = 2\tilde{h}_{3/2}\left(\frac{x+\ln(2\zeta(2))}{2}\right).$$

■

We see that the multiple-arm threshold of Theorem 7 has $h^{-1}(1+x) > x$ where Corollary 31 has just x. This additional blowup is the overhead that our approach incurs for controlling multiple arms by means of a "Cramér-Chernoff" approach.

Appendix C. Refined Deviation Inequalities for Gaussian and Gamma Distributions

Theorem 9 and Theorem 10 follow from a similar martingale construction, that could actually be used for other one-dimensional exponential families with divergence function $d(\cdot,\cdot)$, provided one is able to construct a continuous prior satisfying a general assumption given below.

Assumption 32 *For every $\lambda \in]0,1[$, $\mu \in \mathcal{I}$, there exists a family of functions $(p_t^{\lambda,\mu})_{t\geq 1}$ such that, for every $t \geq 1$,*

$$\forall x \in \mathcal{I}, \quad \int p_t^{\lambda,\mu}(\eta) e^{\eta t x - \phi_\mu(\eta) t} \, \mathrm{d}\eta = e^{\lambda t d(x,\mu)}. \tag{27}$$

Moreover, for every $1 \leq n_1 \leq n_2$ and every $\eta \in \mathbb{R}$,

$$p_{n_1}^{\lambda,\mu}(\eta) \;\geq\; \sqrt{\frac{n_1}{n_2}} p_{n_2}^{\lambda,\mu}(\eta). \tag{28}$$

In words, this assumption implies that for all a and $t \geq 1$ there exists a prior distribution for which the corresponding mixture martingale exactly attains $e^{\lambda t d(\hat{\mu}_a(t),\mu_a)}$ and such that one can control the variation of the prior corresponding to two different time steps. Under this assumption, we are able to prove the following.

Theorem 33 *Assume that Assumption 32 is satisfied and let*

$$C_0(t,\lambda) := \sup_{\mu \in \mathcal{I}} \int p_t^{\lambda,\mu}(\eta) \, \mathrm{d}\eta.$$

Fix $\xi > 0$, $c > 1$ and define

$$g_0(\lambda,\xi,c) = \ln\left[\sum_{i=1}^{\infty} \frac{1}{i^{\lambda c}\zeta(\lambda c)} C_0\left((1+\xi)^{i-1},\lambda\right)\right].$$

The stochastic process $X_a(t) = N_a(t) d(\hat{\mu}_a(t),\mu_a) - c\ln\big(\ln(1+\xi) + \ln N_a(t)\big)$ is $g_{\xi,c}$-VCC where

$$\begin{array}{rccl} g_{\xi,c} : (c^{-1},1] & \longrightarrow & \mathbb{R}^+ & \\ \lambda & \mapsto & g_0(\lambda,\xi,c) + \frac{1}{2}\ln(1+\xi) + \lambda c \ln\left(\frac{1}{\ln(1+\xi)}\right) + \ln\zeta(\lambda c). \end{array}$$

Theorem 33 directly provides a deviation inequality using Lemma 4. It thus remains to find sequences of priors satisfying Assumption 32, which we were able to do for two particular examples, Gaussian and Gamma distributions. One can note that finding functions $p_t^{\lambda,\mu}$ is closely related to computing a (bilateral) inverse Laplace transform. Indeed, if q is the inverse Laplace transform of $e^{\lambda t d(x,\mu)}$, meaning that $\forall x : \int_{-\infty}^{\infty} q(s) e^{-sx} \, \mathrm{d}s = e^{\lambda t d(x,\mu)}$, the assumption is satisfied for $p_t^{\lambda,\mu}(\eta) = tq(-\eta t)e^{\phi_\mu(\eta)t}$. However, computing such inverse Laplace transforms is not easy beyond Gaussian or Gamma distributions.

Proof of Theorem 33 For $i = 1, 2, \ldots$ we introduce grid points $T_i = (1+\xi)^{i-1}$ with prior weights $\gamma_i = \frac{1}{i^{\lambda c}\zeta(\lambda c)}$ and define the (un-normalized) martingale

$$\tilde{M}_a^{\lambda}(t) \;:=\; \sum_{i=1}^{\infty} \gamma_i \int p_{T_i}^{\lambda,\mu_a}(\eta) e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)} \, \mathrm{d}\eta,$$

that satisfies $\tilde{M}_a^{\lambda}(0) \leq \exp(g_0(\lambda,\xi,c))$.

