Informational Confidence Bounds for Self-Normalized Averages and Applications

Aurélien Garivier

Institut de Mathématiques de Toulouse - Université Paul Sabatier

Thursday, September 12th 2013

Outline

1. Two Different Problems
   - Context Tree Estimation
   - Stochastic Bandit Problems

2. Confidence Bounds for Self-Normalized Averages
   - Sub-gaussian Case
   - Beyond the Sub-gaussian Case

Variable Length Memory

- Linguistics, lossless compression:

t r y i n g _ v a n i l l a _ q u i e t

- Music, biology, genomics. . .

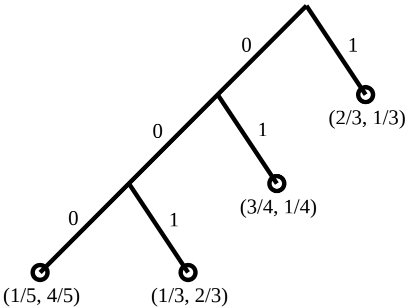
$\Longrightarrow$ memory structure as fingerprint of the source.
Example: Brazialian and European Portuguese.

Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\begin{aligned}
&\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10) \\
= \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) \\
\times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) \\
\times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) \\
\times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) \\
\times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011)
\end{aligned}$$



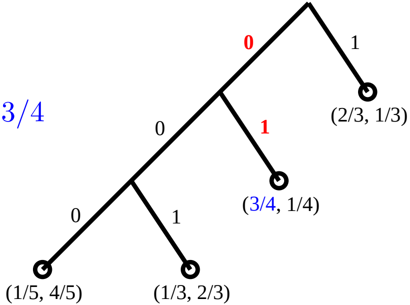
Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned} = \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) \qquad 3/4 \\ \times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) \\ \times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) \\ \times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) \\ \times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) \end{aligned}$$



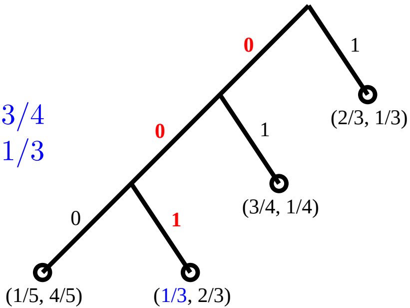
Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{lll} = & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & \\ \times & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & \\ \times & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & \end{array}$$



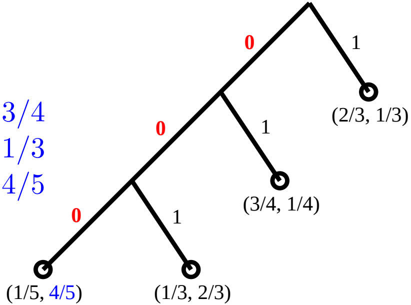
Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned}
= \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) && 3/4 \\
\times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) && 1/3 \\
\times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) && 4/5 \\
\times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) && \\
\times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) &&
\end{aligned}$$



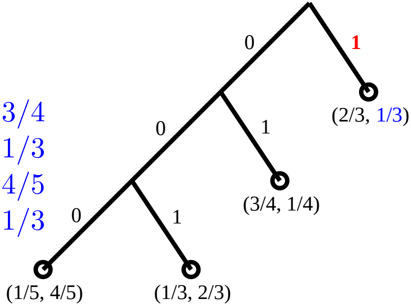
Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{aligned} = \quad & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times \quad & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times \quad & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times \quad & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\ \times \quad & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & \end{aligned}$$



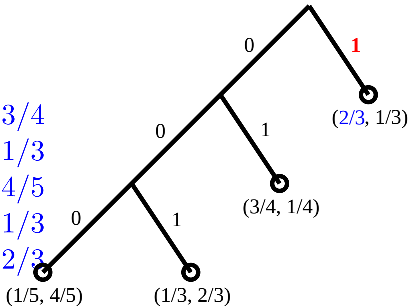
Context Tree Sources

A Context Tree Source (CTS) or Variable Length Markov Chain (VLMC) is a Markov Chain whose order may vary with the value of the past.

