LARGE DEVIATIONS FOR SUMS OF PARTLY DEPENDENT RANDOM VARIABLES

SVANTE JANSON

ABSTRACT. We use and extend a method by Hoeffding to obtain strong large deviation bounds for sums of dependent random variables with suitable dependency structure. The method is based on breaking up the sum into sums of independent variables. Applications are given to U-statistics, random strings and random graphs.

1. INTRODUCTION

Many random variables can be written as a sum

$$X = \sum_{\alpha \in \mathcal{A}} Y_\alpha. \tag{1.1}$$

of simpler random variables Y_α, with α ranging over some index set. For example, each Y_α may be an indicator variable taking the values 0 and 1 only, i.e. $Y_\alpha \sim \mathrm{Be}(p_\alpha)$ for some $p_\alpha \in [0,1]$.

We are interested in situations where the variables Y_α may be dependent, but there is a large amount of independence among them. A typical situation is the sum

$$X = \sum_{(i_1,\dots,i_d)\in\mathcal{A}} f_{i_1\cdots i_d}(\xi_{i_1},\dots,\xi_{i_d}), \tag{1.2}$$

for some functions $f_{i_1\cdots i_d}$ and independent random variables $\xi_1,\dots,\xi_n$, and some set $\mathcal{A} \subseteq [n]^d_<$, where $[n]^d_<$ is the set of all d-tuples $(i_1,\dots,i_d)$ with $1 \le i_1 < \cdots < i_d \le n$. Here d and n are some positive integers; typically d is small (perhaps only 2 or 3) and n is large.

One example of such sums (1.2) is the family of *U-statistics* [13], which is the symmetric case obtained by taking $\xi_1,\dots,\xi_n$ i.i.d., all $f_{i_1\cdots i_d}$ equal to some symmetric function f, and $\mathcal{A} = [n]^d_<$. More generally, if we in this situation sum over a subset $\mathcal{A} \subset [n]^d_<$, we obtain an *incomplete U-statistic.* Also *two-sample U-statistics* are of the general type (1.2), but now the ξ_i are of two different types.

Another example of (1.2), now with a non-symmetric f, is the problem on random strings described in Example 4.2; this problem was the direct motivation to write this paper. Further examples, for random graphs and hypergraphs, are also given in Section 4.

The purpose of this paper is to prove small bounds for the probability of large deviations of a variable (1.1). Such bounds have a long history. In the case of independent summands, some of the most important contributions are

Date: September 19, 2002; updated October 9, 2003.

Bernstein [4], Cramér [9], Feller [10], Chernoff [8], Okamoto [19], Bennett [3], and Hoeffding [14]. For dependent summands, there are many results proved by many authors using different methods under various assumptions; for a few of these, see the surveys in [16] and [18].

The present paper is based on Hoeffding [14], which besides the independent case also studies several dependent cases, among them U-statistics. Hoeffding's method for this case is based on breaking up the sum (1.2) into several parts, each part being a sum of independent variables. The same idea has been used in a somewhat different form (see Remark 5.2) by, among others, [21], [18], [20]. We will show how Hoeffding's original method, with only minor modifications, extends to general sums (1.2) and (1.1).

The main results are stated in Section 2, proved in Section 3 and applied in Section 4. Section 5 contains some further remarks.

Acknowledgements. I thank Wojtek Szpankowski for stimulating discussions that led to this research. Part of the research was carried out at the Newton Institute in Cambridge. I thank Ravi Kannan, Andrzej Ruciński and Peter Winkler for helpful comments.

2. Results

Definitions. Given $\mathcal{A}$ and $\{Y_\alpha\}$, $\alpha \in \mathcal{A}$, we make the following definitions.

- A subset $\mathcal{A}'$ of $\mathcal{A}$ is *independent* if the corresponding random variables $\{Y_\alpha\}_{\alpha\in\mathcal{A}'}$, are independent.
- A family $\{\mathcal{A}_j\}_j$ of subsets of $\mathcal{A}$ is a *cover* of $\mathcal{A}$ if $\bigcup_j \mathcal{A}_j = \mathcal{A}$.
- A family $\{(\mathcal{A}_j, w_j)\}_j$ of pairs $(\mathcal{A}_j, w_j)$, where $\mathcal{A}_j \subseteq \mathcal{A}$ and $w_j \in [0,1]$ is a *fractional cover* of $\mathcal{A}$ if $\sum_j w_j \mathbf{1}_{\mathcal{A}_j} \geq \mathbf{1}_{\mathcal{A}}$, i.e. $\sum_{j:\alpha\in\mathcal{A}_j} w_j \geq 1$ for each $\alpha \in \mathcal{A}$.
- A (fractional) cover is *proper* if each set $\mathcal{A}_j$ in it is independent.
- $\chi(\mathcal{A})$ is the size of the smallest proper cover of $\mathcal{A}$, i.e. the smallest m such that $\mathcal{A}$ is the union of m independent subsets.
- $\chi^*(\mathcal{A})$ is the minimum of $\sum_j w_j$ over all proper fractional covers $\{(\mathcal{A}_j, w_j)\}_j$.

Note that, in spite of our notation, $\chi(\mathcal{A})$ and $\chi^*(\mathcal{A})$ depend not only on $\mathcal{A}$ but also on the family $\{Y_\alpha\}_{\alpha\in\mathcal{A}}$.

A cover can be regarded as a fractional cover with every $w_j = 1$. Hence

$$\chi^*(\mathcal{A}) \leq \chi(\mathcal{A}). \tag{2.1}$$

We can thus replace $\chi^*(\mathcal{A})$ by $\chi(\mathcal{A})$ in all results below, and the reader that prefers it may consider only covers and ignore the fractional ones.

