Informational Confidence Bounds for Self-Normalized Averages and Applications

Aurélien Garivier

To cite this version:

Informational Confidence Bounds for Self-Normalized Averages and Applications

Aurélien Garivier
Institut de Mathématiques de Toulouse
Université Paul Sabatier
118 route de Narbonne 31062 Toulouse cedex 9
Email: aurelien.garivier@math.univ-toulouse.fr

***Abstract*—We present deviation bounds for self-normalized averages and applications to estimation with a random number of observations. The results rely on a peeling argument in exponential martingale techniques that represents an alternative to the method of mixture. The motivating examples of bandit problems and context tree estimation are detailed.**

I. Introduction

Contrary to a very usual assumption in statistics, some situations require parameter estimation based on samples of *random size*. Let us first briefly present two probabilistic models of that kind which motivated the derivation of the results presented below.

A. Motivating examples

1) Bandit Problems: Estimation is sometimes used as a intermediate step in a decision process, and the results can influence the presence of further observations. Paradigmatic of this situation are *bandit problems*, named in reference to the archetypal situation of a gambler facing a row of slot-machines and sequentially deciding which one to choose in order to maximize her gains. The basic model is the following: an agent sequentially chooses actions in a finite set of possible options. Each action leads to an independent stochastic reward whose distribution is unknown. What dynamic allocation rule should she choose so as to maximize her cumulated reward? Originally motivated by medical trials, this simple model dates back to the 1930s ; it has recently raised a renewed interest because of computer-driven applications, from computer experiments to recommender systems and Big Data, and numerous variants have been considered (see [1] for a recent survey, and [2] for a related model).

One possible solution consists in constructing, at time t, a confidence interval based on all past observations for the expected reward associated to each action, then choosing the action with highest upper confidence bounds (UCB). This rule, popularized by [3], was recently improved and shown to have some optimality properties [4], [5]. Obviously, the number of observations used to construct the confidence interval strongly depends on the value of these observations, and standard formulas for fixed-size samples do not apply directly. The key element in the recent improvements of this algorithm was the introduction of the informational self-normalized deviation inequalities presented below.

2) Context Tree Estimation: Context tree models, introduced by Jorma Rissanen in [6] as efficient tools in Information Theory, have been successfully studied and used since then in many fields of Probability and Statistics, including Bioinformatics, Universal Coding, Mathematical Statistics or Linguistics. Sometimes also called Variable Length Markov Chain, a context tree process is informally defined as a Markov chain whose memory length depends on past symbols. This property makes it possible to represent the set of memory sequences as a tree, called the *context tree* of the process.

A remarkable tradeoff between flexibility and parsimony explains this success: no more difficult to handle than Markov chains, they include memory only where necessary. Not only do they provide more efficient models for fitting the data: it appears also that, in many applications, the shape of the context tree has a natural and informative interpretation. In Bioinformatics, they have been used to test the relevance of protein families databases [7] and in Linguistics, tree estimation highlights structural discrepancies between Brazilian and European Portuguese [8].

Of course, practical use of context tree models requires the possibility of constructing efficient estimators of the model generating the data. Despite the multiplicity of candidate trees, several procedures have been proposed and proved to be consistent, including pruning algorithms [6], and *Penalized Maximum Likelihood* estimators (see [9], [10] and references therein, see also [11]). These apparently different ideas are in fact closely related [12], the key point being an efficient estimation of the conditional transition probabilities. But for a given sample size, the number of transitions observed from a given context is random, and depends on the values of these transitions. Hence, again, sharp deviation bounds for random-sized averages are required in order to obtain efficient memory estimators.

B. Self-Normalized Process

Several approaches have been proposed to address this problem. The most obvious is to use a simple union bound on all the possible values of the sample size (as for instance in [3]), but this appears to be most often overly pessimistic and significantly sub-optimal. A more refined treatment consists, when possible, in first lower-bounding the size of the sample, thus upper-bounding the variance of the estimator, and then in using this upper-bound of the variance, for example with Bernstein's maximal inequality for martingales (see e.g. [13] for an example on the consistency of a Stochastic Block Model estimator, or [14] on prediction with expert advice).