For $N_a(t) \in [T_i, T_{i+1}[$, we first bound the martingale from below by one of its terms, and then make use of Assumption 32.

$$\begin{aligned}
\tilde{M}_a^\lambda(t) &\geq \gamma_i \int p_{T_i}^{\lambda,\mu_a}(\eta) e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)} \,\mathrm{d}\eta \\
&\geq \sqrt{\frac{T_i}{N_a(t)}} \gamma_i \int p_{N_a(t)}^{\lambda,\mu_a}(\eta) e^{\eta S_a(t) - \phi_{\mu_a}(\eta) N_a(t)} \,\mathrm{d}\eta \\
&= \sqrt{\frac{T_i}{N_a(t)}} \gamma_i \exp\left(\lambda N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right)\right) \\
&\geq \sqrt{\frac{1}{1+\xi}} \gamma_i \exp\left(\lambda N_a(t) d\left(\hat{\mu}_a(t), \mu_a\right)\right),
\end{aligned}$$

where the last inequality uses $N_a(t) \leq T_{i+1}$ and $T_i/T_{i+1} = 1/(1+\xi)$, due to the geometric grid.

Introducing the normalised martingale $M_a^\lambda(t) = \tilde{M}_a^\lambda(t)/\tilde{M}_a^\lambda(0)$ and further using the expression of γ_i yields, for all t such that $N_a(t) \in [T_i, T_{i+1}[$,

$$M_a^\lambda(t) \;\geq\; \tilde{M}_a^\lambda(t) e^{-g_0(\lambda,\xi,c)} \;\geq\; e^{\lambda N_a(t) d(\hat{\mu}_a(t),\mu_a) - g_0(\lambda,\xi,c) - \frac{1}{2}\ln(1+\xi) - \ln\zeta(\lambda c) - \lambda c \ln(i)}.$$

Finally, using that $i \leq 1 + \ln(N_a(t))/\ln(1+\xi)$ yields the desired

$$M_a^\lambda(t) \;\geq\; \exp\left(\lambda X_a(t) - g_{\xi,c}(\lambda)\right).$$

It remains to check the case $N_a(t) = 0$. Then $X_a(t) = -\infty$, so clearly $M_a^\lambda(t) = 1 > e^{-\lambda\infty}$.

□

C.1 Application to Gaussian Distributions

In the Gaussian case, direct computations show that Assumption 32 holds for the choice

$$p_t^{\lambda,\mu}(\eta) = \frac{1}{\sqrt{1-\lambda}} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{\eta^2}{2\sigma_t^2}\right),$$

where $\sigma_t^2 = \frac{\lambda}{t(1-\lambda)}$. As a consequence $C_0(t,\lambda) = \frac{1}{\sqrt{1-\lambda}}$ and $g_0(\lambda,\xi,c) = -\frac{1}{2}\ln(1-\lambda)$. Note that the inequality (28) is actually an equality. We are now ready to prove Theorem 9.

Proof of Theorem 9 By Theorem 33, picking $c = 2$, for every $\xi > 0$ and $\lambda \in]1/2, 1[$ there exists a test martingale $M_a^{\lambda,\xi}(t)$ such that

$$\forall t \in \mathbb{N},\; M_a^{\lambda,\xi}(t) \geq e^{\lambda\left[N_a(t) d(\hat{\mu}_a(t),\mu_a) - f_\xi(N_a(t))\right] - g_\xi(\lambda)}$$

with

$$\begin{aligned}
f_\xi(s) &= 2\ln(\ln(1+\xi) + \ln(s)) \\
g_\xi(\lambda) &= \frac{1}{2}\ln(1+\xi) + 2\lambda \ln\left(\frac{1}{\ln(1+\xi)}\right) + \ln\zeta(2\lambda) - \frac{1}{2}\ln(1-\lambda)
\end{aligned}$$