$$T = \{1, 10, 100, 000\}$$

$$\mathbb{P}(X_1^4 = 00110 | X_{-1}^0 = 10)$$

$$\begin{array}{llr} = & \mathbb{P}(X_1 = 0 | X_{-1}^0 = 10) & 3/4 \\ \times & \mathbb{P}(X_2 = 0 | X_{-1}^1 = 100) & 1/3 \\ \times & \mathbb{P}(X_3 = 1 | X_{-1}^2 = 1000) & 4/5 \\ \times & \mathbb{P}(X_4 = 1 | X_{-1}^3 = 10001) & 1/3 \\ \times & \mathbb{P}(X_5 = 0 | X_{-1}^4 = 100011) & 2/3 \end{array}$$



The Context Algorithm (Rissanen '81)

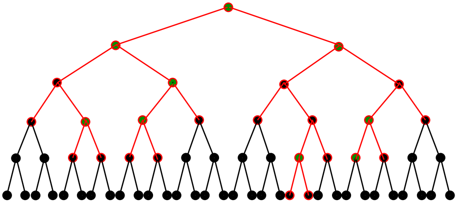
- Same principle as CART algorithm.
- For every possible node $s \in A^*$, compute a distortion such as:

$$\delta(s) = \max_{a \in A} \left\| \hat{P}(\cdot|s), \hat{P}(\cdot|as) \right\| .$$

- Keep the nodes $s \in A^*$ such that

$$\delta(s) \geq \epsilon(s, n)$$

and their ancestors as internal nodes of the estimated tree $\hat{T}_C$.



Penalized Maximum Likelihood

- Estimator

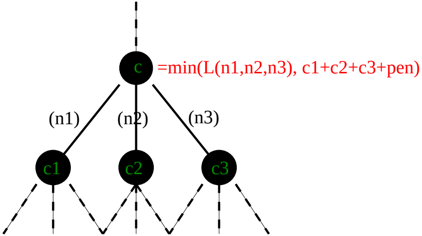
$$T_{PML} = \arg\max_{T} \log \hat{P}_T(x_1^n | x_{-\infty}^0) + \text{pen}(n,T),$$

where $\text{pen}(n,T)$ = penalty function, grows with n and $|T|$.

- MDL - BIC Penalty

$$\text{pen}(n,T) = \frac{|T|(|A|-1)}{2} \log n.$$

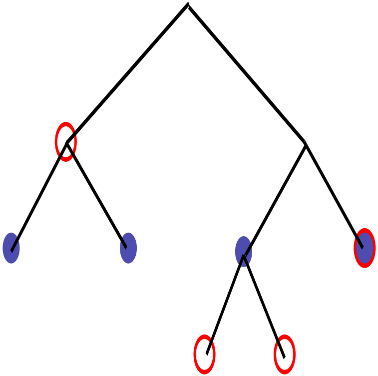
Effective computation:
recursive algorithm "Context Tree Maximization"



Under- and Over-estimation

Two possible types of estimation errors:

1. under-estimation:
   $\exists s \in T_0 : s \notin \hat{T}$
   $\Longrightarrow$ "easily" avoided (large deviation regime) at exponential rate
2. over-estimation:
   $\exists s \in \hat{T} : s \notin T_0$
   $\Longrightarrow$ more delicate, no exponential rates



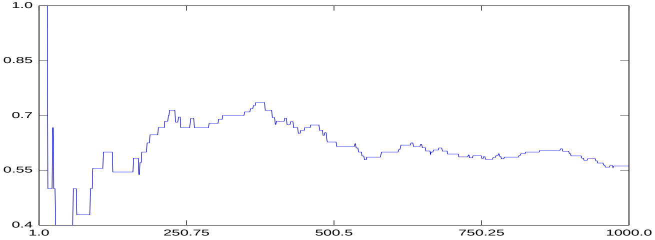
Asymptotic consistency results: [Bühlmann& Wyner '99, Csiszar&Talata '06, G. '06]

Non asymptotic study [G. & Leonardi '11]

- For both algorithms, need to control the 'distance' between $\hat{P}_t(\cdot|s)$ and $P(\cdot|s)$.

$\implies$ the number $N_s(t)$ of summands is random: deviations of

$$Z_t = \frac{1}{N_s(t)} \sum_{u=1}^{t} (\mathbb{1}_{\{X_u = a} - P(a|s)\}) \mathbb{1}_{\{X_{u-|s|}^{u-1} = s\}}$$



Non asymptotic study [G. & Leonardi '11]

- For both algorithms, need to control the 'distance' between $\hat{P}_t(\cdot|s)$ and $P(\cdot|s)$.

