Chernoff–Hoeffding Bounds for Applications with Limited Independence*

Jeanette P. Schmidt†
Computer Science Department
Polytechnic University
Brooklyn, NY 11201

Alan Siegel‡
Computer Science Department
Courant Institute, New York University
New York, NY 10012

Aravind Srinivasan§
Department of Computer Science
Cornell University
Ithaca, NY 14853

Abstract

Chernoff–Hoeffding bounds are fundamental tools used in bounding the tail probabilities of the sums of bounded and independent random variables. We present a simple technique which gives slightly better bounds than these, and which more importantly requires only *limited independence* among the random variables, thereby importing a variety of standard results to the case of limited independence for free. Additional methods are also presented, and the aggregate results are sharp and provide a better understanding of the proof techniques behind these bounds. They also yield improved bounds for various tail probability distributions and enable improved approximation algorithms for jobshop scheduling. The "limited independence" result implies that a reduced amount of randomness and weaker sources of randomness are sufficient for randomized algorithms whose analyses use the Chernoff–Hoeffding bounds, *e.g.,* the analysis of randomized algorithms for random sampling and oblivious packet routing.

1 Introduction

The most fundamental tools used in bounding the tail probabilities of the sums of bounded and independent random variables, are based on techniques initiated by Chernoff [11] and generalized by Hoeffding [17] more than thirty years ago. They are frequently used in the design and analysis of randomized algorithms, derandomization, and in the probabilistic method. We present a simple method which generalizes, somewhat, the classical method for proving the Chernoff-Hoeffding bounds, in the case of bounded random variables confined to the interval $[0, 1]$. More importantly, this approach requires only *limited independence* among the random variables, and thereby imports a variety of standard results to the case of limited independence for free. This and related bounds

*A preliminary version of this work appeared in the *ACM-SIAM Symposium on Discrete Algorithms*, 1993.

†Partially supported by NSF grant CCR-9110255 and CCR-9305873, and the New York State Science and Technology Foundation through its Center for Advanced Technology.

‡Partially supported by NSF grants CCR-8906949 and CCR-8902221.

§Partially supported by the U. S. Army Research Office through the Mathematical Sciences Institute of Cornell University, and by NSF PYI award CCR–89–96272 with matching support from UPS and Sun Microsystems. Current address: DIMACS Center, Rutgers University, Piscataway, NJ 08855 and School of Mathematics, The Institute for Advanced Study, Princeton, NJ 08540.

lead to a variety of applications ranging from improved bounds for tail probability distributions to new algorithmic results.

The "limited independence" result implies that sources of randomness that are weaker than the standard model of unbiased and independent bits, are sufficient for any algorithm whose analysis uses the Chernoff–Hoeffding bounds. It also provides a better understanding of the proof techniques behind these bounds, and gives improved bounds for various tail probability distributions. Via standard techniques, it leads to a simple analysis of algorithms for such classical problems as random sampling. The formulation also leads to approximation algorithms with better approximation guarantees for certain problems.

Given n random variables $X_1, X_2, \ldots, X_n$, suppose we want to upper bound the "upper tail" probability $Pr(X \geq a)$, where $X \doteq \sum_{i=1}^{n} X_i$, $\mu \doteq E[X]$, $a = \mu(1+\delta)$ and $\delta > 0$. The classical idea behind the Chernoff–Hoeffding bounds (see, for instance, Chernoff [11], Hoeffding [17], Raghavan [35] and Alon, Spencer, & Erdős [3]) is as follows. For any fixed $t > 0$, $Pr(X \geq a) = Pr(e^{tX} \geq e^{at}) \leq \frac{E[e^{tX}]}{e^{at}}$, by Markov's inequality. Computing an upper bound $u(t)$ on $E[e^{tX}]$ and minimizing $\frac{u(t)}{e^{at}}$ over $t > 0$ gives an upper bound for $Pr(X \geq a)$.

An important situation in computation is the one in which $X_i \in \{0, 1\}$, $i = 1, 2, \ldots, n$. For this case, we construct a class of functions of X that is as easy to analyze, and which includes the class $\{e^{tX} : t > 0\}$ and do the above minimization over this class. In the process, we discover that $X_1, X_2, \ldots, X_n$ `need only be` $h(n, \mu, \delta)$`-wise independent` for a suitably defined function $h(\cdot,\cdot,\cdot)$, which is typically much less than n for many algorithms; recall that a set of random variables V exhibit k–wise independence if any subset of k or fewer random variables from V are jointly independent, which is to say that their joint probability distribution function is just the product of the individual distributions. One reason for the use of the e^{tX} function in the classical methods is that $E[e^{tX}]$ generates all higher moments of X; using only a constant number of higher moments, for instance, gives weak bounds. However, in the binary case, the first n moments are sufficient to generate *all* higher moments, which motivates our method. Interestingly, this formulation also can be applied to general X_i that take arbitrary values in the interval $[0, 1]$, even though it is *not* true that the first n moments of $X = \sum_{i=1}^{n} X_i$ determine all higher ones.

The results have many applications to tail probability distributions. They imply similar "limited independence" results when $X_1, X_2, \ldots, X_n$ take values in the interval $[0, 1]$; this can be extended to bounded random variables, by scaling their ranges to $[0, 1]$. In the case of the hypergeometric distribution (sampling without replacement), it provides an elementary mechanism to attain slightly better bounds than those implied in [17] and by Chvátal [13]. The method also yields good upper bounds for the tail probabilities of the sums of random variables with limited independence.

These constructions also provide pointers to further improvement of the independence bounds. For example, we will take the liberty of redirecting, somewhat, the estimation method as appropriate, when attaining improved tools for analyzing the behavior of the sum of k–wise independent random variables. The results simplify and sharpen some of the analyses done in [39] and [40]. In particular, we derive good upper bounds on $E[((\sum_{i=1}^{n} X_i) - E[\sum_{i=1}^{n} X_i])^k]$, where $X_1, X_2, \ldots, X_n$ are k–wise independent random variables, each of which lies in the interval $[0, 1]$; this leads to better independence bounds than our $h(n, \mu, \delta)$ when $\delta < 1$. We also prove good bounds on the probability of *exactly* r successes in a sequence of k–wise independent Bernoulli trials, which shows that even with modest independence, probabilities and conditional probabilities are close to the fully independent case, in situations like hashing.

The sufficiency of limited independence has several computational applications. First, it means that any random process whose analysis uses the Chernoff–Hoeffding bounds can be simulated with a weaker random source than one which outputs unbiased and independent bits. Next, via known constructions of random variables with limited independence using fewer random bits (Joffe [19], Carter & Wegman [9], Mehlhorn & Vishkin [26], Alon, Babai & Itai [1], Siegel [43]), we can reduce the randomness required for certain algorithms. One simple example is that of `random sampling`: given a universe U and a subset $X \subseteq U$, the problem is to estimate the fraction of objects of type X in U, such that the absolute error of the output is at most δ with probability at least $1-\epsilon$, for given error parameters δ and ϵ. The new constructions imply that if R independent samples are required to yield the desired bound, then it in fact suffices for those R samples to be k^*-wise independent, for $k^* = O(\log(\frac{1}{\epsilon}))$. These samples can be generated by $O(\log(\frac{1}{\epsilon}))$ random samples from U, using standard methods. Note that the above construction is not optimal with regards to the number of random bits used (see Bellare, Goldreich & Goldwasser [6] for an optimal construction), but is extremely simple. It is also easily parallelizable, while it is not known how to parallelize other schemes for reducing randomness, *e.g.*, random walks on expanders. It has come to our attention that via weaker bounds on the kth moment, essentially the same bounds for the random sampling problem have been obtained by Bellare & Rompel [7]. We believe that there should be additional applications yielding reduced randomness. A spectrum of explicit constructions of oblivious routing algorithms on the butterfly with varying time–randomness parameters is among the results of Peleg & Upfal [32]; our "limited independence" result directly matches these bounds on the hypercube and, we believe, should extend to other interconnection networks.

Finally, we combine the method of conditional probabilities [34, 44] with the new construction to obtain two results. We get a much faster implementation of the sequential jobshop scheduling algorithm of Shmoys, Stein & Wein [41]. It is comparable in time complexity to the speedups due to Plotkin, Shmoys & Tardos [33] and Stein [45] but importantly, the approximation bound it presents is better than the ones of [33] and [45]. Here, we show that a problem can be derandomized directly, thereby avoiding the bottleneck step of solving a huge linear program. We also prove an "exact partition" result for set discrepancy, and derive a polynomial–time algorithm for it.

The organization of the paper is as follows. Section 2 presents the new formulation and its applications to tail probabilities, and Section 3 presents applications of these results to computation.

2 The Basic Method, and Applications to Tail Probabilities

In this section, we introduce the method, discuss its implications to the tail probabilities of various distributions, and analyze some related approaches. We also prove probability bounds for exactly r successes in a sequence of Bernoulli trials under limited independence. As discussed in Section 1, the basic idea used in the Chernoff–Hoeffding (henceforth CH) bounds is as follows. Given n random variables (henceforth "r.v."s) $X_1, X_2, \ldots, X_n$, we want to upper bound the upper tail probability

$Pr(X \geq a)$, where $X \doteq \sum_{i=1}^{n} X_i$, $\mu \doteq E[X]$, $a = \mu(1+\delta)$ and $\delta > 0$. For any fixed $t > 0$,

$$Pr(X \geq a) = Pr(e^{tX} \geq e^{at}) \leq \frac{E[e^{tX}]}{e^{at}};$$

by computing an upper bound $u(t)$ on $E[e^{tX}]$ and minimizing $\frac{u(t)}{e^{at}}$ over $t > 0$, we can upper bound $Pr(X \geq a)$. When $X_1, X_2, \ldots, X_n$ are binary, we construct a class of functions of X that includes the class $\{e^{tX} : t > 0\}$ and do the minimization over this class; in the process, we discover that $X_1, X_2, \ldots, X_n$ `need only be` $h(n, \mu, \delta)$`-wise independent` for a function $h(\cdot,\cdot,\cdot)$ that will be defined in equation 3 of the next section.
Notation. If x is real and r is a positive integer, then $\binom{x}{r}$ will denote, as usual, $\frac{x(x-1)\cdots(x-r+1)}{r!}$ with $\binom{x}{0} \doteq 1$.

2.1 Estimating tail probabilities of binary random variables

The CH bounds are frequently used when the r.v.'s $X_1, X_2, \ldots X_n$ are binary and `independent`. In this section, we first assume that $X_1, X_2, \ldots X_n$ are 0–1 independent r.v.'s with $Pr(X_i = 1) = p_i$, $1 \leq i \leq n$; `the independence assumption will be relaxed later, and the results will be extended to r.v.'s` X_i `with` $0 \leq X_i \leq 1$`, in Section 2.2.` Let $X \doteq \sum_{i=1}^{n} X_i$, and $\mu \doteq E[X] = \sum_{i=1}^{n} p_i$. We want good upper bounds on $Pr(X \geq \mu(1+\delta))$, for $\delta > 0$. Chernoff [11] implicitly showed that for identically distributed 0–1 variables $X_1, X_2, \ldots X_n$ and for $a > \mu$,

$$min_t \frac{E[e^{tX}]}{e^{at}} \leq L(n, \mu, a) = (\frac{\mu}{a})^a (\frac{n-\mu}{n-a})^{n-a}.$$

Hoeffding [17] extended this by showing that $L(n, \mu, a)$ is an upper bound for the above minimum even if the X_i's are not identically distributed and range between 0 and 1. Replacing a with $\mu(1+\delta)$ in the Hoeffding estimate $L(\cdot,\cdot,\cdot)$ gives, for $\delta > 0$,

$$Pr(X \geq \mu(1+\delta)) \leq F(n, \mu, \delta) \doteq \frac{(1 + \frac{\mu\delta}{(n-\mu(1+\delta))})^{n-\mu(1+\delta)}}{(1+\delta)^{\mu(1+\delta)}}.$$

Since $L(n, \mu, a)$ is symmetric with respect to (a, μ) and $(n-a, n-\mu)$, the Hoeffding estimate also shows that

$$Pr(X \leq \mu(1-\delta)) = Pr(n - X \geq n - \mu(1-\delta)) \leq F(n, \mu, -\delta) \doteq \frac{(1 - \frac{\mu\delta}{(n-\mu(1-\delta))})^{n-\mu(1-\delta)}}{(1-\delta)^{\mu(1-\delta)}}.$$

The following simple upper bounds for $F(n, \mu, \delta)$ and $F(n, \mu, -\delta)$ are sufficient to derive most of the useful approximations that have appeared in the literature [17, 4, 35, 3].

$$F(n, \mu, \delta) \leq G(\mu, \delta) \doteq (\frac{e^\delta}{(1+\delta)^{1+\delta}})^\mu \text{ (see, for example, [35]);}$$

$$\text{for } \delta < 1, \quad G(\mu, \delta) \leq e^{-\delta^2\mu/3} \text{ [4];} \qquad \text{for } \delta > 2e - 1, \quad G(\mu, \delta) \leq 2^{-(1+\delta)\mu} \text{ [35], and}$$

$$F(n, \mu, -\delta) \leq G(\mu, -\delta) \leq e^{-\mu\delta^2/2} \quad [17,\ 4,\ 35,\ 3].$$

At the heart of these estimates are the simple calculations associated with the multiplicative nature of $E[e^{\Sigma X_i}]$. Recall that $e^{tX} = \sum_{i=0}^{\infty} \frac{t^i}{i!} X^i$. Consider X^2, for instance. $X^2 = (X_1 + X_2 + \cdots +$

$X_n)^2 = \sum_{i=1}^{n} X_i^2 + 2\sum_{1\leq i_1<i_2\leq n} X_iX_j = \sum_{i=1}^{n} X_i + 2\sum_{1\leq i_1<i_2\leq n} X_iX_j$, since $X_i^2 = X_i$ for $X_i \in \{0,1\}$. Similarly, other higher powers of the X_i's are unnecessary, implying that a form simpler and more useful than functions of the form $\{e^{tX} : t > 0\}$ might exist. There are many ways to formalize this. We define, for $z = (z_1, z_2, \ldots, z_n) \in \Re^n$, a family of symmetric multilinear polynomials $S_j(z), j = 0, 1, \ldots, n$, where $S_0(z) \equiv 1$, and for $1 \leq j \leq n$, $S_j(z) \doteq \sum_{1\leq i_1<i_2\cdots<i_j\leq n} z_{i_1}z_{i_2}\cdots z_{i_j}$. We start with the simple

Lemma 1 *Suppose $z_1, z_2, \ldots, z_n$ take on binary values. Then for any positive integer j, $(z_1 + z_2 \cdots + z_n)^j \equiv \sum_{i=1}^{min(j,n)} a_iS_i(z_1, z_2, \ldots, z_n)$, for some* `non-negative` *integers $a_1, a_2, \ldots, a_{min(j,n)}$.*

The proof of Lemma 1 is trivial and is omitted. The converse of Lemma 1 also is true; if $z = (z_1, z_2, \ldots, z_n) \in \{0,1\}^n$, then for any j, $j = 0, 1, \ldots, n$,

$$\forall(u_0, \ldots, u_j) \in \Re^{j+1}\ \exists(v_0, \ldots, v_j) \in \Re^{j+1} : \sum_{i=0}^{j} u_iS_i(z) \equiv \sum_{i=0}^{j} v_i(z_1 + z_2 + \cdots + z_n)^i.$$

So, the two forms: polynomial of $z_1 + z_2 + \cdots + z_n$ and linear combination of $S_0(z), S_1(z), \ldots, S_n(z)$ are equivalent. Note that if the binary random variables $X_1, X_2, \ldots, X_n$ are independent, then $E[S_i(X_1, X_2, \ldots, X_n)]$ is explicitly available: $E[S_i(X_1, X_2, \ldots, X_n)] = S_i(p_1, p_2, \ldots, p_n)$, where $p_j = Pr(X_j = 1)$. This explains our preference for the S_i's.