It can be checked that the choice of ξ leading to the smallest g_ξ function is $\ln(1+\xi) = 4\lambda$. Denoting by $\xi^*(\lambda)$ this value, it holds that

$$g_G(\lambda) \;=\; g_{\xi^*(\lambda)}(\lambda) \quad = \quad 2\lambda - 2\lambda \ln(4\lambda) + \ln \zeta(2\lambda) - \frac{1}{2}\ln(1-\lambda)\,.$$

For every $\lambda \in]1/2, 1[$, observe that $f_{\xi^*(\lambda)}(s) \leq 2\ln(4 + \ln s)$. Hence, there exists a test martingale $M_a^\lambda(t) = M_a^{\lambda,\xi^*(\lambda)}(t)$ such that

$$\forall t \in \mathbb{N},\; M_a^\lambda(t) \quad \geq \quad e^{\lambda[N_a(t) d(\hat{\mu}_a(t), \mu_a) - 2\ln(4+\ln(N_a(t)))] - g_G(\lambda)},$$

which concludes the proof.

C.2 Application to Gamma Distributions

A Gamma distribution with shape parameter α and mean μ has density at $z > 0$ given by

$$f_{\alpha,\mu}(z) = \frac{e^{-\frac{\alpha z}{\mu}} \left(\frac{\alpha z}{\mu}\right)^\alpha}{z\Gamma(\alpha)}.$$

We recover the Exponential distribution for $\alpha = 1$. More generally, the set of Gamma distributions with a known shape α form a one-parameter exponential family for which

$$d(\mu, \mu') = \alpha\left(\frac{\mu}{\mu'} - 1 - \ln\frac{\mu}{\mu'}\right) \quad \text{and} \quad \phi_\mu(\eta) = \alpha \ln\left(\frac{\alpha}{\alpha - \mu\eta}\right) \text{ for } \eta < \alpha/\mu.$$

Next we show that the family of functions

$$p_t^\lambda(\eta) \;\coloneqq\; \frac{\mu}{\alpha}\frac{(\alpha t/e)^{\lambda\alpha t}}{\Gamma(\lambda\alpha t)}\left(1 - \frac{\eta\mu}{\alpha}\right)^{-\alpha t}\left(\lambda - \frac{\eta\mu}{\alpha}\right)_+^{\lambda\alpha t - 1}. \tag{29}$$

leads to suitable "priors".

Proposition 34 *The family of functions defined in* (29) *satisfies Assumption 32.*

Proof Proving (27) is equivalent to checking that for all $x > 0$,

$$\frac{\mu}{\alpha}\left(\frac{\alpha t x}{\mu}\right)^{\lambda\alpha t}\frac{1}{\Gamma(\lambda\alpha t)}\int_{-\infty}^{\frac{\lambda\alpha}{\mu}}\left(\lambda - \frac{\eta\mu}{\alpha}\right)^{\lambda\alpha t - 1} e^{\eta t x}\,\mathrm{d}\eta = e^{\frac{\lambda\alpha t x}{\mu}}$$

which can be done using change of variables to $y = tx\left(\frac{\alpha\lambda}{\mu} - \eta\right)$ and the definition of the Gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,\mathrm{d}x$. Now let us check condition (28). The condition is trivially satisfied for $\eta \geq \frac{\lambda\alpha}{\mu}$, as both sides are zero. So assume η is smaller. Then

$$\begin{aligned}
\ln\frac{p_{n_1}^\lambda(\eta)}{p_{n_2}^\lambda(\eta)} \;&=\; \ln\frac{\frac{\mu}{\alpha}\frac{(\alpha n_1/e)^{\lambda\alpha n_1}}{\Gamma(\lambda\alpha n_1)}\left(1-\frac{\eta\mu}{\alpha}\right)^{-\alpha n_1}\left(\lambda - \frac{\eta\mu}{\alpha}\right)^{\lambda\alpha n_1 - 1}}{\frac{\mu}{\alpha}\frac{(\alpha n_2/e)^{\lambda\alpha n_2}}{\Gamma(\lambda\alpha n_2)}\left(1-\frac{\eta\mu}{\alpha}\right)^{-\alpha n_2}\left(\lambda - \frac{\eta\mu}{\alpha}\right)^{\lambda\alpha n_2 - 1}} \\
&=\; \ln\frac{\Gamma(\lambda\alpha n_2)(\alpha n_2/e)^{-\lambda\alpha n_2}}{\Gamma(\lambda\alpha n_1)(\alpha n_1/e)^{-\lambda\alpha n_1}} + \alpha(n_2 - n_1)\left(\ln\left(1-\frac{\eta\mu}{\alpha}\right) - \lambda\ln\left(\lambda - \frac{\eta\mu}{\alpha}\right)\right) \\
&\geq\; \frac{1}{2}\ln\left(\frac{n_1}{n_2}\right) + \alpha(n_2 - n_1)\left(\lambda\ln\lambda + \ln\left(1-\frac{\eta\mu}{\alpha}\right) - \lambda\ln\left(\lambda - \frac{\eta\mu}{\alpha}\right)\right) \\
&\geq\; \frac{1}{2}\ln\left(\frac{n_1}{n_2}\right).
\end{aligned}$$

