Missing Mass, and Optimal Discovery

based on a joint work with Sébastien Bubeck and Damien Ernst

Aurélien Garivier

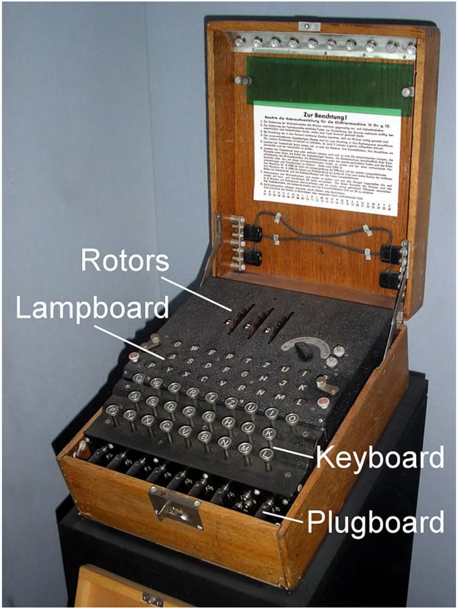
Indian Institute of Science Bengalore
July 11th 2023

Table of contents

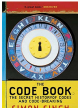
1. Estimating the Unseen
2. Discovering dangerous contigencies in electrical systems
3. The Good-UCB algorithm
4. Optimality results

Estimating the Unseen

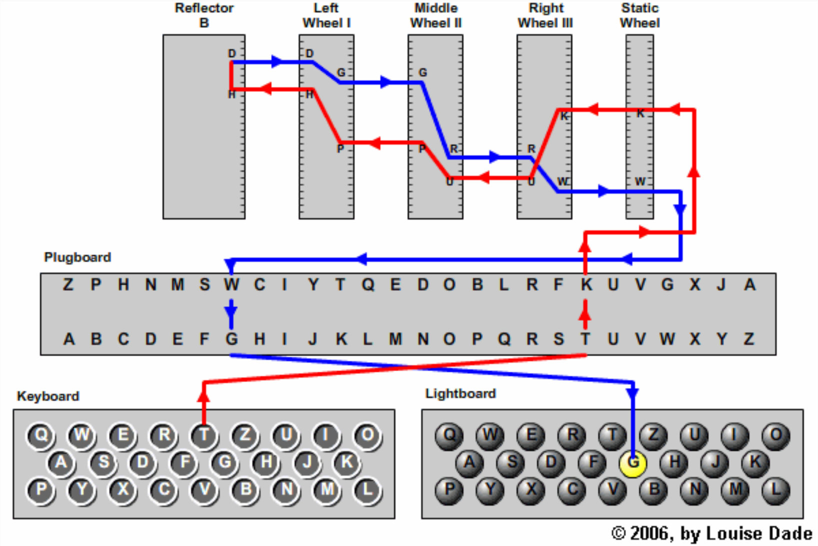
Enigma



- Electro-mechanical rotor cipher machines, 26 characters
- Invented at the end of WW1 by Arthur Scherbius
- Commercial use, then German Army during WW2
- First cracked by Marian Rejewski in the 1930s (Bomb), then improved to 3.10^{114} configurations
- Read Simon Singh, *The Code Book*

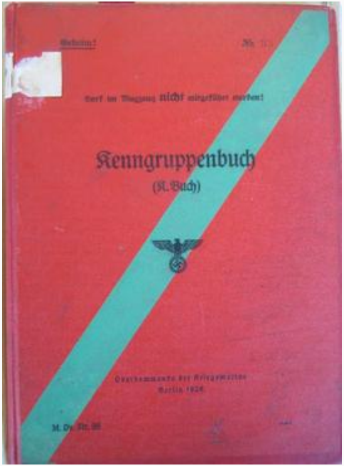


Enigma



Src: `http://enigma.louisedade.co.uk/`

Battle of the Atlantic



- Massively used by the German Kriegsmarine and Luftwaffe
- weakness: 3-letters setting to initiate communication, taken from the *Kenngruppenbuch*
- Government Code and Cypher School: Bletchley Park (on the train line between Cambridge and Oxford)
- Colossus (first programmable computers) in 1943

Estimating probabilities

- Discrete alphabet A.
- Unknown probability P on A
- Sample $X_1, \dots, X_n$ of independent draws of P.
- Goal : use the sample estimate $P(a)$ for all $a \in A$.