Since the expansion $e^{tZ} = \sum_{i=0}^{\infty} \frac{t^i}{i!}Z^i$ converges for all t and Z and since all the coefficients $\frac{t^i}{i!}$ are positive if $t > 0$, we get

Corollary 1 *Let $z_1, z_2, \ldots, z_n$ take on binary values. Then for any $t > 0$, there exist* `non-negative` *reals $a_0, a_1, \ldots, a_n$ such that $e^{t(z_1+z_2+\cdots z_n)} \equiv \sum_{i=0}^{n} a_iS_i(z_1, z_2, \ldots, z_n)$.*

One reason for the use of the function e^{tX} in the CH bounds is the need for higher moments of X. In particular, the `moment generating function` of X is defined to be $E[e^{tX}] = \sum_{i=0}^{\infty} \frac{t^i}{i!}E[X^i]$; its derivatives generate all higher moments of X. Moreover, the use of moment generating functions embed the problem of attaining probability estimates in a space rich with algebraic structure and convex inequalities. (More about the computational aspects of such an alternative approach can be found in [42].) The need for higher moments is due to the fact that a direct application of Markov's or Chebyshev's inequality to upper bound $Pr(X \geq E[X] \cdot (1 + \delta))$ leads to weak bounds. Higher moments and exponentials give dramatically better estimates. But when X is the sum of random bits $X_1, X_2, \ldots, X_n$, Lemma 1 and Corollary 1 imply that *all* the higher moments of X can be linearly generated by $\{E[S_i(X_1, X_2, \ldots, X_n)] : i = 0, 1, \ldots, n\}$. Equivalently, they are also generated linearly by *any* n higher moments of X.

So, we now consider functions of the form $\sum_{i=0}^{n} y_iS_i(X_1, X_2, \ldots, X_n)$ where $y_0, y_1, \ldots, y_n \geq 0$, instead of restricting ourselves to those of the form e^{tX}, for some $t > 0$. Indeed, by Corollary 1, we will be considering a class of functions which includes the class $\{e^{tX} : t > 0\}$. For any $y = (y_0, y_1, \ldots, y_n) \in \Re_+^{n+1}$ and $z = (z_1, z_2, \ldots z_n) \in \Re^n$, define $f_y(z) \doteq \sum_{i=0}^{n} y_iS_i(z)$. With this notation, we can restate Corollary 1 as $\forall t > 0\ \exists y \in \Re_+^{n+1} : f_y(X_1, X_2, \ldots, X_n) = e^{tX}$. Let $a \doteq \mu(1 + \delta)$ be assumed to be integral. Note that for any non–negative integer m, $X = m$ iff

$f_y(X_1, X_2, \ldots, X_n) = \sum_{i=0}^{m} y_i \binom{m}{i}$ and hence,

$$Pr(X \geq a) = Pr\left(f_y(X_1, \ldots, X_n) \geq \sum_{i=0}^{a} y_i \binom{a}{i}\right) \leq \frac{E[f_y(X_1, \ldots, X_n)]}{\sum_{i=0}^{a} y_i \binom{a}{i}} = \frac{\sum_{i=0}^{n} y_i S_i(p_1, p_2, \ldots, p_n)}{\sum_{i=0}^{a} y_i \binom{a}{i}}. \tag{1}$$

So, our goal now is to minimize this upper bound over $(y_0, y_1, \ldots y_n) \in \Re_+^{n+1}$. To do this, note that $y_{a+1}, y_{a+2}, \ldots, y_n$ must all be set to 0 since they contribute non–negative terms to the numerator and nothing to the denominator. Next, note that the r.h.s. of inequality(1) is minimized by setting $y_i = 1$ if $i = j^*$ and 0 otherwise, where j^* is the integer at which $S_i(p_1, p_2, \ldots, p_n)/\binom{a}{i}$ is minimized, over the range $i = 0, 1, \ldots, a$. To get a better handle on this minimum, we need

Lemma 2 *For any $i > 0$ and $s \geq 0$, $S_i(z_1, z_2, \ldots, z_n)$ is maximized by setting $z_1 = z_2 \cdots = z_n = \frac{s}{n}$, when subject to the constraints that $(z_1, z_2, \ldots, z_n) \in \Re_+^n$ and $\sum_{j=1}^{n} z_j = s$.*

Proof. Suppose $z_p < z_q$ for some vector z satisfying the constraints $(z_1, z_2, \ldots, z_n) \in \Re_+^n$ and $\sum_{j=1}^{n} z_j = s$. Then, set $z'_p = z_p + \epsilon$ and $z'_q = z_q - \epsilon$ for any $\epsilon < z_q - z_p$, and set $z'_j = z_j$ for all indices j, $j \notin \{p, q\}$. It is easy to verify that z' satisfies the above constraints, and that $S_i(z') > S_i(z)$. Hence $S_i(z)$ is maximized at $z = z^*$, where $z_j^* = \frac{s}{n}$ for $j = 1, 2, \ldots, n$. □

Inequality(1) and Lemma 2 imply that if $p \doteq \frac{\sum_{i=1}^{n} p_i}{n} = \frac{\mu}{n}$, then for any $y \in \Re_+^{n+1}$,

$$Pr(X \geq a) \leq \frac{\sum_{i=0}^{n} y_i \binom{n}{i} p^i}{\sum_{i=0}^{a} y_i \binom{a}{i}}. \tag{2}$$

Since $\frac{\binom{n}{i+1} p^{i+1} / \binom{a}{i+1}}{\binom{n}{i} p^i / \binom{a}{i}} = \frac{(n-i)p}{a-i}$ which is less than, equal to, or greater than 1 according as i is less than, equal to, or greater than $\frac{a-np}{1-\mu/n}$, $\binom{n}{i} p^i / \binom{a}{i}$ is minimized at

$$i^* = h(n, \mu, \delta) \doteq \lceil \frac{a - \mu}{1 - \mu/n} \rceil = \lceil \frac{\mu\delta}{1 - \mu/n} \rceil. \tag{3}$$

So, the r.h.s. of inequality(2) is minimized at $y = y^*$, where $y_i^* = 1$ if $i = i^*$, and 0 otherwise. Hence, we get

$$Pr(X \geq \mu(1+\delta)) \leq U_1(n, p_1, \ldots, p_n, \delta) \doteq \frac{S_{j^*}(p_1, p_2, \ldots, p_n)}{\binom{\mu(1+\delta)}{j^*}} \leq U_2(n, \mu, \delta) \doteq \frac{\binom{n}{i^*} \left(\frac{\mu}{n}\right)^{i^*}}{\binom{\mu(1+\delta)}{i^*}}.$$

$U_1(n, p_1, \ldots, p_n, \delta)$ is guaranteed to be better than any estimate based on the CH method, since we have considered a larger class of functions. Also, the upper bound $U_2(n, \mu, \delta)$ on $U_1(n, p_1, \ldots, p_n, \delta)$ is better than any such estimate which depends only on μ and which is oblivious to the actual values of $p_1, p_2, \ldots, p_n$; this includes $F(n, \mu, \delta)$ and $G(\mu, \delta)$.

But most importantly, note that `these new bounds will hold even if` $X_1, X_2, \ldots X_n$ `are only` $h(n, \mu, \delta)$`-wise independent`. This is because each term in $S_k(X_1, X_2, \ldots, X_n)$ is of the form $X_{i_1} X_{i_2} \ldots X_{i_k}$ for any integer k, and hence, $E[S_k(X_1, X_2, \ldots X_n)]$ will be the same for k–wise independent $X_1, X_2, \ldots X_n$ as for completely independent $X_1, X_2, \ldots X_n$. Since $\mu(1+\delta) \leq n$, $h(n, \mu, \delta) \leq n$; in typical algorithmic situations, $h(n, \mu, \delta) \ll n$. This will be seen to be of great use later on.

Theorem 1 *Let bits $X_1, X_2, \ldots, X_n$ be random with $Pr(X_i = 1) = p_i$, $X = \sum_{i=1}^n X_i$ and $\mu = E[X] = \sum_{i=1}^n p_i$. Suppose further, that the X_i are k–wise independent for $k \geq h(n, \mu, \delta)$. Then for any $\delta > 0$,*

$$Pr(X \geq \mu(1+\delta)) \leq U_1(n, p_1, \ldots, p_n, \delta) \leq U_2(n, \mu, \delta).$$

Furthermore, $U_2(n, \mu, \delta) \leq F(n, \mu, \delta) \leq G(\mu, \delta)$, i.e., `the CH upper bounds hold even if the` X_i`s are only` $h(n, \mu, \delta)$`-wise independent.`

Our results also imply upper tail bounds for r.v.'s with smaller independence than $h(n, \mu, \delta)$.

Lemma 3 *Let $X_1, X_2, \ldots, X_n$ be binary r.v.'s with $X \doteq \sum_{i=1}^n X_i$ and $\mu \doteq E[X]$. Then for any $\delta > 0$, $Pr(X \geq \mu(1+\delta)) \leq \binom{n}{k}(\mu/n)^k / \binom{\mu(1+\delta)}{k}$, if the X_i's are k–wise independent for any $k < h(n, \mu, \delta)$.*

Proof. Set $y_i = 0$ for $i \neq k$ and $y_k = 1$ in inequality(2). □

It turns out that $U_2(n, \mu, \delta)$ is almost the same as $F(n, \mu, \delta)$.

Theorem 2 *Given n random bits $X_1, X_2, \ldots, X_n$, let $X = \sum_{i=1}^n X_i$, $\mu = E[X]$, and $p \doteq \mu/n$. Then for any $\delta > 0$,*

1. *If the X_i's are $\lceil \mu\delta \rceil$–wise independent, then $Pr(X \geq \mu(1+\delta)) \leq G(\mu, \delta)$, where:*
$$G(\mu, \delta) = \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \leq \begin{cases} e^{-\mu\delta^2/3}, & \text{if } \delta < 1; \\ e^{-\mu\delta \ln(1+\delta)/2} \leq e^{-\mu\delta/3}, & \text{if } \delta \geq 1. \end{cases}$$
2. *If the X_i's are $\lceil \frac{\mu\delta}{1-p} \rceil$–wise independent, then $Pr(X \geq \mu(1+\delta)) \leq F(n, \mu, \delta)$.*
3. *If the X_i's are $\lceil \frac{\mu\delta}{p} \rceil = \lceil n\delta \rceil$–wise independent, then $Pr(X \leq \mu(1-\delta)) \leq F(n, \mu, -\delta)$, where:*
$$F(n, \mu, -\delta) \leq \begin{cases} e^{-\mu\delta^2/(2(1-p))}, & \text{if } p \leq 1/2; \\ e^{-2p\mu\delta^2}, & \text{if } p > 1/2. \end{cases}$$

Proof. The first claim follows by setting $k = \lceil \mu\delta \rceil \leq h(n, \mu, \delta)$ in Lemma 3. The only interesting case is that $k < h(n, \mu, \delta)$. We apply the inequality $\binom{n}{k}(\mu/n)^k / \binom{a}{k} \leq Q(n, k, a) \doteq (\frac{n}{a})^k (\mu/n)^k (\frac{n}{n-k})^{n-k} (\frac{a-k}{a})^{a-k}$, valid for any $a < n$; this inequality follows by induction on k, combined with the fact that the function $(1 - \frac{1}{x})^{x-1}$ is nonincreasing for $x > 1$. Let $a = (1+\delta)\mu$ and $k' = \mu\delta$. Then,

$$Q(n, k', a) = \frac{(n/(n-\mu\delta))^{n-\mu\delta}}{(1+\delta)^{\mu(1+\delta)}} = \left(\frac{1}{1+\delta}\right)^{\mu(1+\delta)} \left(1 + \frac{\mu\delta}{n-\mu\delta}\right)^{n-\mu\delta} \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu = G(\mu, \delta).$$

It is easy to verify that $Q(n, x, a)$ is nonincreasing for $x \leq h(n, \mu, \delta)$, and hence the bound $G(\mu, \delta)$ established for k' also holds for $k = \lceil \mu\delta \rceil$. The upper bounds for $G(\mu, \delta)$ are either straightforward or have been established in [4, 35, 3].

The second claim follows immediately from Theorem 1, while the third claim follows by obtaining lower tail bounds from the upper tail of $\sum_{i=1}^n (1 - X_i)$, and importing the upper tail bounds established in [17]. By Theorem 1, these bounds hold with independence $h(n, \mu, \delta)$. □

As we will show in Section 2.3, bounds almost as good as $G(\mu, \delta)$ and $F(n, \mu, -\delta)$ hold with the much smaller independence $k = \lfloor \mu\delta^2 \rfloor$, when $\delta < 1$.

