Combinatorics, Probability and Computing (1998) 7, 57–63. Printed in the United Kingdom
© 1998 Cambridge University Press

A Large Deviation Inequality for Functions of Independent, Multi-Way Choices

D. A. GRABLE†

Institut für Informatik, Humboldt-Universität zu Berlin,
Unter den Linden 6, D-10099 Berlin, Germany
(e-mail: grable@informatik.hu-berlin.de)

Received 4 March 1996; revised 6 July 1996

1. Introduction

Often when analysing randomized algorithms, especially parallel or distributed algorithms, one is called upon to show that some function of many independent choices is tightly concentrated about its expected value. For example, the algorithm might colour the vertices of a given graph with two colours and one would wish to show that, with high probability, very nearly half of all edges are monochromatic.

The classic result of Chernoff [3] gives such a large deviation result when the function is a sum of independent indicator random variables. The results of Hoeffding [5] and Azuma [2] give similar results for functions which can be expressed as martingales with a bounded difference property. Roughly speaking, this means that each individual choice has a bounded effect on the value of the function. McDiarmid [9] nicely summarized these results and gave a host of applications. Expressed a little differently, his main result is as follows.

We are given an underlying probability space generated by a finite set of mutually independent choices, indexed by $i \in I$, where choice i is from a set A_i, and a random variable Y on this probability space. (Equivalently, Y is a function of the underlying choices.) Let c_i be such that changing choice i (but keeping all else the same) can change Y by at most c_i. We call c_i the *effect* of i. The result is then that

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > \sqrt{\varphi \sum_i c_i^2/2}\right] \leqslant 2e^{-\varphi}$$

† Supported by Deutsche Forschungsgemeinschaft project number Pr 296/4-1.

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

for any $\varphi > 0$. Note that the martingale is not explicitly mentioned in the statement of this result. We will follow this example and systematically conceal the martingale in our own results.

Kim [8] gave an improved large deviation inequality for the situation where all of the choices are YES/NO choices and the deviation is not too large. Basically, he obtains an additional factor of $8p(1-p)$ under the radical, resulting in a generalization of the bounds of the tails of the binomial distribution which follow from the DeMoivre–Laplace limit theorem. Note, however, that this result for the case when all $c_i = 1$ was implied already by Lemma 7.14 of [9]. In recent work, Alon, Kim and Spencer [1] further refine this result by introducing the idea of a query game. The basic idea here is that while some choices might have an effect, they don't always have an effect.

The purpose of this note is to point out a few improvements in the Alon–Kim–Spencer result and, more significantly, to prove a version of the result in which multi-way choices are allowed. This last point is of particular importance in the analysis of, for example, colouring algorithms. To illustrate this, we also present a sample application.

2. Results

This situation is quite common: we have a probability space generated by independent choices, choice $i \in I$ from the set A_i, and a random variable Y of interest. Consider a query game in which we determine the value of Y by making queries of the form 'What was choice i?' to a truthful oracle. Our *strategy* can be expressed as a decision tree whose internal nodes designate queries to be made and whose leaves designate values of Y.

Define the *variance* of a query (internal node) q concerning choice i to be

$$v_q = \sum_{a \in A_i} p_{i,a} \mu_{q,a}^2,$$

where

$$p_{i,a} = \Pr[\text{choice } i \text{ was } a]$$

and

$$\mu_{q,a} = \mathrm{Ex}[Y \mid \text{choice } i \text{ was } a \text{ and all previous queries}] - \mathrm{Ex}[Y \mid \text{all previous queries}].$$

The quantity $\mu_{q,a}$ is not as mysterious as it appears. It simply measures the amount which our expectation changes when the answer to query q is revealed to be a. Also define the *maximum effect* of query q as

$$c_q = \max_{a,b \in A_i} |\mu_{q,a} - \mu_{q,b}|.$$

As an upper bound on c_q we often take the maximum amount which Y can change if choice i is changed but all other choices remain the same.

A *line of questioning* is a path in the decision tree from the root to a leaf and the *variance* of a line of questioning is the sum of the variances of the queries along it. Finally, the *variance* of a strategy is the maximum variance over all lines of questioning. The use of the term variance is meant to be suggestive: the variance of a strategy for determining Y is an upper bound on the variance of Y.

