Basic Research in Computer Science

BRICS RS-95-48 D. P. Dubhashi: Simple Proofs of Occupancy Tail Bounds

Simple Proofs of Occupancy Tail Bounds

Devdatt P. Dubhashi

BRICS Report Series RS-95-48

ISSN 0909-0878 September 1995

Copyright © 1995, BRICS, Department of Computer Science
University of Aarhus. All rights reserved.

Reproduction of all or part of this work
is permitted for educational or research use
on condition that this copyright notice is
included in any copy.

See back inner page for a list of recent publications in the BRICS Report Series. Copies may be obtained by contacting:

BRICS
Department of Computer Science
University of Aarhus
Ny Munkegade, building 540
DK - 8000 Aarhus C
Denmark

Telephone: +45 8942 3360
Telefax: +45 8942 3255
Internet: BRICS@daimi.aau.dk

BRICS publications are in general accessible through WWW and anonymous FTP:

`http://www.brics.aau.dk/BRICS/`
`ftp ftp.brics.aau.dk (cd pub/BRICS)`

Simple Proofs of Occupancy Tail Bounds

Devdatt Dubhashi
BRICS[*],
Department of Computer Science,
University of Aarhus,
Ny Munkegade,
DK-8000 Aarhus C, Denmark
Email: `dubhashi@daimi.aau.dk`

September 15, 1995

Abstract

We give short proofs of some occupancy tail bounds using the method of bounded differences in expected form and the notion of negative association.

1 Introduction

The purpose of this note is to give short, simple and natural proofs of some tail bounds on occupancy problems in [5]. The proofs also serve as advertisements for some very useful, but apparently not very well–known concepts and techniques:

- The *method of bounded differences* in the *expected* form [6, Cor. 6.10].
- A concept of negative dependence called *negative association* from the theory of multivariate probability inequalities, [3, 4, 9, 10].

*Basic Research in Computer Science,
Centre of the Danish National Research Foundation.

The first of these yields the Occupancy Bound 1 of [5, Theorem 2] by a direct plug–in substitution. The second gives a short and enlightening calculation–free proof of the (Chernoff) Occupancy Bound 2 in [5, Theorem 3].

The setting is the classical probabilistic experiment of throwing m balls independently and uniformly [1] into n bins (for positive integers m, n). The random variables of interest are defined as follows: for $i \in [n]$ [2], let Z_i be the indicator variable which is 1 if bin i is empty and 0 otherwise. Set $Z := \sum_i Z_i$ to be the number of empty bins. We are interested in tail bounds on the distribution of Z.

2 Bounded Differences

The "method of bounded differences" is usually stated and used in the following form [6, Lemma 1.2]

Proposition 1 (McDiarmid) *Let $X_1, \ldots, X_n$ be independent random variables, variable X_i taking values in a finite set A_i for each $i \in [n]$, and suppose the function $f : \prod_i A_i \to \mathsf{R}$ satisfies the following "bounded difference" condition: For each $i \in [n]$, there is a constant c_i such that*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_i,$$

whenever the vectors $\mathbf{x}, \mathbf{x}'$ differ only in the kth co–ordinate. Then

$$\Pr[|f(\mathbf{X}) - E[f(\mathbf{X})]| > t] < 2\exp(-2t^2/\sum_i c_i^2).$$

This lemma has an attractive packaged form, but can be too weak for some applications. In this case, we may resort to the "method of bounded differences" in the expected form [6, Cor. 6.10]:

Proposition 2 (McDiarmid) *Let $X_1, \ldots, X_n$ be independent random variables, variable X_i taking values in a finite set A_i for each $i \in [n]$, and suppose the function f satisfies the following "bounded difference" conditions: For*

[1] The techniques apply equally well even if the uniformity assumption is dropped, but we retain it for simplicity.

