Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals

Emilie Kaufmann, Wouter M. Koolen

▶ To cite this version:

Emilie Kaufmann, Wouter M. Koolen. Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals. 2018. hal-01886612v2

HAL Id: hal-01886612
https://hal.science/hal-01886612v2

Preprint submitted on 27 Nov 2018 (v2), last revised 7 Dec 2021 (v3)

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

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

Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals

Emilie Kaufmann
CNRS & ULille, UMR 9189 (CRIStAL), Inria SequeL, Lille, France EMILIE.KAUFMANN@UNIV-LILLE.FR

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

Editor:

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 may take into account several 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 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

$$\sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \mu_a). \tag{1}$$

Here $N_a(t)$ and $\hat{\mu}_a(t)$ are the observation count and empirical mean of arm a 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 deviate from the true means, the larger the self-normalised sum. Note that the self-normalised sum equals the log likelihood ratio $\ln \frac{\ell(X_1,\ldots,X_t;\hat{\boldsymbol{\mu}}(t))}{\ell(X_1,\ldots,X_t;\boldsymbol{\mu})}$, where $\ell(X_1, \ldots, X_t; \boldsymbol{\lambda})$ is the likelihood of the observations under a bandit model whose vector of means is $\boldsymbol{\lambda}$.

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 main result is the construction of explicit *threshold functions* $\mathcal{T}(x) = x + o(x)$ (we obtain different constants under different assumptions) 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 is with high probability bounded by

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

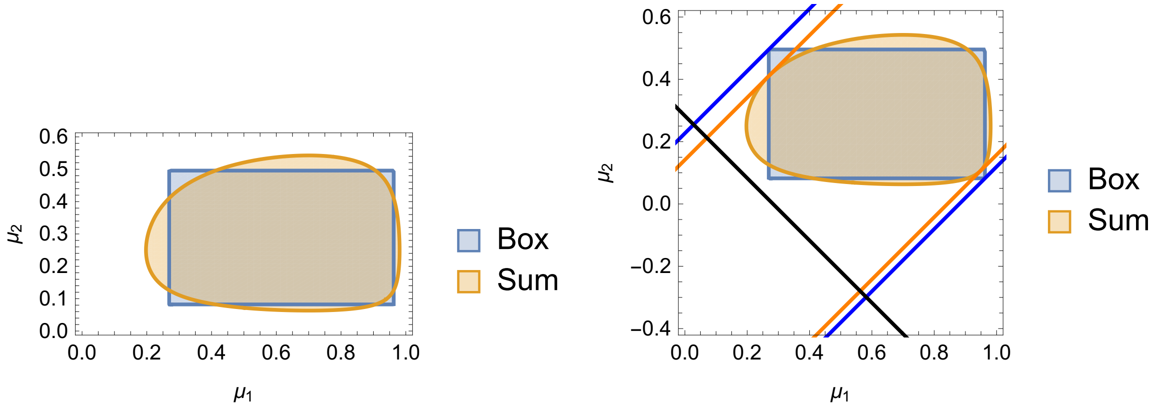
The salient features of this result are that it is uniform in time, respects the information geometry (KL) intrinsic to the exponential family, and combines in the strong summation sense the evidence from multiple arms. Moreover, at the moderate price of a weighted union bound we may apply the bound to any arbitrary 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 (at least in the Gaussian case) that, upon proper normalisation, the self-normalised deviation $\limsup_{N_a(t)\to\infty} \frac{N_a(t)d(\hat{\mu}_a(t),\mu_a)}{\ln\ln N_a(t)}$ is a universal constant a.s., whence the correction in the sum. Moreover, Wilk's phenomenon (see de la Peña et al., 2009, Chapter 17) informs us that twice the self-normalised sum (1) is asymptotically pivotal, with χ^2_K distribution. The K degrees of freedom are reflected in the perspective scaling of the threshold in (2). In this work we obtain essentially tight threshold functions by building suitable martingales. We will show that a threshold function $\mathcal{T}$ 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 $\lambda \in (0, 1)$. Our results will be obtained by leveraging some particular martingales called *mixture martingales* that have this property.

On the applications side, deviation inequalities of the form (2) allow us to analyze a stopping rule based on a Generalized Likelihood Ratio statistic for generic sequential identification problems. We notably show that under some assumptions on the identification problem itself, such stopping rules combined with a suitable sampling rule are (asymptotically) optimal in terms of sample complexity. Moreover, we provide refined stopping criteria for some particular tests that replace the number of arms K by a new notion of rank. Then, the sum form of the left-hand quantity in the above result 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. The intuition behind these ideas is visualised in Figure 1.

1.1 Related Work

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 $X_1, \ldots, X_t$ are seen as reward and the goal is to find a sequential strategy to maximize the (expected) cumulated reward, which is equivalent to



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

(b) Confidence intervals for a linear function of $\boldsymbol{\mu}$, obtained by projecting the confidence regions on the black normal vector. The interval between the blue tangents (for Box) strictly contains that between the orange tangents (for Sum).

Figure 1: Visual two-arm comparison of confidence regions and implied confidence intervals. 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".

minimizing some notion of regret. Several algorithms exist for this problem and we refer to Bubeck and Cesa-Bianchi (2012) for a survey.

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, the goal 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 attain this optimal sample complexity. Extensions of this best arm identification problem in which one should decide quickly and accurately *something* about the means of the arms have been studied recently Huang et al. (2017); Chen et al. (2017). In this work, we propose new stopping rules for those general adaptive decision making problems, as well as a generalization of the Tracking rule to attain optimal sample complexity.