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

Theorem. *If there is a strategy for determining Y with variance at most V then*

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > 2\sqrt{\varphi V}\right] \leqslant 2e^{-\varphi}$$

for every $0 \leqslant \varphi \leqslant V / \max c_q^2$.

This is more-or-less an analogue of the Alon–Kim–Spencer result for multi-way choices. The main difference is that they consider the effect and variance of choices rather than queries. The way we've done things allows a choice to have different levels of effect depending upon the situation when the query is made, rather than just 'all or nothing'.

Note that we can improve the constants a bit, but this basic form is all that is usually needed for the sorts of applications in which we're interested. In the improved form, we can show that $\Pr\left[|Y - \mathrm{Ex}[Y]| > \sqrt{2(1+\varepsilon)\varphi V}\right] \leqslant 2e^{-\varphi}$ for every $\varepsilon > 0$ and $0 \leqslant \varphi \leqslant \varepsilon^2(1+\varepsilon)V/2\max c_q^2$.

One further improvement is possible. It can happen that certain very low probability situations occur that force certain choices to have a very large effect. In other words, there might be just a few, very low probability, vectors of choices (outcomes) which lead to lines of questioning with very large variance. This of course drives up the variance of the entire strategy, possibly ruining the usefulness of the result. We'd like to be able to ignore these *exceptional* outcomes and the corresponding lines of questioning. The following easy corollary of the theorem says that essentially we can.

Corollary. *Let m and M be, respectively, the minimum and maximum values taken by Y over all possible outcomes. Let $\mathscr{C}$ be a set of exceptional outcomes. Consider a strategy for determining Y, assuming that the actual outcome is not in $\mathscr{C}$. If the variance of this strategy is at most V then*

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > 2\sqrt{\varphi V} + (M-m)\Pr[\mathscr{C}]\right] \leqslant 2e^{-\varphi} + \Pr[\mathscr{C}],$$

for every $0 \leqslant \varphi \leqslant V / \max c_q^2$.

When applying the theorem or the corollary, to determine the variance of each query, we must bound the sum $\sum p_{i,a}\mu_{q,a}^2$. There are two 'standard' ways to do this. If we only know that the μs differ by at most c_q, the best we can say is that

$$v_q = \sum_{a \in A_i} p_{i,a}\mu_{q,a}^2 \leqslant c_q^2/4. \tag{2.1}$$

If we know, in addition, that the μs take on two values, one corresponding to choices with total probability p and the other to choices with total probability $1-p$, then we know that

$$v_q = \sum_{a \in A_i} p_{i,a}\mu_{q,a}^2 \leqslant p(1-p)c_q^2. \tag{2.2}$$

These bounds follow from elementary, but non-trivial computations, which we omit. Using only the first bound gives us up to a constant the result of Hoeffding. Using only the second bound, restricted to YES/NO choices, gives us the Alon–Kim–Spencer result.

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

3. An application

Before proving the theorem, we present a sample application. Our application is the analysis of one step of a randomized, distributed (or parallel) vertex colouring algorithm similar to Algorithm 2 of Kim [7]. Given a d-regular triangle-free graph G, we tentatively colour each vertex of G independently at random from identical palettes of c colours. If a vertex's tentative colour is unique among its neighbours' tentative colours, we make the colour permanent.

Fix a vertex v. Let Y be the number of v's neighbours which successfully colour themselves. It is not too hard to compute Y's expectation. The probability that vertex w colours itself is

$$\left(1-\frac{1}{c}\right)^d \sim \exp\left\{-\frac{d}{c}\right\},$$

so the expected number of successful neighbours is

$$\mathrm{Ex}[Y] \sim d\exp\left\{-\frac{d}{c}\right\}.$$

But how concentrated is Y? We are especially interested in proving sharp concentration results when the number of colours c is $\Theta(d/\ln d)$ with some small constant factor. When we iterate this procedure (with the necessary modifications as the vertex palettes evolve), we might then be able to show that G is $(1+o(1))(d/\ln d)$-colourable, leading to a simplified proof of Kim [7] and Johansson [6] as well as demonstrating the effectiveness of such a vertex colouring algorithm. See Grable and Panconesi [4] for a related analysis of an edge colouring algorithm.

Here is a strategy for determining Y: first query v (*i.e.* ask what vertex v's tentative colour is); then query all neighbours of v; finally, query all neighbours of the neighbours which chose a different colour than v (the ones that chose the same colour are already known to have failed).