2.2 Tail probabilities of bounded random variables

We now show that almost the same results hold for arbitrary r.v.'s which take values in $[0,1]$. Analogous bounds for bounded r.v.'s that are constrained to lie in other intervals can be obtained by a linear transformation of their ranges to $[0,1]$. Given arbitrary r.v.'s X_i such that $0 \le X_i \le 1$, $i = 1, 2, \ldots, n$, we wish to upper bound $Pr(X \ge \mu(1+\delta))$, where $X = \sum_{i=1}^n X_i$, $\mu = E[X]$ and $\delta > 0$. Hoeffding [17] has proved upper tail bounds for bounded random variables, assuming full independence among the X_i's; the main point of interest here again is that partial independence suffices, giving almost as good bounds as Hoeffding's. Almost all of the work is done by

Lemma 4 *Let z_i be real numbers, with $0 \le z_i \le 1$, $i = 1, 2, \ldots, n$, and suppose that $a \ge 0$, $j \le \lfloor a \rfloor$ and $\sum_{i=1}^n z_i \ge a$. Then,*

$$S_j(z_1, z_2, \ldots, z_n) \ge \binom{a}{j}.$$

Proof. We will just consider the case $\sum_{i=1}^n z_i = a$; then the upper bound will directly follow if $\sum_{i=1}^n z_i > a$. If $0 < z_p \le z_q < 1$ for $p \ne q$, then $S_j(z)$ decreases if we set $z_p := z_p - \epsilon$ and $z_q := z_q + \epsilon$, for any $\epsilon < \min(z_p, 1 - z_q)$. So, if $S_j(z)$ is minimized at z^* in the domain $[0,1]^n$ under the constraint that $\sum_{i=1}^n z_i = a$, then $0 < z_i^* < 1$ for at most one i, $1 \le i \le n$.

If a is integral, then $z_i^* \in \{0,1\}$, $i = 1, 2, \ldots, n$, and hence $S_j(z^*) = \binom{a}{j}$. Otherwise, suppose a is non–integral; let $a_1 = \lfloor a \rfloor$ and $a_2 = a - a_1$. Hence, $z_p^* = a_2$ for some index p, and $z_i^* \in \{0,1\}$ for $i \in \{1, 2, \ldots, n\} - \{p\}$; so,

$$S_j(z^*) = \binom{a_1}{j} + a_2 \cdot \binom{a_1}{j-1},$$

and we need to show that this is at least $\binom{a}{j}$; *i.e.*, that

$$[a_1]_j + a_2 j \cdot [a_1]_{j-1} \ge [a_1 + a_2]_j = [a]_j,$$

where $[x]_r$ denotes $x(x-1)\cdots(x-r+1)$ and $[x]_0 \doteq 1$. This is easily seen by induction on j, as follows. Equality clearly holds for $j = 1$. For $j > 1$, $[a_1]_j + a_2 j \cdot [a_1]_{j-1} = [a_1]_j + (j-1)a_2[a_1]_{j-1} + a_2[a_1]_{j-1}$. Since $a_2 < 1$, $a_2[a_1]_{j-1} \ge a_2[a_1 + a_2 - 1]_{j-1}$ and hence,

$$\begin{aligned}
[a_1]_j + a_2 j \cdot [a_1]_{j-1} &\ge a_1([a_1 - 1]_{j-1} + (j-1)a_2[a_1 - 1]_{j-2}) + a_2[a_1 + a_2 - 1]_{j-1} \\
&\overset{ind\,hyp.}{\ge} a_1[a_1 + a_2 - 1]_{j-1} + a_2[a_1 + a_2 - 1]_{j-1} \\
&= [a_1 + a_2]_j.
\end{aligned}$$

□

By essentially the same analysis as before, we get

Theorem 3 *Given n arbitrary r.v.'s $X_1, X_2, \ldots, X_n$ with $0 \le X_i \le 1$ and $E[X_i] = p_i$, let $X \doteq \sum_{i=1}^n X_i$ and $\mu \doteq E[X]$. Then if $X_1, X_2, \ldots, X_n$ are k–wise independent for $k \ge h(n, \mu, \delta)$, then $Pr(X \ge \mu(1+\delta)) \le U_1(n, p_1, \ldots, p_n, \delta) \le U_2(n, \mu, \delta)$, for any $\delta > 0$.*

PROOF. From Lemma 4, we have that for any $a > 0$ and for *non–negative* $y_0, y_1, \ldots, y_n$,

$$Pr(X \geq a) \leq Pr(f_y(X_1, \ldots, X_n) \geq \sum_{i=0}^{\lfloor a \rfloor} y_i \binom{a}{i}) \leq \frac{E[f_y(X_1, \ldots, X_n)]}{\sum_{i=0}^{\lfloor a \rfloor} y_i \binom{a}{i}},$$

and the rest of the proof follows as before. □

Remark. The methods of Section 2.1 were motivated by the fact that if X is the sum of n 0–1 random variables, then any n higher moments of X linearly generate all the higher moments of X. However, note that if random variables $X_1, X_2, \ldots, X_n$ take arbitrary values in the interval $[0, 1]$ and if $X = \sum_{i=1}^{n}$, then such a result is *not* true: in fact, no bound can be put on the number of higher moments needed to generate all the moments of X. However, the intuition gained from Section 2.1 has helped us obtain a large deviation bound for X, which is as good as the known bound [17]. This is despite the fact that we have not considered all the higher moments of X; one of the original motivations for Chernoff to consider $E[e^{tX}]$ was that it generates all the higher moments of X. A possible interpretation of our result of this subsection is that it pinpoints the "crucial" higher moments.

2.3 Redirecting the method

Recall that in Section 2.1 we introduced the class of functions $S_0(z), S_1(z), \ldots, S_n(z)$ and generalized Chernoff's idea by working with `non-negative` linear combinations of these functions. A natural generalization of this is to allow arbitrary linear combinations, but the corresponding optimization problem, described below, seems hard to analyze.

Suppose we have n *binary* random variables $X_1, X_2, \ldots, X_n$ with $Pr(X_i = 1) = p_i$ and with $X = \sum_{i=1}^{n} X_i$, and we want good upper bounds on $Pr(X \geq a)$ where $a > E[X]$, when the X_i's are k–wise independent. As before, let

$$f_y(X_1, X_2, \ldots, X_n) = \sum_{i=0}^{n} y_i S_i(X_1, X_2, \ldots, X_n),$$

with the further restriction that $y_i = 0$ for $i \geq k + 1$ to capture the idea of k–wise independence; note that $f_y(X_1, X_2, \ldots, X_n)$ is a function of X,

$$f_y(X_1, X_2, \ldots, X_n) = g_y(X) \doteq \sum_{i=0}^{min(k,X)} y_i \binom{X}{i}.$$

Now, if $g(t) \geq 0$ for $t = 0, 1, \ldots, n$ (so that Markov's inequality can be applied) and if $g_y(b) \geq g_y(a)$ for $b \geq a$, then

$$Pr(X \geq a) \leq Pr(g_y(X) \geq g_y(a)) \leq \frac{E[g_y(X)]}{g_y(a)} = \frac{\sum_{i=0}^{k} y_i S_i(p_1, p_2, \ldots, p_n)}{g_y(a)}.$$

We can scale the y_i's so that $g_y(a) = 1$ and thus, we get the following linear program with $y_0, y_1, \ldots, y_k$ being arbitrary real variables.

$LP(a, k, p_1, p_2, \ldots, p_n)$:
Minimize $\sum_{i=0}^{k} y_i S_i(p_1, p_2, \ldots, p_n)$ subject to

- $g_y(j) \geq 0$, $j = 0, 1, \ldots, n$.
- $g_y(a) = 1$, and
- $g_y(b) \geq 1$, $b = a+1, a+2, \ldots, n$.

We unfortunately have been unable to analytically compute the optimum of this linear program. However, we now consider an important case where some of the multipliers are negative, and which is a feasible solution to the above LP; our results generalize a result of [22, 8, 27]. We use the kth moment inequality

$$Pr(|X - E[X]| \geq \delta E[X]) \leq \frac{E[|X - E[X]|^k]}{(\delta E[X])^k},$$

which is attributable, in various formulations and generalizations, to Chebyshev, Markov, and Loéve [18], and has been used to attain probability deviation estimates for over a century. Note that if $X = X_1 + X_2 + \cdots + X_n$ where the X_i's are random bits, then $(X - E[X])^k$ is a linear combination of $S_0(X_1, X_2, \ldots, X_n)$, $S_1(X_1, X_2, \ldots, X_n)$, $\ldots$, $S_k(X_1, X_2, \ldots, X_n)$, with some of the multipliers being negative. We derive good upper bounds on $E[|X - E[X]|^k]$, where $X = \sum_{i=1}^n X_i$, with the X_i's being k–wise independent random variables which satisfy $|X_i - E[X_i]| \leq 1$, yielding bounds that are better than those given Theorem 1 and Lemma 3, when $k \ll h(n, \mu, \delta)$ and $\delta < 1$. Moreover, the large deviation bounds derived in Theorem 5 for k-wise independent random variables agree with the simple exponential forms of the large deviation bounds most often cited for sequences of fully independent Bernoulli trials.

Theorem 4 is similar in spirit and proof to Lemma 4.19 of [22] for identically distributed X_i and constant k, but the present result is somewhat tighter even in the case of identically distributed X_i, especially if $X = \sum_{i=1}^n X_i$ has small variance. A slightly weaker form of a special case of one of the inequalities proven in Theorem 4 was also obtained in [8] and some related formulae were given in [27]. The proof of Theorem 4, as well as related proofs presented elsewhere, is based upon estimates for the k-th moment of X. Estimates related to ours, but for a more general class of random variables, were established in [28]. That formulation however, is considerably more complicated than ours, and is not as tight for the cases specifically considered here. In particular, Theorem 5 cannot be derived from the bounds in [28] for the k-th moment. Other related work was done by Gladkov ([15], with later improvements in [16]). He shows that if $Y_1, Y_2, \ldots, Y_n$ are *independent* r.v.'s with Y_i having the same distribution as X_i and with $Y \doteq Y_1 + Y_2 + \cdots + Y_n$, then as $n \to \infty$, the convergence of Y to the normal distribution implies a comparable convergence for X, provided k is sufficiently large.

Theorem 4 *Let $X_1, \ldots, X_n$ be a sequence of k-wise independent random variables, that satisfy $|X_i - \mathrm{E}[X_i]| \leq 1$. Let $X = \sum_{i=1}^n X_i$ with $E[X] = \mu$, and let $\sigma^2[X]$ denote the variance of X, so that $\sigma^2[X] = \sum_{i=1}^n \sigma^2[X_i]$ (provided $k \geq 2$, which we require.) Then the following hold for any even k.*

(I) For[1] $C \geq \sigma^2[X]$, $$Pr(|X - \mu| \geq T) \leq \sqrt{2}\cosh\left(\sqrt{\frac{k^3}{36C}}\right)\left(\frac{kC}{eT^2}\right)^{k/2}.$$

[1]Recall that $\cosh(x) = \frac{e^x + e^{-x}}{2}$. Throughout this manuscript, e denotes, as usual, the base of the natural logarithm.

(II) *For* $2 \le k \le 3(\sigma^2[X])^{1/3}$, $\quad Pr(|X-\mu| \ge T) \le 2\left(\frac{k\sigma^2[X]}{eT^2}\right)^{k/2}$.

(III) *For* $C \ge max\{k, \sigma^2[X]\}$, $\quad Pr\left(|X-\mu| \ge T\right) \le \left(\frac{kC}{e^{2/3}T^2}\right)^{k/2}$.

PROOF. We use the kth moment inequality

$$Pr(|X-\mu| \ge T) \le \frac{\mathrm{E}\left[|X-\mu|^k\right]}{T^k}. \tag{4}$$

For even k, $\mathrm{E}\left[|X-\mu|^k\right] = \mathrm{E}\left[(X-\mu)^k\right]$. Most of our effort will therefore be invested in estimating the k-th moment of $X-\mu$. Let $p_i = \mathrm{E}[X_i]$, for $1 \le i \le n$. Then

$$\mathrm{E}\left[(X-\mu)^k\right] = \mathrm{E}\left[\left(\sum_{i=1}^{n}(X_i-p_i)\right)^k\right] = \sum_{i_1+i_2\cdots+i_n=k}\binom{k}{i_1,\ldots,i_n}\mathrm{E}[(X_1-p_1)^{i_1}]\cdots\mathrm{E}[(X_n-p_n)^{i_n}]. \tag{5}$$

Clearly, $\mathrm{E}[X_i-p_i]=0$ and any term in (5) that has some $i_j=1$ must be zero. More generally, $|\mathrm{E}[X_i-p_i]^\ell| \le \sigma^2[X_i]$ for any $\ell \ge 2$, since $|X_i-p_i| \le 1$ and therefore $|X_i-p_i|^\ell \le |X_i-p_i|^2$, hence $|\mathrm{E}[X_i-p_i]^\ell| \le |\mathrm{E}[X_i-p_i]^2| = \sigma^2[X_i]$. Thus,

$$\begin{aligned}
\mathrm{E}\left[(X-\mu)^k\right] &= \sum_{\ell=0}^{k/2-1}\sum_{\substack{j_1+j_2+\cdots+j_{k/2-\ell}=k\\ j_i\ge 2}}\binom{k}{j_1,j_2,\ldots,j_{k/2-\ell}}\sum_{i_1<i_2<\cdots<i_{k/2-\ell}}\prod_{r=1}^{k/2-\ell}\mathrm{E}[X_{i_r}-p_{i_r}]^{j_r}\\
&\le \sum_{\ell=0}^{k/2-1}\sum_{\substack{j_1+j_2+\cdots+j_{k/2-\ell}=k\\ j_i\ge 2}}\binom{k}{j_1,j_2,\ldots,j_{k/2-\ell}}\sum_{i_1<i_2<\cdots<i_{k/2-\ell}}\prod_{r=1}^{k/2-\ell}\sigma^2[X_{i_r}]\\
&\le \sum_{\ell=0}^{k/2-1}\sum_{\substack{j_1+j_2+\cdots+j_{k/2-\ell}=k\\ j_i\ge 2}}\binom{k}{j_1,j_2,\ldots,j_{k/2-\ell}}\frac{(\sigma^2[X])^{k/2-\ell}}{(k/2-\ell)!}.
\end{aligned} \tag{6}$$

Estimate (6) comes about because the summation

$$\sum_{i_1<i_2<\cdots<i_{k/2-\ell}}\prod_{r=1}^{k/2-\ell}\sigma^2[X_{i_r}]$$

is maximized when all the $\sigma^2[X_{i_c}]$ are equal, by Lemma 2, and hence is at most

$$\binom{n}{k/2-\ell}\left(\frac{\sigma^2[X]}{n}\right)^{k/2-\ell} \le \frac{(\sigma^2[X])^{k/2-\ell}}{(k/2-\ell)!}.$$