Note further that $\chi^*(\mathcal{A}) \geq 1$ (unless $\mathcal{A} = \emptyset$) and that $\chi^*(\mathcal{A}) = 1$ if and only if the variables Y_α, $\alpha \in \mathcal{A}$, are independent. It is often convenient to consider a *dependency graph* for $\{Y_\alpha\}$. This is a graph Γ with vertex set $\mathcal{A}$ such that if $\mathcal{B} \subset \mathcal{A}$ and $\alpha \in \mathcal{A}$ is not connected by an edge to any vertex in $\mathcal{B}$, then Y_α is independent of $\{Y_\beta\}_{\beta\in\mathcal{B}}$. A typical example is for sums of type (1.2), where $\mathcal{A} \subseteq [n]^d_<$; we define Γ to have an edge between every pair of d-tuples $\alpha, \beta \in [n]^d_<$ such that α and β have an index in common. (See further Remark 5.3 and the examples in Section 4.) It follows that a subset of $\mathcal{A}$ that is independent in Γ in the graph theory sense (no edges inside the subset) is independent in the sense above. (The converse is not necessarily true, see Remark 5.3.) Hence, $\chi(\mathcal{A}) \le \chi(\Gamma)$ and $\chi^*(\mathcal{A}) \le \chi^*(\Gamma)$, where $\chi(\Gamma)$ and $\chi^*(\Gamma)$ are the usual chromatic and fractional chromatic numbers of Γ.

We further let $\Delta(\Gamma)$ denote the maximum degree of Γ and let (for convenience) $\Delta_1(\Gamma) := \Delta(\Gamma) + 1$. It is well-known (and easily seen by a greedy colouring) that $\chi(\Gamma) \le \Delta_1(\Gamma)$, see e.g. [6]. When Γ is a dependency graph for $\{Y_\alpha\}$, we thus have

$$\chi^*(\mathcal{A}) \le \chi^*(\Gamma) \le \chi(\Gamma) \le \Delta_1(\Gamma), \tag{2.2}$$

so $\chi^*(\mathcal{A})$ can be replaced by any of $\chi^*(\Gamma)$, $\chi(\Gamma)$, $\Delta_1(\Gamma)$ in our results; in many applications this causes no significant loss.

We now can state our main results. The first extends [14, Theorem 2], which is the case $\chi^*(\mathcal{A}) = 1$ (i.e. independent variables).

Theorem 2.1. *Suppose that X is as in (1.1) with $a_\alpha \le Y_\alpha \le b_\alpha$ for every $\alpha \in \mathcal{A}$ and some real numbers a_α and b_α. Then, for $t > 0$,*

$$\mathbb{P}(X \ge \mathbb{E}\, X + t) \le \exp\Bigl(-2\frac{t^2}{\chi^*(\mathcal{A}) \sum_{\alpha\in\mathcal{A}} (b_\alpha - a_\alpha)^2}\Bigr).$$

The same estimate holds for $\mathbb{P}(X \le \mathbb{E}\, X - t)$.

Corollary 2.2. *Suppose that X is as in (1.1) with $Y_\alpha \sim \mathrm{Be}(p_\alpha)$ for some $p_\alpha \in (0,1)$ and all $\alpha \in \mathcal{A}$. Then, for $t \ge 0$,*

$$\mathbb{P}(X \ge \mathbb{E}\, X + t) \le \exp\Bigl(-2\frac{t^2}{\chi^*(\mathcal{A})|\mathcal{A}|}\Bigr).$$

The same estimate holds for $\mathbb{P}(X \le \mathbb{E}\, X - t)$. $\square$

These results, while useful in many situations, can be improved (again by Hoeffding's methods) when the summands have variances that are substantially smaller than the upper bound $(b_\alpha - a_\alpha)^2/4$. Note that we assume a one-sided bound on Y_α in Theorem 2.3 (and Theorem 2.5) below; as in Theorem 2.3 the obtained estimate holds for $\mathbb{P}(X \le \mathbb{E}\, X - t)$ too (by considering $-X$), but only if the boundedness assumption is reversed to $Y_\alpha - \mathbb{E}\, Y_\alpha \ge -b$. We define

$$\varphi(x) := (1+x)\ln(1+x) - x. \tag{2.3}$$

Theorem 2.3. *Suppose that X is as in (1.1) with $Y_\alpha - \mathbb{E}\, Y_\alpha \le b$ for some $b > 0$ and all $\alpha \in \mathcal{A}$. Then, with $\varphi(x)$ as in (2.3) and $S := \sum_{\alpha\in\mathcal{A}} \operatorname{Var} Y_\alpha$, for $t \ge 0$,*

$$\mathbb{P}(X \ge \mathbb{E}\, X + t) \le \exp\Bigl(-\frac{S}{b^2\chi^*(\mathcal{A})}\varphi\Bigl(\frac{4bt}{5S}\Bigr)\Bigr) \le \exp\Bigl(-\frac{8t^2}{25\chi^*(\mathcal{A})(S + bt/3)}\Bigr).$$

Corollary 2.4. *Suppose that X is as in* (1.1) *with $Y_\alpha \sim \mathrm{Be}(p)$ for some $p \in (0,1)$ and all $\alpha \in \mathcal{A}$. Let $N := |\mathcal{A}|$. Then, for $t \geq 0$,*

$$\mathbb{P}(X \geq \mathbb{E}\,X + t) \leq \exp\Bigl(-\frac{Np}{(1-p)\chi^*(\mathcal{A})}\varphi\Bigl(\frac{4t}{5Np}\Bigr)\Bigr) \tag{2.4}$$

$$\leq \exp\Bigl(-\frac{8t^2}{25\chi^*(\mathcal{A})(Np+t/3)}\Bigr); \tag{2.5}$$

$$\mathbb{P}(X \leq \mathbb{E}\,X - t) \leq \exp\Bigl(-\frac{N(1-p)}{p\chi^*(\mathcal{A})}\varphi\Bigl(\frac{4t}{5N(1-p)}\Bigr)\Bigr); \tag{2.6}$$

$$\mathbb{P}(X \leq \mathbb{E}\,X - t) \leq \exp\Bigl(-\frac{8t^2}{25\chi^*(\mathcal{A})Np}\Bigr). \tag{2.7}$$

If we are given a dependency graph Γ, we may by (2.2) replace $\chi^*(\mathcal{A})$ by $\Delta_1(\Gamma)$. Actually, with Δ_1, which is often convenient for applications, we may improve the bounds a little as follows in the important case when all Y_α have the same distribution. (The improvement amounts typically to a factor $\approx 25/16$ in the exponent. For many applications this is irrelevant.)