The first query has variance 0, since by itself v's colour choice does not affect the expectation of Y. When we query each neighbour of v, there are two types of outcome: either the neighbour chose the same tentative colour as v or it didn't. In any case, the expectation of Y can change by at most 1. Using (2.2), we see therefore that each of these queries has variance at most $\frac{1}{c}(1-\frac{1}{c})1^2 < \frac{1}{c}$.

When it comes to bounding the variance of the neighbours of the neighbours of v, there is a potential problem. If, as in Kim [7], G also contains no cycles of length 4, things are very easy since no vertex has more than one neighbour in common with v. Let u be a neighbour of v and w a neighbour of u. When we query w, there are only two sorts of outcome. Either w chose the same colour as u (which means u was not successful) or not. Thus we can use (2.2) to bound the variance of the query: $v_w \leqslant \frac{1}{c}\left(1-\frac{1}{c}\right)1^2 < \frac{1}{c}$. Since there are $d(d-1)$ neighbours of neighbours, the total variance of this strategy is at most $V = d^2/c$. The theorem therefore implies that *when G contains no 4-cycles*

$$\Pr\left[|Y-\mathrm{Ex}[Y]| > 2\sqrt{\varphi d^2/c}\right] \leqslant 2e^{-\varphi}$$

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

for every $0 \leqslant \varphi \leqslant d^2/c$. In particular, for $c = \Theta(d/\ln d)$, we get

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > \Theta\left(\sqrt{\varphi d \ln d}\right)\right] \leqslant 2e^{-\varphi}$$

for every $0 \leqslant \varphi \leqslant \Theta(d \ln d)$. Furthermore, we usually only need inverse polynomially small bounds on the probability of deviation, so we usually only need $\varphi = \Theta(\ln d)$. Thus Y is almost always within $\Theta\left(\sqrt{d}\ln d\right)$ of its expectation.

Return to the general case, where G may contain 4-cycles. Because the structure of G is beyond our control, it is possible that a single vertex w shares many common neighbours with v. The extreme example is when G is $K_{d,d}$. In this case, there is the (unlikely) possibility that many of v's neighbours choose the same tentative colour. This results in w having a large potential effect.

This is where we make use of the corollary's ability to handle exceptional choice vectors. Of course, the expected number of neighbours of v tentatively choosing any particular colour γ is d/c. Since the number of vertices tentatively choosing γ is binomially distributed, the Chernoff bounds give that for any $\delta > 2e-1$, the probability that more than $(1+\delta)d/c$ vertices choose γ is less than $2^{-(1+\delta)d/c}$. Thus the probability that some colour is used more than $r_\delta = (1+\delta)d/c$ times is less than $c2^{-(1+\delta)d/c}$. Let $\mathscr{C}$ be the set of outcomes in which this is the case – that is, those outcomes where more than r_δ of the neighbours of v choose the same colour.

Let w be a vertex sharing exactly x neighbours with v and, for each colour a, let x_a be the number of common neighbours tentatively coloured a. Thus, $|\mu_{w,a}| \leqslant x_a$ since w can make x_a of v's neighbours unsuccessful by choosing a. Thus,

$$v_w = \sum p_{i,a}\mu_{w,a}^2 \leqslant \frac{1}{c}\sum x_a^2.$$

Since each $x_a \leqslant r_\delta$ and the sum of squares is maximized at the extreme point, when there are x/r_δ terms each equal to r_δ^2,

$$v_w \leqslant \frac{1}{c}\frac{x}{r_\delta}r_\delta^2 = \frac{xr_\delta}{c} = \frac{(1+\delta)xd}{c^2}.$$

Summing the variances along the (non-exceptional) lines of questioning, we get

$$0 + \frac{d}{c} + d(d-1)\frac{(1+\delta)d}{c^2}.$$

(The last term follows since, counting with multiplicity, there are $d(d-1)$ neighbours of neighbours. Each one which shares x neighbours with v has an amortized variance of $(1+\delta)d/c^2$ per adjacency. Of course, we only query once per vertex!) We can bound this variance by $V = (1+\delta)d^3/c^2$ as long as $c \leqslant (1+\delta)d$.