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. Such self-normalized deviation inequalities have been mostly obtained by carefully using martingales, either with a so-called peeling trick (see, e.g. Cappé et al. 2013) or with the "method of mixtures" that has been popularized by de la Peña et al. (2004, 2009). Mixture martingales have indeed been used to obtain self-normalized deviation inequalities, e.g. by Abbasi-Yadkori et al. (2011); Howard et al. (2018) (see also our detailed discussion in Section 2.3). In this work we propose new prior constructions, as well as a central assumption under which deviation inequalities can be obtained.

The self-normalized deviation inequalities that we propose in this paper generalize in several directions existing results from the literature. First, the particular case of Gaussian distributions and a subset of size 1 has been treated by Robbins (1970); Robbins and Siegmund (1970), also building on mixture martingales. Using an appropriate (complicated) continuous prior, they obtain a threshold that is shown to have the right asymptotic rate in t, $\ln(1+\ln(t))$ which is compatible with the Law of the Iterated Logarithm. More recently, time-uniform inequalities for the one-armed Gaussian case have also been obtained independently by Jamieson et al. (2014), Kaufmann et al. (2016) and Zhao et al. (2016). Those inequalities also have the right $\ln(1+\ln(t))$ dependency in t.

Beyond Gaussian distributions, Garivier and Cappé (2011) and Magureanu et al. (2014) propose deviation inequalities expressed in KL-divergence that are uniform over a fixed time interval $t \in \{1,\dots,n\}$, respectively for a single arm and for the subset $\mathcal{S} = \{1,\dots,K\}$. Our results provide uniform deviations over the whole time range ($t \in \mathbb{N}$). Moreover, a detailed comparison in Section 4 shows that our bounds are essentially tighter in the presence of multiple arms.

1.2 Outline

The paper is structured as follows. In Section 2 we set forth our general method to obtain deviation inequalities in bandit models and formally introduce mixture martingales. We then present two different mixture-martingale constructions that yield threshold functions for the Gaussian and Gamma special cases (Section 3) and for general exponential families (Section 4). We integrate these results with the Track-and-Stop strategy to obtain an asymptotically optimal algorithm for generic sequential identification problems (Section 5). We then develop refined applications to stopping rules for sequential testing (Section 6) and for projected confidence intervals (Section 7).

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 $b : \Theta \to \mathbb{R}$ (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 $\nu^{\mu_1},\dots,\nu^{\mu_k}$ that belong to some one-dimensional canonical exponential family $\mathcal{P}$: it can be fully parametrized by the vector of means $\boldsymbol{\mu} = (\mu_1,\dots,\mu_K) \in \mathcal{I}^K$. 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 $S_a(t) = \sum_{s=1}^t X_t \mathbb{1}_{(A_s=a)}$ 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 $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ be the filtration generated by the observations gathered after 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

$$X_a(t) = N_a(t)d(\hat{\mu}_a(t),\mu_a) - c\ln(d + \ln N_a(t)) \tag{3}$$

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 marginals of $\boldsymbol{X}(t)$ by combining the Doob inequality for martingales with the Cramér-Chernoff method. For this reason, we call such processes g-DCC (in reference to the Doob-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-DCC 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*

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

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

Remark 2 *To calibrate what to expect for g, we can use knowledge of the asymptotic distribution of the $X_a(t)$ given in (3). In our applications, Wilk's 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. For $2Y \sim \chi^2$, we have $e^{\lambda Y} = (1-\lambda)^{-1/2}$. 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-DCC stochastic process $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$, we provide a general deviation inequality for the sum of the marginals $X_a(t)$ 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-DCC deviation inequality.

Lemma 4 *Fix $\mathcal{S} \subseteq \{1,\ldots,K\}$. Let $\boldsymbol{X}(t) = \{X_a(t)\}_{a=1}^K$ be a g-DCC 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-DCC (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 Doob's inequality that

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a\in\mathcal{S}} X_a(t) > u\right) \leq e^{-[\lambda u - |\mathcal{S}| g(\lambda)]} \tag{4}$$

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

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

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

If the function g is strictly convex (which will be the case for all the examples studied later in this paper), it is also possible to compute C^g directly (either in closed form or numerically using e.g. binary search) by using the following observation.

Proposition 6 *If g is C^1 and strictly convex, the derivative of $G(\lambda) = \frac{g(\lambda)+x}{\lambda}$ has at most one zero, given by the solution to*

$$\lambda g'(\lambda) - g(\lambda) = x. \tag{6}$$

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

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

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, the *mixture martingale*

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

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* $Z_{\mathcal{S}}^{\boldsymbol{\pi}}(t) := \prod_{a\in\mathcal{S}} Z_a^{\pi_a}(t)$ 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-DCC 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. By exhibiting two different types of hierarchical priors, we first prove deviation inequalities for Gaussian and Gamma distributions in Section 3, followed by a broader result that applies to any exponential family in Section 4 .

Related work. The first use of such a mixture martingale can be traced back to the work of Robbins (1970) which considers 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^π for a single standard Gaussian arm a. It is used 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 $\sqrt{2t\ln\ln(t)}$. This "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 more general stochastic processes. It has later been used by Balsmubramani (2015) who propose time-uniform Hoeffding or Bernstein deviation inequalities and by Abbasi-Yadkori et al. (2011) who propose a self-normalized deviation inequality for a vector-valued martingale applied to the linear bandit problem. 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 some hierarchical priors. The recent work by Howard et al. (2018) is also of note, as it studies in great detail the power of elementary martingales for bounding the probability of crossing linear thresholds. We develop mixture martingale methods for obtaining curved thresholds, as hinted at in (Howard et al., 2018, Section 4.3).