The most satisfying approach, however, is to consider the associated *self-normalized process*. From the estimator's point of view, as the size of the sample grows, something changes only when a new observation appears. At those (random) times, the internal clock of the estimator increases by 1. When the nth observation has been reached, the internal clock has a random value which is at most equal to n, and on which the variance of the estimator depends. The confidence interval must be constructed accordingly, by taking into account the maximal deviations of the self-normalized deviation process.

This paper focuses on the case where a sequence of non-asymptotic confidence intervals $[a_t, b_t]$ is required for the common expectation μ of independent, real-valued random variables $(X_t)_{t\geqslant 1}$, and where all the confidence intervals are required to be jointly valid over an epoch $t \in \{1, \dots, n\}$. In other words, for all positive α, the goal is to construct $\sigma(X_1, \dots, X_t)$-measurable random variables a_t and b_t so as to ensure that the event $\bigcap_{t\leqslant n} \{\mu \in [a_t, b_t]\}$ has probability at least $1 - \alpha$.

In order to obtain sub-gaussian deviation bounds, the *method of mixture* (see [15], [16] and references therein) provides a powerful and elegant tool, recently used in [17] for bandit problems. The results presented below follow a different path. Rather than a mixture, they rely on a *peeling* device: the possible numbers of observations are divided into exponentially growing slices, which are treated independently. On each slice, a Cramer-type bound is obtained by a maximal inequality for martingales.

These results can be considered, in some sense, as non-asymptotic counterparts of the Law of the Iterated Logarithm for martingales. They are presented so as to clearly emphasize the cost of the randomness of the sample size: namely, a logarithmic factor of n in front of the exponential Cramer bound (instead of a factor n for the union bound). The proof method is generic enough to apply to a large variety of situations, and in particular not only to the sub-gaussian case.

C. Informational Confidence Bounds

If the bounds of these confidence intervals are classically chosen to be symmetric around the empirical mean $\bar{X}_t$, so that $\bar{X}_t - a_t = b_t - \bar{X}_t = c/\sqrt{t}$ for a given constant c, then the above discussion shows that one needs to control the following supremum of the self-normalized process:

$$\sup_{t\leqslant n} \sqrt{t} \left|\bar{X}_t - \mu\right| \ .$$

This choice, however, is often sub-optimal and was not sufficient in the applications mentioned above. The approach used below is somewhat different: the deviations of $\bar{X}_t$ are not measured in absolute value, but using a information deviation measure, leading to possibly asymmetric confidence bounds. Let us recall it briefly: suppose that, for all possible values of the expectation μ, the following Cramer-type inequality with rate function $I(\cdot, \mu)$ is satisfied:

$$\forall x_t \geqslant \mu, P(\bar{X}_t \geqslant x_t) \leqslant \exp(-tI(x_t; \mu)) \ .$$

For a concrete example, one may think about i.i.d. Bernoulli variables with $I(x; \mu) = \mathrm{kl}(x, \mu) = x\log(x/\mu) + (1 - x)\log((1 - x)/(1 - \mu))$. As the function $I(\cdot; \mu)$ increases on $[\mu, +\infty[$, this bound can be rewritten $P(I(\bar{X}_t; \mu) \geqslant I(x_t; \mu), \bar{X}_t \geqslant \mu) \leqslant \exp(-t\, I(x_t; \mu))$ or, defining $\delta = tI(x_t; \mu)$, $P(t\, I(\bar{X}_t; \mu) \geqslant \delta, \bar{X}_t \geqslant \mu) \leqslant \exp(-\delta)$; proceeding similarly on the other side of μ, one obtains

$$P\left(t\, I(\bar{X}_t; \mu) \geqslant \delta\right) \leqslant 2\exp(-\delta) \ .$$

Consequently, one is tempted to choose, as a confidence interval of risk α, a neighborhood of $\bar{X}_t$ in the sense of the pseudo-distance I:

$$[a_t, b_t] = \left\{\mu : t\, I(\bar{X}_t; \mu) \leqslant \log\frac{2}{\alpha}\right\} \ .$$