$\Longrightarrow$ the number $N_s(t)$ of summands is random: deviations of

$$Z_t = \frac{1}{N_s(t)} \sum_{u=1}^{t} (\mathbb{1}_{\{X_u=a} - P(a|s)\}) \mathbb{1}_{\{X_{u-|s|}^{u-1}=s\}}$$

- The right 'distance' to consider is:

$$KL\left(\hat{P}_t(\cdot|s), P(\cdot|s)\right).$$

$\Longrightarrow$ the quantity of interest is:

$$W_t = N_s(t)\, KL\left(\hat{P}_t(\cdot|s); P(\cdot|s)\right).$$

Outline

Example: sequential clinical trials

- patients visit the medical center *one after another* for a given disease
- they are prescribed one of the (say) 5 treatments available
- the treatments are not equally efficient. . .
- . . . but nobody knows which one is the best: they only observe the effect of the prescribed treatment on each patient

⇒ What is the best allocation strategy?

- Prescriptions may be chosen using only the previous observations
- The goal is not to estimate each treatment's efficiency precisely, but to heal as many patients as possible

The (stochastic) Multi-Armed Bandit Model

Environment K arms with parameters $\theta = (\theta_1, \dots, \theta_K)$ such that for any possible choice of arm $a_t \in \{1, \dots, K\}$ at time t, one receives the reward

$$X_t = X_{a_t,t}$$

where, for any $1 \leq a \leq K$ and $s \geq 1$, $X_{a,s} \sim \nu_a$, and the $(X_{a,s})_{a,s}$ are independent.

Reward distributions $\nu_a \in \mathcal{F}_a$ parametric or not.

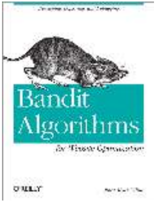
Example Bernoulli rewards: $\theta \in [0,1]^K$, $\nu_a = \mathcal{B}(\theta_a)$

Strategy The agent's actions follow a dynamical strategy $\pi = (\pi_1, \pi_2, \dots)$ such that

$$A_t = \pi_t(X_1, \dots, X_{t-1})$$

Real challenges

- Randomized clinical trials
  - original motivation since the 1930's
  - dynamic strategies can save resources
- Recommender systems:
  - advertisement
  - website optimization
  - news, blog posts, . . .



- Computer experiments
  - large systems can be simulated in order to optimize some criterion over a set of parameters
  - but the simulation cost may be high, so that only few choices are possible for the parameters
- Games and planning (tree-structured options)

Performance Evaluation, Regret

Cumulated Reward $S_T = \sum_{t=1}^{T} X_t$

Our goal Choose π so as to maximize

$$\mathbb{E}[S_T] = \sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{E}\left[\mathbb{E}[X_t \mathbb{1}\{A_t = a\} | X_1, \ldots, X_{t-1}]\right]$$
$$= \sum_{a=1}^{K} \mu_a \mathbb{E}[N_a^\pi(T)]$$

where $N_a^\pi(T) = \sum_{t \leq T} \mathbb{1}\{A_t = a\}$ is the number of draws of arm a up to time T, and $\mu_a = E(\nu_a)$.

Regret Minimization equivalent to minimizing

$$R_T = T\mu^* - \mathbb{E}[S_T] = \sum_{a:\mu_a < \mu^*} (\mu^* - \mu_a)\mathbb{E}[N_a^\pi(T)]$$

where $\mu^* \in \max\{\mu_a : 1 \leq a \leq K\}$

Upper Confidence Bound Strategies

UCB [Lai&Robins '85; Agrawal '95; Auer&al '02]

- Construct an upper confidence bound for the expected reward of each arm:

$$u_a(t) = \underbrace{\frac{S_a(t)}{N_a(t)}}_{\text{estimated reward}} + \underbrace{\sqrt{\frac{c\log(t)}{2N_a(t)}}}_{\text{exploration bonus}}$$

- Choose the arm with the highest UCB

- It is an *index strategy* [Gittins '79], easily interpretable and intuitively appealing.

