Toward a usable theory of Chernoff Bounds for heterogeneous and partially dependent random variables

Alan Siegel

Courant Institute

Abstract

Let X be a sum of real valued random variables and have a bounded mean $\mathrm{E}[X]$. The generic Chernoff-Hoeffding estimate for large deviations of X is: $Pr\{X - \mathrm{E}[X] \geq a\} \leq \min_{\lambda \geq 0} e^{-\lambda(a+\mathrm{E}[X])}\mathrm{E}[e^{\lambda X}]$, which applies with $a \geq 0$ to random variables with very small tails. At issue is how to use this method to attain sharp and useful estimates. We present a number of Chernoff-Hoeffding bounds for sums of random variables that may have a variety of dependent relationships and that may be heterogeneously distributed.

AMS classifications 60F10, Large deviations, 68Q25 Analysis of algorithms, 62E17, Approximations to distributions (nonasymptotic), 60E15, Inequalities.

Key words: Hoeffding bounds, Chernoff bounds, dependent random variables, Bernoulli trials.

This research was supported, in part, by grants NSF-CCR-8902221, NSF-CCR-8906949, and NSF-CCR-9204202.

Summary

In the analysis of probabilistic algorithms, some of the following problems may arise, possibly in complex combinations.

1) **Instance**: A collection of n random variables, that are partitioned among k sets. Random variables within a set are mutually independent, but worst case dependencies may exist among members of different sets. In the case of Bernoulli trials, only a bound on k, the count n, and the expectation of the sum of the n trials might be known, but nothing else.

Question: Give a good large deviation bound for the sum of the random variables.

2) **Instance**: $\{X_1, X_2, \ldots, X_k\}$ is a collection of dependent random variables.

Question: Find an effective large deviation bound for the sum of the random variables when the bound from 1) turns out to be weak, due to the heterogeneity among the probability distributions. Perhaps X_i is the sum of n_i mutually independent Bernoulli trials with $\bar{p}_i = \mathrm{E}[X_i]/n_i$, and only $\sum_{i=1}^{k}(1-\bar{p}_i)^2$ and $\sum_{i=1}^{k}\sqrt{\bar{p}_i(1-\bar{p}_i)n_i}$ are known.

Subquestion: What approximate Chernoff-Hoeffding bound for the case of mutually independent Bernoulli trials is most suitable for this question?

3) **Instance**: S is a collection of n possibly dependent random variables, which are not necessarily identically distributed. L_k is a family of real valued functions defined on k-item subsets of S. Perhaps all k-element subsets are equally likely to be selected. Perhaps not. Suppose L_k is monotone in k: $Pr\{L_k \geq a\} \leq Pr\{L_m \geq a\}$, for $k < m$.

Question: What is an effective large deviation bound for L_k?

Subquestion: Characterize a large class of functions that yield suitable L_k.

4) **Instance**: $X = x_1 + x_2 + \cdots + x_n$ is the sum of n mutually independent Bernoulli trials with probabilities of success p_i. The p_i are not identical, and our known parameters are $\bar{p} = \mathrm{E}[X]/n$, and $\sigma^2 = \sum_i p_i(1-p_i)/n$. Perhaps $\sigma^2 << \bar{p}(1-\bar{p})$.

Question: What is an effective large deviation bound for X?

Among the results we prove are the following, which, apart from 2.2 and 2.3, are all Chernoff-tight.

1) Let $X = \sum_{i=1}^{k} X_i$, where X_i, $i = 1, 2, \ldots, k$, are arbitrary possibly dependent real valued random variables. A valid Chernoff-Hoeffding bound can be attained by pretending the X_i's are independent and taking the k'th root of the resulting estimate. Formally, let $Y = \sum_{i=1}^{k} Y_i$, where the Y_i are mutually independent with Y_i equal to X_i in distribution: $Pr\{Y_i \leq x\} = Pr\{X_i \leq x\}$, for $i = 1, 2, \ldots, k$. Let B be a Chernoff-Hoeffding estimate for $Pr\{Y - \mathrm{E}[Y] \geq a\}$. Then we have the Chernoff-Hoeffding estimate

$$Pr\{X - \mathrm{E}[X] \geq a\} \leq B^{1/k}.$$

This inequality gives the exactly appropriate Chernoff-Hoeffding bound when X_1 is X_2 is $\cdots$ is X_k. Interestingly, the bound can fail to hold for $B = Pr\{Y - \mathrm{E}[Y] \geq a\}$.