[2] We denote $[n] := \{1, \ldots, n\}$

each $i \in [n]$, there is a constant c_i such that for any $x_k \in A_k, k \in [i-1]$ and for any $x_i, x_i' \in A_i$

$$\begin{aligned} |E[f(\mathbf{X}) \mid X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i] - \\ E[f(\mathbf{X}) \mid X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i']| \quad \leq c_i \quad . \end{aligned} \tag{1}$$

Then

$$\Pr[|f(\mathbf{X}) - E[f(\mathbf{X})]| > t] < 2\exp(-2t^2 / \sum_i c_i^2).$$

We shall illustrate the "method of bounded differences" in the above form by applying it to study the occupancy statistics in the classical balls and bins experiment. First, we have by simple calculations, $E[Z_i] = (1 - \frac{1}{n})^m$ and $E[Z] = \sum_i E[Z_i] = m(1 - \frac{1}{n})^m$.

To get a tail probability estimate, regard $Z = Z(B_1, \ldots, B_m)$ where the random variables B_k take values in the set $[n]$ indicating which bin ball k occupies, for each $k \in [m]$. If one were to employ the "method of bounded differences" in the form of Proposition 1, it is easy to see that one must take $c_i := 1$ for each $i \in [m]$. Then one gets the tail probability bound:

$$\Pr[|Z - E[Z]| > t] < 2\exp(-2t^2 / \sum_i c_i^2) = 2\exp(-2t^2/m).$$

This can be quite weak for large m and small t; compare the bound below.

An improved bound can be obtained by applying the expected version of Proposition 2. Fix some $i \in [m]$ and let us compute

$$\begin{aligned} |E[Z \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] - \\ E[Z \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i']| \quad , \end{aligned}$$

for fixed $b_1, \ldots, b_{i-1}, b_i, b_i' \in [n]$. Set $b := b_i \neq b_i' =: b'$. Let $I := \{b_1, \ldots, b_{i-1}\} \subseteq [n]$. Of course for $j \in I$,

$$\begin{aligned} E[Z_j \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] &= 0 = \\ E[Z_j \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i'] & \quad . \end{aligned} \tag{2}$$

Also,

$$\begin{aligned} E[Z_b \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] &= 0 \\ E[Z_{b'} \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i'] &= 0. \end{aligned} \tag{3}$$

For $j \in [m] \setminus (I \cup \{b, b'\})$, we have

$$\begin{aligned} E[Z_j \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] &= (1 - \frac{1}{n})^{m-i} = \\ E[Z_j \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b'_i] & . \end{aligned} \tag{4}$$

Now suppose $b \in I$ but $b' \notin I$. Then with we have

$$E[Z_b \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b'_i] = 0, \tag{5}$$

whereas

$$E[Z_{b'} \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] = (1 - \frac{1}{n})^{m-i}. \tag{6}$$

Finally (apart from the symmetric case of the previous one) if $b, b' \notin I$ then,

$$\begin{aligned} E[Z_{b'} \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] &= (1 - \frac{1}{n})^{m-i}. \\ E[Z_b \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b'_i] &= (1 - \frac{1}{n})^{m-i}. \end{aligned} \tag{7}$$

Comparing (2) through (7), for $i \in [m]$,

$$\begin{aligned} |E[Z \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b_i] - & \\ E[Z \mid B_1 = b_1, \ldots, B_{i-1} = b_{i-1}, B_i = b'_i]| &\leq c_i, \end{aligned}$$

where

$$c_i = (1 - 1/n)^{m-i},$$

Thus we have from Theorem 2,

$$\Pr[|Z - E[Z]| > t] < 2\exp(-2t^2 / \sum_i c_i^2).$$

Since

$$\sum_i c_i^2 = \frac{n^2 - \mu^2}{2n - 1},$$

we get the occupancy bound Theorem 2 in [5]:

$$\Pr[|Z - \mu| \geq \theta\mu] \leq 2\exp\left(-\frac{\theta^2\mu^2(n - 1/2)}{n^2 - \mu^2}\right).$$