Natural idea:

$$\hat{P}(a) = \frac{N(a)}{n}, \quad \text{where } N(a) = \#\{i : X_i = a\}$$

Safari preparation

Observe animal sample

1 giraffe, 2 elephants, 3 zebras

Probability estimation?

Empirical frequency

Species	Probability
giraffes	1/6
elephants	2/6
zebras	3/6

Problem?

[Src: Alon Orlitsky, https://ic.epfl.ch/files/content/sites/ic/files/Inka/Orlitsky%20Talk%202016.pdf]

Bigram Model for NLP

[Src: https://nlp.stanford.edu/~wcmac]

Learning set:

john read moby dick

mary read a different book

she read a book by cher

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)}$$

$$P(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$$\begin{array}{cccccc}
P(& john & read & a & book &) \\
= & P(john|\cdot) & P(read|john) & P(a|read) & P(book|a) & P(\cdot|book) \\
= & \frac{c(\cdot\ john)}{\sum_w c(\cdot\ w)} & \frac{c(john\ read)}{\sum_w c(john\ w)} & \frac{c(reada)}{\sum_w c(read\ w)} & \frac{c(a\ book)}{\sum_w c(a\ w)} & \frac{c(book\ \cdot)}{\sum_w c(book\ w)} \\
= & \frac{1}{3} & \frac{1}{1} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} \\
\approx & 0.06 & & & &
\end{array}$$

Bigram Model for NLP

[Src: `https://nlp.stanford.edu/~wcmac`]

Learning set:

john read moby dick

mary read a different book

she read a book by cher

$$P(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)}$$

$$P(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

$$\begin{array}{cccccc}
P(& cher & read & a & book &) \\
= & P(cher|\cdot) & P(read|john) & P(a|read) & P(book|a) & P(\cdot|book) \\
= & \frac{c(\cdot\ cher)}{\sum_w c(\cdot\ w)} & \frac{c(cher\ read)}{\sum_w c(cher\ w)} & \frac{c(reada)}{\sum_w c(read\ w)} & \frac{c(a\ book)}{\sum_w c(a\ w)} & \frac{c(book\ \cdot)}{\sum_w c(book\ w)} \\
= & \frac{0}{3} & \frac{0}{1} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} \\
= & 0 & & & &
\end{array}$$

$\Longrightarrow$ useless, the unseen **must** be treated correctly.

Bayesian Approach: Laplace Estimator

Pierre-Simon de Laplace (1749-1827), Thomas Bayes (1702-1761)

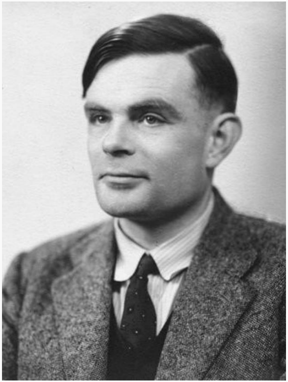
Will the sun rise tomorrow?

$$\hat{P}(a) = \frac{N(a) + 1}{n + |A|}$$

- good for small alphabets and many samples
- very bad when lots of items seen once (ex: DNA sequences)
- $|A|$ can be very large (or even infinite), but P concentrated on few items

$\Longrightarrow$ not a satisfying solution to the problem

Alan Turing



1912-1954

student of Godfrey Harold Hardy in Cambridge

PhD from Princeton with Alonzo Church

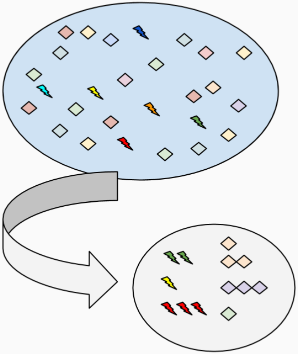
Irving John Good

1916-2009

Graduated in Cambridge

Academic carrer in Bayesian statistics in Manchester and then in the University of Virginia (USA)