Observe that $\mu \in [a_t, b_t]$ if and only if $t\, I(\bar{X}_t; \mu) \leqslant \log\frac{2}{\alpha}$. For a *sequential* confidence intervals of this kind, where $P\left(\bigcap_{t\leqslant n} \{\mu \in [a_t, b_t]\}\right)$ needs to be controlled, one is thus led to study

$$\sup_{t\leqslant n} t\, I\big(\bar{X}_t; \mu\big) \ . \tag{1}$$

In Section II, deviation bounds for (1) are presented. The generic result of Theorem 1 is refined, under some additional hypotheses, in Theorem 2 and Equation (2). Theorem 3 contains a variant that does not require an upper-bound on the sample size. A subgaussian inequality is given for the discounted case in Equation (4). In Section III, these results are applied to estimation in various models: one-parameter canonical exponential famillies, bounded variables, and multinomial distributions.

II. Self-Normalized Deviation Inequalities

For an increasing filtration $(\mathcal{F}_t)_{t\geqslant 0}$ on some probability space, consider an adapted, real-valued discrete time process $(S_t)_{t\geqslant 0}$ such that $S_0 = 0$. Further assume that the increments $X_t = S_t - S_{t-1}$ are bounded as follows: there exist $\lambda_1 \in [-\infty, 0[$, $\lambda_2 \in]0, +\infty]$, and a function $\phi :]\lambda_1, \lambda_2[\to \mathbb{R}$ such that for all $\lambda \in]\lambda_1, \lambda_2[$ and for all $t \geqslant 1$:

$$\mathbb{E}\left[\exp(\lambda X_t) \,|\mathcal{F}_{t-1}\right] \leqslant \exp\left(\phi(\lambda)\right) \ .$$

In other words, the function ϕ dominates the logarithmic moment-generating function (lmgf) of the increments $(X_t)_t$ that are assumed to share the same finite expectation μ. If the increments X_t are identically distributed, ϕ can be chosen as the common lmgf, but it proves useful to consider more general cases. Nevertheless, ϕ will be supposed to satisfy all usual properties of a lmgf (see [18], Chapter 2) : ϕ is convex and smooth over $]\lambda_1, \lambda_2[$, $\phi(\mu) = 0$; its Legendre transform $I(\cdot; \mu)$,defined on $\mathbb{R}$ as

$$I(x; \mu) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \phi(\lambda)\} \ ,$$

is a convex rate function whose domain is included in $\mathbb{R}^+ \cup \{+\infty\}$; it is finite and smooth on an open interval $\mathcal{D}_I \subset \mathbb{R}$ containing 0, such that $I(\mu,\mu) = 0$. For all x such that $I(x) < \infty$, there exists a unique real number $\lambda(x) \in]\lambda_1, \lambda_2[$ such that

$$\phi'(\lambda(x)) = x \quad \text{and } I(x;\mu) = \lambda(x)x - \phi(\lambda(x)) \ .$$

$I(x;\mu)$ tends to infinity with x, and can be equal to $+\infty$ outside of some interval (x_-, x_+) where it is finite: it holds that $P(X_t \in [x_-, x_+]) = 1$, and the limit of $I(\cdot,\mu)$ when tends to x_+ is denoted I_+. Under those assumptions, the following result holds:

Theorem 1: For every $\delta > 0$,

$$P\left(\exists t \in \{1,\dots,n\} : t\, I(\bar{X}_t;\mu) \geqslant \delta\right) \\ \leqslant 2e \lceil \delta \log(n) \rceil \exp(-\delta) \ .$$

A. Short Proof of Theorem 1

The proof of this result, short enough to be sketched here, is inspired by the proof of the Law of the Iterated Logarithm for martingales that can be found in [19]. The epoch $\{1,\dots,n\}$ is divided into “slides” $\{t_{k-1}+1,\dots,t_k\}$ of exponentially increasing sizes: let $t_0 = 0$, let $\eta > 0$ and, for every positive integer k, let $t_k = \left\lfloor (1+\eta)^k \right\rfloor$. Denoting $D = \lceil \log(n)/\log(1+\eta) \rceil$ the smallest integer such that $t_D \geqslant n$, the union bound yields :

$$P\left(\bigcup_{t=1}^{n} \left\{ tI\left(\bar{X}_t;\mu\right) \geqslant \delta \right\}\right) \leqslant \sum_{k=1}^{D} P(A_k) \ ,$$

where $A_k = \bigcup_{t=t_{k-1}+1}^{t_k} \left\{ tI\left(\bar{X}_t;\mu\right) \geqslant \delta \right\}$. Denote by s the smallest integer such that $\delta/(s+1) \leqslant I_+$: for $t \leqslant s$, obviously $P(t\, I(\bar{X}_t;\mu) \geqslant \delta, \bar{X}_t > \mu) = 0$ and thus $P(A_k) = 0$ if $t_k \leqslant s$.