3 Negative Association

A very useful and robust notion of negative dependence between random variables called *negative association* was introduced by Joag–Dev and Proschan [4]:

Definition 3 (Negative Association) *Let* $\mathbf{X} := (X_1, \ldots, X_n)$ *be a vector of random variables.*

$(-A)$ *The random variables,* $\mathbf{X}$ *are* **negatively associated** *if for every two disjoint index sets,* $I, J \subseteq [n]$*,*

$$E[f(X_i, i \in I)g(X_j, j \in J)] \leq E[f(X_i, i \in I)]E[g(X_j, j \in J)],$$

for functions $f : \mathsf{R}^{|I|} \rightarrow \mathsf{R}$ *and* $g : \mathsf{R}^{|J|} \rightarrow \mathsf{R}$ *that are both non–decreasing (or both non–increasing) with respect to the usual co–ordinatewise ordering of Euclidean spaces.*

The next lemma list some properties that facilitate proofs of negative association. The properties themselves are immediate from the definition.

Lemma 4 *1. If* $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy* $(-A)$ *and are mutually independent, then the augmented vector* $(\mathbf{X}, \mathbf{Y}) = (X_1, \cdots, X_n, Y_1, \cdots, Y_m)$ *satisfies* $(-A)$*.*

2. Let $\mathbf{X} := (X_1, \cdots, X_n)$ *satisfy* $(-A)$*. Let* $I_1, \cdots, I_k \subseteq [n]$ *be disjoint index sets, for some positive integer* k*. For* $j \in [k]$*, let* $h_j : \mathsf{R}^{|I_k|} \rightarrow \mathsf{R}$ *be non–decreasing (or non–increasing) functions, and define* $Y_j := h_j(X_i, i \in I_j)$*. Then the vector* $\mathbf{Y} := (Y_1, \cdots, Y_k)$ *also satisfies* $(-A)$*. That is, non–decreasing (or non–increasing) functions of disjoint subsets of negatively associated variables are also negatively associated.*

Proposition 5 *The random variables* $Z_1, \ldots, Z_n$ *are negatively associated.*

Sketch of Proof. Introduce the indicator variables for $i \in [n], k \in [m]$,

$$B_{i,k} := \begin{cases} 1, & \text{if ball } k \text{ goes into bin } i; \\ 0, & \text{otherwise.} \end{cases}$$

For each $k \in [m]$, it is easy to show that the variables $(B_{i,k} \mid i \in [n])$ are negatively associated (observe that their joint distribution is simply a permutation distribution on $0, 0, \ldots, 0, 1$). Since each ball is thrown independently of the others, we conclude from Lemma 4(1) that the full set of variables $(B_{i,k} \mid i \in [n], k \in [m])$ are negatively associated. Finally observe that (using the Iverson symbol $[P]$ which is 1 if the boolean property P is true and 0 otherwise) $Z_i = [\sum_k B_{i,k} = 0]$ is a non–increasing function of $B_{i,k}, k \in [m]$. Hence the result follows from Lemma 4(2). ∎

Remark 6 In [2], a much fuller discussion of negative dependence in the balls and bins experiment can be found.

Proposition 7 ([2]) *The Chernoff–Hoeffding bound applies to sums of negatively associated variables.*

Sketch of Proof. From the definition of negative association, it follows by induction that if $X_1, \ldots, X_n$ are negatively associated, then

$$E[\prod_{i \in [n]} f_i(X_i)] \leq \prod_{i \in [n]} E[f_i(X_i)],$$

for any non–decreasing functions $f_i : \mathsf{R} \to \mathsf{R}$. Now we apply the usual proof of the Chernoff–Hoeffding bound (see for instance [1, 7]) using the above inequality to replace the equality $E[\prod_{i \in [n]} e^{X_i}] = \prod_{i \in [n]} E[e^{X_i}]$ (which holds when the variables are independent) by the above inequality with each $f_i(x) := e^x$. ∎

This yields directly, in a calculation–free manner, the (Chernoff) Occupancy Bound 2 in [5, Theorem 3].