2.1) Let $X = X_1 + X_2 + \cdots + X_k$ be the sum of k dependent random variables. A strengthening of 1 gives: Let $a = a_1 + a_2 + \cdots + a_k$ be partitioned so that Chernoff-Hoeffding estimates for $Pr\{X_1 - \mathrm{E}[X_1] \geq a_1\}$, $Pr\{X_2 - \mathrm{E}[X_2] \geq a_2\}$, $\cdots$, and $Pr\{X_k - \mathrm{E}[X_k] \geq a_k\}$ are all bounded by the value C. Then

$$Pr\{X - \mathrm{E}[X] \geq a\} \leq C.$$

Equivalently,

$$Pr\{X - \mathrm{E}[X] \geq a\} \leq \inf_{\substack{a_1 + a_2 + \cdots + a_k = a \\ a_1, a_2, \ldots, a_k \geq 0}} \max_i H(X_i, a_i),$$

where $H(X_i, a_i)$ is the Chernoff-Hoeffding estimate for $Pr\{X_i - \mathrm{E}[X_i] \geq a_i\}$. This bound significantly improves a related bound of Hoeffding [Ho-63].

2.2) Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables, where $0 \leq x_i \leq 1$, and $p = \mathrm{E}[X]/n$. Then for $0 \leq a \leq 1 - p$, we have the expressive fairly strong estimate

$$Pr\{X - \mathrm{E}[X] \geq an\} \leq e^{-\frac{a^2 n}{2p(1-p) + 2a(1-2p)/3 - 2a^2/9}}.$$

There are no extra restrictions on a. (For $a > 1 - p$, the probability, of course, is zero.) This estimate, which follows simply from Hoeffding's bound, is sharper, for almost all ranges of a and p than the more expressive approximations commonly used, and implies, via trivial maximization, the simpler but weaker estimates $Pr\{X - \mathrm{E}[X] \geq an\} \leq e^{-2a^2 n}$ [Ho-63]; $Pr\{X - \mathrm{E}[X] \geq \epsilon \mathrm{E}[X]\} \leq e^{-\epsilon^2 \mathrm{E}[X]/3}$, for $\epsilon < 1$ [AV-79]; $Pr\{X - \mathrm{E}[X] \leq -\epsilon \mathrm{E}[X]\} \leq e^{-\epsilon^2 \mathrm{E}[X]/2}$ [ASE-91].

2.3) Let $X = X_1 + \cdots + X_k$, where $\mathrm{E}[X_i]/n_i = \bar{p}_i$, and X_i is the sum of n_i mutually independent random variables, each belonging to $[0,1]$. The X_i's may exhibit arbitrary mutual dependencies. Combining 2.1 with a weak version of 2.2 shows, for $a > 0$:

$$Pr\{X - \mathrm{E}[X] \geq a\} < e^{-\frac{a^2}{8\left(\sum_i \sqrt{\bar{p}_i(1-\bar{p}_i)n_i}\right)^2}} + e^{-\frac{3a}{4\sum_i (1-\bar{p}_i)^2}}.$$

For sufficient heterogeneity, this estimate improves the $B^{1/k}$ bound stated in 1) by much more than a constant factor in the exponent.

3) Let $F(x_1, x_2, \ldots, x_n)$ be increasing in each variable (e.g. as in a measure of work), and let X_i be nonnegative random variables. Let $b_1, b_2, \ldots, b_n$ be mutually independent Bernoulli trials that are independent of the X_i. Let $m = \lfloor \mathrm{E}[\sum_i b_i] \rfloor$. Then

$$Pr\{F(X_1 \cdot b_1, \ldots, X_n \cdot b_n) > a \mid \sum_i b_i = m\} \leq 2Pr\{F(X_1 \cdot b_1, \ldots, X_n \cdot b_n) > a\}.$$

Thus deviation bounds where selection of k events occurs can be attained from a simpler, possibly independent process. The function F need not be symmetric, and the $\mathrm{E}[b_i]$ can vary.

4) Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli trials. Let $\bar{p} = \mathrm{E}[X]/n$, $\sigma^2 = (\mathrm{E}[X^2] - \mathrm{E}^2[X])/n$, and put $\hat{p} = \dfrac{\sigma^2}{1 - \frac{\sigma^2}{1-\bar{p}}}$. Then for $a > 0$, we attain the estimate

$$Pr\{X - n\bar{p} \geq na\} \leq \left(\frac{\hat{p}}{\hat{p}+a}\right)^{(a+\hat{p})n} \left(\frac{1-\bar{p}}{1-\bar{p}-a}\right)^{(1-\bar{p}-a)n}.$$

This bound sharpens a prior result of Hoeffding, by formulating the inequality as a function of both mean and variance, but holds only for Bernoulli trials, and not for sums of arbitrary random variables in $[0,1]$. The inequality is Chernoff-tight for all possible σ^2 and $\bar{p}$. When the random variables are identically distributed, $\hat{p} = \bar{p}$, whence the original Hoeffding estimate reappears. Reformulations can admit expressive approximations.