Performance of the UCB algorithm

Non-asymptotic regret bound
[Auer, Cesa-Bianchi, Freund and Schapire '02]

$$N_a(T) \leq \frac{16\log(T)}{2(\mu^* - \mu_a)^2} + 4\,.$$

- The lower-bound for Bernoulli arms is

$$N_a(T) \geq \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu^*)}\big(1 - o(1)\big)$$

  where $\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

- Thinking of Pinsker's inequality $\mathrm{kl}(\mu_a, \mu^*) \geq 2(\mu^* - \mu_a)^2$:

  $\Longrightarrow$ remove factor 16;

  $\Longrightarrow$ replace $2(\mu^* - \mu_a)^2$ by $\mathrm{kl}(\mu_a, \mu^*)$.

How to construct the UCB?

The analysis shows that:

- $u_a(t)$ must upper-bound μ_a with probability $\geq 1 - 1/t$ for all a and $t \leq T$.
- Better UCB $\implies$ better regret (both in theory and in practice)

UCB algorithm:

$$u_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{c \log(t)}{2N_a(t)}}$$

Reminiscent of Hoeffding's inequality ($c = 1$) but the number of terms $N_a(t)$ is random.

In both examples:

- We have a large total number of observations. . .
- . . . but they are splitted into sub-samples of (possibly small) random size
- we need to confidence regions for the parameters on each sub-sample
- the diameter of the confidence regions may be data-driven

$\Longrightarrow$ Goal: do as if the number of obersations where not random

In both examples:

- We have a large total number of observations. . .
- . . . but they are splitted into sub-samples of (possibly small) random size
- we need to confidence regions for the parameters on each sub-sample
- the diameter of the confidence regions may be data-driven
- WARNING: Law of Iterated Logarithm !

$\Longrightarrow$ Goal: do as if the number of obersations where not random

In both examples:

- We have a large total number of observations. . .
- . . . but they are splitted into sub-samples of (possibly small) random size
- we need to confidence regions for the parameters on each sub-sample
- the diameter of the confidence regions may be data-driven
- WARNING: Law of Iterated Logarithm !

$\Longrightarrow$ Goal: do almost as if the number of obersations where not random

Outline

Simple framework

Observations: $(X_t)_t$ iid with expectation μ.

Optionnal skipping: ϵ_t is $\{0,1\}$-valued and $\sigma\big(X_t, \dots, X_{t-1}\big)$-measurable.

Total number of observations: $N_t = \sum_{s=1}^{t} \epsilon_t$.

Estimator of μ: $\hat{\mu}(t) = N_t^{-1} \sum_{k=1}^{n} \epsilon_t X_t$.

Cumulated deviation: $S_t = \sum_{k=1}^{n} \epsilon_t \big(X_t - \mu\big)$ satisfies $\forall\lambda$,

$$\mathbb{E}\left[e^{\lambda S_t - N_t \phi(\lambda)}\right] \leq 1$$

where $\phi(\lambda) = \log \mathbb{E}\left[e^{\lambda X_1}\right]$.

Sub-gaussian case: $\phi(\lambda) = \sigma^2\lambda^2/2$.

Can be generalized: X_t martingale increments,. . . .

Idea 0: Plain union bound

Union bound (sub-gaussian case)

With probability at least $1 - \delta$,

$$|\hat{\mu}(t) - \mu| \leq \sigma \sqrt{\frac{2 \log \frac{2t}{\delta}}{N_t}} .$$

Proof: Union bound + Chernoff

$$P\left(S_t > \sigma \sqrt{2N_t \log \frac{2t}{\delta}}\right) = \mathbb{P}\left(\bigcup_{n=1}^{t} \left\{e^{\lambda_n S_t} > e^{\sigma \lambda_n \sqrt{2N_t \log \frac{2t}{\delta}}}\right\} \cap \{N_t = n\}\right)$$

$$\leq \sum_{n=1}^{t} e^{-\sigma \lambda_n \sqrt{2n \log \frac{2t}{\delta}} + \frac{\sigma^2 \lambda_n^2}{2} n}$$

$$\leq \sum_{n=1}^{t} \frac{\delta}{2t} \qquad \text{for the choice } \lambda_n = \frac{1}{\sigma} \sqrt{\frac{2 \log \frac{2t}{\delta}}{n}} .$$

Idea 1: Method of Mixture [see De la Peña et al. '04, '07]