Let k be such that $t_k > s$, and $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$. For every $t \in \{\tilde{t}_{k-1}+1,\dots,t_k\}$, there exists $x_t \in [\mu, x_+]$ such that $t\, I(x_t;\mu) = \delta$. Let $\lambda_k = \lambda(x_{t_k})$, so that $I(x_{t_k};\mu) = \lambda_k x_{t_k} - \phi(\lambda_k)$, and consider the super-martingale $(W_t^k)_t$ defined by $W_0^k = 1$ and, for every $t \geqslant 1$, $W_t^k = \exp\left(\lambda_k S_t - t\phi(\lambda_k)\right)$. A maximal inequality ensures that, for all positive real c,

$$P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{ W_t^k \geqslant c \right\}\right) \leqslant \frac{1}{c} \ .$$

Let us deduce an upper-bound for $P(A_k)$. As $t\, I(x_t;\mu) = \delta$, it holds that

$$I(x_{t_k};\mu) \leqslant I(x_t;\mu) < I(x_{t_k};\mu)\,(1+\eta) \ .$$

As $I(\cdot;\mu)$ is increasing on the right side of μ, $x_t \geqslant x_{t_k}$ and

$$\lambda_k x_t - \phi(\lambda_k) \geqslant \lambda_k x_{t_k} - \phi(\lambda_k) = I(x_{t_k};\mu) \geqslant \frac{I(x_t;\mu)}{1+\eta} \ .$$

Hence, if $tI\left(\bar{X}_t;\mu\right) \geqslant \delta$ and $\bar{X}_t \geqslant \mu$, then $\lambda_k \bar{X}_t - \phi(\lambda_k) \geqslant \lambda_k x_t - \phi(\lambda_k) \geqslant \frac{\delta}{t(1+\eta)}$ and $\lambda_k S_t - t\phi(\lambda_k) \geqslant \frac{\delta}{1+\eta}$, and thus $W_t^k \geqslant \exp\left(\frac{\delta}{1+\eta}\right)$. This entails that

$$P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{ tI\left(\bar{X}_t;\mu\right) \geqslant \delta \right\} \cap \left\{ \bar{X}_t > \mu \right\}\right) \\ \leqslant P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{ W_t^k \geqslant \exp\left(\frac{\delta}{1+\eta}\right) \right\}\right) \\ \leqslant \exp\left(-\frac{\delta}{1+\eta}\right) \ .$$

The case $\bar{X}_t < \mu$ can be treated similarly, and the first claim of the theorem follows. The second claim is a consequence of the inequality $\log(1+1/(\delta-1)) \geqslant 1/\delta$, applied with the approximately optimal choice $\eta = \delta/(\delta-1)$.

Remark that the simple bound of Theorem 1 highlights the cost of time uniformity: a factor $2e\lceil \delta \log(n) \rceil$, instead as the factor n given by the union bound. The fact that this cost is sub-polynomial in n appears (especially in [4], [20], [12], [5]) to be crucial in the analysis of some algorithms and estimators.