For the first inequality we used that the approximation error $\ln(\Gamma(x)) - x\ln(x) + x - \frac{1}{2}\ln\left(\frac{2\pi}{x}\right)$ is a decreasing function of $x \in \mathbb{R}_+$ (as can be easily verified by a plot), so that in particular

$$\ln\frac{\Gamma(\lambda\alpha n_2)}{\Gamma(\lambda\alpha n_1)} \;\geq\; \frac{1}{2}\ln\left(\frac{n_1}{n_2}\right) + \lambda\alpha n_2 \ln(\lambda\alpha n_2/e) - \lambda\alpha n_1 \ln(\lambda\alpha n_1/e).$$

For the second inequality we use that the expression above switches from decreasing to increasing at $\eta = 0$, and is hence minimised there. Plugging in the value $\eta = 0$ gives the result. ■

We are now ready to prove Theorem 10, as a consequence of Theorem 33.

Proof of Theorem 10 In order to evaluate the function $g_0(\lambda, \xi, c)$ featured in Theorem 33, we first compute

$$C_0(t,\lambda) \;=\; \frac{\Gamma((1-\lambda)\alpha t)}{\Gamma(\alpha t)}(\alpha t/e)^{\lambda\alpha t}(1-\lambda)^{-(1-\lambda)\alpha t}.$$

To see this, perform the variable substitution $z = \frac{\alpha\lambda - \eta\mu}{\alpha - \eta\mu} \in [0,1]$ to render this a standard Beta integral

$$\begin{aligned}
C_0(t,\lambda) \;&=\; \frac{(\alpha t/e)^{\lambda\alpha t}}{\Gamma(\lambda\alpha t)} \int_{-\infty}^{\frac{\lambda\alpha}{\mu}} \left(1 - \frac{\eta\mu}{\alpha}\right)^{-\alpha t} \left(\lambda - \frac{\eta\mu}{\alpha}\right)^{\lambda\alpha t - 1} \frac{\mu}{\alpha}\,\mathrm{d}\eta \\
&=\; \frac{(\alpha t/e)^{\lambda\alpha t}}{\Gamma(\lambda\alpha t)} \int_0^1 \left(1 - \frac{\lambda - z}{1 - z}\right)^{-\alpha t} \left(\lambda - \frac{\lambda - z}{1 - z}\right)^{\lambda\alpha t - 1} \frac{1-\lambda}{(1-z)^2}\,\mathrm{d}z \\
&=\; \frac{(\alpha t/e)^{\lambda\alpha t}}{\Gamma(\lambda\alpha t)} (1-\lambda)^{-(1-\lambda)\alpha t} \int_0^1 z^{\lambda\alpha t - 1} (1-z)^{(1-\lambda)\alpha t - 1}\,\mathrm{d}z \\
&=\; (\alpha t/e)^{\lambda\alpha t} (1-\lambda)^{-(1-\lambda)\alpha t} \frac{\Gamma((1-\lambda)\alpha t)}{\Gamma(\alpha t)}
\end{aligned}$$

Proposition 35 *$C_0(t,\lambda)$ is decreasing in $t \in \mathbb{R}_+$.*

Proof Let $\psi^{(0)}(x) = \frac{\partial \ln\Gamma(x)}{\partial x}$. The derivative of $\ln C_0(t,\lambda)$ w.r.t. t is negative iff

$$(1-\lambda)\psi^{(0)}((1-\lambda)\alpha t) - (1-\lambda)\ln((1-\lambda)\alpha t) \;<\; \psi^{(0)}(\alpha t) - \ln(\alpha t).$$