3. Deviation Inequalities for Gaussian and Gamma Distributions

We first propose a general assumption for an exponential family under which a deviation inequality for a sum over multiple arms of the quantities $N_a(t)d(\hat{\mu}_a(t), \mu_a)$ can be obtained though Lemma 4. 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.

Assumption 7 *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 $t \geq 1$,*

$$\forall x \in \mathcal{I}, \quad \int p_t^{\lambda,\mu}(\eta) e^{\eta t x - \phi_\mu(\eta) t} \,\mathrm{d}\eta = e^{\lambda t d(x,\mu)}. \tag{9}$$

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

$$p_{n_1}^{\lambda,\mu}(\eta) \;\geq\; \sqrt{\frac{n_1}{n_2}} p_{n_2}^{\lambda,\mu}(\eta). \tag{10}$$

Theorem 8 *Assume that Assumption 7 is satisfied and let*

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

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

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

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 $g_{\eta,c}$-DCC where

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

Theorem 8 directly provides a deviation inequality using Lemma 4. It thus remains to find sequences of priors satisfying Assumption 7. We now discuss two examples, Gaussian and Gamma distributions, for which we were able to exhibit such priors. 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 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

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

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

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 7.

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

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 $N_a(t) \in [T_i, T_{i+1}[$,

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

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 $M_a^\lambda(t) = 1 > e^{-\lambda\infty}$. ■

Remark 9 *We use as our correction function $c\ln(4+\ln N_a(t))$, 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 $c\ln(4+\ln(1+N_a(t)))$.*

3.1 Application to Gaussian Distributions

In the Gaussian case, direct computations show that Assumption 7 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 (10) is actually an equality. Using Theorem 8, one can prove the following.

Corollary 10 *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-DCC where*

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

Hence for every subset $\mathcal{S}$ and $x>0$,

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

Proof of Corollary 10 By Theorem 8, 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

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

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

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

which concludes the proof.

3.2 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

$$p_t^\lambda(\eta) \coloneqq \frac{\mu}{\alpha} \frac{(\alpha t/e)^{\lambda\alpha t}}{\Gamma(\lambda\alpha t)} \left(1 - \frac{\eta\mu}{\alpha}\right)^{-\alpha t} \left(\lambda - \frac{\eta\mu}{\alpha}\right)_+^{\lambda\alpha t - 1}. \tag{11}$$

leads to suitable "priors".

Proposition 11 *The family of functions defined in* (11) *satisfies Assumption* 7*.*

Proof Proving (9) 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 (10). 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. ∎

Corollary 12 *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-DCC 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 for every subset $\mathcal{S}$ and $x > 0$,

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

Proof of Corollary 12 In order to evaluate the function $g_0(\lambda, \xi, c)$ featured in Theorem 8, 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 13 *$C_0(t,\lambda)$ 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

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

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 8 (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 10, 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-DCC (see Definition 1) for the function $g_\Gamma(\lambda) = 2\lambda - 2\lambda\ln(4\lambda) + \ln\zeta(2\lambda) - \ln(1-\lambda)$.

4. General Deviation Inequalities for Exponential Families

Define $d^+(u,v) = d(u,v)\mathbb{1}_{(u \leq v)}$ and $d^-(u,v) = d(u,v)\mathbb{1}_{(u \geq v)}$. In this section we will provide a deviation result that holds for any one-dimensional exponential family and can also accommodate *one-sided deviations*. We 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 $\forall t, X_a(t) = Y_a(t)$ or $\forall t, X_a(t) = Y_a^+(t)$ or $\forall t, X_a(t) = Y_a^-(t)$.

4.1 Main result

We provide in Theorem 14 a new self-normalized deviation inequality featuring a threshold function $\mathcal{T}$. As can be seen in the proof given below, this results follows by exhibiting *a family of functions* g_ξ such that $\boldsymbol{X}(t)$ is g_ξ-DCC, applying Lemma 4 and then optimizing the parameters to obtain the best possible threshold. The family of associated martingales will still be mixture martingales, that rely on different types of hierarchical priors.

To state the main result 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

$$\tilde{h}_z(x) \;=\; \begin{cases} e^{1/h^{-1}(x)} h^{-1}(x) & \text{if } x \geq h^{-1}(1/\ln z), \\ z(x - \ln \ln z) & \text{o.w.} \end{cases} \tag{12}$$

Next we state our main deviation inequality, making precise (2), in terms of this function.

Theorem 14 *Let $\mathcal{T} : \mathbb{R}^+ \to \mathbb{R}^+$ be the function defined by*

$$\mathcal{T}(x) \;=\; 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x) + \ln(2\zeta(2))}{2}\right) \tag{13}$$

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

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

Proposition 15 below (proved in Appendix C) gives a tight bound on the inverse function h^{-1}, which yields an upper bound on the threshold function $\mathcal{T}$ featured in Theorem 14. On can easily see that $\mathcal{T}(x) \sim x$ when x tends to infinity. For $x \geq 5$, a good approximation of the threshold is $\mathcal{T}(x) \simeq x + 4\ln(1 + x + \sqrt{2x})$, which is slightly larger than the approximation $\simeq x + \ln(x)$ that is added for comparison to Figure 2.