Missing mass estimation



$X_1, \ldots, X_n$ independent draws of $P \in \mathfrak{M}_1(A)$.

$$O_n(x) = \sum_{m=1}^{n} \mathbb{1}\{X_m = x\}$$

How to 'estimate' the total mass of the *unseen* items

$$R_n = \sum_{x \in A} P(x)\, \mathbb{1}\{O_n(x) = 0\} \text{ ?}$$

The Good-Turing Estimator

See [I.J. Good, 1953], credits idea to A. Turing

Idea: in order to estimate the mass of the unseen

$$R_n = \sum_{x \in A} P(x) \, \mathbb{1}\{O_n(x) = 0\} ,$$

use the number of **hapaxes** = items seen only once (linguistic)

$$\hat{R}_n = \frac{U_n}{n}, \quad \text{where } U_n = \sum_{x \in A} \mathbb{1}\{O_n(x) = 1\}$$

Lemma [Good '53]: For *every* distribution P,

$$0 \leq \mathbb{E}\big[\hat{R}_n\big] - \mathbb{E}\big[R_n\big] \leq \frac{1}{n}$$

Completely non-parametric: no assumption on P

Bias of the Good-Turing Estimator

$$
\begin{aligned}
\mathbb{E}\big[\hat{R}_n\big] - \mathbb{E}\big[R_n\big] &= \frac{1}{n} \sum_{x \in A} \mathbb{P}\big(O_n(x) = 1\big) - \sum_{x \in A} P(x)\, \mathbb{P}\big(O_n(x) = 0\big) \\
&= \frac{1}{n} \sum_{x \in A} n\, P(x)\big(1 - P(x)\big)^{n-1} - \sum_{x \in A} P(x)\big(1 - P(x)\big)^{n} \\
&= \sum_{x \in A} P(x)\big(1 - P(x)\big)^{n-1} \Big(1 - \big(1 - P(x)\big)\Big) \\
&= \frac{1}{n} \sum_{x \in A} P(x) \times n\, P(x)\big(1 - P(x)\big)^{n-1} \\
&= \frac{1}{n} \sum_{x \in A} P(x)\, \mathbb{P}\big(O_n(x) = 1\big) \\
&= \frac{1}{n}\, \mathbb{E}\left[\sum_{x \in A} P(x) \mathbb{1}\{O_n(x) = 1\}\right] \quad \in \left[0, \frac{1}{n}\right]
\end{aligned}
$$

Jackknife interpretation

If we had additionnal samples, we would estimate R_n by the proportion of unseen elements in $X_{n+1}, X_{n+2}, \ldots$

We have no additionnal samples, **but** we keep every observation as a "test", pretending that the samples was made of everything else:

$$\begin{aligned}
\hat{R}_n &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{x_i \notin \{x_j : j \neq i\}\right\} \\
&= \frac{1}{n}\sum_{i=1} \mathbb{1}\left\{O_n(x_i) = 1\right\} \\
&= \frac{1}{n}\sum_{x\in A} \mathbb{1}\left\{O_n(x) = 1\right\}
\end{aligned}$$

Remark: jackknife is a **resampling method**, related to **bootstrap** and **crossvalidation** (of great use in Machine Learning).

Deviation Bounds

Proposition: With probability at least $1-\delta$ for *every* P,

$$\hat{R}_n - \frac{1}{n} - (1+\sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}} \leq R_n \leq \hat{R}_n + (1+\sqrt{2})\sqrt{\frac{\log(4/\delta)}{n}}$$

See [McAllester and Schapire '00, McAllester and Ortiz '03]:

- deviations of $\hat{R}_n$: McDiarmid's inequality
- deviations of R_n: negative association

Other tool: Poissonization [see Optimal Probability Estimation with Applications to Prediction and Classification, by Acharya, Jafarpour, Orlitsky Suresh, Colt 2013]