Let $T_0, T_1, \ldots T_{k/2-1}$ denote the $k/2$ terms in summation (6), hence:

$$T_\ell = \sum_{\substack{j_1+j_2+\cdots+j_{k/2-\ell}=k\\ j_i\ge 2}}\binom{k}{j_1,j_2,\ldots,j_{k/2-\ell}}\frac{(\sigma^2[X])^{k/2-\ell}}{(k/2-\ell)!}, \tag{7}$$

and

$$T_0 \;=\; \binom{k}{2,\ 2, \ldots, 2}\frac{(\sigma^2[X])^{k/2}}{(k/2)!}. \tag{8}$$

There are exactly $\binom{k/2+\ell-1}{2\ell}$ terms (*i.e.*, possible sets of assignments for $j_1, j_2, \ldots, j_{k/2-\ell}$) in T_ℓ, since $\sum j_i = k$ and $j_i \geq 2$. For each such assignment of values, $\sum_{k=1}^{k/2-\ell}(j_k - 2) = 2\ell$, hence $\prod_{k=1}^{k/2-\ell} j_k! \geq 2^{k/2-\ell}3^{2\ell}$ (and equality holds only when $\ell \leq k/6$ and exactly 2ℓ of the $k/2-\ell$ values equal 3, while the remaining values equal 2). Hence $\binom{k}{j_1, j_2, \ldots, j_{k/2-\ell}} \leq (2/9)^\ell \binom{k}{2,\ 2, \ldots, 2}$, and

$$T_\ell \leq \binom{k/2+\ell-1}{2\ell}\left(\frac{2}{9\sigma^2[X]}\right)^\ell \frac{(k/2)!}{(k/2-\ell)!}T_0.$$

Since $\binom{k/2+\ell-1}{2\ell}\frac{(k/2)!}{(k/2-\ell)!} < \frac{(k/2)^{3\ell}}{(2\ell)!}$, we have $\mathrm{E}\left[(X-\mu)^k\right] \leq T_0 \sum_{\ell=0}^{k/2-1} \frac{k^{3\ell}}{(36\sigma^2[X])^\ell (2\ell)!}$. Using Taylor's Series for $\cosh(x) = \frac{e^x+e^{-x}}{2} = \sum_j \frac{x^{2j}}{(2j)!}$, we see that $\sum_{\ell=0}^{k/2-1} \frac{\left(\frac{k^3}{36\sigma^2[X]}\right)^{\frac{2\ell}{2}}}{(2\ell)!} \leq \cosh\left(\sqrt{\frac{k^3}{36\sigma^2[X]}}\right)$. Consequently

$$\mathrm{E}\left[(X-\mu)^k\right] \leq \cosh\left(\sqrt{\frac{k^3}{36\sigma^2[X]}}\right) T_0. \tag{9}$$

T_0 is readily bounded by expanding (8) to get $T_0 = \dfrac{k!}{2^{k/2}(k/2)!}(\sigma^2[X])^{k/2}$. We may apply a strong version of Stirling's Formula [37]:

$$(r/e)^r\sqrt{2\pi r}e^{1/(12r+1)} \leq r! \leq (r/e)^r\sqrt{2\pi r}e^{1/(12r)},$$

which is valid for all $r \geq 1$, to bound both $k!$ and $(k/2)!$. This yields $T_0 \leq \sqrt{2}\left(\frac{k\sigma^2[X]}{e}\right)^{k/2}$. Substituting for T_0 in (9) gives

$$\mathrm{E}\left[(X-\mu)^k\right] \leq \sqrt{2}\cosh\left(\sqrt{\frac{k^3}{36\sigma^2[X]}}\right)\left(\frac{k\sigma^2[X]}{e}\right)^{k/2}, \tag{10}$$

which establishes the desired bound for $\mathrm{E}\left[(X-\mu)^k\right]$.

Now, estimate (6) is an increasing function of $\sigma^2[X]$, and the estimate in (10) exceeds (6). Therefore $\sigma^2[X]$ can be replaced by any $C \geq \sigma^2[X]$ in (10). The proof of (I) is completed by applying this estimate to (4).

All other bounds are special cases of (I). When $k \leq 3(\sigma^2[x])^{1/3}$, we use $C = \sigma^2[X]$ in Theorem 4.I and overestimate $\cosh\left(\sqrt{\frac{k^3}{36\sigma^2[X]}}\right)$ by $\cosh\left(\sqrt{\frac{3}{4}}\right) < \sqrt{2}$.

(III) is easily verified for $k = 2$, by applying Chebyshev's inequality:

$$Pr(|X-\mu| \geq T) \leq \frac{\sigma^2[X]}{T^2} \leq \frac{2\sigma^2[X]}{e^{2/3}T^2}$$

For $k \geq 4$ we may replace C by $max\{\sigma^2[X], k\}$ in (I) and overestimate $\cosh\left(\sqrt{\frac{k^3}{36C}}\right)$ by $\cosh(k/6)$. Since $\cosh(x) < e^x/\sqrt{2}$ for $x \geq 1/2$, we get $\cosh(k/6) \leq e^{k/6}/\sqrt{2}$, and hence

$$Pr\left(|X-\mu| \geq T\right) \leq \left(\frac{kC}{e^{2/3}T^2}\right)^{k/2}.$$

This concluses the proof of estimate (III). $\square$

We now combine the results of Theorem 4 and Theorem 2 to establish Chernoff-like bounds [11, 17], where the independence k might even be much smaller than the deviation we wish to bound.

Theorem 5 *If X is the sum of k–wise independent r.v.'s, each of which is confined to the interval $[0,1]$ with $\mu = E[X]$, then:*

(I) For $\delta \le 1$,
(a) if $k \le \left\lfloor \delta^2 \mu e^{-1/3} \right\rfloor$, then $Pr\left(|X-\mu| \ge \delta\mu\right) \le e^{-\lfloor k/2 \rfloor}$.
(b) if $k = \left\lfloor \delta^2 \mu e^{-1/3} \right\rfloor$, then $Pr\left(|X-\mu| \ge \delta\mu\right) \le e^{-\left\lfloor \delta^2\mu/3 \right\rfloor}$.
(II) For $\delta \ge 1$,
(a) if $k \le \left\lfloor \delta \mu e^{-1/3} \right\rfloor$, then $Pr\left(|X-\mu| \ge \delta\mu\right) \le e^{-\lfloor k/2 \rfloor}$.
(b) if $k = \left\lfloor \delta \mu e^{-1/3} \right\rfloor$, then $Pr\left(|X-\mu| \ge \delta\mu\right) \le e^{-\lfloor \delta\mu/3 \rfloor}$.
(III) For $\delta \ge 1$ and $k = \lceil \delta\mu \rceil$, then

$$Pr\left(|X-\mu| \ge \delta\mu\right) \le \quad G(\mu,\delta) \le e^{\frac{-\delta \ln(1+\delta)\mu}{2}} < e^{\frac{-\delta\mu}{3}}.$$

Proof. (I). To prove that (Ia) holds we apply Theorem 4.III with $C = \mu$, $T = \delta\mu$ and $k = \left\lfloor \delta^2\mu/e^{1/3} \right\rfloor$, which is permissible since $\mu \ge k$ and $\mu \ge \sigma^2[X]$ for variables in the range $[0,1]$. When $k = \left\lfloor \delta^2\mu/e^{1/3} \right\rfloor$ and k is even, this gives a bound of

$$Pr(|X-\mu| \ge \delta\mu) \le \left(\frac{k}{e^{2/3}\delta^2\mu}\right)^{k/2} \le e^{-k/2} \le e^{-\left\lfloor \delta^2\mu/3 \right\rfloor}, \text{ since } 2e^{1/3} < 3.$$

If k is odd, we follow a calculation similar to that above, but only use independence $k-1$. This gives

$$Pr(|X-\mu| \ge \delta\mu) \le \left(\frac{k-1}{e^{2/3}\delta^2\mu}\right)^{k-1/2} \le \left(\frac{k-1}{ek}\right)^{(k-1)/2} \le e^{-(k-1)/2} e^{-(k-1)/2k}.$$

Since $k \ge 2$ for the above bound to give anything less than 1, $e^{-(k-1)/2k} \le e^{-1/3}$ and hence,

$$Pr(|X-\mu| \ge \delta\mu) \le e^{-(k-1)/2+1/3)} \le e^{-\left\lfloor \delta^2\mu/3 \right\rfloor}.$$

In part (II) we follow the same iterations as in part (I), but set $C = \delta\mu$ and $k \le \lfloor \delta\mu/e^{1/3} \rfloor$, for (IIa); in (IIb) we use $k = \lfloor \delta\mu/e^{1/3} \rfloor$ or $\lfloor \delta\mu/e^{1/3} \rfloor - 1$, depending on the parity of $\lfloor \delta\mu/e^{1/3} \rfloor$. In part (III) we reiterate the result of Theorem 2, combined with Theorem 3. $\square$

Remark. The proofs of parts (I) and (III) of Theorem 5 also point out the relative merits of the basic method (Sections 2.1 and 2.2) versus its redirection of this subsection. The basic method of using non–negative linear combinations of the symmetric polynomials S_i gives better probability bounds when δ, the relative deviation from the mean, is greater than 1: it yields the probability bound of $exp(-\Theta(\delta \ln(1+\delta)\mu))$ in this case. On the other hand, the formulation involving the kth moment inequality gives a much smaller bound on the amount of independence needed, when $\delta < 1$.

2.4 Probability Bounds for Exactly r Successes under Limited Independence

Some applications require estimates for the probability that **exactly** r successes occur in cases where the occurrence of at least r successes is not too improbable. The following theorem shows how and when this can be done. It also provides relative errors, which can be useful for estimating conditional probabilities.

Theorem 6 *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be Bernoulli trials with probabilities of success* $\mathrm{E}[X_i] = \mathrm{E}[Y_i] = p_i$*. We let the Y_i's be independent, but only require the X_i's to be k-wise independent. Let $p(r) = Pr(\sum_i Y_i = r)$, and $p_k(r) = Pr_k(X = r)$, where the subscript k denotes the k-wise independent trials. Let $P(r) = \sum_{\ell \geq r} p(\ell)$, and $P_k(r) = \sum_{\ell \geq r} p_k(\ell)$.*

(I) If $r \leq k$, then $\quad |p_k(r) - p(r)| \leq \binom{n}{k}\binom{k}{r}\left(\frac{\mu}{n}\right)^k.$

(II) If $k \geq e\mu + \ln(1/p(0)) + r + D$, then $\quad |p_k(r) - p(r)| \leq e^{-D} p(r).$

(III) If $k \geq e\mu + \ln(1/p(0)) + r + D$, then $\quad |P_k(r) - P(r)| \leq (1 - P(r))e^{-D}.$

(IV) If $r \geq (1+\delta)\mu + k$, then $\quad P_k(r) \leq (1+\delta)^{-k}$ *and* $P(r) \leq (1+\delta)^{-k}$,
and hence $\quad |P_k(r) - P(r)| \leq (1+\delta)^{-k}.$

Although (IV) holds for all values of k it is meant to be used for $k \ll \lceil \delta\mu \rceil$, indeed:

(V) If $r \geq (1+\delta)\mu + k$ and $k \geq \lceil \delta\mu \rceil$, then $\quad P_k(r) \leq G(\mu, \delta)$ *and* $P(r) \leq G(\mu, \delta)$,
and hence $\quad |P_k(r) - P(r)| \leq G(\mu, \delta).$

Proof. For an arbitrary event A, we may use standard inclusion-exclusion to estimate the probability of the event $\left[A \wedge [\bigwedge_{\ell \notin \{i_1,\ldots,i_r\}} (X_\ell = 0)]\right]$. The probability $p(r)$ can be expressed in terms of events $A = [\bigwedge_{j \in \{i_1,\ldots,i_r\}} (X_j = 1)]$, which admits a simple estimation as follows.

$$\begin{aligned}
p_k(r) &= \sum_{i_1 < i_2 < \cdots < i_r} Pr_k\left(\left(\bigwedge_{j\in\{i_1,\ldots,i_r\}} (X_j = 1)\right) \wedge \left(\bigwedge_{\ell \notin \{i_1,\ldots,i_r\}} (X_\ell = 0)\right)\right) \\
&= \sum_{i_1 < i_2 < \cdots < i_r} \sum_{l=0}^{n-r} \sum_{\substack{i_{r+1} < \cdots < i_{r+l} \\ i_{r+1},\ldots,i_{r+l} \notin \{i_1,\ldots,i_r\}}} (-1)^l Pr_k\left(\bigwedge_{j \in \{i_1,\ldots,i_{r+l}\}} (X_j = 1)\right) \\
&= \sum_{l=0}^{n-r} \sum_{i_1 < \cdots < i_{r+l}} (-1)^l \binom{r+l}{r} Pr_k\left(\bigwedge_{j\in\{i_1,\ldots,i_{r+l}\}} (X_j = 1)\right).
\end{aligned}$$

Truncating the outer summation at $l = k - r$ introduces an error that is bounded by the last term of the truncated sum. Let $p_k^T(r)$ and $p^T(r)$ denote these truncated sums, in the respective cases of k-wise and full independence. Since the first $k - r$ terms in the outer summation are the same for both fully and k-wise independent random variables, $p_k^T(r) = p^T(r)$. Furthermore,

$Pr_k\left(\bigwedge_{j\in\{i_1,\ldots,i_k\}}(X_j=1)\right)=\prod_{j=1}^k p_{i_j}$. Hence,

$$p_k(r)=p^T(r)-(-1)^{k-r}\delta_k\binom{k}{r}\sum_{i_1<\ldots<i_k}\prod_{j=1}^k p_{i_j}$$

for some $\delta_k\in[0,1]$, and an identical inequality holds without the k subscripts. Hence,

$$|p_k(r)-p(r)|\le\binom{k}{r}\sum_{i_1<\ldots<i_k}\prod_{j=1}^k p_{i_j}. \tag{11}$$

$\sum_{i_1<\ldots<i_k}\prod_{j=1}^k p_{i_j}$ is maximized when all p_{i_j} are equal (Lemma 2), and hence,

$$|p_k(r)-p(r)|\le\binom{k}{r}\binom{n}{k}\frac{(p_1+p_2+\ldots+p_n)^k}{n^k}=\binom{n}{k}\binom{k}{r}(\mu/n)^k,$$

and (I) now follows.