Proposition 15 *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,*

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

Remark 16 *It is perfectly reasonable to have each arm come from its own specific exponential family. Theorem 14 applies, now with each arm's deviation measured in the associated divergence $d(\cdot, \mu_a)$.*

4.2 Comparison and Positioning of our Results

The three deviation inequalities given in Corollaries 10 and 12 and Theorem 14 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 threshold 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 threshold function $\mathcal{C}$, which can be set to C^{g_G} for Gaussian distributions, to C^{g_Γ} for Gamma distributions and to $\mathcal{T}$ for general exponential families.

Those three threshold functions are hard to compare at first as they have no closed-form expressions. Equation (13) provides an explicit expression for $\mathcal{T}$ but that still requires to numerically inverse the function h, while C^{g_G} and C^{g_Γ} can be numerically approximated by using Proposition 6. In Figure 2 we compare those three thresholds to the "ideal" threshold $C^{g_{\chi^2}}$ where $g_{\chi^2}(\lambda) = -\frac{1}{2}\ln(1-\lambda)$ (see Remark 2). We see that that this idealized threshold satisfies $C^{g_{\chi^2}}(x) \simeq x + \ln(x)$ and that the thresholds obtained for Gaussian and Gamma distributions are very close to it. The threshold function $\mathcal{T}$ seems to be off by an additive term of order 10.

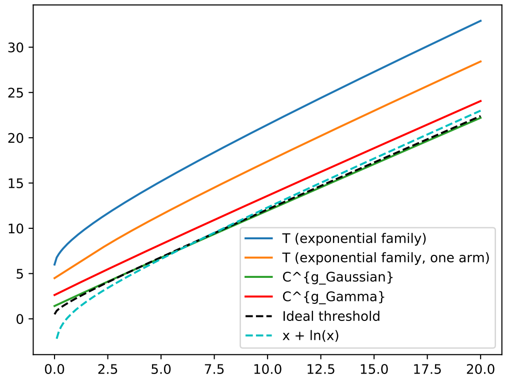


Figure 2: Several threshold functions $\mathcal{C}(x)$ as a function of x.

Despite this slightly larger threshold, our general exponential family result is interesting for the following reasons. First, obviously it covers more distributions like Bernoulli distributions that are often relevant for applications of multi-armed bandits. Then, one can note that Theorem 14 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): $\mathcal{T}$ can be replaced by a slightly smaller threshold in that case, as mentioned below in the proof of Theorem 14, by choosing a prior supported only on positive or negative values. However, the method discussed in Section 3 cannot be adapted to obtain better results for one-sided deviations. Finally, the presence of the positive part in the definition of $Y_a(t)^\pm$ lead to the following improved result:

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

Our results generalize in several directions existing results from the literature. As mentioned in the Introduction, the one-armed Gaussian case has been extensively studied Robbins (1970); Jamieson et al. (2014); Kaufmann et al. (2016); Zhao et al. (2016), but few results are available for more general exponential families and/or subset of size larger than one. We review them now and provide a detailed comparison with our results.

For general one-dimensional exponential families, the only available results are uniform over a *bounded time interval* $\{1, \ldots, n\}$. Garivier and Cappé (2011) provide a first result for a subset of size one that can be rephrased in the following way:

$$\mathbb{P}\Big(\exists t \leq n : N_a(t)d^+(\hat{\mu}_a(t), \mu_a) \geq h^{-1}\left(1 + \ln\ln(n) + x\right)\Big) \leq e^{-x}. \tag{14}$$

This result was later extended by Magureanu et al. (2014) for Bernoulli distributions and a subset of size K (although their analysis actually extends easily to one-dimensional exponential families and an arbitrary subset $\mathcal{S}$). More precisely, Theorem 2 in Magureanu et al. (2014) can be rephrased as follows, introducing the function $\tilde{f}(u) = u - 2\ln(u)$ for $u \geq 2$:

$$\mathbb{P}\left(\exists t \leq n : \sum_{a \in \mathcal{S}} N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq |\mathcal{S}| \tilde{f}^{-1}\left(1 + \ln\ln(n) + \frac{x+1}{|\mathcal{S}|}\right)\right) \leq e^{-x}. \tag{15}$$

Theorem 14 can also be used to obtain deviations that are uniform over a bounded time interval, for example for general exponential families:

$$\mathbb{P}\left(\exists t \leq n : \sum_{a \in \mathcal{S}} N_a(t) d^+(\hat{\mu}_a(t), \mu_a) \geq 3\ln(1 + \ln(n)) + |\mathcal{S}| \mathcal{T}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x} \tag{16}$$

and with the corresponding improved thresholds in the Gaussian and Gamma case. Numerically, it appears that the threshold featured in (16) is smaller than the threshold in (15), as illustrated in Figure 3. However, in the particular case $|\mathcal{S}| = 1$, (14) is the tightest result. Compared to those two related works in exponential families, note that our work is the only one that makes use of mixture martingales.