Application to Classification: minimax optimality

[Optimal Probability Estimation with Applications to Prediction and Classification, by Acharya, Jafarpour, Orlitsky Suresh, Colt 2013]

- P_1, P_2 probability distributions on A
- Given: two samples $(X_1^1, \dots, X_n^1)$ of P_1 and $(X_1^2, \dots, X_n^2)$ of P_2
- Goal: if $I = 1, 2$ with probability $1/2$ and if $X \sim P_I$, build a classifier $\phi_n : A \to \{1, 2\}$ so that $P\big(\phi_n(X) = I\big)\big)$ is as large as possible
- Maximal risk :

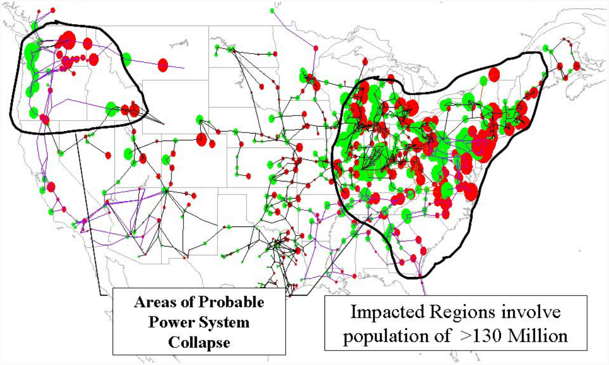
$$\bar{R}_n(\phi) = \max_{P_1, P_2} \mathbb{P}\big(\phi(X) \neq I\big)\big)$$

- **Prop**: if $\phi_n^{\mathrm{ML}}(x) = \arg\max_i \#\big\{j : X_j^i = x\big\}$ then there exists $c > 0$ such that for all $n \geq 1$, $\quad \bar{R}_n(\phi_n^{\mathrm{ML}}) \geq \min_\phi R_n(\phi) + c$.
- **Theorem**: there exists a Good-Turing based classifier ϕ_n^{GT} such that for all $n \geq 1$, $\quad \bar{R}_n(\phi_n^{\mathrm{GT}}) \leq \min_\phi R_n(\phi) + O\big(n^{-1/5}\big)$.

Discovering dangerous contigencies in electrical systems

The problem

Power system security assessment

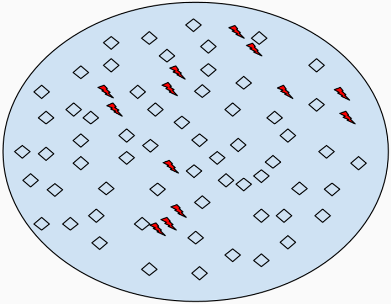


By Mark MacAlester, Federal Emergency Management Agency [Public domain], via Wikimedia Commons

Damien Ernst (Electrical Engineering, Liège): How to identify quickly contingencies/scenarios that could lead to unacceptable operating conditions (dangerous contingencies) if no preventive actions were taken?

The model

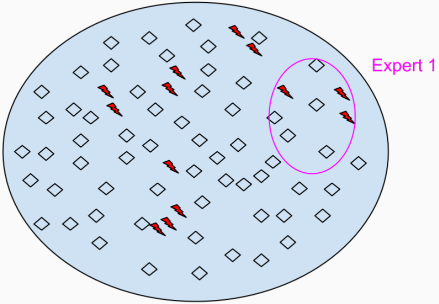
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

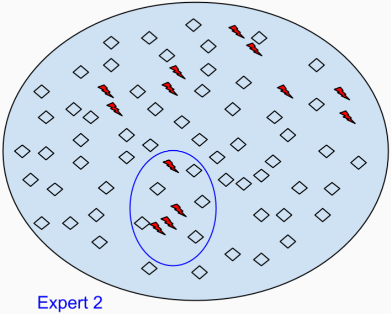
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

