Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals

Emilie Kaufmann, Wouter M. Koolen

Journal of Machine Learning Research 22 (2021) 1-45 Submitted 11/18; Revised 9/21; Published 11/21

Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals

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

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

Editor: Csaba Szepesvári

Abstract

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

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

1.1 Novelty of our Concentration Results

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

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

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

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

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

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

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

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

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

1.2 Applications to Sequential Learning

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

1.2.1 Related Work on Bandits

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

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

1.2.2 Our Contributions

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

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

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

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

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

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

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

2. Martingales and Deviation Inequalities for Exponential Family Bandit Models

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

2.1 Exponential Family Bandit Models

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