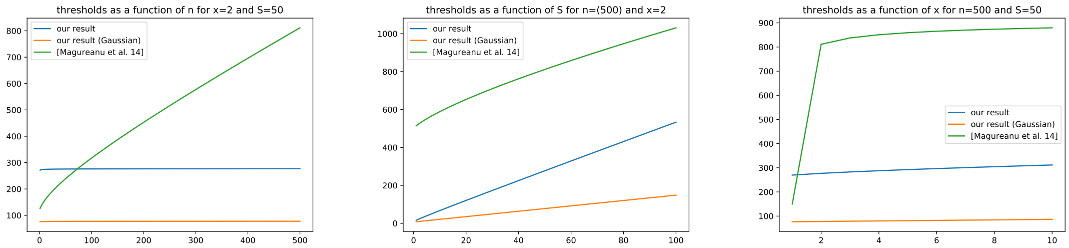


Figure 3: Thresholds that follow from Corollary 10 and Theorem 14 compared to that obtained by Magureanu et al. (2014)

To the best of our knowledge, we provide the first deviations results that hold uniformly for $t \in \mathbb{N}$ for multiple arms and beyond Gaussian distributions. As we shall see in the next section, those types of results are very useful for analyzing sequential tests, that involve *random stopping*.

4.3 Proof of Theorem 14

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 $(*)$. We first derive the inequality in Theorem 14 based on Lemma 17 below and later prove this result. As will be seen in the proof of Lemma 17, 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. This improvement yields the same statement as Theorem 14 with the following slightly smaller threshold (omitting the factor 2):

$$\mathcal{T}(x) \;=\; 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x)+\ln(\zeta(2))}{2}\right).$$

Lemma 17 *For $\xi \in [0,1/2]$, $\boldsymbol{X}$ is g_ξ-DCC (see Definition 1).*

Using Lemma 4, one can obtain a deviation inequality expressed with the threshold function C^{g_ξ} or the conjugate function g^*_ξ. The proof is completed by applying Lemma 18 below, proved in Appendix C.2, to compute the optimal tuning of $\xi \in [0,1/2]$.

Lemma 18 *Let $C(\xi) = \frac{2\zeta(2)}{(\ln(1+\xi))^2}$. Fix $z \in [0, e-1]$ and $x \geq 0$. Then*

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

Proof of Lemma 17: building the martingale Lemma 19 below shows that the deviations of $X_a(t)$ can be related to the deviations of a well-chosen mixture martingale $Z^\pi_a(t)$, where π has a discrete support. The proof of Lemma 19 is given in Appendix A.

Lemma 19 (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 (see (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

$$\left\{X_a(t) - (1+\xi)\ln\left(\frac{\zeta(2)}{(\ln(1+\xi))^2}\right) \geq x\right\} \subseteq \left\{Z_a^{\pi(x)}(t) \geq e^{\frac{x}{1+\xi}}\right\}. \tag{17}$$

Let us continue with the proof of Lemma 17. A consequence of Lemma 19 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,

$$W_a^{z,\lambda}(t) \geq \mathbb{1}_{\left(e^{\lambda(X_a(t)-(1+\xi)\ln C(\xi))} \geq z\right)}. \tag{18}$$

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 (18) 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 (17) 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.

5. Asymptotically Optimal Adaptive Sequential Testing

We now 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 $\forall \boldsymbol{\mu} \in \mathcal{O}$, $\mathbb{P}_{\boldsymbol{\mu}}\left(\boldsymbol{\mu} \in \mathcal{O}_{\hat{\imath}}\right) \geq 1-\delta$. 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 propose below the *extended GLR* stopping rule. 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.

5.1 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 partitioning to which $\boldsymbol{\mu}$ belongs; in particular $\boldsymbol{\mu} \in \mathcal{O}_{i^*(\boldsymbol{\mu})}$ and $\mathrm{Alt}(\boldsymbol{\mu}) = \mathcal{O}\backslash\mathcal{O}_{i^*(\boldsymbol{\mu})}$. 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 20 *The* extended GLR statistic *is defined as*

$$\hat{\Lambda}_t = \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{K} N_a(t) d\left(\hat{\mu}_a(t), \lambda_a\right). \tag{19}$$

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

$$\tau_\delta := \inf\left\{t \in \mathbb{N} : \hat{\Lambda}_t > \hat{c}_t(\delta)\right\}. \tag{20}$$

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

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

where $X_1, \ldots, X_t$ are some observations whose likelihood $\ell(X_1, \ldots, X_t; \mu)$ 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 extended GLR statistic $\hat{\Lambda}_t$ can thus be interpreted as a statistic for testing $\mathcal{H}_0 : (\boldsymbol{\mu} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t)))$ 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, hence the denomination "extended" GLR. Still, large values $\hat{\Lambda}_t$ tend to reject $(\boldsymbol{\mu} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t)))$: hypothesis $\hat{\imath}(t)$ must be true.

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*

$$\mathcal{C}_t(\delta) := \left\{\boldsymbol{\lambda} : \sum_{a=1}^{K} N_a(t) d(\hat{\mu}_a(t), \lambda_a) \leq \hat{c}_t(\delta)\right\}. \tag{21}$$

The extended GLR stopping rule (20) 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 partitioning. Building on Theorem 14, Proposition 21 below provides a choice of thresholds for which the extended GLR stopping rule yields a δ-correct algorithm. We provide a choice of thresholds for which the extended 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 21 *Let $\mathcal{T}$ be the threshold function defined in Theorem 14. The sequence of thresholds*

$$\hat{c}_t(\delta) = 3\sum_{a=1}^{K} \ln(1 + \ln N_a(t)) + K\mathcal{T}\left(\frac{\ln(1/\delta)}{K}\right) \tag{22}$$

is such that, for every sampling rule*,*

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

Proof Using Theorem 14 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 21. ■

5.2 An Asymptotically Optimal Adaptive Testing Procedure

Proposition 21 provides a threshold for which the extended GLR stopping rule (20) 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 trick.