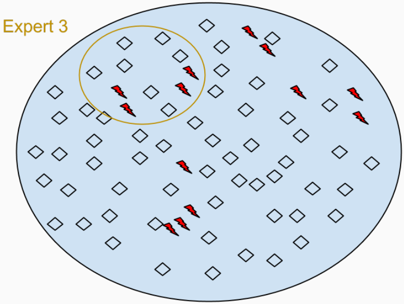
- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \leq i \leq K}$: sequential independent draws



Goal: discover rapidly the elements of A

The model

- Subset $A \subset \mathcal{X}$ of important items
- $|\mathcal{X}| \gg 1$, $|A| \ll |\mathcal{X}|$
- Access to $\mathcal{X}$ only by probabilistic experts $(P_i)_{1 \le i \le K}$: sequential independent draws



Goal: discover rapidly the elements of A

Goal

At each time step $t = 1, 2, \ldots$:

- pick an index $I_t = \pi_t\big(I_1, Y_1, \ldots, I_{s-1}, Y_{s-1}\big) \in \{1, \ldots, K\}$ according to past observations
- observe $Y_t = X_{I_t, n_{I_t,t}} \sim P_{I_t}$, where

$$n_{i,t} = \sum_{s \leq t} \mathbb{1}\{I_s = i\}$$

Goal: design the strategy $\pi = (\pi_t)_t$ so as to maximize the number of important items found after t requests

$$F^\pi(t) = \Big| A \cap \big\{ Y_1, \ldots, Y_t \big\} \Big|$$

Assumption: non-intersecting supports

$$A \cap \mathrm{supp}(P_i) \cap \mathrm{supp}(P_j) = \emptyset \text{ for } i \neq j$$

Is it a Bandit Problem ?

It looks like a bandit problem…

- sequential choices among K options
- want to maximize cumulative rewards
- exploration vs exploitation dilemma

… but it is **not a bandit problem** !

- rewards are not i.i.d.
- destructive rewards: no interest to observe twice the same important item
- all strategies eventually equivalent

The oracle strategy

Proposition: Under the non-intersecting support hypothesis, the greedy oracle strategy

$$I_t^* \in \arg\max_{1 \le i \le K} P_i \left(A \setminus \{ Y_1, \dots, Y_t \} \right)$$

is optimal: for every possible strategy π, $\mathbb{E}\big[F^\pi(t)\big] \le \mathbb{E}\big[F^*(t)\big]$.

Remark: the proposition is false if the supports may intersect

$\Longrightarrow$ estimate the "*missing mass* of important items"!

The Good-UCB algorithm

Our solution and analysis

Journal of Machine Learning Research 14 (2013) 601-623 Submitted 10/11; Revised 11/12; Published 2/13

Optimal Discovery with Probabilistic Expert Advice: Finite Time Analysis and Macroscopic Optimality

Sébastien Bubeck SBUBECK@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ, 08544, USA

Damien Ernst DERNST@ULG.AC.BE
Department of Electrical Engineering and Computer Science
University of Liège, Institut Montefiore, B28
B-4000 Liège, Belgium

Aurélien Garivier AURELIEN.GARIVIER@MATH.UNIV-TOULOUSE.FR
Institut de Mathématiques de Toulouse
Université Paul Sabatier
118, route de Narbonne
F-31062 Toulouse Cedex 9, France

Editor: Nicolo Cesa-Bianchi

Abstract

We consider an original problem that arises from the issue of security analysis of a power system and that we name optimal discovery with probabilistic expert advice. We address it with an algorithm based on the optimistic paradigm and on the Good-Turing missing mass estimator. We prove two different regret bounds on the performance of this algorithm under weak assumptions on the probabilistic experts. Under more restrictive hypotheses, we also prove a macroscopic optimality result, comparing the algorithm both with an oracle strategy and with uniform sampling. Finally, we provide numerical experiments illustrating these theoretical findings.