Now this follows from the fact that $x\psi^{(0)}(x) - x\ln x$ can be checked to be an increasing function of $x \in \mathbb{R}_+$. ■

We find that $C_0(t,\lambda)$ decreases from $\frac{1}{1-\lambda}$ at $t \to 0$ to $\frac{1}{\sqrt{1-\lambda}}$ for $t \to \infty$. For the former, we use

$$\begin{aligned}
C_0(t,\lambda) \;&=\; \frac{\Gamma((1-\lambda)\alpha t)}{\Gamma(\alpha t)}(\alpha t/e)^{\lambda\alpha t}(1-\lambda)^{-(1-\lambda)\alpha t} \\
&=\; \frac{1}{1-\lambda}\frac{((1-\lambda)\alpha t)\Gamma((1-\lambda)\alpha t)}{(\alpha t)\Gamma(\alpha t)}(\alpha t/e)^{\lambda\alpha t}(1-\lambda)^{-(1-\lambda)\alpha t} \\
&=\; \frac{1}{1-\lambda}\frac{\Gamma(1+(1-\lambda)\alpha t)}{\Gamma(1+\alpha t)}(\alpha t/e)^{\lambda\alpha t}(1-\lambda)^{-(1-\lambda)\alpha t}
\end{aligned}$$

The claimed limit for $t \to 0$ now follows by taking the limit of each factor, using $\Gamma(1) = 1$ and $t^t \to 1$. For the latter, the first-order Stirling's approximation $\Gamma(z) \sim \sqrt{2\pi} e^{-z} z^{z-\frac{1}{2}}$ yields

$$C_0(t,\lambda) \sim \frac{1}{\sqrt{1-\lambda}} \text{ when } t \to \infty$$

Finally, we have that for all $\lambda \in (0,1)$ and $t \in \mathbb{N}$,

$$C_0(t,\lambda) \in \left[\frac{1}{\sqrt{1-\lambda}}; \frac{1}{1-\lambda}\right].$$

It follows that for all $\xi > 0$, $-\frac{1}{2}\ln(1-\lambda) \leq g_0(\lambda,\xi,c) \leq -\ln(1-\lambda)$. We might be able to show that g_0 is actually closer to $-\frac{1}{2}\ln(1-\lambda)$ as the Stirling approximation is known to be good for moderate values of t. However using Theorem 33 (and picking $c = 2$) one can already prove that for every $\xi > 0$ and $\lambda \in (c^{-1}, 1)$, there exists a test martingale $M_a^{\lambda,\xi}(t)$ such that

$$\forall t \in \mathbb{N},\ M_a^{\lambda,\xi}(t) \geq e^{\lambda\left[N_a(t)d(\hat{\mu}_a(t),\mu_a) - f_\xi(N_a(t))\right] - g_\xi(\lambda)}$$

with

$$\begin{aligned} f_\xi(s) &= 2\ln(\ln(1+\xi) + \ln(s)) \\ g_\xi(\lambda) &= \frac{1}{2}\ln(1+\xi) + 2\lambda \ln\left(\frac{1}{\ln(1+\xi)}\right) + \ln\zeta(2\lambda) - \ln(1-\lambda). \end{aligned}$$

Just like in the proof of Corollary 9, the function g is optimised in ξ at $\ln(1+\xi) = 4\lambda$. We conclude similarly that $X_a(t) = N_a(t)d(\hat{\mu}_a(t),\mu_a) - 2\ln(4 + \ln(N_a(t))$ is g_Γ-VCC (see Definition 1) for the function $g_\Gamma(\lambda) = 2\lambda - 2\lambda\ln(4\lambda) + \ln\zeta(2\lambda) - \ln(1-\lambda)$.

□

Appendix D. Optimal Sample Complexity: Proof of Theorem 17

The first ingredient of the proof is a (deterministic) property of the Tracking sampling rule, that reformulates Lemma 8 in Garivier and Kaufmann (2016).