B. Improvements and Variants

This result can be significantly improved under some additional assumptions on the function $I(\cdot;\mu)$:

Theorem 2: Let $\delta > 0$. If the function $I(\cdot;\mu)$ is log-concave, then for every $\eta > 0$

$$P\left(\exists t \in \{1,\dots,n\} : t\, I(\bar{X}_t;\mu) \geqslant \delta\right) \\ \leqslant 2\left\lceil \frac{\log n}{\log\left(1+\eta\right)} \right\rceil \exp\left(-\left(1 - \frac{\eta^2}{8}\right)\delta\right) \ .$$

In particular, for $\eta = 2/\sqrt{\delta}$, one obtains:

$$P\left(\exists t \in \{1,\dots,n\} : t\, I(\bar{X}_t;\mu) \geqslant \delta\right) \\ \leqslant 2\sqrt{e}\left\lceil \frac{\sqrt{\delta}}{2}\log(n) \right\rceil \exp(-\delta) \ .$$

The law of the Iterated Logarithm suggests that such a result is hardly improvable: in a sub-gaussian setting where $I(x;\mu) \geqslant (x-\mu)^2/(2\sigma^2)$, it implies indeed that for all $c > 1$:

$$P\left(\sup_{t \leqslant n} \frac{S_t - t\mu}{\sqrt{2\sigma^2 t \log\log(n)}} > c\right) \\ \leqslant P\left(\sup_{t \leqslant n} tI\left(\bar{X}_t;\mu\right) > c^2 \log\log(n)\right) \to 0$$

when n tends to infinity. Observe that the log-concavity of $I(\cdot,\mu)$, although not always satisfied (even for bounded variables), is reasonable at least locally around μ if one thinks to the gaussian regime.

Let us mention that in the quadratic (gaussian) case $I(x;\mu) = 2(x-\mu)^2/K^2$, the bound can be slightly improved:

$$P\left(\exists t \in \{1,\dots,n\} : t\, I(\bar{X}_t;\mu) \geqslant \delta\right) \\ \leqslant 2\left\lceil \frac{\log n}{\log\left(1+\eta\right)} \right\rceil \exp\left(-\left(1 - \frac{\eta^2}{16}\right)\delta\right) \ . \quad (2)$$

Finally, the method can be adapted in order to obtain non-asymptotic that hold for all $t \geqslant 1$ in the spirit of the Law of the Iterated Logarithm:

Theorem 3: For all $\delta > 1$ and all $c > 1$,

$$P\left(\exists t \geqslant 1 : t\, I(\bar{X}_t;\mu) \geqslant \frac{\delta c}{\delta - 1}\log\log t + \delta\right) \leqslant \frac{2\,\mathrm{e}\, c\delta^c}{c-1}\exp(-\delta)\ .$$

In particular, for $c = 1 + 1/\log(\delta)$, one obtains:

$$P\left(\exists t \geqslant 1 : t\, I(\bar{X}_t;\mu) \geqslant \frac{\delta(1+\log\delta)}{(\delta-1)\log\delta}\log\log t + \delta\right) \leqslant 2e^2\delta\exp(-\delta)\ .$$

C. Self-Normalized Form

In the applications mentioned above, the necessity to guarantee the joint validity of the confidence intervals over the entire epoch comes from the fact that the variables X_t are observed only episodically in a predictable way: there exists, for each $t \in \{1,\ldots,n\}$, a $\{0,1\}$-valued, $\mathcal{F}_{t-1}$-measurable random variable $\varepsilon_t \in \{0,1\}$ such that the current estimate at time n is

$$\bar{X}(n) = S(n)/N(n) \tag{3}$$

where $S(n) = \sum_{t=1}^n \varepsilon_t X_t$ and $N(n) = \sum_{t=1}^n \varepsilon_t$. Theorem 1 yields:

$$P\left(I\left(\bar{X}(n);\mu\right) \geqslant \frac{\delta}{N(n)}\right) \leqslant 2e\left\lceil\delta\log(n)\right\rceil\exp(-\delta)\ .$$

In the definition (3), $S(n)$ is written as a martingale transform or, equivalently, a discrete stochastic integral. Continuous-time variants of Theorem 1 can be obtained following the same lines (using the same peeling trick) for stochastic integrals.