Keywords: optimal discovery, probabilistic experts, optimistic algorithm, Good-Turing estimator, UCB

The Good-UCB algorithm

Estimator of the missing important mass for expert i:

$$\hat{R}_{i,n_{i,t-1}} = \frac{1}{n_{i,t-1}} \sum_{x \in A} \mathbb{1}\left\{ \sum_{s=1}^{n_{i,t-1}} \mathbb{1}\{X_{i,s} = x\} = 1 \text{ and } \sum_{j=1}^{K} \sum_{s=1}^{n_{j,t-1}} \mathbb{1}\{X_{j,s} = x\} = 1 \right\}$$

Good-UCB algorithm:

1: For $1 \le t \le K$ choose $I_t = t$.
2: **for** $t \ge K+1$ **do**
3: Choose $I_t = \arg\max_{1 \le i \le K} \left\{ \hat{R}_{i,n_{i,t-1}} + C\sqrt{\frac{\log(4t)}{n_{i,t-1}}} \right\}$
4: Observe Y_t distributed as P_{I_t}
5: Update the missing mass estimates accordingly
6: **end for**

Optimality results

Classical analysis

Theorem: For any $t \geq 1$, *under the non-intersecting support assumption*, Good-UCB (with constant $C = (1+\sqrt{2})\sqrt{3}$) satisfies

$$\mathbb{E}\left[F^*(t) - F^{UCB}(t)\right] \leq 17\sqrt{Kt\log(t)} + 20\sqrt{Kt} + K + K\log(t/K)$$

Remark: Usual result for bandit problem, but not-so-simple analysis

Sketch of proof

1. On a set $\tilde{\Omega}$ of probability at least $1-\sqrt{\frac{K}{t}}$, the "confidence intervals" hold true simultaneously all $u \geq \sqrt{Kt}$
2. Let $\bar{I}_u = \arg\max_{1\leq i\leq K} R_{i,n_{i,u-1}}$. On $\tilde{\Omega}$,

$$R_{I_u,n_{I_u,u-1}} \geq R_{\bar{I}_u,n_{\bar{I}_u,u-1}} - \frac{1}{n_{I_u,u-1}} - 2(1+\sqrt{2})\sqrt{\frac{3\log(4u)}{n_{I_u,u-1}}}$$

3. But one shows that $\mathbb{E}F^*(t) \leq \sum_{u=1}^{t} \mathbb{E}R_{\bar{I}_u,n^{\pi}_{\bar{I}_u,u-1}}$
4. Thus

$$\begin{aligned}
&\mathbb{E}\left[F^*(t) - F^{UCB}(t)\right] \\
&\leq \sqrt{Kt} + \mathbb{E}\left[\sum_{u=1}^{t} \frac{1}{n_{I_u,u-1}} + 2(1+\sqrt{2})\sqrt{\frac{3\log(4t)}{n_{I_u,u-1}}}\right] \\
&\leq \sqrt{Kt} + K + K\log(t/K) + 4(1+\sqrt{2})\sqrt{3Kt\log(4t)}
\end{aligned}$$

Experiment: restoring property

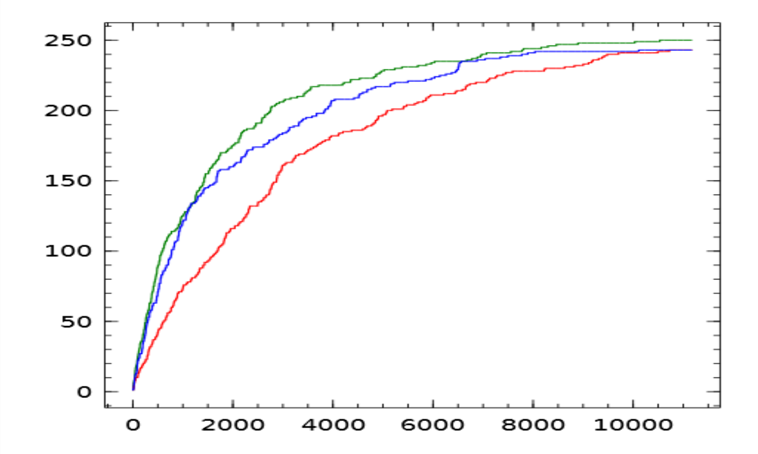


Figure 1: green: oracle, blue: Good-UCB, red: uniform sampling

Another analysis of Good-UCB

For $\lambda \in (0,1)$, $T(\lambda)$ = time at which missing mass of important items is smaller than λ on all experts:

$$T(\lambda) = \inf\left\{t : \forall i \in \{1,\ldots,K\}, P_i\big(A \setminus \{Y_1,\ldots,Y_t\}\big) \le \lambda\right\}$$