Remark 8 It should be mentioned that [5] fail to mention [8] where this result is also obtained by arguments similar to theirs.

References

[1] N. Alon, P. Erdös, J. Spencer, *The Probabilistic Method*, John Wiley, 1992.

[2] D. Dubhashi, D. Ranjan, "Balls and Bins: A Study in Correlations", Seventh Annual Conference on Random Structures and Algorithms, Atlanta 1995.

[3] M.L. Eaton, *Lectures on Topics in Probability Inequalities*, CWI Tracts in Mathematics, 35. 1982.

[4] K. Joag-Dev and F. Proschan, "Negative Association of Random Variables with Applications", *Annals of Statistics*, 11:4, pp. 286–295, 1983.

[5] A. Kamath, R. Motwani, K. Palem and P. Spirakis, " Tail Bounds for Occupanvy Problems and the Satisfiability Threshold Conjecture", *Random Structures and Algorithms*, 7:1, pp. 59–80, 1995.

[6] C. McDiarmid, "On the Method of Bounded Differences", in J. Siemons (ed) *Surveys in Combinatorics*, London mathematical Society lecture Note Series 141, 1989.

[7] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[8] A. Panconesi and A. Srinivasan, "Fast Randomized Algorithms for Distributed Edge Coloring", in *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 1992, pp. 251–262.

[9] R. Szekeli, *Stochastic Ordering and Dependence in Applied probability*, Lecture Notes in Statistics 97, Springer–Verlag 1995.

[10] Y.L. Tong, *Probability Inequalities in Multivariate Distributions*, Academic Press, 1980.

Recent Publications in the BRICS Report Series

RS-95-48 Devdatt P. Dubhashi. *Simple Proofs of Occupancy Tail Bounds*. September 1995. 7 pp.

RS-95-47 Dany Breslauer. *The Suffix Tree of a Tree and Minimizing Sequential Transducers*. September 1995. 15 pp.

RS-95-46 Dany Breslauer, Livio Colussi, and Laura Toniolo. *On the Comparison Complexity of the String Prefix-Matching Problem*. August 1995. 39 pp. Appears in Leeuwen, editor, *Algorithms - ESA '94: Second Annual European Symposium proceedings*, LNCS 855, 1994, pages 483–494.

RS-95-45 Gudmund Skovbjerg Frandsen and Sven Skyum. *Dynamic Maintenance of Majority Information in Constant Time per Update*. August 1995. 9 pp.

RS-95-44 Bruno Courcelle and Igor Walukiewicz. *Monadic Second-Order Logic, Graphs and Unfoldings of Transition Systems*. August 1995. 39 pp. To be presented at CSL '95.

RS-95-43 Noam Nisan and Avi Wigderson. *Lower Bounds on Arithmetic Circuits via Partial Derivatives (Preliminary Version)*. August 1995. 17 pp. To appear in *36th Annual Conference on Foundations of Computer Science*, FOCS '95, IEEE, 1995.

RS-95-42 Mayer Goldberg. *An Adequate Left-Associated Binary Numeral System in the λ-Calculus*. August 1995. 16 pp.

RS-95-41 Olivier Danvy, Karoline Malmkjær, and Jens Palsberg. *Eta-Expansion Does The Trick*. August 1995. 23 pp.

RS-95-40 Anna Ingólfsdóttir and Andrea Schalk. *A Fully Abstract Denotational Model for Observational Congruence*. August 1995. 29 pp.

RS-95-39 Allan Cheng. *Petri Nets, Traces, and Local Model Checking*. July 1995. 32 pp. Full version of paper appearing in Proceedings of AMAST '95, LNCS 936, 1995.

RS-95-38 Mayer Goldberg. *Gödelisation in the λ-Calculus*. July 1995. 7 pp.