Lemma 36 *Under the Tracking rule for each $a \in \{1,\dots,K\}$, $N_a(t) \geq (\sqrt{t} - K/2)_+ - 1$. Moreover, for all $\epsilon > 0$, for all t_0, there exists $t_\epsilon \geq t_0$ such that*

$$\sup_{t \geq t_0} \max_{a \in \{1,\dots,K\}} |w_a^*(\hat{\boldsymbol{\mu}}(t)) - w_a^*(\boldsymbol{\mu})| \leq \epsilon \quad \Rightarrow \quad \sup_{t \geq t_\epsilon} \max_{a \in \{1,\dots,K\}} \left| \frac{N_a(t)}{t} - w_a^*(\boldsymbol{\mu}) \right| \leq 3(K-1)\epsilon\,.$$

To ease the notation, we fix $\boldsymbol{\mu} \in \mathcal{O}_1$. From the continuity of $\boldsymbol{w}^*$ in $\boldsymbol{\mu} \in \mathcal{O}_1$, there exists $\xi = \xi(\epsilon, \boldsymbol{\mu})$ such that

$$\mathcal{I}_\epsilon := [\mu_1 - \xi, \mu_1 + \xi] \times \cdots \times [\mu_K - \xi, \mu_K + \xi]$$

is included in $\mathcal{O}_1$ and is such that for all $\boldsymbol{\mu}' \in \mathcal{I}_\epsilon$,

$$\max_{a \in \{1,\dots,K\}} |w_a^*(\boldsymbol{\mu}') - w_a^*(\boldsymbol{\mu})| \leq \epsilon.$$

In particular, whenever $\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon$, it holds that $\hat{\imath}(t) = 1$.

Let $T \in \mathbb{N}$ and define the "good tail" event

$$\mathcal{E}_T(\epsilon) = \bigcap_{t=T^{1/4}}^{T} (\hat{\boldsymbol{\mu}}(t) \in \mathcal{I}_\epsilon) .$$

By Lemma 36, under the Tracking rule each arm is drawn at least of order $\sqrt{t}$ times at round t. This permits to establish the following concentration result, stated as Lemma 19 in Garivier and Kaufmann (2016).

Lemma 37 *There exist two constants B, C (that depend on $\boldsymbol{\mu}$ and ϵ) such that*

$$\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_T^c(\epsilon)) \leq BT \exp(-CT^{1/8}).$$

Using Lemma 36, there exists a constant T_ϵ such that for $T \geq T_\epsilon$, it holds that on $\mathcal{E}_T(\epsilon)$,

$$\forall t \geq \sqrt{T}, \max_{a \in \{1,\dots,K\}} \left| \frac{N_a(t)}{t} - w_a^*(\mu) \right| \leq 3(K-1)\epsilon$$

On the event $\mathcal{E}_T(\epsilon)$, for $t \geq T^{1/4}$ it holds that $\hat{\imath}(t) = 1$, thus $\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t)) = \mathrm{Alt}(\boldsymbol{\mu})$ and $\hat{\Lambda}_t = t\hat{M}(t)$ where

$$\hat{M}(t) := \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in \{1,\dots,K\}} \frac{N_a(t)}{t} d\left(\hat{\mu}_a(t), \lambda_a\right).$$

One can rewrite

$$\hat{M}(t) \quad = \quad g\left(\hat{\boldsymbol{\mu}}(t), \left(\frac{N_a(t)}{t}\right)_{a \in \{1,\dots,K\}}\right),$$

with g a mapping defined on $\mathcal{O}_1 \times [0,1]^K$ by

$$g(\boldsymbol{\mu}', \boldsymbol{w}') = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a \in \{1,\dots,K\}} w_a' d\left(\mu_a', \lambda_a\right).$$

As the mapping $(\boldsymbol{\lambda}, \boldsymbol{\mu}', \boldsymbol{w}') \mapsto \sum_{a \in \{1,\dots,K\}} w_a' d\left(\mu_a', \lambda_a\right)$ is jointly continuous and the constraint set $\mathrm{Alt}(\boldsymbol{\mu})$ doesn't depend on $(\boldsymbol{\mu}', \boldsymbol{w}')$, it follows from the application of Berge's maximum theorem (Berge, 1963) that g is continuous.