Theorem 2.5. *Suppose that X is as in* (1.1) *and that all Y_α have the same distribution with $Y_\alpha - \mathbb{E}\,Y_\alpha \leq b$ for some $b > 0$. Suppose further that Γ is a dependency graph for $\{Y_\alpha\}_{\alpha\in\mathcal{A}}$. Let $N := |\mathcal{A}|$, $S := N\operatorname{Var} Y_\alpha$ and $\Delta_1 := \Delta_1(\Gamma)$. Then, with $\varphi(x)$ as in* (2.3), *for $t \geq 0$,*

$$\mathbb{P}(X \geq \mathbb{E}\,X + t) \leq \exp\Bigl(-\frac{S}{b^2\Delta_1}\varphi\Bigl(\frac{bt}{S(1+\Delta_1/8N)}\Bigr)\Bigr)$$

$$\leq \exp\Bigl(-\frac{t^2(1-\Delta_1/4N)}{2\Delta_1(S+bt/3)}\Bigr).$$

Corollary 2.6. *Under the assumptions of Corollary 2.4, and with $\Delta_1 = \Delta_1(\Gamma)$, where Γ is a dependency graph for $\{Y_\alpha\}$, for $t \geq 0$,*

$$\mathbb{P}(X \geq \mathbb{E}\,X + t) \leq \exp\Bigl(-\frac{Np}{(1-p)\Delta_1}\varphi\Bigl(\frac{t}{(N+\Delta_1/8)p}\Bigr)\Bigr) \tag{2.8}$$

$$\leq \exp\Bigl(-\frac{t^2(1-\Delta_1/4N)}{2\Delta_1(Np+t/3)(1-p)}\Bigr); \tag{2.9}$$

$$\mathbb{P}(X \leq \mathbb{E}\,X - t) \leq \exp\Bigl(-\frac{N(1-p)}{\Delta_1 p}\varphi\Bigl(\frac{t}{(N+\Delta_1/8)(1-p)}\Bigr)\Bigr); \tag{2.10}$$

$$\mathbb{P}(X \leq \mathbb{E}\,X - t) \leq \exp\Bigl(-\frac{t^2(1-\Delta_1/4N)}{2\Delta_1 Np}\Bigr). \tag{2.11}$$

3. Proofs

The reader is advised to first consider the case of covers only (every $w_i = 1$) for simplicity.

We say that a fractional cover $\{(\mathcal{A}_j, w_j)\}_j$ is *exact* if $\sum_j w_j \mathbf{1}_{\mathcal{A}_j} = \mathbf{1}_{\mathcal{A}}$.

Lemma 3.1. *If $\{(\mathcal{A}_j, w_j)\}_j$ is an exact fractional cover of $\mathcal{A}$, and x_α, $\alpha \in \mathcal{A}$, are any numbers, then*

$$\sum_{\alpha\in\mathcal{A}} x_\alpha = \sum_j w_j s_j,$$

where $s_j := \sum_{\alpha\in\mathcal{A}_j} x_\alpha$. In particular, $|\mathcal{A}| = \sum_j w_j |\mathcal{A}_j|$.

Proof.

$$\sum_j w_j s_j = \sum_j w_j \sum_{\alpha\in\mathcal{A}} \mathbf{1}_{\mathcal{A}_j}(\alpha) x_\alpha = \sum_{\alpha\in\mathcal{A}} \sum_j w_j \mathbf{1}_{\mathcal{A}_j}(\alpha) x_\alpha = \sum_{\alpha\in\mathcal{A}} x_\alpha.$$

The final claim follows by taking each $x_\alpha = 1$. □

The following simple lemma will enable us to consider only exact (fractional) covers, which will simplify proofs. In particular, it implies that χ^* is the minimum of $\sum_j w_j$ over all exact proper fractional covers.

Lemma 3.2. *If $\{(\mathcal{A}_j, w_j)\}_j$ is a fractional cover, we can replace each $(\mathcal{A}_j, w_j)$ by one or several $(\mathcal{A}_{jk}, w_{jk})$ with $\mathcal{A}_{jk} \subseteq \mathcal{A}_j$ and $\sum_k w_{jk} = w_j$ such that $\{(\mathcal{A}_{jk}, w_{jk})\}_{jk}$ is an exact fractional cover.*

Proof. Take an element $\alpha \in \mathcal{A}$. It is easily seen that $\{(\mathcal{A}_j, w_j)\}_j$ can be replaced by a family $\{(\mathcal{A}_{jk}, w_{jk})\}_{jk}$ with each $\mathcal{A}_{jk} = \mathcal{A}_j$ or $\mathcal{A}_j \setminus \{\alpha\}$, $\sum_k w_{jk} = w_j$, and $\sum w_{jk} \mathbf{1}_{\mathcal{A}_{jk}}(\alpha) = 1$. Repeat the same procedure for each $\alpha \in \mathcal{A}$. □

Proof of Theorem 2.1. We follow the method of Hoeffding [14]. Subtracting the means, we may assume that $\mathbb{E}\, Y_\alpha = 0$ for all α, and thus $\mathbb{E}\, X = 0$. Then, by [14, (4.16)], for every real h,

$$\mathbb{E} \exp\bigl(hY_\alpha\bigr) \le \exp\bigl(\tfrac{1}{8} h^2 (b_\alpha - a_\alpha)^2\bigr). \tag{3.1}$$

Let $\{(\mathcal{A}_j, w_j)\}_j$ be an exact proper fractional cover of $\mathcal{A}$, and let $X_j := \sum_{\mathcal{A}_j} Y_\alpha$. Then $X = \sum_j w_j X_j$ by Lemma 3.1. Let p_j be any positive numbers with $\sum_j p_j = 1$. By Jensen's inequality, for any real u,

$$\exp(uX) = \exp\Bigl(\sum_j p_j \frac{uw_j}{p_j} X_j\Bigr) \le \sum_j p_j \exp\Bigl(\frac{uw_j}{p_j} X_j\Bigr).$$