To get multiplicative error bounds, let $Q_r=\sum_{i_1<i_2<\cdots<i_r}\prod_{\ell=1}^r\left(\frac{p_{i_\ell}}{1-p_{i_\ell}}\right)$, and define the summation in the error estimate of equation (11) by $R_k=\sum_{i_1<\ldots<i_k}\prod_{j=1}^k p_{i_j}$. Observe that R_k is the expected number of size k sets of successes among n trials, so that z successes total accounts for $\binom{z}{k}$ such sets. In the fully independent case, $p(r)=p(0)\sum_{i_1<i_2<\cdots<i_r}\prod_{\ell=1}^r\left(\frac{p_{i_\ell}}{1-p_{i_\ell}}\right)=p(0)Q_r$. Furthermore, $R_r\le Q_r$, and $\binom{k}{r}R_k\le R_r\times R_{k-r}$. It follows that

$$|p_k(r)-p(r)|\le\binom{k}{r}R_k\le R_r\times R_{k-r}\le Q_rR_{k-r}\le\frac{p(r)\mu^{k-r}}{p(0)(k-r)!},$$

since $R_{k-r}\le\binom{n}{k-r}\left(\frac{\mu}{n}\right)^{k-r}\le\frac{\mu^{k-r}}{(k-r)!}$.

For $k\ge e\mu-\log(p(0))+r+D$, the factor $\frac{(\mu)^{k-r}}{p(0)(k-r)!}$ is bounded by $\left(\frac{k+\log(p(0))-D-r}{k-r}\right)^{k-r}e^{-\log(p(0))}\le e^{-D}$, which establishes (II).

Part (III) is immediate, since

$$P_k(r)=1-\sum_{\ell<r}p_k(\ell)\text{ and hence }|P_k(r)-P(r)|\le\sum_{\ell<r}|p_k(\ell)-p(\ell)|\le\sum_{\ell<r}p(\ell)e^{-D}=(1-P(r))e^{-D}.$$

Finally, suppose that $r=(1+\delta)\mu+k$. Then by Lemma 3,

$$\begin{aligned}P_k(r)&\le\frac{\mu^k}{k!\binom{r}{k}}\le\frac{\mu^k}{r(r-1)(r-2)\ldots(r-k+1)}\\&\le(1+\delta)^{-k}.\end{aligned}\tag{12}$$

Part (IV) is completed by observing that (12) also holds with $P(r)$ substituted for $P_k(r)$. Part (V) is an immediate consequence of Theorem 2. This concludes the proof of Theorem 6. □

It is worth pointing out that parts (I) through (III) of Theorem 6 are not strong when $\mu>\sqrt{n}$, since it follows from the work of Linial & Nisan [24] that $Pr(X=\ell)=Pr(Y=\ell)(1+$

$O(e^{-2(k-\ell)/\sqrt{n}})$) independently of μ, which gives a much sharper bound in this case. Also, independent of our work, a result similar to part(I) of Theorem 6 has been proven by Even, Goldreich, Luby, Nisan & Veličković [14]: they show that $|p_k(0) - p(0)| \leq 2^{-\Omega(k)}$.

Theorem 6, in fact, achieves its greatest strength when μ is small, say $\mu = o(n)$, or even $\mu = O(1)$. Such instances are not unusual when pseudorandom integers are being generated uniformly in the range $[0, n]$ and a successful trial corresponds to just a few different values. This is precisely the usual circumstance in, for instance, hashing [40, 51]. As an example, consider the (uniformly distributed) random placement of n balls among n slots. The expected number of items in slot 1 is just 1. The probability $p(0)$ that no item lands in a given slot is about $\frac{1}{e}$. Theorem 6 shows that if the independence k is $e+1+r+D$, the probability that exactly r items land in that slot is the same in the k-wise independent case as in the fully independent case, up to a multiplicative factor of $(1+e^{-D})$ or less. Suppose that, during the placements of balls l through m, exactly r balls fall into slot j for $1 \leq l \leq m < n$, $r \leq m - l + 1$. Let $\chi^r_{[l,m]}$ denote this event (with the dependence upon j understood). The conditional probability that, under k-wise independence, ball $m+1$ also falls into slot j is $Pr_k(\chi^1_{[m+1,m+1]}|\chi^r_{[l,m]})$. If we use one degree of freedom for the $m+1$-st ball, it will be uniformly distributed while the previous m balls will enjoy $(k-1)$-wise independence. We may estimate the conditional probability as $Pr_k(\chi^r_{[l,m]}|\chi^1_{[m+1,m+1]}) \times \frac{Pr(\chi^1_{[m+1,m+1]})}{Pr_{k-1}(\chi^r_{[l,m]})}$. Since both $Pr_{k-1}(\chi^r_{[l,m]})$ and $Pr_k(\chi^r_{[l,m]}|\chi^1_{[m+1,m+1]})$ can be estimated by $Pr(\chi^r_{[l,m]})(1 \pm err_{k-1})$ where err_{k-1} is the relative error that results from the limited independence, we see that $Pr_k(\chi^1_{[m+1,m+1]}|\chi^r_{[l,m]})$ is very close to $1/n$, with a relative error that is approximately $2err_{k-1}$. For $k \geq 1+e+1+r+D$, the resulting accuracy is about $1 \pm 2e^{-D}$. Thus even with modest independence, this process behaves "as expected" much of the time; that is, the corresponding conditional probabilities for k-wise independence are very close to the ones for full independence, in many cases.

2.5 How close to optimal are our results?

It is known that the standard Chernoff–Hoeffding bounds are optimal in general to within a constant factor in the exponent, since we know by the Central Limit Theorem that as $n \to \infty$, the tail of the scaled sum of i.i.d. r.v.'s tends to the tail of a normal distribution, and hence we cannot significantly improve the tail probabilities presented by Theorem 5. However, what about the independence we get? Can it be reduced further to get the same tail probabilities?

To answer this, we note that the tail probabilities presented by Theorem 5 for k–wise independent r.v.'s are of the form $e^{-c \cdot k}$. However, n k–wise independent r.v.'s require a sample space of size at least

$$\sum_{i=0}^{\lfloor k/2 \rfloor} \binom{n}{i} \approx (O(n/k))^{\lfloor k/2 \rfloor},$$

as shown for binary unbiased r.v.'s by Chor, Goldreich, Håstad, Friedman, Rudich & Smolensky [12] and for general r.v.'s by Alon, Babai & Itai [1]. Noting next that any nonzero probability in a sample space of size t is at least $1/t$, we see that to get a tail probability of the form $e^{-c \cdot k}$, we need at least $\Omega(\frac{k}{\log(n/k)})$–wise independence. Thus, the independence we get cannot in general be reduced by more than a factor of $O(\log n)$.

However, by using results from the newly developing theory of *approximating probability distri-*

butions (Naor & Naor [29], Azar, Motwani & Naor [5], Alon, Goldreich, Håstad & Peralta [2], Even, Goldreich, Luby, Nisan & Veličković [14] and Chari, Rohatgi & Srinivasan [10]), we get optimal results in the case where the X_i's are binary with $Pr(X_i = 1) = 1/2$. A sample space X for n–bit vectors was defined to be k–wise ϵ–biased by Naor & Naor [29] (see also Vazirani [49]) if

$$\forall S \subseteq \{1, 2, \ldots, n\}, 1 \leq |S| \leq k, |Pr(\bigoplus_{i \in S} x_i = 1) - Pr(\bigoplus_{i \in S} x_i = 0)| \leq \epsilon,$$

where $\bigoplus$ denotes the XOR function and x_i denotes the ith bit of an n–bit string x picked uniformly at random from X. One property of such a sample space is that $\forall \ell, \ell = 1, 2, \ldots, k, \forall \{i_1, i_2, \ldots i_\ell\} \subseteq \{1, 2, \ldots, n\}, \forall b_1 b_2 \ldots b_\ell \in \{0, 1\}^\ell$,

$$|Pr(x_{i_1} = b_1, x_{i_2} = b_2, \ldots, x_{i_\ell} = b_\ell) - \frac{1}{2^\ell}| \leq \epsilon. \tag{13}$$

X is ϵ–biased if it is n–wise ϵ–biased. Constructions of k–wise ϵ–biased sources of size $poly(k, \log n, \frac{1}{\epsilon})$ were presented in [29, 2]. Such sample spaces have been shown to have several applications to explicit constructions and to derandomization, mainly since probabilistic analyses may be expected to be robust under small perturbations of the probabilities. Now, it is easy to see how our methods can be used to derive large deviation bounds for $x_1 + x_2 + \cdots + x_n$; from inequality (13), it follows that for a k–wise ϵ–biased source X,

$$\forall \ell \leq k \;\; E[S_\ell(x_1, x_2, \ldots, x_n)] \leq \binom{n}{\ell} \cdot \left(\frac{1}{2^\ell} + \epsilon\right)$$

and hence by picking $\epsilon \leq \frac{1}{2^\ell}$, this quantity can at most be double its unbiased value of $\binom{n}{\ell} \cdot \frac{1}{2^\ell}$. Thus, for a k–wise ϵ–biased random source with $k = h(n, n/2, \delta) = n\delta$ and with $\epsilon = 2^{-k}$,

$$Pr(\sum_{i=1}^{n} x_i \geq \frac{n}{2}(1 + \delta)) \leq 2 \cdot U_2(n, n/2, \delta). \tag{14}$$

Since such a source can be generated using $poly(k, \log n)$ random bits, we see that this result is optimal as long as $k = \Omega(\log n)$; if $k = O(\log n)$, then the probability space is polylogarithmic in size and should in most cases be dispensable, via brute-force search of the sample space. Similar results hold when the X_i's are binary with their probabilities of being one being the same negative power of two (not necessarily 1/2), using identical methods.

2.6 Upper Tail Bounds for some other Distributions

Suppose we have random bits $X_1, X_2, \ldots, X_n$ with some arbitrary distribution. Let $X = \sum_{i=1}^n X_i$, and let $\mu = E[X]$. Then, for any $a > \mu$, the methods of Section 2.1 yield

$$Pr(X \geq a) \leq \frac{\sum_{i=0}^{a} y_i E[S_i(X_1, X_2, \ldots, X_n)]}{\sum_{i=0}^{a} y_i \binom{a}{i}}, \;\forall (y_0, y_1, \ldots y_a) \in \Re_+^{a+1}.$$

The following theorem is immediate.

Theorem 7 *Given n random bits $X_1, X_2, \ldots, X_n$ with $X = \sum_{i=1}^n X_i$ and $\mu = E[X]$, suppose z_j is an upper bound on $E[S_j(X_1, X_2, \ldots, X_n)]$, $j = 1, 2, \ldots, n$. Then, if $a = \mu(1 + \delta)$ for $\delta > 0$,*

- $Pr(X \geq a) \leq \frac{\sum_{i=0}^{a} y_i z_i}{\sum_{i=0}^{a} y_i \binom{a}{i}}$, $\forall (y_0, y_1, \ldots y_a) \in \Re_+^{a+1}$.

- *If* $X_1, X_2, \ldots, X_n$ *are* k*–wise independent, then* $Pr(X \geq a) \leq min_{i=1,2,\ldots,k} \frac{z_i}{\binom{a}{i}}$.

As an example of a distribution which benefits from the above, consider the `self-weakening random variables` defined and used in [31]: random bits $X_1, X_2, \ldots, X_n$ are defined to be self–weakening with parameter λ in [31] if for all j and for all distinct indices $X_{i_1}, X_{i_2}, \ldots, X_{i_j}$, $E[\Pi_{\ell=1}^{j} X_{i_\ell}] \leq \lambda \Pi_{\ell=1}^{j} E[X_{i_\ell}]$; note that $z_j \leq \lambda \binom{n}{j} (\frac{\mu}{n})^j$ in this case. Hence, Theorem 7 directly implies one of the main theorems of [31], which states that if $X_1, X_2, \ldots, X_n$ are self–weakening random bits with parameter λ with $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$, then for any $\delta > 0$, $Pr(X \geq \mu(1+\delta))$ is at most λ times any Chernoff–Hoeffding type upper bound on the corresponding probability had the X_i been independent, with the same individual distributions. Indeed, it was the work of [31] which mainly motivated the methods of Section 2.1. Further, the applications sketched in Section 3.2 use Theorem 7.

Theorem 7 helps improve the known upper tail probability bounds for the `hypergeometric distribution`, an important source of self-weakening random variables. Suppose n balls are picked at random `without replacement` from an urn containing M red balls and $N - M$ balls of other colors. Let X be the number of red balls picked in the random sample, and let $p \doteq M/N$. Then for $\delta > 0$, a special case of a result of Hoeffding [17] (see Chvátal [13] for another proof) implies that

$$Pr(X \geq np(1+\delta)) \leq F(n, np, \delta). \tag{15}$$

We prove the following strengthened version of inequality (15).

Lemma 5 *Suppose a random set of* n *balls is picked from an urn containing* M *red balls and* $N - M$ *balls of other colors. If* X *denotes the number of red balls picked,* $p = M/N$, $\delta > 0$ *and if* $k \doteq h(n, np, \delta) = o(N)$, *then*

$$Pr(X \geq np(1+\delta)) \leq U_2(n, np, \delta) e^{-\Theta(k^2/M)} \leq F(n, np, \delta) e^{-\Theta(k^2/M)}.$$

Proof. (Sketch) Number the balls picked as $1, 2, \ldots, n$, and let X_i be the indicator r.v. for the event that ball i was red. Then $X = \sum_{i=1}^{n} X_i$ and

$$Pr(X \geq a) \leq U_3(n, np, \delta) \doteq \frac{E[S_k(X_1, X_2, \ldots, X_n)]}{\binom{a}{k}},$$

where $a \doteq \lceil np(1+\delta) \rceil$. For distinct indices $i_1, i_2, \ldots, i_k$, $E[\Pi_{j=1}^{k} X_{i_j}] = Pr(\bigwedge_{j=1}^{k} X_{i_j} = 1) = \prod_{i=0}^{k-1} \frac{M-i}{N-i}$; hence,

$$\frac{U_3(n, np, \delta)}{U_2(n, np, \delta)} = (\frac{N}{M})^k \cdot \prod_{i=0}^{k-1} (\frac{M-i}{N-i}) = \Pi_{i=1}^{k-1} (1 - (\frac{N}{M} - 1) \frac{i}{N-i}) \leq e^{-(\frac{N}{M}-1) \sum_{i=1}^{k-1} \frac{i}{N-i}}$$

which is $e^{-\Theta(k^2/M)}$, if $k = o(N)$. □

Remark. Note that sampling without replacement produces r.v.'s which are self–weakening with parameter 1. Lemma 5 gives good improvements over inequality (15) in many interesting cases, *e.g.*, consider the case p = constant, δ = constant, and $\Omega(M^{0.5+\epsilon}) \leq n = o(N)$, for any fixed $\epsilon > 0$.