The corollary then tells us that, for $\delta > 2e-1$ and $0 \leqslant \varphi \leqslant V/r_\delta^2 = d/(1+\delta)$,

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > 2\sqrt{\varphi(1+\delta)\frac{d^3}{c^2}} + dc2^{-(1+\delta)d/c}\right] \leqslant 2e^{-\varphi} + c2^{-(1+\delta)d/c}.$$

Setting $c = \Theta(d/\ln d)$ and $\varphi = \Theta(\ln d)$, this says that Y deviates from its expectation by more than $\Theta\left(\sqrt{d\ln^3 d}\right)$ with inverse polynomially small probability.

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

4. Proofs

The proof of the theorem is similar in spirit to the Alon–Kim–Spencer proof. For simplicity, replace Y by $Y - \mathrm{Ex}[Y]$, so that we may assume that Y has zero mean. By symmetry, we bound only the upper tail of Y.

The theorem follows easily from the following lemma.

Lemma. $\mathrm{Ex}\left[e^{\lambda Y}\right] \leqslant e^{\lambda^2 V}$ *holds for any* $0 \leqslant \lambda \leqslant 1/\max c_q$.

Simply set $\lambda = \sqrt{\varphi/V}$. The side condition of the theorem implies the side condition of the lemma. The lemma together with Markov's inequality then implies that

$$\Pr\left[Y > 2\sqrt{\varphi V}\right] = \Pr\left[e^{\lambda Y} > e^{2\lambda\sqrt{\varphi V}}\right] < e^{-2\lambda\sqrt{\varphi V}}\mathrm{Ex}\left[e^{\lambda Y}\right] \leqslant e^{-2\varphi}e^{\varphi} = e^{-\varphi}.$$

The proof of the lemma is by induction on the depth of the decision tree. In the base case, $Y = 0$, so the claim is trivially true. Otherwise, consider the root node (query) q. It is of the form 'what was choice i?' for some choice i. Because Y has expectation zero, $\mu_{q,a} = \mathrm{Ex}[Y \mid \text{choice } i \text{ was } a]$ for each a in A_i. As before, the variance is $v_q = \sum p_{i,a}\mu_{q,a}^2$ and the maximum effect is $c_q = \max|\mu_{q,a} - \mu_{q,b}|$.

Since $(Y - \mu_{q,a} \mid \text{choice } i \text{ was } a)$ is a random variable with zero mean and the total variation of the a-subtree, which concerns itself with this random variable, is at most $V - v_q$, we have by induction that

$$\mathrm{Ex}\left[e^{\lambda(Y-\mu_{q,a})} \mid \text{choice } i \text{ was } a\right] \leqslant e^{\lambda^2(V-v_q)}.$$

Therefore,

$$\begin{aligned}\mathrm{Ex}\left[e^{\lambda Y}\right] &= \sum p_{i,a}e^{\lambda\mu_{q,a}}\mathrm{Ex}\left[e^{\lambda(Y-\mu_{q,a})} \mid \text{choice } i \text{ was } a\right] \\ &\leqslant e^{\lambda^2(V-v_q)}\sum p_{i,a}e^{\lambda\mu_{q,a}}.\end{aligned}$$

Furthermore, if we take the series expansion of the latter sum, recall the definition of v_q, and observe that the side condition implies that $\lambda\mu_{q,a} \leqslant \lambda c_q \leqslant 1$, we see that

$$\begin{aligned}\sum p_{i,a}e^{\lambda\mu_{q,a}} &= \sum p_{i,a}\left(1 + \lambda\mu_{q,a} + \lambda^2\mu_{q,a}^2/2 + \lambda^3\mu_{q,a}^3/6 + \cdots\right) \\ &= 1 + 0 + \sum p_{i,a}\lambda^2\mu_{q,a}^2/2 + \sum p_{i,a}\lambda^3\mu_{q,a}^3/6 + \cdots \\ &= 1 + \lambda^2 v_q\sum_{i\geqslant 2}(\lambda\mu_{q,a})^{i-2}/i! \\ &\leqslant 1 + \lambda^2 v_q\sum_{i\geqslant 2}1/i! \\ &< 1 + \lambda^2 v_q \quad \leqslant \quad e^{\lambda^2 v_q}.\end{aligned}$$

Thus,

$$\mathrm{Ex}\left[e^{\lambda Y}\right] \leqslant e^{\lambda^2(V-v_q)}e^{\lambda^2 v_q} = e^{\lambda^2 V},$$

which proves the lemma and therefore the theorem.