Taking the expectations we find using (3.1), since X_j is a sum of independent variables Y_α, $\alpha \in \mathcal{A}_j$,

$$\begin{aligned}\mathbb{E} \exp(uX) &\le \sum_j p_j \,\mathbb{E} \exp\Bigl(\frac{uw_j}{p_j} X_j\Bigr) = \sum_j p_j \prod_{\alpha\in\mathcal{A}_j} \mathbb{E} \exp\Bigl(\frac{w_j u}{p_j} Y_\alpha\Bigr) \\ &\le \sum_j p_j \prod_{\alpha\in\mathcal{A}_j} \exp\Bigl(\frac{w_j^2 u^2}{8p_j^2}(b_\alpha - a_\alpha)^2\Bigr) = \sum_j p_j \exp\Bigl(\frac{w_j^2 u^2 c_j}{8p_j^2}\Bigr),\end{aligned} \tag{3.2}$$

where $c_j := \sum_{\alpha\in\mathcal{A}_j} (b_\alpha - a_\alpha)^2$. We choose $p_j = w_j c_j^{1/2}/T$ with $T := \sum_j w_j c_j^{1/2}$, and find

$$\mathbb{E} \exp(uX) \le \exp\bigl(\tfrac{1}{8} T^2 u^2\bigr), \qquad u \in \mathbb{R}.$$

Hence, for $u \ge 0$,

$$\mathbb{P}(X \ge t) \le e^{-ut}\, \mathbb{E}\, e^{uX} \le \exp\bigl(\tfrac{1}{8} T^2 u^2 - ut\bigr),$$

and the optimal choice $u := 4t/T^2$ yields

$$\mathbb{P}(X \geq t) \leq \exp\bigl(-2t^2/T^2\bigr). \tag{3.3}$$

By the Cauchy–Schwarz inequality and Lemma 3.1,

$$T^2 = \Bigl(\sum_j w_j c_j^{1/2}\Bigr)^2 \leq \sum_j w_j \cdot \sum_j w_j c_j = \sum_j w_j \cdot \sum_{\alpha\in\mathcal{A}} (b_\alpha - a_\alpha)^2$$

and the result follows from (3.3) by choosing $\{(\mathcal{A}_j, w_j)\}_j$ with $\sum_j w_j = \chi^*(\mathcal{A})$.

The result for $\mathbb{P}(X \leq \mathbb{E}\, X - t)$ follows by considering $-X$. $\square$

Remark 3.3. In some cases (when the Cauchy–Schwarz inequality above is far from sharp), (3.3) may be substantially better than the theorem.

We will derive Theorems 2.3 and 2.5 from the following more technical estimate.

Theorem 3.4. *Suppose that X is as in (1.1) with $Y_\alpha - \mathbb{E}\, Y_\alpha \leq b$ for some $b > 0$ and all $\alpha \in \mathcal{A}$. Suppose further that $\{(\mathcal{A}_j, w_j)\}_j$ is a proper fractional cover of $\mathcal{A}$. Let $S := \sum_{\alpha\in\mathcal{A}} \operatorname{Var} Y_\alpha$ and $\sigma_j^2 := \sum_{\alpha\in\mathcal{A}_j} \operatorname{Var} Y_\alpha$. Then, with $\varphi(x)$ as in (2.3),*

$$\mathbb{P}(X \geq \mathbb{E}\, X + t) \leq \exp\Bigl(-\frac{S}{b^2 W}\varphi\Bigl(\frac{btW}{US}\Bigr)\Bigr), \qquad t \geq 0, \tag{3.4}$$

where $W := \sum_j w_j$ and

$$U := \sum_j w_j \max\Bigl(1, \frac{\sigma_j W^{1/2}}{S^{1/2}}\Bigr). \tag{3.5}$$

Proof. This is similar to the proof of Theorem 2.1, but we use a (usually) sharper estimate than (3.1).

We may by Lemma 3.2 assume that the fractional cover $\{(\mathcal{A}_j, w_j)\}_j$ is exact; note that the modification in Lemma 3.2 preserves W and preserves or decreases U. We may further assume that $b = 1$ and, for each α, $\mathbb{E}\, Y_\alpha = 0$; the general case follows by replacing Y_α by $(Y_\alpha - \mathbb{E}\, Y_\alpha)/b$. We thus have $\mathbb{E}\, Y_\alpha = 0$ and $Y_\alpha \leq 1$ for all α. Hence, with $\sigma_\alpha^2 := \operatorname{Var} Y_\alpha$, by [3, p. 42], see also [14, Lemma 2],

$$\mathbb{E}\exp\bigl(hY_\alpha\bigr) \leq \frac{1}{1+\sigma_\alpha^2}e^{-h\sigma_\alpha^2} + \frac{\sigma_\alpha^2}{1+\sigma_\alpha^2}e^h, \qquad h \geq 0. \tag{3.6}$$

We rewrite this as

$$\mathbb{E}\exp\bigl(hY_\alpha\bigr) \leq \exp\bigl(f(\sigma_\alpha^2; h)\bigr), \qquad h \geq 0,$$

where

$$f(s; h) := \ln\Bigl(\frac{1}{1+s}e^{-hs} + \frac{s}{1+s}e^h\Bigr).$$

It is proved in [14, Lemma 3 and p. 24] that f is a concave function of $s \geq 0$ with $f(0) = 0$, and thus

$$f(s; h) \leq sg(h), \qquad s \geq 0,$$

where

$$g(h) := \frac{\partial}{\partial s} f(0;h) = e^h - 1 - h.$$

Consequently,

$$\mathbb{E}\exp\bigl(hY_\alpha\bigr) \le \exp\bigl(\sigma_\alpha^2 g(h)\bigr), \qquad h \ge 0. \tag{3.7}$$

As in (3.2), for any $u \ge 0$ and positive numbers p_j with $\sum_j p_j = 1$, now using (3.7),

$$\begin{aligned}\mathbb{E}\exp(uX) &\le \sum_j p_j \prod_{\alpha\in\mathcal{A}_j} \mathbb{E}\exp\Bigl(\frac{w_j u}{p_j} Y_\alpha\Bigr)\\ &\le \sum_j p_j \prod_{\alpha\in\mathcal{A}_j} \exp\Bigl(\sigma_\alpha^2 g\Bigl(\frac{w_j u}{p_j}\Bigr)\Bigr) = \sum_j p_j \exp\Bigl(\sigma_j^2 g\Bigl(\frac{w_j u}{p_j}\Bigr)\Bigr).\end{aligned} \tag{3.8}$$