Also, the CH bounds [11, 17, 35, 3] depend only on μ and not on the actual values of p_i, and give the upper bound $F(n, \mu, \delta) \geq U_2(n, \mu, \delta)$. We know for any $\delta > 0$ that $Pr(X \geq \mu(1+\delta)) \leq U_1(n, p_1, \ldots, p_n, \delta) = S_k(p_1, \ldots p_n)/\binom{\mu(1+\delta)}{k}$, where $k = h(n, \mu, \delta)$. By Lemma 2, this is maximized by setting $p_i = \mu/n \ \forall i$, subject to the constraint that $E[X] = \mu$. But if the values p_i are rather different, we get bounds formally superior to $U_2(n, \mu, \delta)$ and $F(n, \mu, \delta)$; suppose, for example, that $\mu = n/2$, $p_i = \epsilon$, $i = 1, 2, \ldots, n/2$, and $p_i = 1 - \epsilon$, $i = n/2 + 1, \ldots, n$, where $0 < \epsilon \leq 1/2$. Then,

$$\frac{U_1(n, p_1, p_2, \ldots, p_n, \delta)}{U_2(n, \mu, \delta)} \leq f(\epsilon) \doteq 2^k (\sum_{i=0}^{k} \binom{n/2}{i} \binom{n/2}{k-i} \epsilon^i (1-\epsilon)^{k-i}) / \binom{n}{k},$$

where $k = h(n, \mu, \delta)$. Note that $f(0) \leq e^{-\Theta(k^2/n)}$ by Lemma 5 and that $f(\epsilon)$ can get arbitrarily close to $f(0)$ since $f(\cdot)$ is continuous.

Remark. This particular result can only increase the constant factor in the exponent of $U_2(n, \mu, \delta)$. But, it is a small step towards better understanding of the dependence of $Pr(X \geq \mu(1+\delta))$ on $n, p_1, \ldots, p_n$, and δ. Similar improvements can also be made in the case of non–binary r.v.'s. An alternative approach might be to derive Chernoff-Hoeffding bounds for a sum of Bernoulli trials as a function of the variance as well as μ, a, and n, as in [42].

A final application is to the *semi-random* source introduced by Santha & Vazirani [38]. A random source which outputs bits $X_1, X_2, \ldots, X_n$ is defined to be ϵ-semirandom in [38] if

$$\forall i \ 1/2 - \epsilon \leq Pr(X_i = 1 | X_1, X_2, \ldots, X_{i-1}) \leq 1/2 + \epsilon,$$

i.e., the random bits can be correlated, but only to a limited extent, independent of the past history. Despite its seemingly weak nature, such a model has been shown to be able to simulate complexity classes such as RP (Vazirani & Vazirani [50]), and the study of a generalization of this model due to Zuckerman [53] has led to rich results recently (Nisan & Zuckerman [30], Wigderson & Zuckerman [52]). Noting that for such a source,

$$E[\prod_{j=1}^{k} X_{i_j}] \leq (1/2 + \epsilon)^k$$

for all $k \geq 1$ and for all distinct indices $i_1, i_2, \ldots, i_k$, we see that

$$Pr(\sum_{i=1}^{n} X_i \geq n(1/2 + \epsilon)(1 + \delta)) \leq U_2(n, n(1/2 + \epsilon), \delta), \ \forall \delta > 0,$$

for an ϵ-semirandom source.

Inequality (14) shows another application of our techniques.

3 Applications to Computation

The most striking point of Theorems 1 and 5 in our opinion is that bounds as good as the CH bounds can be obtained with small independence. This implies, for any analysis that relies on the CH bounds, much weaker requirements on the random sources used. We now present some further computational applications of the new results.

3.1 Reduced independence for randomized algorithms

There are known constructions of r.v.'s with limited independence using a small number of random bits; for example, the construction of [19] and the use of universal hash functions [9] to generate $|F|$ many k–wise independent random elements from a finite field F using $O(k \log |F|)$ random bits, and the result of [1] using coding techniques [25], which gives a polynomial (in n) time algorithm to construct n k–wise independent and unbiased random bits, given $O(k \log n)$ independent unbiased bits for any k, $k \leq n$. Combining these with our result on reduced independence for the CH bounds, we get a major reduction in the amount of randomness needed for several randomized algorithms.

3.1.1 Reduced randomness for random sampling

In random sampling, we have a huge finite universe U and a subset $W \subseteq U$, and we want to estimate the fraction $f^* = |W|/|U|$. Given error parameters $\delta, \epsilon > 0$, the method used is to pick a random sample S of size $N(\delta, \epsilon)$ from U and output the fraction f of type W elements in S; $N(\delta, \epsilon)$ must be such that $Pr(|f^* - f| \geq \delta) \leq \epsilon$. This is required, for instance, in PAC learning [47] and in running BPP algorithms. What was known so far is that $N(\delta, \epsilon) = O(\frac{1}{\delta^2} \log(\frac{1}{\epsilon}))$ with all the samples being independent. We can improve this to

Theorem 8 *Given a universe U, a subset $W \subseteq U$, and error parameters $\delta, \epsilon > 0$, suppose a set S of $O(\frac{1}{\delta^2} \log(\frac{1}{\epsilon}))$ random samples with $O(\log(\frac{1}{\epsilon}))$–wise independence are picked from U. Then, if f^* and f are the respective fractions of type W elements in U and S, $Pr(|f^* - f| \geq \delta) \leq \epsilon$ will hold.*

Proof. Consider a randomized algorithm which looks at a random set of samples S from U, and outputs the ratio f of type W elements in S. We now look at the random properties of S which are required for the claim $Pr(|f^* - f| \geq \delta) \leq \epsilon$ to hold.

Let $|S| = n$. In the notation of Theorem 5, we want to claim that $Pr(|X - \mu| \geq \delta' \mu) \leq \epsilon$, where $\mu = E[X]$, $X = nf$ and $\delta' = \delta / f^*$. We apply Theorem 5 with $X = fn$ and $\mu = f^* n$, and choose k^* consistent with (Ia) and (IIa). For such a suitable choice of independence k^* among the elements of S, $Pr(|X - \mu| \geq \delta' \mu)$ can be bounded by $e^{-\lfloor k^*/2 \rfloor}$. Hence, it suffices to choose $k^* \geq 2\lceil \ln(\frac{1}{\epsilon}) \rceil$, for the above probability to be bounded by ϵ. To compute the required sample size n, we distinguish between the two cases $\delta' \leq 1$ and $\delta' > 1$. Then, from part (Ia) of Theorem 5, the probability that $|X - \mu|$ is greater than $\delta' \mu$ is bounded by $e^{-\lfloor k^*/2 \rfloor}$ provided that $k^* \leq \lfloor \delta'^2 \mu e^{-1/3} \rfloor$, and $k^* \geq 2\lceil \ln(\frac{1}{\epsilon}) \rceil$ for this probability to be bounded by ϵ. It therefore suffices to choose $\delta'^2 \mu e^{-1/3} \geq 2\lceil \ln(\frac{1}{\epsilon}) \rceil$. This will hold if $n\delta^2 e^{-1/3}/(2f^*) \geq \lceil \ln(\frac{1}{\epsilon}) \rceil$. Since $e^{-1/3}/2 > 1/3$, it suffices to choose $n \geq N_1(\delta, \epsilon) = \frac{3f^*}{\delta^2} \lceil \ln(\frac{1}{\epsilon}) \rceil$. If $\delta' > 1$, then a similar analysis using Theorem 5.IIa implies that $N_2(\delta, \epsilon) = \frac{3}{\delta} \lceil \ln(\frac{1}{\epsilon}) \rceil$ many samples with $2\lceil \ln(\frac{1}{\epsilon}) \rceil$–wise independence suffice to satisfy the error bounds.

Note that since both f^* and δ are clearly bounded by 1, the number of samples and independence needed in both the above cases can be upper bounded by $N_3(\delta, \epsilon) = \frac{3}{\delta^2} \lceil \ln(\frac{1}{\epsilon}) \rceil$ and $k^* = 2\lceil \ln(\frac{1}{\epsilon}) \rceil$. Note further though that by Theorem 5 the choice for the independence that minimizes the error bound is an increasing function of the sample size, increasing the sample space size when given a fixed independence will reduce the error probability; a proof of this claim follows. In the proof of Theorem 5, parts(I) and (II) were derived from part(III) of Theorem 4 with $C = \mu$. Note that in the current problem, $C = nf^*$ and $T = n\delta$ where n is the number of random samples picked, in the notation of Theorem 4. When the independence k is fixed, the bound given by part(III) of Theorem 4 decreases with n for these values of C and T, and the claim follows. □

Theorem 4 also allows an estimate for the required size of a sample space with k–wise independent variables, in case $k < 2\lceil \ln(\frac{1}{\epsilon}) \rceil$.

Theorem 9 *Given a universe U, a subset $W \subseteq U$, and error parameters $\delta, \epsilon > 0$, suppose that S is a sample space of U whose elements are k-wise independent, for some even k. Then, if f^* and f are the respective fractions of of type W elements in U and S, then for $Pr(|f^* - f| \geq \delta) \leq \epsilon$ to hold, it is sufficient to choose $|S| \geq \frac{ck}{\delta^2 \epsilon^{(2/k)}}$, for some constant c.*

Proof. Use part(III) of Theorem 4 by setting $C = n$, where $n = |S|$. □

Theorems 8 and 9 imply "reduced randomness" results for random sampling, if the universe U has some properties. For instance, if U is a finite field and if the field operations can be done in polynomial (in $\frac{1}{\delta}$ and $\log(\frac{1}{\epsilon})$) time, then any number of k–wise independent samples from U can be generated from k independent random samples [19, 9]. Also, via weaker bounds on the kth moment, it has been independently shown in [7] that essentially the same bounds as those given in Theorem 8 can be obtained for random sampling; they also show how iterated sampling can be used to decrease the number of random bits, at the expense of a controlled increase in the sample size.

The above constructions are not optimal with regards to the minimum number of random bits used. Using random walks on expander graphs to generate the random bits, it is shown in [6] that $O(\log(|U|) + \log(\frac{1}{\delta}))$ random bits are necessary and sufficient for this problem. Our construction has the advantage of being elementary and parallelizable.

3.1.2 Reduced randomness for oblivious permutation routing

We now show how our results directly imply bounds that match the explicit constructions of algorithms with reduced randomness due to Peleg & Upfal [32], for *oblivious permutation routing on fixed interconnection networks* (see also [20, 35, 36, 46, 48]).

Given some interconnection network with N nodes and a permutation $\sigma : \{1, 2, \ldots, N\} \to \{1, 2, \ldots, N\}$, the problem is to route a packet ν_i residing at each node i, to its destination $\sigma(i)$ so that the total time taken is small. Further, the routing must be *oblivious* in that the path $P_\sigma(x)$ chosen for a packet x must be "independent" of the path $P_\sigma(y)$ chosen for any other packet y (see [32] for a precise definition when randomized routing protocols are allowed). Explicit constructions of algorithms with a spectrum of time–randomness parameters are among the results proved in [32] for the degree–4 butterfly network; these are also extendible to other networks (see Karloff & Raghavan [20] for a protocol for the hypercube with slightly weaker bounds). Here, we show how our results of Section 2.1 directly imply the bounds of [32] for the hypercube; we believe that similar results should hold for other interconnection networks.

Consider the implementation of Valiant's two–phase scheme [46] (see also Valiant and Brebner [48]) on a hypercube with $N = 2^n$ nodes: (I) Each vertex i picks a random $\rho(i) \in \{1, 2, \ldots, N\}$ as an *intermediate destination* for ν_i, and routes ν_i there; (II) Each packet ν_i is routed to its final destination $\sigma(i)$. We now follow the discussion of the standard aspects of this from [35]. Assume FIFO queues at each edge, and that phase(I) routes ν_i from i to $\rho(i)$ by "correcting" its bits from left to right assuming that the nodes of the hypercube are indexed by n bits, and that phase(II) "corrects" bits right to left. So, phase(II) is like "running phase(I) backwards", and so we consider phase(I) alone here. It is shown in [35] that the time taken for packet ν_i in phase(I) is at most

$$n + \sum_{j=1}^{N} H_{ij}, \tag{16}$$

where $H_{ij} = 1$ if the paths $< i, \rho(i) >$ and $< j, \rho(j) >$ share an edge in phase(I), and 0 otherwise (recall that $n = \log_2 N$). It is also shown in [35] that if each $\rho(i)$ is uniformly distributed in $\{1, 2, \ldots, N\}$, then $\forall i$, $E[\sum_{j=1}^{N} H_{ij}] \leq n$. Here is the theorem that matches the explicit construction of [32].

Theorem 10 *There are explicit constructions of oblivious routing algorithms on the hypercube which, for any T, $c \log N \leq T \leq \sqrt{N}$ ($c > 4$ is a constant):*

1. *use $O(\frac{\log(N/Q)}{\log(T/\log N)} \log N)$ random bits and terminate in T steps with probability at least $1 - Q$ for any $0 < Q \leq 1$;*

2. *use $O(\frac{\log^2 N}{\log(T/\log N)})$ random bits and terminate in expected time at most T.*

Proof. (Sketch.) Consider any packet ν_i; the probability that it takes more than $T/2$ steps in phase(I) is at most $Pr(\sum_{j=1}^{N} H_{ij} \geq T/2 - \log N)$. If the $\rho(i)$s are picked uniformly and in k–wise independent fashion, then the H_{ij} are $(k-1)$-wise independent, while $E[\sum_{j=1}^{N} H_{ij}] \leq \log N$, as before. It follows from our discussion of Section 2.1 and from Lemma 3 that, if $T > E[\sum_{j=1}^{N} H_{ij}]$, (*i.e.,* if $T > 4 \log N$), then

$$Pr(\sum_{j=1}^{N} H_{ij} \geq T/2 - \log N) \leq \frac{\binom{N}{k-1}(\frac{\log N}{N})^{k-1}}{\binom{T/2 - \log N}{k-1}} \leq \frac{(\log N)^{k-1}}{\Pi_{j=0}^{k-2}(T/2 - \log N - j)}.$$

By picking $k = \Theta(\frac{\log(N/Q)}{\log(T/\log N)})$, we can ensure that $Pr(\sum_{j=1}^{N} H_{ij} \geq T/2 - \log N) \leq Q/(2N)$ holds. Arguing similarly for phase(II) and summing up over all i, we get (1) above.