Proposition 22 *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*

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

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

$$A_{t+1} \in \begin{cases} \underset{a \in \mathcal{U}_t}{\text{argmin}}\ N_a(t) & \text{if } \mathcal{U}_t \neq \emptyset \quad \textit{(forced exploration)} \\ \underset{a \in [K]}{\text{argmax}}\ t\, w_a^*(\hat{\boldsymbol{\mu}}(t)) - N_a(t) & \text{o.w.} \quad\quad \textit{(tracking the plug-in estimate)} \end{cases} \tag{23}$$

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, if $\boldsymbol{\mu} \in \mathcal{O}$ 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 23 *Assume that the following assumptions are satisfied:*

1. *For every $\boldsymbol{\mu}$, 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 extended GLR stopping rule (20) with thresholds $\hat{c}_t(\delta)$. The Tracking rule (23) ensures

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

The proof of Theorem 23 is given in Appendix B. Combining this result with Proposition 21 yields that an adaptive sequential test using the Tracking rule and the extended GLR stopping rule with thresholds (22) is δ-correct for every $\delta \in (0, 1]$ and its sample complexity is asymptotically matching the lower bound of Proposition 22, 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 (22) can be used in some cases.

Of the two assumptions in Theorem 23, we believe that the continuity assumption 2 is very mild in practice. Continuity of the highly related oracle regret problem for the structured multi-armed bandit problem in the fixed-budget setting, was recently proved by Combes et al. (2017, Lemma 1) under a unique optimiser assumption similar to 1. Analogous methods will undoubtedly yield continuity for pure identification problems under mild assumptions on $\{\mathcal{O}_i\}_i$.

6. Smaller Thresholds for Better Sequential Tests

A stylized form of (two-sided) deviation inequalities obtained in this paper (in Corollaries 10 and 12 and Theorem 14) is the following. For any subset of arms $\mathcal{S} \subseteq \{1, \dots, K\}$, for all x large enough,

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \mu_a) > \sum_{a \in \mathcal{S}} c \ln(d + \ln(N_a(t))) + |\mathcal{S}| \mathcal{C}\left(\frac{x}{|\mathcal{S}|}\right)\right) \leq e^{-x} \tag{24}$$

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 (24) with a weighted union bound, one obtains in Lemma 24 below a deviation inequality that is uniform over subsets belonging to the support of the "prior" $\tilde{\pi}$.

Lemma 24 (weighted union bound) *Assume* (24) *holds. Let $\tilde{\pi}$ be a probability distribution over subsets: $\sum_{S\subseteq\{1,\dots,K\}} \tilde{\pi}(\mathcal{S}) = 1$. Then for all $x > 0$*

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

We now explain how this result can serve to tighten the analysis of the extended GLR stopping rule for some particular sequential testing problems, to allow for the use of smaller threshold functions. We later discuss in Section 7 the impact of this result on the design of confidence regions.

6.1 Improved Stopping Rules for Best Arm Identification

The (fixed-confidence) Best Arm Identification problem is a particular sequential identification problem as defined in Section 5 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 extended GLR statistic (19) rewrites to

$$\hat{\Lambda}_t = \min_{b\neq\hat{\imath}(t)} \min_{\lambda_b > \lambda_{\hat{\imath}(t)}} \left[N_{\hat{\imath}(t)}(t) d(\hat{\mu}_{\hat{\imath}(t)}(t), \lambda_{\hat{\imath}(t)}) + N_b(t) d(\hat{\mu}_b(t), \lambda_b)\right] \tag{25}$$

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 (22) 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:

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

From Theorem 14 and a union bound over the $K - 1$ subsets $\{1, 2\}, \dots, \{1, K\}$ (Lemma 24 with a prior $\tilde{\pi}(\{1, a\}) = 1/(K - 1)$ for $a \neq 1$) it holds that

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

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

$$\hat{c}_t(\delta) = 6\ln\left(\ln\left(\frac{t}{2}\right) + 1\right) + 2\mathcal{T}\left(\frac{\ln\frac{K-1}{\delta}}{2}\right). \tag{26}$$

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 23.

Remark 25 *The improved threshold (26) yields a δ correct stopping rule, however the corresponding confidence interval (21) does not satisfy $\mathbb{P}\left(\forall t \in \mathbb{N} : \boldsymbol{\mu} \in \mathcal{C}_t(\delta)\right) \geq 1 - \delta$. There is no equivalence between the improved δ-correct stopping rule and improved δ-valid confidence regions. We will discuss the implications of Lemma 24 for confidence regions in Section 7.*

6.2 Smaller Thresholds for More General Tests

The reason why we are able to propose a smaller threshold for the BAI problem is that for it the extended GLR statistic (25) 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 (22) but on a smaller *effective number* of arms.

Definition 26 *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 $i \in \{1, \dots, M\}$ we can write*

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

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 extended 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 (25) in the BAI case.