Choose this time $p_j := w_j \max(1, \frac{\sigma_j W^{1/2}}{S^{1/2}})/U$, cf. (3.5). If $\sigma_j^2 \le S/W$, then $p_j = w_j/U$ and

$$\sigma_j^2 g\Bigl(\frac{uw_j}{p_j}\Bigr) \le \frac{S}{W} g(uU). \tag{3.9}$$

If $\sigma_j^2 \ge S/W$, then $p_j = w_j\sigma_j W^{1/2}/(S^{1/2}U)$ and

$$\sigma_j^2 g\Bigl(\frac{uw_j}{p_j}\Bigr) = \sigma_j^2 g\Bigl(\frac{uS^{1/2}U}{\sigma_j W^{1/2}}\Bigr) \le \sigma_j^2 \frac{S}{\sigma_j^2 W} g(uU),$$

because $g(0) = g'(0) = 0$ and g'' is increasing, and thus $x \mapsto g(x)/x^2$ is increasing. Hence, (3.9) holds in this case too, and (3.8) yields

$$\mathbb{E}\exp(uX) \le \exp\Bigl(\frac{S}{W} g(uU)\Bigr), \qquad u \ge 0. \tag{3.10}$$

Hence,

$$\mathbb{P}(X \ge \mathbb{E}\, X + t) \le \exp\Bigl(\frac{S}{W} g(uU) - ut\Bigr), \qquad u \ge 0. \tag{3.11}$$

We optimize by choosing $u = U^{-1}\ln\bigl(1 + tW/SU\bigr)$, which yields (3.4). □

To prove Theorems 2.3 and 2.5, it remains to estimate W and U for suitable proper fractional covers.

Proof of Theorem 2.3. By the inequality $y \le 1 + y^2/4$ and Lemma 3.1, we always have

$$U \le \sum_j w_j\Bigl(1 + \frac{\sigma_j^2 W}{4S}\Bigr) = \sum_j w_j + \sum_j w_j\sigma_j^2\frac{W}{4S} = W + S\frac{W}{4S} = \frac{5}{4}W.$$

The first inequality in Theorem 2.3 now follows from Theorem 3.4 if we choose an optimal proper fractional cover $\{(\mathcal{A}_j, w_j)\}_j$ such that $W = \chi^*(\mathcal{A})$.

The second inequality follows by the elementary inequality

$$\varphi(x) \ge \frac{x^2}{2(1 + x/3)}, \qquad x \ge 0, \tag{3.12}$$

see e.g. [3] or [16, proof of Theorem 2.1]. □

Proof of Corollary 2.4. For (2.4)–(2.5), apply Theorem 2.3 with $b = 1 - p$ and $S = Np(1-p)$.

For (2.6), replace Y_α by $1 - Y_\alpha$ and apply (2.4).

Finally, (2.7) is trivial for $t > \mathbb{E}\, X = Np$, and for $t \le Np$ it follows from (2.6) and (3.12), using $(1-p)(1 + 4t/15N(1-p)) \le 1 - p + t/N \le 1$. □

Proof of Theorem 2.5. For Theorem 2.5, we use instead the Hajnal–Szemerédi Theorem [12], [5], which says that $\mathcal{A}$ can be partitioned into Δ_1 independent sets, each of size equal to either $\lceil |\mathcal{A}|/\Delta_1 \rceil$ or $\lfloor |\mathcal{A}|/\Delta_1 \rfloor$. We use these sets as our $\mathcal{A}_j$, with all weights $w_j = 1$. If $N = k\Delta_1 + l$, with k and l integers and $0 \le l < \Delta_1$, there are necessarily $\Delta_1 - l$ sets of size k and l of size $k+1$, and thus $W = \Delta_1$ and, by (3.5),

$$\begin{aligned} U &= \Delta_1 - l + l\Big(\frac{(k+1)\Delta_1}{N}\Big)^{1/2} = \Delta_1 + l\bigg(\sqrt{1 + \frac{\Delta_1 - l}{N}} - 1\bigg) \\ &\le \Delta_1 + \frac{l(\Delta_1 - l)}{2N} \le \Delta_1\Big(1 + \frac{\Delta_1}{8N}\Big). \end{aligned}$$

The result follows from Theorem 3.4 and (3.12). □

Proof of Corollary 2.6. This follows from Theorem 2.5 by the arguments for Corollary 2.4. □

Remark 3.5. The idea to use the Hajnal–Szemerédi Theorem in this context is due to Ruciński [16, 18] and Pemmaraju [20] (independently). Pemmaraju [20] further explores the possibility of improving the bound by finding by other means, for specific Γ, a partition of Γ into fewer than Δ_1 independent sets of (almost) the same size. Note that such arguments can beat our general bound in Theorem 2.3 by at most a factor 25/16 in the exponent. Indeed, the proof of Theorem 2.3 uses the general estimate $U/W \le 5/4$ proved above, while Theorem 2.5 and [20] exploits better estimates of U/W in special situations; however, $U/W \ge 1$ by (3.5), so the room for improvement is limited.

4. Examples

Example 4.1 (U-statistics)**.** Consider a sum

$$X = \sum_{1 \le i_1 < \dots < i_d \le n} f_{i_1 \cdots i_d}(\xi_{i_1}, \dots, \xi_{i_d}) \tag{4.1}$$

where $\xi_1, \dots, \xi_n$ are independent random variables. This is the special case of (1.2) with $\mathcal{A} = [n]^d_<$, and includes U-statistics.

We define, as in Section 2, a dependency graph Γ with vertex set $[n]^d_<$ by declaring that α and β are joined by an edge if $\alpha \cap \beta \ne \emptyset$ (regarding the d-tuples α and β as sets).