Furthermore, this approach can be adapted to non-stationary contexts: assume for simplicity that the variables $(X_t)_t$ are independent, of expectation μ_t respectively, and that their absolute value is almost-surely bounded by B. If μ_t does not change too fast (or too often) with t, one may consider the discounted estimator $\bar{X}_\gamma(n)$ of μ_n defined by

$$\bar{X}_\gamma(n) = \frac{S_\gamma(n)}{N_\gamma(n)}\ ,$$

where $\gamma \in]0,1[$, $S_\gamma(n) = \sum_{t=1}^n \gamma^{n-t}\varepsilon_t X_t$ and $N_\gamma(n) = \sum_{t=1}^n \gamma^{n-t}\varepsilon_t$. The difference between $\bar{X}_\gamma(n)$ and μ_n can be decomposed into a term of bias (which is not discussed here) and a fluctuation term $\bar{X}_\gamma(n) - M_\gamma(n)/N_\gamma(n)$, where $M_\gamma(n) = \sum_{t=1}^n \gamma^{n-t}\varepsilon_t\mu_t$. This fluctuation term can be controlled by the adapting the martingale techniques above: one obtains that

$$P\left(\frac{S_\gamma(n) - M_\gamma(n)}{\sqrt{N_{\gamma^2}(n)}} \geqslant \delta\right) \leqslant \left\lceil\frac{\log\nu_\gamma(n)}{\log(1+\eta)}\right\rceil\exp\left(-\frac{2\delta^2}{B^2}\left(1-\frac{\eta^2}{16}\right)\right)\ , \tag{4}$$

with $\nu_\gamma(n) = \sum_{t=1}^n \gamma^{n-t} = (1-\gamma^n)/(1-\gamma) < \min\{(1-\gamma)^{-1}, n\}$. This results permits to analyze the *Discounted-UCB* algorithm [20] earlier proposed by Kocsis Szepesvári [21].

III. Application to Estimation

Let us now show briefly how these inequalities may be used in the analysis of some stochastic algorithms. The key point is that Theorem 1 allows the construction of a sequence of confidence intervals $([a_t,b_t])_{1\leqslant t\leqslant n}$ for μ that are simultaneously valid with high probability. The interval

$$[a_t,b_t] = \left\{\mu \in [x_-,x_+] : tI\left(\bar{X}_t;\mu\right) \leqslant \delta\right\}$$

contains all the values in a neighborhood of $\bar{X}_t$ in the sense of the pseudo-distance defined by I. By Theorem 1,

$$P\left(\bigcap_{t=1}^n \{\mu \in [a_t,b_t]\}\right) \geqslant 1 - 2e\left\lceil\delta\log(n)\right\rceil\exp(-\delta)\ .$$

Similarly, one obtains obtains confidence intervals for the case presented in Equation (3). This framework applies as well in bandit problems, where only the reward of the chosen arm is observed, that the estimation of Markovian models where, at each time, only the estimates relative to the current past observations are updated. Of course, in these examples, the identity of the variable(s) observed at time t is absolutely not independent of the past observations. By choosing δ such that $2e\left\lceil\delta\log(n)\right\rceil\exp(-\delta) \leqslant \alpha$, one obtains the confidence interval $\left\{\mu : I\left(\bar{X}(n);\mu\right) \leqslant \delta/N(n)\right\}$ of risk at most α.

A. One-Parameter Exponential Model and Bounded Variables

In this section, we assume that the variables $(X_t)_t$ are independent and identically distributed, and that their distribution P_{θ_0} belongs to a canonical exponential model of the form $\{P_\theta : \theta \in \Theta\}$, where Θ is an real interval and where P_θ has, with respect to some reference measure, the density $p_\theta : \mathbb{R} \to \mathbb{R}$ defined by:

$$p_\theta(x) = \exp\left(x\theta - b(\theta) + c(x)\right)\ .$$

Here, c is a real function and the log-partition function b is supposed to be twice differentiable. It is well-known that, by denoting $\mu(\theta) = \dot{b}(\theta)$ the expectation of P_θ, one defines a one-to-one, differentiable mapping μ. In this case, one easily shows that the rate function I is directly related to the Kullback-Leibler divergence (which is here a Bregman divergence for b) as follows: for every $\beta,\theta \in \Theta$,

$$\mathrm{KL}(P_\beta;P_\theta) = I(\mu(\beta);\mu(\theta)) = b(\theta) - b(\beta) - \dot{b}(\beta)(\theta-\beta)\ .$$

Hence, a sequence $(R_t)_{t\geqslant 1}$ of confidence intervals for the parameter θ_0 jointly valid with probability $1 - 2e\left\lceil\delta\log(n)\right\rceil\exp(-\delta)$ is obtained by choosing:

$$\begin{aligned} R_t &= \left\{\theta : \mathrm{KL}\left(P_{\mu^{-1}(\bar{X}_t)};P_\theta\right) \leqslant \frac{\delta}{t}\right\} \\ &= \left\{\theta : I\left(\bar{X}_t;\mu(\theta)\right) \leqslant \frac{\delta}{t}\right\}\ . \end{aligned}$$

This applies in particular to usual families of distributions like Poisson, Exponential, Gamma (with fixed shape parameter)...