Proposition 27 *Fix an identification problem of rank R. Then the extended GLR stopping rule (19) is δ-correct with threshold*

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

Proof Fix $\boldsymbol{\mu} \in \mathcal{O}$. For each $i \neq i^*$, $\boldsymbol{\mu} \in \mathcal{O} \setminus \mathcal{O}_i$, thus from Definition 26 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

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

where the last inequality follows from Theorem 14 and a union bound over $M-1$ subsets (Lemma 24 with a prior $\tilde{\pi}(\{k_1^{i,q_i}, \dots, k_R^{i,q_i}\}) = 1/(M-1)$ for $i \neq i^*$) together with the concavity of $s \mapsto \ln(1 + \ln(s))$ that ensures

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

■

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 extended 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 27 the extended GLR stopping rule (20) is δ-correct with the threshold

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

Remark 28 *For Largest Profit Identification the oracle weights $\boldsymbol{w}^*(\boldsymbol{\mu})$, which are needed for implementing the asymptotically optimal procedure of Section 5.2, 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).*

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 numbers of actions of the first and second player. This is much smaller than the number of leaves, which is $K \cdot L$. Assuming the weights (which are only numerically computable) satisfy the continuity assumption of Theorem 23, the extended GLR rule with a rank $L + 1$ threshold is asymptotically optimal in combination with the Tracking rule.

7. 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 24 in combination with our deviation results allows to build such confidence regions. Indeed for any prior $\tilde{\pi}$ over subsets, the following confidence interval is δ-uniformly valid (with c and d as given by the Lemma):

$$\mathcal{C}_t^{\tilde{\pi}}(\delta) := \left\{\boldsymbol{\lambda} : \forall \mathcal{S}, \sum_{a \in \mathcal{S}} N_a(t) d(\hat{\mu}_a(t), \lambda_a) \leq c \sum_{a \in \mathcal{S}} \ln(d + \ln N_a(t)) + |\mathcal{S}| \mathcal{T}\left(\frac{\ln(1/(\tilde{\pi}(\mathcal{S})\delta))}{|\mathcal{S}|}\right)\right\}. \quad (27)$$

A natural question is thus which prior $\tilde{\pi}$ 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 $f(\boldsymbol{\mu}) \in \mathcal{I}_t(\delta) := \{f(\boldsymbol{\lambda}) | \boldsymbol{\lambda} \in \mathcal{C}_t\}$. 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.

7.1 Linear functions

In this section we consider an arbitrary linear function $f(\boldsymbol{\mu}) = \boldsymbol{c}^\intercal \boldsymbol{\mu}$ where $\boldsymbol{c} \in \mathbb{R}^K$. We will derive our results in the Gaussian case because it admits revealing and explicit closed-form expressions. In that case the confidence region (27) is δ-uniformly valid for $c = 2$ and $d = 4$ and $g = g_G$, as licensed by Corollary 10. The following two confidence intervals on $\boldsymbol{c}^\intercal \boldsymbol{\mu}$ follow from two extreme prior choices: a prior supported on all the singleton sets or on the full set.

Proposition 29 (Box) *The following is a δ-uniformly valid confidence interval on $\boldsymbol{c}^\intercal \boldsymbol{\mu}$*

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

Proof Simple algebra show that $\mathcal{I}_t(\delta) = \{\boldsymbol{c}^T\boldsymbol{\lambda}, \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi}}(\delta)\}$ where $\tilde{\pi}$ is uniform of 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 30 (Ellipse) *The following is a δ-uniformly valid confidence interval on $\boldsymbol{c}^\intercal \boldsymbol{\mu}$*

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

Proof We show that $\mathcal{I}_t(\delta) = \{\boldsymbol{c}^T\boldsymbol{\lambda}, \boldsymbol{\lambda} \in \mathcal{C}_t^{\tilde{\pi}}(\delta)\}$ where $\tilde{\pi}$ is a point-mass on the whole set: $\tilde{\pi}(\{1, \ldots, K\}) = 1$. Letting $C = \sum_{a=1}^K \ln(1 + \ln N_a(t)) + K\mathcal{T}(\ln(1/\delta)/K)$, computing the upper bound of this confidence interval requires to compute

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

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

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

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

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

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

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

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

We see that the box bound depends on the one-norm of $\boldsymbol{c}$, whereas the ellipse bound depends on the two-norm of $\boldsymbol{c}$, 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 $N_a(t) = t\frac{|c_a|}{\sum_a |c_a|}$, 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 $\sqrt{K}$.

7.2 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

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

that are both δ-uniformly valid by Lemma 24. It follows that $\mathbb{P}\left\{\forall t \in \mathbb{N} : \min_a \mu_a \le \mathrm{U}_t^{\tilde{\pi}}(\delta)\right\} \ge 1-\delta$ and $\mathbb{P}\left\{\forall t \in \mathbb{N} : \min_a \mu_a \ge \mathrm{L}_t^{\tilde{\pi}}(\delta)\right\} \ge 1-\delta$. We investigate in each case the tightest possible confidence bound that can be obtained by optimising the choice of the prior $\tilde{\pi}$.

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

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

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

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

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

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

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 $\min_a \lambda_a$. ■

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 32 how to compute $\mathrm{U}_t^{\tilde{\pi}}$ for a general prior $\tilde{\pi}$. We then show that empirically a prior supported on *all subsets* can be helpful.

Proposition 32 *Let $\theta(t)$ be the solution in θ to the equation*

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

Then $\mathbb{P}\left\{\forall t \in \mathbb{N}, \min_a \mu_a \le \theta(t)\right\} \ge 1-\delta$.

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

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

What does this tell us about $\min_{a\in[K]} \lambda_a$? Well, consider a candidate value $\theta \ge \min_a \hat{\mu}_a(t)$ 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

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

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

Practical choice of prior 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 priors of the form $\tilde{\pi}(\mathcal{S}) = \pi(|\mathcal{S}|)/\binom{K}{|\mathcal{S}|}$ for some prior π on sizes $[K]$.