For $T \geq T_\epsilon$, introducing the constant

$$C_\epsilon^*(\boldsymbol{\mu}) = \inf_{\substack{\boldsymbol{\mu}' : ||\boldsymbol{\mu}' - \boldsymbol{\mu}|| \leq \xi(\epsilon) \\ \boldsymbol{w}' : ||\boldsymbol{w}' - \boldsymbol{w}^*(\boldsymbol{\mu})|| \leq 3(K-1)\epsilon}} g(\boldsymbol{\mu}', \boldsymbol{w}') ,$$

on the event $\mathcal{E}_T(\epsilon)$ it holds that for every $t \geq \sqrt{T}$, $\hat{M}(t) \geq C_\epsilon^*(\boldsymbol{\mu})$.

Let $T \geq T_\epsilon$. On $\mathcal{E}_T(\epsilon)$,

$$\begin{aligned}\min\left(\tau_\delta^{\text{GLR}}, T\right) &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(\tau_\delta > t)} \leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{\left(t\hat{M}(t) \leq c_t(\delta)\right)} \\ &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbb{1}_{(tC_\epsilon^*(\boldsymbol{\mu}) \leq c_T(\delta))} \leq \sqrt{T} + \frac{c_T(\delta)}{C_\epsilon^*(\boldsymbol{\mu})} \,.\end{aligned}$$

Introducing

$$T_0^\epsilon(\delta) = \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{c_T(\delta)}{C_\epsilon^*(\boldsymbol{\mu})} \leq T\right\},$$

for every $T \geq \max(T_0^\epsilon(\delta), T_\epsilon)$, one has $\mathcal{E}_T(\epsilon) \subseteq (\tau_\delta \leq T)$, therefore

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta > T\right) \leq \mathbb{P}(\mathcal{E}_T^c) \leq BT\exp(-CT^{1/8})$$

and

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \leq T_0^\epsilon(\delta) + T_\epsilon + \sum_{T=1}^{\infty} BT\exp(-CT^{1/8}) \,.$$

We now provide an upper bound on $T_0^\epsilon(\delta)$. For $\xi > 0$ we introduce the constant

$$C(\xi) = \inf\{T \in \mathbb{N} : T - \sqrt{T} \geq T/(1+\xi)\}.$$

Using moreover the upper bound on the threshold yields

$$T_0^\epsilon(\delta) \leq C + C(\xi) + \inf\left\{T \in \mathbb{N} : \frac{\ln\left(\frac{DT}{\delta}\right)}{C_\epsilon^*(\boldsymbol{\mu})} \leq \frac{T}{1+\xi}\right\}.$$

Letting h^{-1} be the function defined in the statement of Theorem 7 which is related to the Lambert function. One has

$$T_0(\delta) \leq C + C(\xi) + \frac{(1+\xi)}{C_\epsilon^*(\boldsymbol{\mu})} h^{-1}\left(\ln\left(\frac{(1+\xi)D}{C_\epsilon^*(\boldsymbol{\mu})\delta}\right)\right).$$

Using Proposition 8, it follows that

$$T_0(\delta) \leq C + C(\xi) \quad + \quad \frac{(1+\xi)}{C_\epsilon(\boldsymbol{\mu})}\left[\ln\left(\frac{(1+\xi)D}{C_\epsilon^*(\boldsymbol{\mu})\delta}\right) + \ln\left(\ln\left(\frac{(1+\xi)D}{C_\epsilon^*(\boldsymbol{\mu})\delta}\right) + \sqrt{2\ln\left(\frac{(1+\xi)D}{C_\epsilon^*(\boldsymbol{\mu})\delta}\right) - 2}\right)\right].$$

From this last upper bound, for every $\xi > 0$ and $\epsilon > 0$,

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}\left[\tau_\delta^{\text{GLR}}\right]}{\ln(1/\delta)} \leq \frac{(1+\xi)}{C_\epsilon^*(\boldsymbol{\mu})}.$$

Letting ξ and ϵ go to zero and using that, by continuity of g and by definition of $\boldsymbol{w}^*(\boldsymbol{\mu})$,

$$\lim_{\epsilon \to 0} C_\epsilon^*(\boldsymbol{\mu}) = T^*(\boldsymbol{\mu})^{-1}$$

yields

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} \leq T^*(\boldsymbol{\mu})$$

To conclude, the lower bound of Proposition 16 implies that this inequality is an equality.