For (2) above, we set $Q = 1/(2N)$ and replace T by $T - 1$ in (1). Let T_{max} be a random variable denoting the time taken by the protocol, *i.e.*, the maximum, over all packets i, of the number of steps taken by packet ν_i to reach $\sigma(i)$. Note that $T_{max} \leq \log N + N$, from the upper bound (16). Also,

$$Pr(T_{max} > (T-1)) \leq Q,$$

from (1) above. Hence,

$$\begin{aligned} E[T_{max}] &\leq (T-1) \cdot Pr(T_{max} \leq (T-1)) + (\log N + N) \cdot Pr(T_{max} > (T-1)) \\ &\leq (T-1)(1 - \frac{1}{2N}) + (\log N + N)\frac{1}{2N} \\ &\leq T. \end{aligned}$$

Note that for any k, k–wise independent $\rho(i)$s can be generated from $k \log N$ random bits using hash functions [9], since the $\rho(i)$s can be thought of as belonging to the field $GF(2^n)$. Hence, we get bounds that match those of [32]. □

The above example typifies the type of application we expect our methods to find, *i.e.*, as direct "plug–in"s in analyses where the CH bounds are normally used.

3.2 The New Formulation and the Method of Conditional Probabilities

The `method of conditional probabilities` [34, 44] is an important technique for the derandomization of algorithms; the reader is referred to [35] for details. We now show how this method can be combined with the formulation of Section 2.6. This will enable us to derive simple and efficient deterministic polynomial–time algorithms from randomized algorithms which can be analyzed using our formulation, *in a unified way.*

Given n random bits $X_1, X_2, \ldots, X_n$, can the conditional expectation

$$E[S_k(X_1, X_2, \ldots, X_n) | X_1 = b_1, X_2 = b_2, \ldots, X_i = b_i]$$

be evaluated, or at least be given a "reasonable" upper bound, for any k, any i, $i = 0, 1, 2, \ldots, n-1$, and any $b_1 b_2 \ldots b_i \in \{0,1\}^i$? If the X_i's are identically distributed, then it is reasonable to assume that an upper bound U_ℓ on $E[\Pi_{r=1}^{\ell} X_{j_r} | X_1 = b_1, X_2 = b_2, \ldots, X_i = b_i]$ is known for all ℓ, $\ell = 1, 2, \ldots, n-i$, and for all distinct indices $X_{j_1}, X_{j_2}, \ldots, X_{j_\ell} \in \{i+1, i+2, \ldots, n\}$; this is sufficient for the two applications shown below. Then, if

$$|\{j \mid (1 \leq j \leq i) \ \wedge \ (b_j = 1)\}| = i_1,$$

we can see that

$$E[S_k(X_1, X_2, \ldots, X_n) | X_1 = b_1, X_2 = b_2, \ldots X_i = b_i] \leq \sum_{r=0}^{min(i_1,k)} \binom{i_1}{r} \binom{n-i}{k-r} U_{k-r}. \qquad (17)$$

We now present two applications where the combination of our formulation and the method of conditional probabilities leads to fast polynomial–time algorithms, via upper bound(17). The first application, to jobshop scheduling, is a "natural" derandomization of the randomized algorithm of [41], faster than the derandomization techniques of [41] and [33]; this is shown in Section 3.2.1. The second application is to discrepancy theory, and is discussed in Section 3.2.2. The "usual" method of conditional probabilities for these problems frequently calls for independence among the random variables corresponding to the bits $X_1, X_2, \ldots, X_n$ seen above; this is not the case for these two problems and in general for many other problems.

3.2.1 Improved algorithms for packet routing and jobshop scheduling

We now present simpler approximation algorithms for `packet routing` (Leighton, Maggs & Rao [23]) and `jobshop scheduling` (Shmoys, Stein & Wein [41]) which provide improved approximation guarantees, by using ideas from above. The *non-preemptive jobshop scheduling* problem is as follows: given n jobs, m machines and a sequence of `operations` for each job where each operation is assigned to a specific machine, construct a schedule to run the jobs so that the time taken to process all the jobs is minimized, subject to: (i) the operations of each job must be done in sequence; (ii) no operation of any job running on any machine can be preempted till it is completed, and (iii) a machine can process at most one operation at a time. One of the results of [23] tackles a special case of this problem; the general case is handled in [41]. Both these papers give polynomial–time algorithms to produce good approximations to an optimal schedule.

Let P_i be the total time needed for job J_i, and let $P_{max} = max_{i \in [1,n]} P_i$. Let Π_j be the total time for which machine M_j is needed, and let $\Pi_{max} = max_{i \in [1,m]} \Pi_i$. Before an actual schedule is constructed in [41], a pseudo–schedule $\mathcal{S}$ is constructed which temporarily assumes that each machine can work on upto D operations simultaneously, where $D > 1$ depends on the input instance. The pseudo–schedule is later used to construct an actual schedule. The only step where randomization is used in [41] is during the construction of the pseudo–schedule and is the following.

An initial random delay $d_i \in \{1, 2, \ldots, \Pi_{max}\}$ is assigned for each job J_i. Suppose that the sequence of operations of job J_i are $O_{i,1}, \ldots O_{i,r_i}$, and that operation $O_{i,r}$ takes time $t_{i,r}$; then, in the pseudo–schedule $\mathcal{S}$, job J_i is scheduled to start at time d_i and runs to completion without interruption, *i.e.*, operation $O_{i,r}$ starts at time $d_i + \sum_{\ell=1}^{r-1} t_{i,\ell}$. We denote the offset $\sum_{\ell=1}^{r-1} t_{i,\ell}$ by $\tau(O_{i,r})$. As shown in [23, 41], if the d_i's are generated uniformly and independently, then with high probability, every machine at every unit of time will have (a congestion of) at most $D(n, m_{max}) \doteq c \cdot \frac{\log(n \cdot m_{max})}{\log\log(n \cdot m_{max})}$ jobs scheduled on it for some constant c, where m_{max} is the maximum number of operations in any job. This step is then derandomized to deterministically compute initial delays leading to a congestion bound of $O(\log(n \cdot m_{max}))$. Linear programming is used for the derandomization, making this step the bottleneck. This step is sped up in [33, 45]. Here, we get a better congestion bound of $D(n, m_{max})$ as opposed to the previously known $O(\log(n \cdot m_{max}))$ bound, with an algorithm which is more direct than the ones of [33, 45], while having time complexities comparable to theirs.

We assign random initial delays $\{d_i \in \{1, 2, \ldots, \Pi_{max}\}\}$ uniformly and independently to the jobs. Suppose that the operations scheduled on machine M_i are $O_1, \ldots O_{m_i}$, which respectively belong to jobs $J_{i_1}, J_{i_2}, \ldots, J_{i_{m_i}}$ and take $t_1, \ldots, t_{m_i}$ units of time. For any machine M_i and time instance $t \in \{1, 2, \ldots, \Pi_{max} + P_{max}\}$, we define $\sum_{\ell \in [1,m_i]} t_\ell = \Pi_i$ many indicator r.v.'s $X_j^i(t)$, $j = 1, 2, \ldots, \Pi_i$, to analyze the congestion on machine M_i at time t in $\mathcal{S}$; each of these r.v.'s is an indicator for the event that a particular unit of time of some operation gets scheduled on M_i at time t in $\mathcal{S}$, as follows. The index j encodes the time unit and operation: let $j_r = \sum_{\ell=1}^{r-1} t_\ell$ for $r = 1, 2, \ldots, m_i$; then if $j_r < j \leq j_{r+1}$, j represents the $j - j_r$th time unit of O_r as follows.

$$X_j^i(t) = \begin{cases} 1 & \text{if } j = j_r + p,\ 1 \leq p \leq t_r, \text{ and the } p\text{th time unit of } O_r \text{ is scheduled for time } t, \\ & \textit{i.e.}, \text{ if } d_{i_r} + \tau(O_r) + p - 1 = t; \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $E[X_j^i(t)] \leq \frac{1}{\Pi_{max}}$, and that for any positive integer k, the probability that machine M_i has congestion at least k at time unit t is

$$Pr(\sum_{r=1}^{\Pi_i} X_r^i(t) \geq k) \leq E[S_k(X_1^i(t), \ldots, X_{\Pi_i}^i(t)) \leq \binom{\Pi_i}{k} \left(\frac{1}{\Pi_{max}}\right)^k.$$

In addition, if $\sum_{r=1}^{\Pi_i} X_r^i(t) \geq k$ holds for some time t, then $\sum_{r=1}^{\Pi_i} X_r^i(t') \geq k$ also holds for some time t', where t' is one of the starting times of the operations scheduled on M_i. Further, the starting time of each operation O_r is uniformly distributed in $[\tau(O_r), \Pi_{max} - 1 + \tau(O_r)]$. Hence, for any k, Pr(some machine has congestion at least k at some time instance) is at most

$$\sum_{i=1}^{m} \sum_{r=1}^{m_i} \sum_{t=\tau(O_r)}^{\Pi_{max}-1+\tau(O_r)} \frac{1}{\Pi_{max}} E[S_{k-1}(X_1^i(t), \ldots, X_{\Pi_i}^i(t))], \tag{18}$$

which is at most

$$\sum_{i=1}^{m} m_i \frac{\Pi_{max}}{\Pi_{max}} \binom{\Pi_i}{k-1} \left(\frac{1}{\Pi_{max}}\right)^{k-1} \leq \sum_{i=1}^{m} \frac{m_i}{(k-1)!}.$$

Clearly $\sum_{i=1}^{m} m_i \leq n \cdot m_{max}$, hence for $k - 1 > k^* = c_1 \frac{\log(n \cdot m_{max})}{\log\log(n \cdot m_{max})}$ for some suitable constant c_1, the above probability estimate is less than one. We may now use the above form as a `pessimistic estimator` [34] to deterministically set the delays d_i for the jobs one–by–one by the method of conditional probabilities [34, 44], to achieve the congestion bound of $D(n, m_{max})$.

Assume inductively that initial delays $d_1 = d_1^*, d_2 = d_2^*, \ldots, d_s = d_s^*$ have been set deterministically for the jobs $J_1, J_2, \ldots, J_s$; the aim is to compute d_{s+1}^* now. Consider any machine M_i on which J_{s+1} has at least one operation; let these operations be $A_{s+1,1}, A_{s+1,2}, \ldots, A_{s+1,a_i}$. Let $B_{s,1}, B_{s,2}, \ldots, B_{s,b_i}$ be the operations which belong to some job in $\{J_1, J_2, \ldots, J_s\}$ and which are scheduled on M_i, and let $t_{s,1}, t_{s,2}, \ldots, t_{s,b_i}$ be the times at which they are scheduled to start on M_i; these times are known, since we know the values of $d_1, d_2, \ldots, d_s$. Let $O_1, O_2, \ldots, O_{c_i}$ be the operations on machine M_i that belong to jobs from the set $\{J_{s+2}, J_{s+3}, \ldots, J_n\}$. We define, for any $t \in \{1, 2, \ldots, P_{max} + \Pi_{max}\}$ and $r \in \{1, 2, \ldots, \Pi_{max}\}$,

$$g(s+1, i, t, r) \doteq E[S_{k^*}(X_1^i(t), X_2^i(t), \ldots, X_{\Pi_i}^i(t)) | d_1 = d_1^*, d_2 = d_2^*, \ldots, d_s = d_s^*, d_{s+1} = r],$$

and $num(i, t, < x_1, x_2, \ldots, x_j >)$ to be the number of operations from jobs $J_1, J_2, \ldots, J_j$ that are scheduled on machine M_i at time t, given that $d_1 = x_1, \ldots, d_j = x_j$.

When conditioned on the event $d_1 = d_1^*, d_2 = d_2^*, \ldots, d_s = d_s^*, d_{s+1} = \Delta$ for any $\Delta \in \{1, 2, \ldots, \Pi_{max}\}$, upper bound(18) becomes

$$f(s+1, \Delta) \doteq \sum_{i=1}^{m} (\sum_{j=1}^{a_i} g(s+1, i, \tau(A_{s+1,j}) + \Delta, \Delta) + \sum_{j=1}^{b_i} g(s+1, i, t_{s,j}, \Delta) + \sum_{j=1}^{c_i} \sum_{t=\tau(O_j)}^{\tau(O_j)+\Pi_{max}-1} \frac{1}{\Pi_{max}} g(s+1, i, t, \Delta)).$$

Recall that the method of conditional probabilities will set $d_{s+1} = d_{s+1}^*$, where d_{s+1}^* is the index at which $f(s+1, \cdot)$ is minimized. Note that $f(s+1, \Delta)$ can be readily computed for any Δ and hence, so can d_{s+1}^*. To make the computation more efficient, we use the following observations.

1. Suppose we need to compute $f(s+1, \Delta)$ for some Δ, using upper bound (17). Then, for any machine M_j which has some operation from job J_{s+1}, the term "$n - i$" in upper bound (17) corresponds to the number of operations of jobs $J_1, J_2, \ldots, J_{s+1}$ on machine M_j, *i.e.*, $a_j + b_j$. Hence, for any $t \in \{1, 2, \ldots, P_{max} + \Pi_{max}\}$, $g(s+1, j, t, \Delta)$ can be computed in $O(k^*)$ time if $num(j, t, < d_1^*, d_2^*, \ldots, d_s^*, \Delta >)$ is known, since upper bound(17) involves a sum over at most k^* terms (recall that k^* is an upper bound on the number of operations scheduled at the same time).

2. Suppose inductively that $num(i, t, < d_1^*, d_2^*, \ldots, d_s^* >)$ is known, for all machines M_i and for all times t. Let ω_{s+1} be the number of operations of job J_{s+1}. We will consider only those machines that have some operation of J_{s+1}; the number of such machines is clearly at most ω_{s+1}. Hence, given the $num(i, t, < d_1^*, d_2^*, \ldots, d_s^* >)$ values, the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, 1 >)$ values need to be updated for at most ω_{s+1} machines and for $P_{max} + \Pi_{max}$ time units, and can be done in $O(\omega_{s+1}(P_{max} + \Pi_{max}))$ time. Given the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, 1 >)$ values, computation of $f(s+1, 1)$ takes $O(\omega_{s+1} k^*(P_{max} + \Pi_{max}))$ time, since $g(\cdot, \cdot, 1)$ can be computed in $O(k^*)$ time, given these values.