First, let's consider the computation of $\theta(t)$ for those priors: 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

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

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 on 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" prior that is uniform on the singletons ($\pi(1) = 1$), a prior supported on the whole set ($\pi(K) = 1$), a prior that is uniform over subset sizes ($\pi(k) = 1/K$) and a "Zipf" prior that gives more weights to smaller subset sizes ($\pi(k) \propto 1/k$). 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 prior (or the Zipf prior which performs almost identically) leads to smaller upper confidence bounds when compared to Box when M increases.

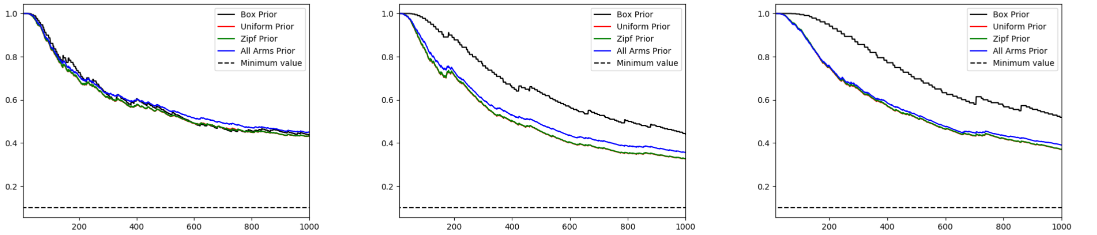


Figure 4: $\mathrm{U}_t^{\tilde{\pi}}(\delta)$ as a function of t for several cardinality-based priors $\tilde{\pi}$ in a presence of $M = 1$ (left), $M = 5$ (middle) and $M = 10$ (right) identical arms with the minimal mean.

This experiment shows that for small values of δ a uniform cardinality-based prior 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.

8. 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 (extended) 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 grant ANR-16-CE40-0002 (project BADASS) 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.

References

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

A. Balsmubramani. 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.

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

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1763–1771. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6773-minimal-exploration-in-structured-stochastic-bandits.pdf.

A. Philip Dawid, Steven de Rooij, Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Insuring against loss of evidence in game-theoretic probability. *Statistics & Probability Letters*, 81(1):157 – 162, 2011. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2010.10.013. URL http://www.sciencedirect.com/science/article/pii/S0167715210002968.

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 Shao Q. *Self-normalized processes. Limit Theory and Statistical applications*. Springer, 2009.

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, 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.

Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2017.

Steve Howard, Aaditya Ramdas, John McAuliffe, and Jasjeet Sekhon. Exponential line-crossing inequalities. *ArXiv e-prints*, August 2018.

Ruitong Huang, Mohammad M. Ajallooeian, Csaba Szepesvári, and Martin 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.

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.

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

Emilie Kaufmann, Wouter M. Koolen, and Aurélien Garivier. Sequential test for the lowest mean: From Thompson to Murphy sampling. Accepted to Advances in Neural Information Processing Systems (NIPS) 31, December 2018. URL https://arxiv.org/pdf/1806.00973.pdf.

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

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.

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.

H. Robbins and D. Siegmund. Boundary crossing probabilities for the wiener process and sample sums. *Annals of Mathematical Statistics*, 41(5):1410–1429, 1970.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir 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.

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

Appendix A. Details for exponential families: proof of Lemma 19

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 33, 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 33 *Fix $i \in \mathbb{N}^*$, $x > 0$ and $\xi > 0$. There exists $\eta_i^+(x, \xi)$ 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

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

where $\eta_i^\pm(x, \xi)$ are defined in Lemma 33. 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 $C(\xi) = \frac{\ln\zeta(2)}{(\ln(1+\xi))^2}$.

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

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, assuming 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 33,

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

Proof of Lemma 33 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

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

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

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

which concludes the proof.

A.1 One-arm bounds

Lemma 19 allows us to directly derive valid thresholds involving only a single arm. Namely, we have

Corollary 34 *Let $\tilde{h}_z(x)$ be as defined in (12). For every arm a and confidence parameter $x \geq 0$*

$$\mathbb{P}\left\{ X_a(t) \geq 2\tilde{h}_{3/2}\left(\frac{x + \ln(2\zeta(2))}{2} \right) \right\} \leq e^{-x}.$$

Proof By Lemma 19, for every $\xi \in [0, 1/2]$,

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

Minimising the threshold w.r.t. ξ using Lemma 39 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 14 has $h^{-1}(1+x) > x$ where Corollary 34 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 B. Optimal sample complexity: Proof of Theorem 23

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 35 *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

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

By Lemma 35, 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 36 *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 35, there exists a constant T_ϵ such that for $T \geq T_\epsilon$, it holds that on $\mathcal{E}_T(\epsilon)$,

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

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

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

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 $\text{Alt}(\boldsymbol{\mu})$ 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 14 which is related to the Lambert function. One has

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

Using Proposition 15, it follows that

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

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

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

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

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

To conclude, the lower bound of Proposition 22 implies that this inequality is an equality.

Appendix C. Technical results

C.1 Proof of Proposition 15

We may write

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

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

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

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

C.2 Tight Tuning: Proof of Lemma 18

In this section we prove Lemma 18, which gives the tightest possible tuning achievable with our method. We first prove two auxiliary lemmas.

Lemma 37 *Let $x \ge 0$. 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

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

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

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

We can now make the connection to (12).

Lemma 39 *Fix $z \in [1,e]$. Then*

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

Proof The objective in Definition 38 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 18) We reorganise, apply Lemma 37 and then Lemma 39 to find

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

■