Idea: as $\forall\lambda, \mathbb{E}\left[e^{\lambda S_t - N_t\phi(\lambda)}\right] \leq 1$:

$$\int_{-\infty}^{+\infty} \mathbb{E}\left[e^{\lambda S_t - N_t \frac{\sigma^2\lambda^2}{2}}\right] \frac{y}{\sqrt{2\pi}} e^{-\frac{\lambda^2 y^2}{2}} d\lambda = \mathbb{E}\left[\frac{y}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\lambda S_t - N_t \frac{\sigma^2\lambda^2}{2} - \frac{\lambda^2 y^2}{2}} d\lambda\right]$$

$$= \mathbb{E}\left[\frac{y}{\sqrt{N\sigma^2 + y^2}} e^{\frac{S_t^2}{2(N\sigma^2+y^2)}}\right] \leq 1$$

and one obtains (for example):

Method of mixture (sub-gaussian case only!)

With probability at least $1-\delta$,

$$|\tilde{\mu}(t) - \mu| \leq \sigma \sqrt{\frac{2\log\frac{1}{\delta}}{N_t+1}\left(1 + \frac{1}{2}\log\left(N_t+1\right)\right)} ,$$

where $\tilde{\mu}(t) = \left(\sum_{t=1}^{n} \epsilon_t X_t\right)/(N_t+1)$.

Succesfully used in [Abbasi-Yadkori, Pál and Szepesvári '11] for linear bandits.

Idea 2: Peeling [G. & Moulines '08, '11]

Decompose the value of N_t by slices as follows: if $N_t > 1$ then

$$N_t \in \bigcup_{k=1}^{\left\lceil \frac{\log(t)}{\log(1+\eta)} \right\rceil} \left](1+\eta)^{k-1}, (1+\eta)^k\right]$$

Treat each slice independently with a unique λ_k (instead of the λ_n)

Control the loss in accuracy

Peeling (sub-gaussian case)

With probability at least $1-\delta$,

$$|\hat{\mu}(t) - \mu| \leq \sigma \sqrt{\frac{2\log\frac{4\log(t)}{\delta} + \log\left(2\log\left(\frac{4\log(t)}{\delta}\right)\right)}{N_t}} .$$

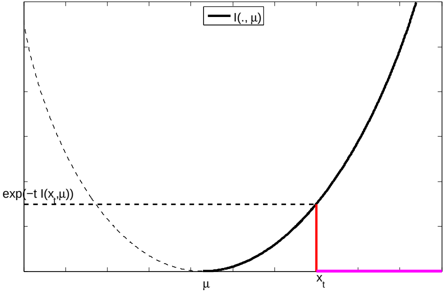
Outline

Rewriting Chernoff's bound

$$\mathbb{E}\left[\exp\left(\lambda S_t - t\phi(\lambda)\right)\right] \leq 1$$

if $\bar{X}_t = \mu + S_t/t$, and $x_t \geq \mu$, yields for $\lambda = \lambda(x_t)$:

$$P(\bar{X}_t \geq x_t) \leq \exp(-tI(x_t;\mu))$$



In other words:

$$P\Big(I(\bar{X}_t;\mu) \geq I(x_t;\mu), \bar{X}_t \geq \mu\Big) \leq \exp(-tI(x_t;\mu))$$

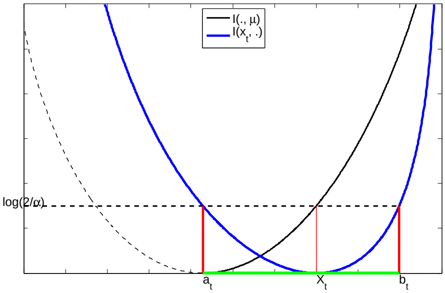
or, denoting $\delta = tI(x_t;\mu)$,

$$P\Big(tI(\bar{X}_t;\mu) \geq \delta, \bar{X}_t \geq \mu\Big) \leq \exp(-\delta)$$

$\implies$ confidence interval of risk at most α : I-neighboorhood of $\bar{X}_t$