The corollary follows easily. First we apply the theorem to the random variable $(Y|\overline{\mathscr{C}})$

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press

with the given strategy, to get

$$\Pr\left[|Y - \mathrm{Ex}[Y|\overline{\mathscr{C}}]| > 2\sqrt{\varphi V} \mid \overline{\mathscr{C}}\right] \leqslant 2e^{-\varphi}$$

for every $0 \leqslant \varphi \leqslant V/\max c_q^2$. But we're neither interested in concentration around $\mathrm{Ex}[Y|\overline{\mathscr{C}}]$, which is quite likely difficult to compute, nor the probability conditioned on $\overline{\mathscr{C}}$.

To get rid of the first difficulty, we show that $\mathrm{Ex}[Y]$ and $\mathrm{Ex}[Y|\overline{\mathscr{C}}]$ must not differ by much: observe that, since

$$\mathrm{Ex}[Y] = \mathrm{Ex}[Y|\overline{\mathscr{C}}](1 - \Pr[\mathscr{C}]) + \mathrm{Ex}[Y|\mathscr{C}]\Pr[\mathscr{C}]$$

and $m \leqslant Y \leqslant M$,

$$\mathrm{Ex}[Y|\overline{\mathscr{C}}] - \big(\mathrm{Ex}[Y|\overline{\mathscr{C}}] - m\big)\Pr[\mathscr{C}] \leqslant \mathrm{Ex}[Y] \leqslant \mathrm{Ex}[Y|\overline{\mathscr{C}}] + \big(M - \mathrm{Ex}[Y|\overline{\mathscr{C}}]\big)\Pr[\mathscr{C}],$$

and therefore that

$$|\mathrm{Ex}[Y] - \mathrm{Ex}[Y|\overline{\mathscr{C}}]| \leqslant (M - m)\Pr[\mathscr{C}].$$

The second difficulty is also easily dealt with: notice that for any quantity A

$$\begin{aligned}\Pr\big[|Y - \mathrm{Ex}[Y]| > A\big] &= \Pr\big[|Y - \mathrm{Ex}[Y]| > A \text{ and } \overline{\mathscr{C}}\big] + \Pr\big[|Y - \mathrm{Ex}[Y]| > A \text{ and } \mathscr{C}\big] \\ &\leqslant \Pr\big[|Y - \mathrm{Ex}[Y]| > A \mid \overline{\mathscr{C}}\big] + \Pr[\mathscr{C}].\end{aligned}$$

Combining these statements with $A = 2\sqrt{\varphi V} + (M - m)\Pr[\mathscr{C}]$ gives the conclusion of the corollary, namely that

$$\Pr\left[|Y - \mathrm{Ex}[Y]| > 2\sqrt{\varphi V} + (M - m)\Pr[\mathscr{C}]\right] \leqslant 2e^{-\varphi} + \Pr[\mathscr{C}],$$

for every $0 \leqslant \varphi \leqslant V/\max c_q^2$.

References

[1] Alon, N., Kim, J. H. and Spencer J. (1997) Nearly perfect matchings in regular simple hypergraphs. *Israel J. Math.* **100** 171–188.

[2] Azuma, K. (1967) Weighted sums of certain dependent random variables. *Tôkuku Math. J.* **19** 357–367.

[3] Chernoff, H. (1952) A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observables. *Ann. Math. Stat.* **23** 493–509.

[4] Grable, D. A. and Panconesi, A. (1997) Nearly optimal distributed edge colouring in $O(\log\log n)$ rounds. *Random Structures and Algorithms* **10** 385–405.

[5] Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[6] Johansson, A. (1996) An improved upper bound on the choice number for triangle free graphs. Preprint.

[7] Kim, J. H. (1995) On Brooks' theorem for sparse graphs. *Combinatorics, Probability and Computing* **4** 97–132.

[8] Kim, J. H. (1995) The Ramsey number $R(3,t)$ has order of magnitude $t^2/\log t$. *Random Structures and Algorithms* **7** 173–207.

[9] McDiarmid, C. (1989) On the method of bounded differences. In *Surveys in Combinatorics: LMS Lecture Notes Series* (Siemons, J., ed.), **141** 148–188.

https://doi.org/10.1017/S0963548397003246 Published online by Cambridge University Press