Algorithms Seminar 2000–2001,
F. Chyzak (ed.), INRIA, (2002), pp. 49–52.

Available online at the URL
http://algo.inria.fr/seminars/.

Tail Bounds for Occupancy Problems

Paul Spirakis
Computer Technology Institute, Patras University (Greece)

November 20, 2000

Summary by Stéphane Boucheron

Abstract

The talk was based on [9] and consisted in a presentation of various tail bounds for occupancy problems and applications to the determination of the conjectured satisfiability threshold in the random k-sat problem.

1. Bins and Balls and Occupancy Problems

In bins and balls games, m balls are placed independently and uniformly at random among n bins. Henceforth, a generic allocation will be denoted by $\omega \in \{1,\dots,n\}^m$: $\omega_k = j$ if the k-th ball is located in the j-th bin. Let $X_n(\omega, m)$ denote the number of empty bins when m balls have been assigned a position. The piecewise constant interpolation is defined by $X_n(\omega, t) = X_n\big(\omega, \lceil tn \rceil\big)$. To alleviate notations, we omit ω when this is not a source of confusion. The behavior of the process $X_n(\cdot)$ as n becomes large has been the subject of many investigations in random combinatorics. The lecture is concerned with different derivations of tail bounds for $X_n(\cdot)$ and their application to the analysis of the threshold phenomenon for the (random) k-satisfiability problem.

1.1. Approaches to random allocations.

There are many approaches to random allocation problems. Many early successes of analytic combinatorics have been reported in the monograph by Kolchin, Sevast'yanov and Chystiakov [11].

Probabilistic (Martingale-theoretical) approaches have been successful as well. Let $\mathcal{F}_t$ denote the σ-algebra generated by the first $\lfloor nt \rfloor$ allocations (we do not mention n to alleviate notations). Then it is straightforward to check the relation

$$\mathbf{E}\left[X_n\left(t + \frac{1}{n}\right) \,\middle|\, \mathcal{F}_t \right] = \left(1 - \frac{1}{n}\right) X_n(t).$$

From this, one immediately deduces that $(1 - \frac{1}{n})^{-\lfloor nt \rfloor} X_n(t)$ is an $\mathcal{F}_t$-Martingale. Moreover it has bounded increments, and its quadratic variation process converges in probability towards $t \mapsto e^t - (1+t)$. Applying Martingale limit theorems [8], one easily deduces:

- a law of large numbers: $X_n(\cdot)/n$ converges in probability towards $t \mapsto e^{-t}$,
- a functional central limit theorem: $t \mapsto \big(X_n(t) - ne^{-t}\big)/\sqrt{n}$ converges towards a rescaled time-changed Brownian motion, namely $t \mapsto e^{-t} B\big[e^t - (1+t)\big]$.

Unfortunately, results on convergence in distribution tell little about asymptotic probability of rare events: the convergence rate cannot be better than $O\big(1/\sqrt{n}\big)$, and probability of rare events are especially relevant to the analysis of extreme values that often constitute the core of applications.

Nevertheless, central limit theorems suggest that the tail probabilities of the empty cell statistics might be Gaussian-like. In computer science, sharp upper bounds on tail probabilities are often desirable.

If instead of throwing a fixed number $\lfloor nt \rfloor$ of balls into the n bins, one first draws N according to a Poisson distribution with parameter $\lfloor nt \rfloor$, and then throws N balls into the n bins, the bin occupancies become independent Bernoulli random variables with success probability $\approx \exp(-t)$. $X_n(t)$ is now distributed according to a binomial random variable with parameters n and $\exp(-t)$. Let $\mathbb{P}$ denote the original probability distribution on allocations and let $\mathbb{Q}$ denote this alternate probability distribution on N and allocations. Note that conditionally on $N = \lfloor nt \rfloor$, the distributions of $X_n(t)$ under $\mathbb{P}$ and $\mathbb{Q}$ are identical (the multinomial distribution is a conditioned Poisson process). Then

$$\mathbb{P}\{X_n(t) \in A\} = \frac{\mathbb{Q}\{X_n(t) \in A \wedge N = \lfloor nt \rfloor\}}{\mathbb{Q}\{N = \lfloor nt \rfloor\}} \leq \sqrt{2\pi nt}\, \mathbb{Q}\{X_n(t) \in A\} \tag{1}$$