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

Rewriting Chernoff's bound



$$\mathbb{E}\left[\exp\left(\lambda S_t - t\phi(\lambda)\right)\right] \le 1$$

if $\bar{X}_t = \mu + S_t/t$, and $x_t \ge \mu$, yields for $\lambda = \lambda(x_t)$:

$$P(\bar{X}_t \ge x_t) \le \exp(-tI(x_t;\mu))$$

In other words:

$$P\Big(I(\bar{X}_t;\mu) \ge I(x_t;\mu), \bar{X}_t \ge \mu\Big) \le \exp(-tI(x_t;\mu))$$

or, denoting $\delta = tI(x_t;\mu)$,

$$P\Big(tI(\bar{X}_t;\mu) \ge \delta, \bar{X}_t \ge \mu\Big) \le \exp(-\delta)$$

$\implies$ confidence interval of risk at most α : I-neighboorhood of $\bar{X}_t$

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

Bounds for random N_t [G. & Leonardi '11, G. & Cappé '11]

General bound:

For all $\delta > 0$,

$$P\left(I\left(\hat{\mu}(t);\mu\right) \geq \frac{\delta}{N(t)}\right) \leq 2e\left\lceil \delta \log(t) \right\rceil \mathrm{e}^{-\delta}$$

Log-concave case

If $I(\cdot;\mu)$ is log-concave

$$P\left(\exists t \in \{1,\dots,n\} : tI\left(\hat{\mu}(t);\mu\right) \geq \delta\right) \leq 2\sqrt{e}\left\lceil \frac{\sqrt{\delta}}{2} \log(t) \right\rceil \mathrm{e}^{-\delta}$$

Remark: the LIL suggests that there is little room for improvements.

Extension: non-stationary observations [G. & Moulines '11]

- Let $(X_t)_t$ be independent rv bounded by B, with expectation μ_t varying slowly (or rarely).
- Discounted estimator: for $\gamma \in]0,1[$,

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

where $S_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t X_t$ and $N_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t$

- Bias-variance decomposition: if $M_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t \mu_t$,

$$\bar{X}_\gamma(n) - \mu_n = \underbrace{\bar{X}_\gamma(n) - \frac{M_\gamma(n)}{N_\gamma(n)}} + \frac{M_\gamma(n)}{N_\gamma(n)} - \mu_n$$

- Fluctuations of the variance term: for all $\eta > 0$,

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

où $\nu_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t} < \min\{(1-\gamma)^{-1}, n\}$.

Multinomial laws [G. & Leonardi '11]

Extension using the simple inequality: for all $P, Q \in \mathfrak{M}_1(\mathcal{A})$,

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

Multinomial KL neighborhoods:

If $X_1,\dots,X_n \sim P_0 \in \mathfrak{M}_1(\mathcal{A})$ are iid, and
$\hat{P}_t(k) = \sum_{s=1}^t \mathbb{1}\{X_s = k\}/t$

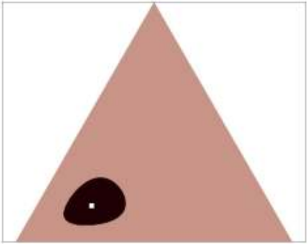
$$P\left(\exists t \in \{1,\dots,n\} : \quad \mathrm{KL}\left(\hat{P}_t;P_0\right) \geq \frac{\delta}{t}\right)$$
$$\leq 2e\left(\delta\log(n) + |\mathcal{A}|\right)\exp\left(-\frac{\delta}{|\mathcal{A}|}\right)$$

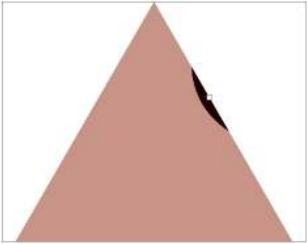
KL-balls [Filippi, G. & Cappé '10]

Sequence $(R_t)_{t\leq n}$ of informational confidence regions for P_0 simultaneously valid with probability at least $1-\alpha$:

$$R_t = \left\{ Q \in \mathfrak{M}_1(\mathcal{A}) : \mathrm{KL}(\hat{P}_t; Q) \leq \frac{\delta}{t} \right\} ,$$

with δ such that $2e\left(\delta\log(n) + |\mathcal{A}|\right)\exp\left(-\delta/|\mathcal{A}|\right) = \alpha$.