Let $k := \lfloor n/d \rfloor$. Let $\mathcal{A}_1, \dots, \mathcal{A}_M$ be a listing of all families of k disjoint d-tuples in $\mathcal{A} = [n]^d_<$. By symmetry, $\sum_{j=1}^M \mathbf{1}_{\mathcal{A}_j} = Mk|\mathcal{A}|^{-1}\mathbf{1}_{\mathcal{A}}$, and thus we have $\sum_j w_j \mathbf{1}_{\mathcal{A}_j} = \mathbf{1}_{\mathcal{A}}$ with $w_j = |\mathcal{A}|/(Mk)$ for all j. Hence,

$$\chi^*(\mathcal{A}) \leq \chi^*(\Gamma) \leq \frac{|\mathcal{A}|}{k} = \frac{\binom{n}{d}}{\lfloor n/d \rfloor}. \tag{4.2}$$

Consequently, Theorem 2.1 shows that if $a \leq f_{i_1 \cdots i_d}(\xi_{i_1}, \dots, \xi_{i_d}) \leq b$ for every $i_1, \dots, i_d$ and two real numbers a and b, then, for $t > 0$,

$$\mathbb{P}\Bigl(X \geq \mathbb{E}\, X + t\binom{n}{d}\Bigr) \leq \exp\Bigl(-2\left\lfloor \frac{n}{d} \right\rfloor t^2/(b-a)^2\Bigr). \tag{4.3}$$

This estimate is due to Hoeffding [14, (5.7)]. (Hoeffding treated U-statistics only; his proof extends to the more general non-symmetric version here.) We have the same estimate for $\mathbb{P}\Bigl(X \leq \mathbb{E}\, X - t\binom{n}{d}\Bigr)$.

Since $d\lfloor n/d \rfloor \geq n - d + 1$, we obtain from (4.2) also

$$\chi^*(\mathcal{A}) \leq \binom{n}{d-1} \tag{4.4}$$

which leads to a slight weakening of (4.3) that may look simpler:

$$\mathbb{P}\bigl(X \geq \mathbb{E}\, X + tn^{d-1/2}\bigr) \leq \exp\Bigl(-2\frac{d!\,(d-1)!}{(b-a)^2}t^2\Bigr). \tag{4.5}$$

In many cases, Theorem 2.3 or Corollary 2.4 yields better results. For example, if $\xi_1, \dots, \xi_n$ are identically distributed and $f_{i_1 \cdots i_d} = f$ does not depend on $i_1, \dots, i_d$ and take the values 0 and 1 only, then Corollary 2.4 yields, with $p := \mathbb{P}\bigl(f(\xi_1, \dots, \xi_d) = 1\bigr)$,

$$\mathbb{P}(X \geq \mathbb{E}\, X + t) \leq \exp\Bigl(-\frac{8t^2}{25\binom{n}{d-1}\bigl(\binom{n}{d}p + t/3\bigr)}\Bigr); \tag{4.6}$$

$$\mathbb{P}(X \leq \mathbb{E}\, X - t) \leq \exp\Bigl(-\frac{8t^2}{25\binom{n}{d-1}\binom{n}{d}p}\Bigr). \tag{4.7}$$

These are better than (4.3) and (4.5) when $p < 0.16$ and (for (4.6)) t is not too large.

In this case, Theorem 2.5 and its corollary are somewhat weaker. We have

$$\Delta(\Gamma) \leq d\binom{n-1}{d-1},$$

and this estimate is asymptotically sharp. Together with (4.4), we thus see that we lose about a factor d in the exponent by using Δ_1 instead of $\chi^*(\mathcal{A})$, which outweighs the gain of a factor 25/16 in Theorem 2.5 compared to Theorem 2.3.

If we have a sparse sum as in (1.2), for example an incomplete U-statistic, we can get better estimates provided we have a reasonable estimate of $\chi^*(\mathcal{A})$ or $\Delta(\mathcal{A})$.

Some other large deviation bounds for U-statistics are given in [1], [2], [7].

Example 4.2 (Patterns in a random string)**.** Let $\xi_1, \dots, \xi_n$ be i.i.d. random letters from a finite alphabet A, and let w be a given word of length d. Flajolet et al. [11] studied the number of subsequences $\xi_{i_1} \cdots \xi_{i_d}$ that equal w. This is a random variable of the form (4.1), with $f(x_1, \dots, x_d) = 1$ if $x_1 \cdots x_d = w$ and 0 otherwise. Hence all bounds in Example 4.1 apply, with $b - a = 1$ and $p = \mathbb{P}(\xi_1 \cdots \xi_d = w) = \prod_{i=1}^d p_{w_i}$, where w_i is the i:th letter of w and $p_x := \mathbb{P}(\xi_1 = x)$. In particular, (4.5) (Theorem 2.1) yields the bound

$$\mathbb{P}(|X - \mathbb{E}\, X| \geq tn^{d-1/2}) < 2\exp\bigl(-2d!(d-1)!\, t^2\bigr).$$

For p small, better results are given by (4.6) and (4.7) (Corollary 2.4).

Flajolet et al. [11] further studied the number of such subsequences with the constraints $i_{j+1} - i_j \leq \ell_j$ on the gaps of the sequence, where $\ell_1, \dots, \ell_{d-1}$ are given numbers (possibly ∞). This is a random variable of the type (1.2), so the results above apply. We now have

$$\mathcal{A} = \{(i_1, \dots, i_d) \in [n]^d_< : 0 < i_{j+1} - i_j \leq \ell_j \text{ for } j = 1, \dots, d-1\}.$$

Let $\mathcal{F} := \{i : \ell_i < \infty\}$ and set $b := d - |\mathcal{F}| = 1 + |\{i : \ell_i = \infty\}|$ and $D := \prod_{i \in \mathcal{F}} \ell_i$. Consider ℓ_j as fixed and let $n \to \infty$. Then $|\mathcal{A}| = \binom{n}{b} D(1 + O(1/n))$, see [11]. Moreover, given any i_k, the gaps $i_{j+1} - i_j$ may be chosen in at most D ways for $j \in \mathcal{F}$ and in at most $\binom{n}{b-1}$ ways for $j \notin \mathcal{F}$. Since any given d-tuple $\alpha \in \mathcal{A}$ has d indices, and any of these may be in (at most) d positions in another d-tuple, it follows that, with the dependency graph Γ defined as in Example 4.1,

$$\Delta_1(\Gamma) \leq d^2 D \binom{n}{b-1} \leq \frac{d^2 D}{(b-1)!} n^{b-1}. \tag{4.8}$$

We thus obtain from Corollary 2.6, for all $t > 0$,

$$\mathbb{P}(X \geq \mathbb{E}\, X + t) \leq \exp\Bigl(-\frac{(b-1)!\, b!}{2d^2 D^2 p} \cdot \frac{t^2}{n^{2b-1}} \Bigl(1 + O\Bigl(\frac{1}{n} + \frac{t}{n^b p}\Bigr)\Bigr)\Bigr)$$

$$\mathbb{P}(X \leq \mathbb{E}\, X - t) \leq \exp\Bigl(-\frac{(b-1)!\, b!}{2d^2 D^2 p} \cdot \frac{t^2}{n^{2b-1}} \Bigl(1 + O\Bigl(\frac{1}{n}\Bigr)\Bigr)\Bigr).$$

Since $\operatorname{Var}(X)$ is of the order n^{2b-1} [11], these estimates are of the subgaussian type $\exp\bigl(-ct^2/\operatorname{Var}(X)\bigr)$, at least for t not too large.