Inequality (1) provides with an easy tail upper bound for rare events under $\mathbb{Q}$, i.e., for large deviations of $X_n(t)$ around its expectation. If $A = \{\, \omega \mid X_n(\omega, t) > ne^{-t} + n\epsilon \,\}$, then

$$\mathbb{P}\{X_n(t) \in A\} \leq \sqrt{2\pi n}\, \exp\Big(-nh\left(e^{-t} + \epsilon, e^{-t}\right)\Big)$$

where $h(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$. It obviously raises two questions: Is the order of the exponent correct? Can we get rid of the $\sqrt{n}$ factor?

1.2. **Known results.** As allocation are performed independently, a very straightforward yet useful bound comes from the Azuma–Mc Diarmid inequality. Namely note that if ω and ω' are two allocation schemes that differ only in one position $\omega_j = \omega_j{}'$ for all $j \leq k = \lfloor tn \rfloor$ except for $j = i$, then $\left|X_n(\omega, t) - X_n(\omega', t)\right| \leq \frac{1}{n}$. As a matter of fact, if the space of allocations is equipped with the Hamming distance, the empty bin statistics is 1-Lipschitz. This implies that

$$\mathbb{P}\Big\{\left|X_n(t) - \mathbf{E}\big[X_n(t)\big]\right| > n\epsilon\Big\} \leq 2\exp\left(-\frac{2n\epsilon^2}{t^2}\right). \tag{2}$$

Inequality (2) is obtained by a Martingale embedding argument. Namely $X_n(t) = \mathbf{E}\big[X_n(t) \mid \mathcal{F}_t\big]$ and the process $M_n(s) = \mathbf{E}\big[X_n(t) \mid \mathcal{F}_s\big]$ is an $\mathcal{F}_s$-martingale, as

$$\mathbf{E}\big[M_n(s+h) \mid \mathcal{F}_s\big] = \mathbf{E}\Big[\mathbf{E}\big[X_n(t) \mid \mathcal{F}_{s+h}\big] \;\Big|\; \mathcal{F}_s\Big] = \mathbf{E}\big[X_n(t) \mid \mathcal{F}_s\big] = M_n(s).$$

One may wonder what the best way to apply Azuma's inequality is.

1.3. **Painless tail bounds.** The first bound presented in [9] is:

$$\mathbb{P}\Big\{\left|X_n(t) - \mathbf{E}\big[X_n(t)\big]\right| > n\epsilon\Big\} \leq 2\exp\left(-\frac{(n - 1/2)n^2\epsilon^2}{n^2 - \mathbf{E}\big[X_n(t)^2\big]}\right). \tag{3}$$

When n becomes large, the exponent on the right-hand side is equivalent to

$$-\frac{n\epsilon^2}{1 - e^{-2t}}.$$

The trivial Poisson estimates (1) clearly shows that this exponent is rather poor as soon as t becomes non-negligible. This is not a denial of the merits of Martingale approach. Indeed, this method provides nearly optimal bounds for smooth Gaussian functionals and for many discrete problems. The apparent flaw in Equation (3) comes from the fact that we did not use tight enough bounds on the quadratic variation process associated with $\mathbf{E}\big[X_n(t) \mid \mathcal{F}_s\big]$.

Next the authors of [9] proceed to establish what they call a Chernof bound for the occupancy problem. It shows that the Poisson tail estimate (1) is correct even if we do not resort to a conditioning argument, i.e., that the $\sqrt{n}$ factor is spurious.