In [4], an example concerning exponential variable detailed: in that case, $I(x,y) = x/y - 1 - \log(x/y)$.

But the case of Bernoulli variables deserves to be highlighted, as it easily extends to general *bounded* variables. Indeed, as observed by Hoeffding [22], the exponential moments of a $[0,1]$-valued variable X with expectation μ are upper-bounded by those of a Bernoulli variable, and for all $\lambda \in \mathbb{R}$ it holds that

$$E\left[\exp(\lambda X)\right] \leqslant 1 - \mu + \mu \exp(\lambda) \ ,$$

with equality if and only if $X \sim \mathcal{B}(\mu)$. Recall that kl denotes the binary entropy function, i.e. the rate function associated to Bernoulli variables. Theorem 1 yields that, for independent variables X_t bounded in $[0,1]$,

$$P\left(\sup_{t\leqslant n} \mathrm{kl}\left(\bar{X}_t, \mu\right) \geqslant \frac{\delta}{t}\right) \leqslant 2e \left\lceil \delta \log(n) \right\rceil \exp(-\delta) \ . \quad (5)$$

Of course, this result together with Pinsker's inequality $\mathrm{kl}(p,q) \geqslant 2(p-q)^2$, yields a self-normalized version of Hoeffding's inequality on the epoch $t \in \{1,\dots,n\}$:

$$P\left(\sup_{t\leqslant n} \left|\bar{X}_t - \mu\right| \geqslant \frac{\delta}{\sqrt{t}}\right) \leqslant 4e \left\lceil \delta^2 \log(n) \right\rceil \exp(-2\delta^2) \ . \quad (6)$$

This bound may seem simpler and easier to use than the previous one. However, the case of bounded bandits (detailed in [4], [5]), as well as the case of context tree estimation (presented in [12]) show that Equation (5) is sometimes really to be preferred, as it leads to significantly more efficient algorithms at the price of an hardly increased computational complexity.

B. Multinomial Distributions

As suggested by Sanov's (asymptotic) Theorem, this kind of inequalities is not limited to real-valued variables. It is also possible to construct informational, self-normalized confidence regions for random vectors; let us detail here the simple case of multinomial laws, as they are required, for example, in order to estimate transition distributions in Markov chains (see [12], [23]). Let P and Q be two elements of the set $\mathcal{S}$ of all probability distributions over a finite set A. By remarking that

$$\begin{aligned}
&-\mathrm{KL}(P;Q) + \sum_{x\in A} \mathrm{kl}\left(P(x);Q(x)\right) \\
&\quad = (|A|-1)\sum_{x\in A} \frac{1-P(x)}{|A|-1} \log\left(\frac{(1-P(x))/(|A|-1)}{(1-Q(x))/(|A|-1)}\right)
\end{aligned}$$

is non-negative, one easily shows that

$$\mathrm{KL}(P;Q) \leqslant \sum_{x\in A} \mathrm{kl}\left(P(x);Q(x)\right) \ .$$

It follows that if $X_1,\dots,X_n$ are i.i.d. variables of law $P_0 \in \mathcal{S}$, and if $\hat{P}_t(k) = \sum_{s=1}^t \mathbb{1}\{X_s = k\}/t$, then

$$\begin{aligned}
&P\Big(\exists t \in \{1,\dots,n\}: \ \mathrm{KL}\left(\hat{P}_t; P_0\right) \geqslant \frac{\delta}{t}\Big) \\
&\qquad \leqslant \sum_{a\in A} P\left(\exists t \in \{1,\dots,n\} : \mathrm{kl}\left(\hat{P}_t(a); P_0(a)\right) \geqslant \frac{\delta}{|A|t}\right) \\
&\qquad \leqslant 2e\left(\delta \log(n) + |A|\right) \exp\left(-\frac{\delta}{|A|}\right) \ . \qquad (7)
\end{aligned}$$