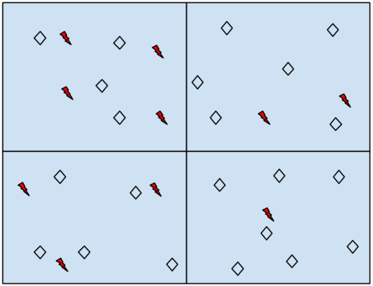
Theorem: Let $c > 0$ and $S \ge 1$. Under the non-intersecting support assumption, for Good-UCB with $C = (1+\sqrt{2})\sqrt{c+2}$, with probability at least $1 - \frac{K}{cS^c}$, for any $\lambda \in (0,1)$,

$$T_{UCB}(\lambda) \le T^* + KS\log\left(8T^* + 16KS\log(KS)\right),$$

$$\text{where} \quad T^* = T^*\left(\lambda - \frac{3}{S} - 2(1+\sqrt{2})\sqrt{\frac{c+2}{S}}\right)$$

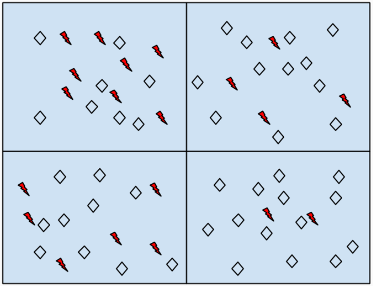
The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \mathrm{supp}(P_i)|/N \to q_i \in (0, 1)$, $q = \sum_i q_i$



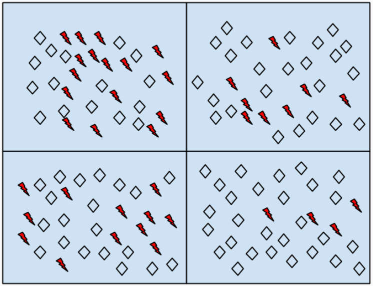
The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \text{supp}(P_i)|/N \to q_i \in (0, 1)$, $q = \sum_i q_i$



The macroscopic limit

- Restricted framework: $P_i = \mathcal{U}\{1, \ldots, N\}$
- $N \to \infty$
- $|A \cap \mathrm{supp}(P_i)|/N \to q_i \in (0, 1)$, $q = \sum_i q_i$



The Oracle behaviour

The limiting discovery process of the Oracle strategy is *deterministic*

Proposition: For every $\lambda \in (0, q_1)$, for every sequence $(\lambda^N)_N$ converging to λ as N goes to infinity, almost surely

$$\lim_{N \to \infty} \frac{T_*^N(\lambda^N)}{N} = \sum_i \left(\log \frac{q_i}{\lambda}\right)_+$$

Oracle vs. uniform sampling

Oracle: The proportion of important items not found after Nt draws tends to

$$q - F^*(t) = I(t)\underline{q}_{I(t)} \exp\left(-t/I(t)\right) \leq K\underline{q}_K \exp(-t/K)$$

with $\underline{q}_K = \left(\prod_{i=1}^K q_i\right)^{1/K}$ the geometric mean of the $(q_i)_i$.

Uniform: The proportion of important items not found after Nt draws tends to $K\bar{q}_K \exp(-t/K)$

$\implies$ Asymptotic ratio of efficiency

$$\rho(q) = \frac{\bar{q}_K}{\underline{q}_K} = \frac{\frac{1}{K}\sum_{i=1}^k q_i}{\left(\prod_{i=1}^k q_i\right)^{1/K}} \geq 1$$

larger if the $(q_i)_i$ are unbalanced

Macroscopic optimality

Theorem: Take $C = (1+\sqrt{2})\sqrt{c+2}$ with $c > 3/2$ in the Good-UCB algorithm.