2. The Large Deviation Approach

The large deviation approach (see [2, 5, 7] for recent presentations) aims at identifying the right exponents for tail probability. It provides the right touchstone for the occupancy problem. Rather than using the martingale structure of the occupancy problem, the large deviation approach relies on the Markovian structure of the occupancy problem: conditionally on $X_n(t)$, $X_n(t+1/n)$ does not depend on $\mathcal{F}_{t-1/n}$. The large deviation principle invoked in [9] comes from a contraction of a functional large deviation principle derived by Azencott and Ruget. The latter shows that asymptotically, the exponent in large deviation probabilities can be represented as a the solution of a variational problem, namely

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\{X_n(t) \geq nx\} = - \inf_{\xi(0)=1,\ \xi(t)=x} \int_0^t h\big(-\dot{\xi}(s), \xi(s)\big)\, ds. \tag{4}$$

The article [9] solves the associated variational problem and provides a closed form for the exponent, confirming the intuition that the exponent obtained by Poissonization is not optimal.

3. Satisfiability Problems

The second part of the paper presents an application of tail bounds for occupancy problems to the analysis of the random 3-sat problem. An instance of the 3-sat problem is a boolean formula in conjunctive normal form, where each clause has at most 3 literals. For each number n of variables, and each problem size k, the set of instances of the 3-sat problem is provided with the uniform probability over the m-tuples of 3-clauses over the n variables. At the time of writing [9], it was conjectured that as n goes to infinity while k/n remains constant, a phase transition occurs. For $k/n < c_3$, random 3-sat formulas are satisfiable with overwhelming probability, while for $k/n > c_3$ random 3-sat formulas are not satisfiable formulas with overwhelming probability.

The paper [9] proposes an upper-bound on the conjectured satisfiability threshold: $c_3 \leq 4.758$. This result came in a series of improvement starting from the straightforward $c_3 \leq 5.19$, through $c_3 \leq 5.08$ [6], $c_3 \leq 4.64$ [3], $c_3 \leq 4.601$ [10], and recently culminating with $c_3 \leq 4.506$ [4].

In the sequel, n and k are supposed to be fixed. F denotes a random 3-sat formula, $\#F$ denotes the number of assignments of the n boolean variables that satisfy F. F is satisfiable if $\#F \geq 1$. $T(F)$ equals 1 if F is satisfiable, 0 otherwise. Let σ denote a generic truth assignment. $F(\sigma)$ equals 1 if σ satisfies F, 0 otherwise. $\mathbf{1}$ denotes the truth assignment where all variables are set to 1. Then, we have

$$\mathbf{E}_F\big[T(F)\big] = \mathbf{E}_F\left[\sum_{\sigma:F(\sigma)=1} \frac{1}{\#F}\right] = \sum_{\sigma} \mathbf{E}_F\left[\frac{F(\sigma)}{\#F}\right] = 2^n \mathbf{E}_F\left[\frac{F(\mathbf{1})}{\#F}\right], \tag{5}$$

where the second equality comes from the fact that the number of formulae that satisfy a particular truth assignment does not depend on the truth assignment. Hence, to get an upper bound on the probability of satisfiability, it is enough to get an upper bound on

$$\left(\frac{7}{8}\right)^{cn} \mathbf{E}_{F'}\left[\frac{1}{\#F}\right],$$

where F is now picked at random among the $\left(\frac{7}{8}\right)^{cn}\binom{n}{3}^{cn}$ formulae that are satisfied by $\mathbf{1}$. This distribution among formulae is a product distribution where each clause is picked uniformly at random among the clauses where at least one literal is not negated.

The main idea of the proof is to establish that conditionally on the fact that it is satisfiable, a 3-sat formula with sufficiently many clauses has exponentially many satisfying truth assignments with overwhelming probability.