The fact that this bound involves directly the Kullback-Leibler divergence between the empirical measure and the true distribution allows, in context tree estimation (see [12]), to suppress unnecessary assumptions that resulted, in previous papers, from the use of Bernstein's inequality. Moreover, the Equation (7) permits to construct a sequence $(R_t)_{t\leqslant n}$ of "Sanov-type" confidence regions for P_0 that are simultaneously valid with probability at least $1-\alpha$, by choosing Kullback-Leibler neighborhoods of the maximum likelihood estimator:

$$R_t = \left\{ Q \in \mathcal{S} : \mathrm{KL}(\hat{P}_t; Q) \leqslant \frac{\delta}{t} \right\} \ ,$$

with δ such that $2e\left(\delta\log(n) + |A|\right)\exp\left(-\delta/|A|\right) = \alpha$. These regions R_t of the simplex have nice geometric properties that are exploited in [23] for reinforcement learning in Markov Decision Process, improving on former results using L^1 regions.

References

[1] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[2] S. Bubeck, D. Ernst, and A. Garivier, "Good-UCB: an optimistic algorithm for discovering unseen data," 2011.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[4] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *23rd Conf. Learning Theory (COLT)*, Budapest, Hungary, 2011.

[5] O. Cappé, A. Garivier, O. Maillard, R. Munos, and G. Stoltz, "Kullback-leibler upper confidence bounds for optimal sequential allocation," *The Annals of Statistics*, in press 2013.

[6] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. 29, no. 5, pp. 656–664, 1983.

[7] J. R. Busch, P. A. Ferrari, A. G. Flesia, R. Fraiman, S. P. Grynberg, and F. Leonardi, "Testing statistical hypothesis on random trees and applications to the protein classification problem," *Annals of applied statistics*, vol. 3, no. 2, 2009.

[8] A. Galves, C. Galves, J. Garcia, N. Garcia, and F. Leonardi, "Context tree selection and linguistic rhythm retrieval from written texts," *Annals of Applied Statistics*, vol. 6, pp. 186–209, 2012.

[9] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

[10] A. Garivier, "Consistency of the unlimited BIC context tree estimator," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4630–4635, 2006.

[11] P. Bühlmann and A. J. Wyner, "Variable length Markov chains," *Ann. Statist.*, vol. 27, pp. 480–513, 1999.

[12] A. Garivier and F. Leonardi, "Context tree selection: A unifying view," *Stochastic Processes and their Applications*, vol. 121, no. 11, pp. 2488–2506, Nov. 2011.

[13] A. Celisse, J.-J. Daudin, , and L. Pierre, "Consistency of maximum likelihood and variational estimators in stochastic block model," *arXiv:1105.3288*, 2010.

[14] N. Cesa-bianchi, G. Lugosi, and G. Stoltz, "Minimizing regret with label efficient prediction," *IEEE Trans. Inform. Theory*, vol. 51, pp. 77–92, 2005.

[15] V. De La Peña, M. Klass, and T. Lai, "Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws," *Annals of Probability*, vol. 32, no. 3, pp. 1902–1933, 2004.

[16] V. Peña, T. Lai, and Q. Shao, *Self-normalized Processes: Limit Theory and Statistical Applications*, ser. Mathematics and Statistics. Springer Berlin Heidelberg, 2009.

[17] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *arXiv:0810.0097*, 2011.

[18] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, ser. Stochastic Modelling and Applied Probability. Berlin: Springer-Verlag, 2010, vol. 38.

[19] J. Neveu, *Martingales à temps discret*. Masson, 1972.

[20] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," in *Algorithmic Learning Theory (ALT)*, ser. Lecture Notes in Computer Science, vol. 6925, 2011.

[21] L. Kocsis and C. Szepesvári, "Discounted UCB," *2nd PASCAL Challenges Workshop*, Venice, Italy, April 2006.

[22] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[23] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and Kullback-Leibler divergence," in *Allerton Conf. on Communication, Control, and Computing*, Monticello, US, 2010.