Example 4.3 (Random graphs)**.** The random graph $G(n,p)$ consists of n (labelled) vertices and edges drawn randomly, with each possible edge appearing with probability p, independently of all other edges.

Let X be the number of triangles in $G(n,p)$. This can be written in the form (1.1), where $\mathcal{A} = [n]^3_<$ and Y_α is the indicator that the three edges between the three vertices in α all are present in $G(n,p)$. (Note that this is not of the form (1.2), since the underlying independent variables are indexed by pairs ij of vertices rather than vertices. X can be regarded as an incomplete U-statistic based on the $\binom{n}{2}$ edge indicators; this will lead to the same results.)

In this case, two variables Y_α and Y_β are independent even if α and β have one vertex (but not two) in common, and we can define an independence graph Γ with vertex set $\mathcal{A}$ such that α and β are joined by an edge if $|\alpha \cap \beta| = 2$. This means fewer edges that in the independence graph in Example 4.1 (for $d = 3$), and thus smaller Δ_1 and χ^*, and better estimates.

We have $\mathbb{E}\, Y_\alpha = \mathbb{P}(Y_\alpha = 1) = p^3$, $|\mathcal{A}| = \binom{n}{3}$ and $\Delta(\Gamma) = 3(n-3)$, whence $\Delta_1(\Gamma) = 3n - 8 < 3n$. It follows from Corollary 2.6 that, for $\varepsilon > 0$,

$$\mathbb{P}(X \geq (1+\varepsilon)\,\mathbb{E}\, X) \leq \exp\Bigl(-\frac{\varepsilon^2}{36+12\varepsilon} n^2 p^3 \bigl(1 + O(1/n)\bigr)\Bigr) \tag{4.9}$$

$$\mathbb{P}(X \leq (1-\varepsilon)\,\mathbb{E}\, X) \leq \exp\Bigl(-\frac{\varepsilon^2}{36} n^2 p^3 \bigl(1 + O(1/n)\bigr)\Bigr). \tag{4.10}$$

If p is not too small, these exponential bounds are rather good. However, for $p < n^{-2/3}$, they are pretty useless. Also for larger p, they are not sharp (unless p is bounded away from 0). Indeed, for this problem the following is known by other methods, assuming for simplicity that $n \to \infty$ and that $p \to 0$ and $np \geq 1$.

As is shown in [15], [16, Chapter 3], the sharp exponent in the lower tail estimate (4.10) is, for any fixed $\varepsilon \leq 1$, really of the order $\min(n^3p^3, n^2p)$.

For the upper tail (4.9), the sharp exponent is of the order n^2p^2 (possibly up to a factor $\log n$) for any fixed $\varepsilon \geq 0$, see [17]. See also partial results and discussions in [22] and [18]; the present method is essential the same as "Break-up" in the latter. (It is a slight improvement, see Remark 5.2.)

We have here for simplicity discussed the number of triangles. Similar results hold for the number of copies of any fixed graph in $G(n,p)$. Again, Corollary 2.6 gives estimates that are exponentially small for certain ranges of p, but we do not obtain optimal results (except in some extreme cases). Cf. [15], [16], [22], [18], [17].

Example 4.4 (Random hypergraphs)**.** Another problem that was used in [18] to compare several different methods was the following: Consider a fixed hypergraph H, for simplicity assumed to be uniform (all hyperedges have the same number of vertices). Delete vertices of H at random, keeping each vertex with probability p, independently of the other vertex, and let X be the number of surviving hyperedges.

As shown in [18] ("Break-up" in Section 3.2), we typically obtain reasonable estimates by the methods of this paper. In some cases, these are the best that we can prove, but in others they are surpassed by other methods.

5. Further remarks

Remark 5.1. The method presented here uses independence of suitable (large) subsets of $\{Y_\alpha\}$, but it does not use any other information on the dependencies. Hence it can be expected to be wasteful and not give optimal results when the dependencies that exist are weak.

Remark 5.2. Several authors, e.g. [21], [16], [18], [20], have used a splitting of (1.1) into sums with independent summands in a slightly different way. In the notation above, the argument runs as follows.

Consider for simplicity Theorem 2.1. Let $\{(\mathcal{A}_j, w_j)\}_j$ be an exact proper fractional cover of $\mathcal{A}$ (usually, only covers are considered), and let again $X_j :=$

$\sum_{\mathcal{A}_j} Y_\alpha$. For any positive p_j with $\sum_j p_j = 1$,

$$\mathbb{P}(X \geq \mathbb{E}\,X+t) = \mathbb{P}\Bigl(\sum_j w_j(X_j-\mathbb{E}\,X_j) \geq \sum_j p_j t\Bigr) \leq \sum_j \mathbb{P}\Bigl(X_j-\mathbb{E}\,X_j \geq \frac{p_j t}{w_j}\Bigr).$$

Since each X_j is a sum of independent variables, the latter probabilities can be estimated by [14, Theorem 2], i.e. our Theorem 2.1 with $\chi^*(\mathcal{A}) = 1$, yielding

$$\mathbb{P}(X \geq \mathbb{E}\,X + t) \leq \sum_j \exp\Bigl(-2\frac{p_j^2}{w_j^2 c_j} t^2\Bigr),$$

where, as in our proof of Theorem 2.1, $c_j := \sum_{\alpha\in\mathcal{A}_j}(b_\alpha - a_\alpha)^2$. Taking again $p_j = w_j c_j^{1/2}/T$ with $T := \sum_j w_j c_j^{1/2}$, we obtain

$$\mathbb{P}(X \geq \mathbb{E}\,X + t) \leq N_1 \exp\bigl(-2t^2/T^2\bigr),$$

where N_1 is the number of sets $\mathcal{A}_j$ in the fractional cover. (For an optimal cover, thus $N_1 = \chi(\mathcal{A})$.) This is the same as (3.3) except for the factor N_1; hence it leads to the estimate in Theorem 2.1 multiplied by N_1 (for an optimal fractional cover).