3. Suppose we have computed the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, \Delta >)$ values and $f(s+1, \Delta)$ for some value Δ, and that we need to compute $f(s+1, \Delta+1)$. We can proceed to first compute the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, \Delta+1 >)$ values and then $f(s+1, \Delta+1)$, as follows. Suppose some operation α of J_{s+1} is scheduled to run on some machine M_i from time t_1 to time t_2, when we set $d_{s+1} = \Delta$. Then, note that
$$num(i, t', < d_1^*, d_2^*, \ldots, d_s^*, \Delta+1 >) = num(i, t', < d_1^*, d_2^*, \ldots, d_s^*, \Delta >), \forall t' \in [t_1+1, t_2].$$
Hence, this operation α leaves at most two of the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, \Delta+1 >)$ values different from the corresponding $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, \Delta >)$ values. Hence, $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, \Delta+1 >)$ can be updated in $O(\omega_{s+1})$ time. It now follows from arguments similar to those used above that $f(s+1, \Delta+1)$ can be computed in $O(k^* \omega_{s+1})$ time.

The above observations imply an efficient algorithm to compute d_{s+1}: inductively maintain the $num(\cdot, \cdot, < d_1^*, d_2^*, \ldots, d_s^*, r >)$ values as r goes from $1, 2, \ldots,$ to Π_{max}, and compute $f(s+1, 1), f(s+1, 2), \ldots, f(s+1, \Pi_{max})$ in that order by sequentially updating the corresponding num values. Since computing d_{s+1}^* takes $O(\omega_{s+1} k^*(P_{max} + \Pi_{max}))$ time, the total time complexity is $O(\omega k^*(P_{max} + \Pi_{max}))$, where ω is the total number of operations. Hence, we have

Theorem 11 *Initial delays $\{d_i : 1 \leq i \leq n\}$ in the range $\{1, 2, \ldots, \Pi_{max}\}$ for each job J_i can be set in $O((P_{max} + \Pi_{max})\omega \frac{\log(n \cdot m_{max})}{\log\log(n \cdot m_{max})})$ time where ω is the total number of operations, such that in the (infeasible) schedule in which every job J_i starts at time d_i and runs without interruption, every machine has at most $O(\frac{\log(n \cdot m_{max})}{\log\log(n \cdot m_{max})})$ jobs scheduled on it at any time.*

We feel that the above is a natural derandomization of the randomized algorithm since it sets the delays one–by–one, as opposed to the more complex ways used before.

3.2.2 Exact Partitions in Set Discrepancy

Set discrepancy problems [3] are combinatorially important, special cases of which can model divide–and–conquer situations; see, *e.g.*, the RNC edge coloring algorithm of Karloff & Shmoys [21]. Given a finite set X and a family of subsets $\mathcal{F} = \{S_1, S_2, \ldots, S_n\}$ of X, the goal is to come up with a "2–coloring" $\chi : X \rightarrow \{0, 1\}$ such that the `discrepancy` $disc(\chi) \doteq max_i\ disc_i(\chi)$, where $disc_i(\chi) \doteq \{|(\sum_{j \in S_i} \chi(j)) - |S_i|/2|\}$, is minimized. It is known that a 2–coloring χ with $disc(\chi) = O(\sqrt{\Delta \log n})$ exists and can be computed in polynomial (in $|X|$ and n) time [3], and that a 2–coloring χ with $disc(\chi) = O(\Delta^{0.5+\epsilon}\sqrt{\log n})$ for any fixed $\epsilon > 0$ can be computed in NC [8, 27, 29], where $\Delta \doteq max_i|S_i|$. Using the ideas of Section 2.1, we can prove

Theorem 12 *Given a finite set X with $|X|$ even and a family of subsets $\mathcal{F} = \{S_1, S_2, \ldots, S_n\}$ of X such that $\Delta = max_i |S_i|$, there exists a 2–coloring $\chi^* : X \rightarrow \{0,1\}$ computable in polynomial (in $|X|$ and n) time, such that: (i) $disc(\chi^*) = O(\sqrt{\Delta \log n})$, and (ii) $|\{y \in X : \chi^*(y) = 0\}| = |\{y \in X : \chi^*(y) = 1\}|$.*

Proof. For the existence proof, we can show that if we pick a random subset $Z \subseteq X$ with $|Z| = |X|/2$ uniformly from the set of all size $|X|/2$ subsets of X and set $\chi_Z(y) = 1$ iff $y \in Z$, then $Pr_Z(disc(\chi_Z) = O(\sqrt{\Delta \log n})) > 0$, as follows. It is well–known [3] and easily checked via the CH bounds that there is a constant $c > 0$ such that if $\chi(y)$ is picked uniformly and independently from $\{0,1\}$, then for any S_i, $Pr(disc_i(\chi) > c\sqrt{\Delta \log n}) < \frac{1}{n}$; hence,

$$Pr(disc(\chi) > c\sqrt{\Delta \log n}) = Pr(\exists i: \ disc_i(\chi) > c\sqrt{\Delta \log n}) < n \cdot \frac{1}{n} = 1.$$

Note that $\{\chi_Z(y) : \ y \in X\}$ is a set of self–weakening random bits with parameter 1, *i.e.*, if Z is picked uniformly at random from the set of $|X|/2$ sized subsets of X, then for any distinct $y_1, y_2, \ldots, y_i \in X$,

$$E[\Pi_{j=1}^{i} \chi_Z(y_j)] = Pr(\chi_Z(y_1) = \chi_Z(y_2) = \cdots = \chi_Z(y_i) = 1) \leq \Pi_{j=1}^{i} Pr(\chi_Z(y_j) = 1) = \Pi_{j=1}^{i} E[\chi_Z(y_j)].$$

Hence, it follows from Section 2.6 that for any S_i, $Pr(disc_i(\chi_Z) > c\sqrt{\Delta \log n}) < \frac{1}{n}$ still holds, concluding the existence proof.

Assume that $|X| = m$ and that $X = \{1, 2, \ldots, m\}$. To use the method of conditional probabilities to derandomize the above construction, note that for $\{i_1, i_2, \ldots, i_j\} \subseteq \{i+1, i+2, \ldots, m\}$,

$$\begin{aligned} E[\Pi_{\ell=1}^{j} \chi_Z(i_\ell) | \chi_Z(1) = b_1, \ldots, \chi_Z(r) = b_r] &= \text{undefined, if } r_1 > m/2 \text{ or } r_2 > m/2 \\ &= 0, \text{ if } r_1 + j > m/2 \\ &= \frac{\binom{m-r-j}{m/2-r_1-j}}{\binom{m-r}{m/2-r_1}}, \text{ otherwise} \end{aligned}$$

where

$$|\{\ell \mid (1 \leq \ell \leq r) \ \wedge \ (b_\ell = 1)\}| = r_1$$

and $r_2 = r - r_1$. This can now be derandomized, by our initial discussion in Section 3.2. □

Acknowledgments

Aravind's sincere thanks to David Shmoys, for his constant encouragement, guidance and perceptive suggestions. Aravind also thanks Suresh Chari, Alessandro Panconesi and Pankaj Rohatgi for valuable discussions. We also thank Mihir Bellare for pointing out references [6] and [7] to us.

References

[1] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7:567–583, 1986.

[2] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k–wise independent random variables. *Random Structures and Algorithms*, 3(3):289–303, 1992.

[3] N. Alon, J. Spencer, and P. Erdős. *The Probabilistic Method.* Wiley–Interscience Series, John Wiley & Sons, Inc., New York, 1992.

[4] D. Angluin and L.G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18:155–193, 1979.

[5] Y. Azar, R. Motwani, and J. Naor. Approximating arbitrary probability distributions using small sample spaces. Manuscript, 1990.

[6] M. Bellare, O. Goldreich, and S. Goldwasser. Randomness in interactive proofs. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 563–573, 1990.

[7] M. Bellare and J. Rompel. Randomness efficient sampling of arbitrary functions. Technical Report MIT/LCS/TM-433.b, Laboratory for Computer Science, Massachusetts Institute of Technology, July 1990.

[8] B. Berger and J. Rompel. Simulating $(\log^c n)$-wise independence in NC. *J. Assoc. Comput. Mach.*, 38(4):1026–1046, 1991.

[9] J. L. Carter and M. N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences*, 18:143–154, 1979.

[10] S. Chari, P. Rohatgi, and A. Srinivasan. Improved algorithms via approximations of probability distributions. In *Proc. ACM Symposium on Theory of Computing*, 1994. To appear.

[11] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[12] B. Chor, O. Goldreich, J. Håstad, J. Friedman, S. Rudich, and R. Smolensky. The bit extraction problem or t–resilient functions. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 396–407, 1985.

[13] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25:285–287, 1979.

[14] G. Even, O. Goldreich, M. Luby, N. Nisan, and B. Veličković. Approximations of general independent distributions. In *Proc. ACM Symposium on Theory of Computing*, pages 10–16, 1992.

[15] B.V. Gladkov. Sums of random variables, any r of which are independent. *Mat. Zametkii*, 32:385–399, 1982.

[16] B.V. Gladkov. A central limit theorem for sums of random variables, any r of which are independent. *Diskretnaia Mat.*, 1:22–28, 1989. English translation by V.A. Vatutin, in Discrete Mathematics and Applications, No. 1 (1991), pages 73–79.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, pages 13–30, 1963.

[18] M. Hofri. *Probabilistic Analysis of Algorithms.* Springer–Verlag, 1987.

[19] A. Joffe. On a set of almost deterministic k–independent random variables. *The Annals of Probability*, 2(1):161–162, 1974.

[20] H. J. Karloff and P. Raghavan. Randomized algorithms and pseudorandom numbers. In *Proc. ACM Symposium on Theory of Computing*, pages 310–321, 1988.

[21] H. J. Karloff and D. B. Shmoys. Efficient parallel algorithms for edge coloring problems. *Journal of Algorithms*, 8:39–52, 1987.

[22] C. Kruskal, L. Rudolph, and M. Snir. A complexity theory of efficient parallel algorithms. *Theoretical Computer Science*, 71:95–132, 1990.

[23] F. T. Leighton, B. Maggs, and S. Rao. Universal packet routing algorithms. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 256–269, 1988.

[24] N. Linial and N. Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.

[25] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error Correcting Codes.* North–Holland, Amsterdam, 1977.

[26] K. Mehlhorn and U. Vishkin. Randomized and deterministic simulations of PRAMs by parallel machines with restricted granularity of parallel memories. *Acta Informatica*, 21:339–374, 1984.

[27] R. Motwani, J. Naor, and M. Naor. The probabilistic method yields deterministic parallel algorithms. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 8–13, 1989.

[28] S.V. Nagaev and I.F. Pinelis. Some inequalities for the distribution of the sums of independent random variables. *Theory of Probability and its Applications*, 22:248–256, 1977.

[29] J. Naor and M. Naor. Small–bias probability spaces: efficient constructions and applications. In *Proc. ACM Symposium on Theory of Computing*, pages 213–223, 1990.

[30] N. Nisan and D. Zuckerman. More deterministic simulation in Logspace. In *Proc. ACM Symposium on Theory of Computing*, pages 235–244, 1993.

[31] A. Panconesi and A. Srinivasan. Fast randomized algorithms for distributed edge coloring. In *Proc. ACM Symposium on Principles of Distributed Computing*, pages 251–262, 1992.

[32] D. Peleg and E. Upfal. A time–randomness tradeoff for oblivious routing. *SIAM J. Comput.*, 19:256–266, 1990.

[33] S. A. Plotkin, D. B. Shmoys, and É. Tardos. Fast approximation algorithms for fractional packing and covering problems. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 495–504, 1991.

[34] P. Raghavan. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37:130–143, 1988.

[35] P. Raghavan. Lecture notes on randomized algorithms. Technical Report RC 15340 (#68237), IBM T.J.Watson Research Center, January 1990. Also available as CS661 Lecture Notes, Technical report YALE/DCS/RR-757, Department of Computer Science, Yale University, January 1990.

[36] A. Ranade. How to emulate shared memory. *Journal of Computer and System Sciences*, 41:307–326, 1991.

[37] H. Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[38] M. Santha and U. V. Vazirani. Generating quasi-random sequences from semi-random sources. *Journal of Computer and System Sciences*, 33:75–87, 1986.

[39] J.P. Schmidt and A. Siegel. On aspects of universality and performance for closed hashing. In *Proc. ACM Symposium on Theory of Computing*, pages 355–366, 1989.

[40] J.P. Schmidt and A. Siegel. The analysis of closed hashing under limited randomness. In *Proc. ACM Symposium on Theory of Computing*, pages 224–234, 1990.

[41] D. B. Shmoys, C. Stein, and J. Wein. Improved approximation algorithms for shop scheduling problems. In *Proc. ACM/SIAM Symposium on Discrete Algorithms*, pages 131–140, 1991.

[42] A. Siegel. Toward a usable theory of Chernoff Bounds for heterogeneous and partially dependent random variables. Manuscript, September 1992.

[43] A. Siegel. On universal classes of fast hash functions, their time-space tradeoff, and their applications. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 20–25, 1989.

[44] J. Spencer. *Ten Lectures on the Probabilistic Method.* SIAM, Philadelphia, 1987.

[45] C. Stein. *Approximation algorithms for multicommodity flow and shop scheduling problems.* PhD thesis, Laboratory for Computer Science, Massachusetts Institute of Technology, 1992. Available as $MIT/LCS/TR-550$.

[46] L. G. Valiant. A scheme for fast parallel communication. *SIAM J. Comput.*, 11:350–361, 1982.

[47] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[48] L. G. Valiant and G. J. Brebner. Universal schemes for parallel communication. In *Proc. ACM Symposium on Theory of Computing*, pages 263–277, 1981.

[49] U. V. Vazirani. *Randomness, Adversaries and Computation.* PhD thesis, EECS, University of California at Berkeley, 1986.

[50] U. V. Vazirani and V. V. Vazirani. Random polynomial time is equal to slightly-random polynomial time. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 417–428, 1985. See also U. V. Vazirani and V. V. Vazirani, Random polynomial time is equal to semi-random polynomial time, Technical Report 88-959, Department of Computer Science, Cornell University, 1988.

[51] M.N. Wegman and J.L. Carter. New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences*, 22:265–279, 1981.

[52] A. Wigderson and D. Zuckerman. Expanders that beat the eigenvalue bound: Explicit construction and applications. In *Proc. ACM Symposium on Theory of Computing*, pages 245–251, 1993.

[53] D. Zuckerman. Simulating BPP using a general weak random source. In *Proc. IEEE Symposium on Foundations of Computer Science*, pages 79–89, 1991.