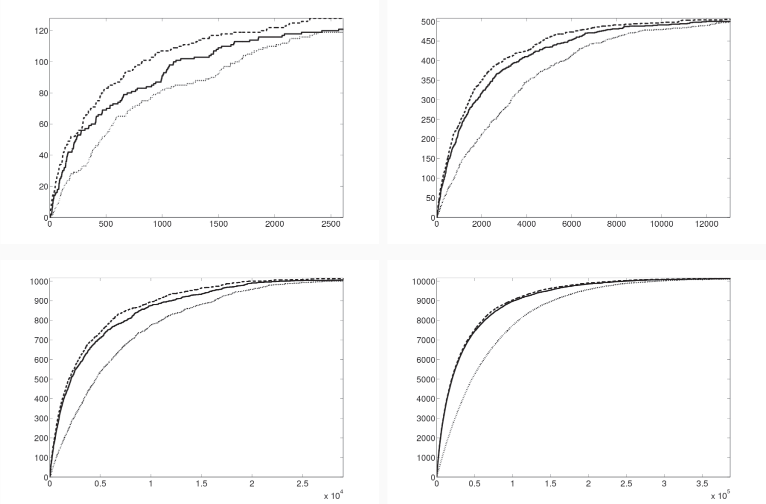
- For every sequence $(\lambda^N)_N$ converging to λ as N goes to infinity, almost surely

$$\limsup_{N\to+\infty} \frac{T^N_{UCB}(\lambda^N)}{N} \leq \sum_i \left(\log \frac{q_i}{\lambda}\right)_+$$

- The proportion of items found after Nt steps $F^{GUCB}(Nt)$ converges *uniformly* to $F^*(Nt)$ as N goes to infinity

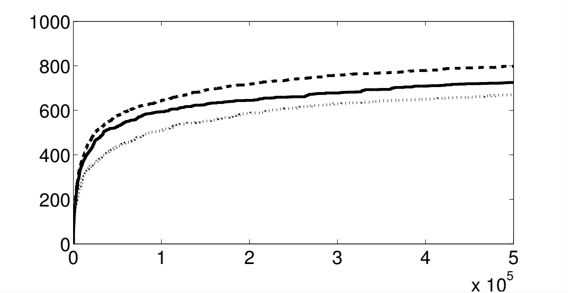
Experiment

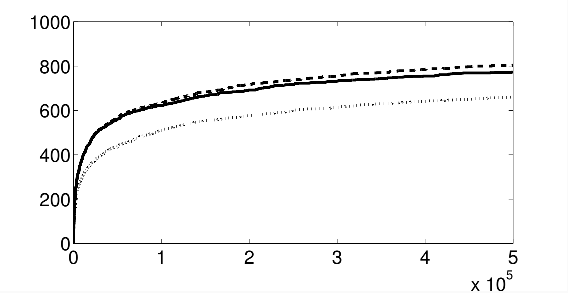
Number of items found by Good-UCB (solid), the OCL (dashed), and uniform sampling (dotted) as a function of time for sizes $N = 128$, $N = 500$, $N = 1000$ and $N = 10000$ in a 7-experts setting.



And when the assumptions are not satisfied?

Number of primes found by Good-UCB (solid), the oracle (dashed) and uniform sampling (dotted) using geometric experts with means 100, 300, 500, 700, 900, for $C = 0.1$ (top) and $C = 0.02$ (bottom).





Conclusion and perspectives

- We propose an algorithm for the optimal discovery with probabilistic expert advice
- We give a standard regret analysis under the only assumption that the supports of the experts are non-overlapping
- We propose a different optimality result, which permits a macroscopic analysis in the uniform case
- Another interesting limit to consider is when the number of important items to find is fixed, but the total number of items tends to infinity (Poisson regime)
- Then, the behavior of the algorithm is not very good: too large confidence bonus because no tight deviations bounds for the Good-Turing estimator when the proportion of important items tends to 0. Improvement by better deviation bounds?

Thank you for your attention!