What is proved in [9] is actually the following. Let $\#F_1$ denote the number of truth assignments σ of F where for each clause in F, there exists a non-negated variable that evaluates to 1 in σ. Obviously $1/\#F \leq 1/\#F_1$. Now to lower bound $\#F_1$, it is enough to determine a minimum family of variables $\mathcal{I}(F)$ such that any truth assignment where all variables in $\mathcal{I}(F)$ evaluates to 1 satisfies the formula F ($\mathcal{I}(F)$ is sometimes called a prime implicant of F). As a matter of fact, we have $\#F_1 \geq 2^{n-\#\mathcal{I}}$, and hence

$$\mathbb{P}\{F \text{ is satisfiable}\} \leq \left(\frac{7}{8}\right)^{cn} \mathbf{E}_{F'}\left[2^{\#\mathcal{I}}\right]. \tag{6}$$

Since the publication of [9], improved upper bounds on c_3 have been derived by refining estimations on the fluctuations of $\#F$ for random formulae. Those estimations still rely on statistics for random allocations. But the empty bins statistics are no more sufficient. The best known upper bounds [4] rely on a statistics that have sometimes been called empirical occupancy measures. As a matter of fact, an allocation ω defines a probability measure on $\mathbb{N}$, $\bar{X}_n(i,t)$ denotes the fraction of bins that contain i balls for $i \in \mathbb{N}$. The large deviations of this measure-valued random variable may be studied in different ways: by resorting to Azencott–Ruget results and projective limit arguments [2], or directly as in [1].

Bibliography

[1] Boucheron (S.), Gamboa (F.), and Léonard (C.). – *Bins and balls: large deviations of the empirical occupancy process.* – Rapport de recherche du LRI n° 1255, Université Paris-Sud, 2000.

[2] Dembo (Amir) and Zeitouni (Ofer). – *Large deviations techniques and applications.* – Springer-Verlag, New York, 1998, second edition, xvi+396p.

[3] Dubois (O.) and Boufkhad (Y.). – A general upper bound for the satisfiability threshold of random r-SAT formulae. *Journal of Algorithms*, vol. 24, n° 2, 1997, pp. 395–420.

[4] Dubois (O.), Boufkhad (Y.), and Mandler (J.). – Typical random 3-sat formulae and the satisability threshold. In *Proceedings of SODA'2000*. ACM, pp. 126–127. – 2000.

[5] Dupuis (Paul) and Ellis (Richard S.). – *A weak convergence approach to the theory of large deviations.* – John Wiley & Sons, New York, 1997, xviii+479p. A Wiley-Interscience Publication.

[6] El Maftouhi (A.) and Fernandez de la Vega (W.). – On random 3-sat. *Combinatorics, Probability and Computing*, vol. 4, n° 3, 1995, pp. 189–195.

[7] Feng (Jin) and Kurtz (Thomas G.). – *Large deviations for stochastic processes.* – 2000. 194 pages. Available from `http://www.math.wisc.edu/~kurtz/feng/ldp.htm`.

[8] Hall (P.) and Heyde (C. C.). – *Martingale limit theory and its application.* – Academic Press, New York, 1980, xii+308p. Probability and Mathematical Statistics.

[9] Kamath (Anil), Motwani (Rajeev), Palem (Krishna), and Spirakis (Paul). – Tail bounds for occupancy and the satisfiability threshold conjecture. *Random Structures & Algorithms*, vol. 7, n° 1, 1995, pp. 59–80.

[10] Kirousis (Lefteris M.), Kranakis (Evangelos), Krizanc (Danny), and Stamatiou (Yannis C.). – Approximating the unsatisfiability threshold of random formulas. *Random Structures & Algorithms*, vol. 12, n° 3, 1998, pp. 253–269.

[11] Kolchin (Valentin F.), Sevast'yanov (Boris A.), and Chistyakov (Vladimir P.). – *Random allocations.* – V. H. Winston & Sons, Washington, D.C., 1978, xi+262p. Translated from the Russian.