Results: context tree estimation

- Context: $\hat{T}_C$ keeps node s if
$$\delta(s) = \sum_b N_n(bs) D\left(\hat{p}_n(\cdot|bs); \hat{p}_n(\cdot|s)\right) \geq \epsilon(n) \ .$$
- Penalized Maximum Likelihood:
$$\hat{T}_{PML} = \arg\max_T \left\{ \log \hat{P}_T(x_1^n | x_{-\infty}^0) + \text{pen}(n, T) \right\} \ .$$
- Assume that $\text{pen}(n, T) = |T|\epsilon(n)$.

Theorem

For every $n \geq 1$ and $\hat{T}(X_1^n) \in \left\{\hat{T}_{PML}(X_1^n), \hat{T}_C(X_1^n)\right\}$ it holds that

$$\mathbb{P}\left(\hat{T}(X_1^n) \preceq T_0\right) \ \geq \ 1 - e\left(\epsilon(n)\log(n) + |A|^2\right) n^2 \exp\left(-\frac{\epsilon(n)}{|A|^2}\right) \ .$$

No unnecessary assumptions like $\forall s, a \in \mathcal{A}, P(a|s) = 0$ ou $P(s; a) > \epsilon$.

Results: an optimal UCB procedure

UCB algorithm with

$$u_a(t) = \sup\left\{\mu \in [0,1] : \quad \mathrm{kl}\big(\hat{\mu}_a(t),\, \mu\big) \le \frac{\log(t) + 3\log\log(t))}{N_a(t)}\right\}.$$

Theorem

$$\mathbb{E}\big[N_a(T)\big] \le \frac{\log(T)}{\mathrm{kl}(\mu_a,\mu^\star)} + \frac{\sqrt{2\pi}\log\left(\frac{\mu^\star(1-\mu_a)}{\mu_a(1-\mu^\star)}\right)}{\big(\mathrm{kl}(\mu_a,\mu^\star)\big)^{3/2}}\sqrt{\log(T) + 3\log\big(\log(T)\big)}$$

$$+\left(4e + \frac{3}{\mathrm{kl}(\mu_a,\mu^\star)}\right)\log\big(\log(T)\big) + \frac{2\left(\log\left(\frac{\mu^\star(1-\mu_a)}{\mu_a(1-\mu^\star)}\right)\right)^2}{\big(\mathrm{kl}(\mu_a,\mu^\star)\big)^2} + 6\,.$$

$\Longrightarrow$ improved logarithmic finite-time regret bound

$\Longrightarrow$ asymptotically optimal in the Bernoulli case

Some References:

1 **[Abbasi-Yadkori&al '11]** Yasin Abbasi-Yadkori, Dávid Pál, Csaba Szepesvári: Online Least Squares Estimation with Self-Normalized Processes: An Application to Bandit Problems CoRR abs/1102.2670: (2011)

2 **[Agrawal '95]** R. Agrawal. Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4) :1054-1078, 1995.

3 **[Audibert&al '09]** J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. Theoretical Computer Science, 410(19), 2009

4 **[Auer&al '02]** P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2) :235-256, 2002.

5 **[Bubeck, Ernst&G. '11]** S. Bubeck, D. Ernst, A. Garivier, Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality Journal of Machine Learning Research vol. 14 Feb. 2013 p.601-623

6 **[De La Pena&al '04]** V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes : exponential inequalities, moment bounds and iterated logarithm laws. Annals of Probability, 32(3) :1902-1933, 2004.

7 **[Filippi, Cappé&Garivier '10]** 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.

8 **[Filippi, Cappé, G.& Szepesvari '10]** S. Filippi, O. Cappé, A. Garivier, and C. Szepesvari. Parametric bandits : The generalized linear case. In Neural Information Processing Systems (NIPS), 2010.

9 **[G.&Cappé '11]** A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In 23rd Conf. Learning Theory (COLT), Budapest, Hungary, 2011.

10 **[Cappé,G.&al '13]** O. Cappé, A. Garivier, O-A. Maillard, R. Munos, G. Stoltz, Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation, Annals of Statistics, vol. 41 (3) Jun. 2013 pp.1516-1541

11 **[G.&Leonardi '11]** A. Garivier and F. Leonardi. Context tree selection : A unifying view. Stochastic Processes and their Applications, 121(11) :2488-2506, Nov. 2011.

12 **[G.&Moulines '11]** A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In Algorithmic Learning Theory (ALT), volume 6925 of Lecture Notes in Computer Science, 2011.

13 **[Lai&Robins '85]** T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1) :4-22, 1985.