Hence Hoeffding's method above, using Jensen's inequality, is better, although the difference usually is negligible in applications.

This version of the splitting argument also applies to the other theorems above. Again, the results above are typically better by an insignificant factor, but exact comparison seems more difficult than for Theorem 2.1.

Remark 5.3. The definition in Section 2 of dependency graphs is not the only one in the literature. In some cases one needs a stronger version, requiring $\{Y_\alpha\}_{\alpha\in\mathcal{A}_1}$ and $\{Y_\alpha\}_{\alpha\in\mathcal{A}_2}$ to be independent of each other whenever $\mathcal{A}_1$ and $\mathcal{A}_2$ are two disjoint subsets of $\mathcal{A}$ with no edges between them. (See e.g. [16].)

In the opposite direction, an even weaker version, still sufficient for our purposes, would be to require only that $\{Y_\alpha\}_{\alpha\in\mathcal{A}_1}$ is a family of independent variables when $\mathcal{A}_1$ is an independent set in the graph. To see that this is strictly weaker than the definition used above, consider the three variables $\xi_1\xi_2$, $\xi_1\xi_3$ and $\xi_2\xi_3$, where ξ_1, ξ_2 and ξ_3 are i.i.d. with $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1/2$. Then Γ is an independence graph in our sense if and only if Γ has at least two edges, while one edge suffices for the weaker definition.

Note further that $|\mathcal{A}| = 3$ and that every proper subset of $\mathcal{A}$ is independent; hence no independence graph describes all independent sets. It is easily checked that $\chi^*(\mathcal{A}) = 3/2$, but $\chi^*(\Gamma) \geq 2$ for every independence graph (for any of the definitions above).

References

[1] A.K. Aleshkyavichene, Probabilities of large deviations for U-statistics and von Mises functionals. (Russian) *Teor. Veroyatnost. i Primenen.* **35** (1990), no. 1, 3–14; English transl. *Theory Probab. Appl.* **35** (1990), no. 1, 1–14 (1991).

[2] M.A. Arcones & E. Giné, Limit theorems for U-processes. *Ann. Probab.* **21** (1993), no. 3, 1494–1542.

[3] G. Bennett, Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** (1962), 33–45.
[4] S. Bernstein, On a modification of Chebyshev's inequality and of the error formula of Laplace. (Russian) *Ann. Sci. Inst. Sav. Ukraine, Sect. Math.* **1** (1924), 38–49.
[5] B. Bollobás, *Extremal Graph Theory.* Academic Press, London, 1978.
[6] B. Bollobás, *Modern Graph Theory.* Springer-Verlag, New York, 1998.
[7] J. Bretagnolle, A new large deviation inequality for U-statistics of order 2. *ESAIM Probab. Statist.* **3** (1999), 151–162.
[8] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics* **23** (1952), 493–507.
[9] H. Cramér, Sur un noveau théorème-limite de la théorie des probabilités. *Les sommes et les fonctions de variables aléatoires*, Actualités Sci. Indust. 736, Hermann et Cie., Paris, 1938, pp. 5–23.
[10] W. Feller, Generalization of a probability limit theorem of Cramér. *Trans. Amer. Math. Soc.* **54** (1943), 361–372.
[11] P. Flajolet, Y. Guivarc'h, W. Szpankowski & B. Vallée, Hidden pattern statistics, *Automata, Languages, and Programming (ICALP'2001)*, eds. F. Orejas et al., Lecture Notes Comp. Sci. 2076, Springer, 2001, pp. 152–165.
[12] A. Hajnal & E. Szemerédi, Proof of a conjecture of Erdős. *Combinatorial Theory and its Applications, Vol. II*, eds. P. Erdős, A. Rényi & V.T. Sós, Colloq. Math. Soc. János Bolyai 4, North-Holland, Amsterdam, 1970, pp. 601–623.
[13] W. Hoeffding, A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** (1948), 293–325.
[14] W. Hoeffding, Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** (1963), 13–30.
[15] S. Janson, T. Łuczak & A. Ruciński, An exponential bound for the probability of nonexistence of a specified subgraph in a random graph. *Random Graphs '87 (Proceedings, Poznań 1987)*, eds. M. Karoński, J. Jaworski & A. Ruciński, John Wiley & Sons, Chichester, 1990, pp. 73–87.
[16] S. Janson, T. Łuczak & A. Ruciński, *Random Graphs*, John Wiley & Sons, New York, 2000.
[17] S. Janson, K. Oleszkiewicz & A. Ruciński, Upper tails for subgraph counts in random graphs. *Israel J. Math.*, to appear.
[18] S. Janson & A. Ruciński, The infamous upper tail. *Random Struct. Alg.* **20** (2002), no. 3, 317-342.
[19] M. Okamoto, Some inequalities relating to the partial sum of binomial probabilities. *Ann. Inst. Statist. Math.* **10** (1958), 29–35.
[20] S.V. Pemmaraju, Equitable coloring extends Chernoff–Hoeffding bounds. *Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques (APPROX 2001 and RANDOM 2001)*, eds. M.X. Goemans et al, Lecture Notes Comp. Sci. 2129, Springer, 2001, pp. 285–296.
[21] V. Rödl & A. Ruciński, Random graphs with monochromatic triangles in every edge coloring. *Random Struct. Alg.* **5** (1994), no. 2, 253–270.
[22] V.H. Vu, A large deviation result on the number of small subgraphs of a random graph. *Combin. Probab. Comput.* **10** (2001), no. 1, 79–94.

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden

E-mail address: `svante.janson@math.uu.se`
URL: `http://www.math.uu.se/~svante/`