These estimates combine quite naturally to give simple effective bounds for complicated problems.

1.0 Background

Probability theory has evolved very expressive notation to capture functional behavior over sample spaces. Let X be a real valued random variable. The probability that an outcome of X is in the interval $[t, t+\epsilon)$ is written $Pr\{X \in [t, t+\epsilon)\}$. This notion includes differentials, so that $Pr\{X \in [t, t+dt)\}$ represents the probability measure for X, which is also denoted by $dPr\{X \leq t\}$ or just dP. This measure would be a just sum of point masses if the random variable were discrete, and a density function if no point masses exist. Formally, $Pr\{X \in [t, t+dt)\}$ represents a positive Riemann-Stieltjes measure of total mass 1, and can include point masses as well as a density function. The mean of the random variable is represented by $\mathrm{E}[X]$, and is computed by

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} t Pr\{X \in [t, t+dt)\}.$$

More generally, for any (measurable) function f,

$$\mathrm{E}[f(X)] = \int_{-\infty}^{\infty} f(t) Pr\{X \in [t, t+dt)\}.$$

Let $X = x_1 + x_2 + \ldots + x_n$ be the sum of n bounded, mutually independent random variables. Large deviation bounds, for such sums, concern estimates for the probability that $Pr\{X \geq \mathrm{E}[X] + a\}$, where a is usually relatively large, so that the probability is quite small. These estimates are often based on the moment generating function[†] G, where $G(\lambda) = \mathrm{E}[e^{\lambda X}]$.

Chernoff [Ch-52] advocated using $\min_{\lambda \geq 0} e^{-(a+\mathrm{E}[X])\lambda} G(\lambda)$ as a good estimate, as explained in Section 1.1, for $Pr\{X \geq a + \mathrm{E}[X]\}$. This estimation method, which applies Chebyshev's inequality to attain a weighted generating function, can be traced as far back as Bernshtein [Be-24].

A typical application would be when $X = x_1 + x_2 + \cdots + x_n$ is the sum of n independent random variables, each confined to some interval. The x_i, for example, might be confined to the range $[0, 1]$, or perhaps be Bernoulli trials, which only take on the values zero and one. Hoeffding not only established the best possible (worst case) results that can be found from this prescription for the sum of bounded random variables such as Bernoulli trials [Ho-63], he also developed related bounds that are sharp in a stochastic sense [Ho-56].

[†] The formulation $\mathrm{E}[z^X]$ often acquires this name, for integer valued X.

To be specific, let X be the sum of n independent Bernoulli trials, so that $X = x_1 + x_2 + \cdots + x_n$, where $Pr\{x_i = 1\} = p_i$, and $Pr\{x_i = 0\} = 1 - p_i$. In addition, let $\bar{p} = \mathrm{E}[X]/n$, and take $A \geq \mathrm{E}[X] + 1$. For this case, Hoeffding [Ho-56] shows that

$$Pr\{X \geq A\} \leq Pr\{B(n, \bar{p}) \geq A\} = \sum_{j \geq A} \binom{n}{j} (\bar{p})^j (1 - \bar{p})^{n-j}, \tag{1}$$

where $B(n, \bar{p})$ is the sum of n independent identically distributed trials, each with probability of success $\bar{p} = \frac{\mathrm{E}[X]}{n}$. Thus as a function of $\mathrm{E}[X]$ alone, this bound is unbeatable, for Bernoulli trials. Now the first term $\binom{n}{A}\bar{p}^A(1 - \bar{p})^{n-A}$ is an obvious underestimate, but if we approximate it with Stirling's formula and ignore the $\sqrt{\frac{n}{2\pi A(n-A)}}$ factor, the resulting expression turns out to be the same as the more general Chernoff-Hoeffding bound [Ho-63] $(\frac{n\bar{p}}{A})^A(\frac{n-n\bar{p}}{n-A})^{n-A}$ expressed as a function of slightly different parameters in (2) below. Thus this bound, while inexact, is never as much as a factor of $\sqrt{\pi n/2}$ too large, for identically distributed Bernoulli trials. It should be noted that estimation methods have been the subject of considerable study. Indeed, for small deviations, approximations by the Gaussian distribution (or possibly the Poisson distribution, depending on $n\bar{p}$) give sharp results, and error estimates have been investigated (c.f. [Pr-53], [Mo-70]). For large deviations, asymptotic expansions have been studied extensively (c.f. [Ba-60], [Ne-83]).

Furthermore, the errors resulting from general Chernoff-Hoeffding estimates have also been a matter of study, and asymptotic expansions have been attained. Among the more general results is the fact that when a Chernoff-Hoeffding bound for the sum of n independent identically distributed random variables gives a bound with exponential decay, $Pr\{X - \mathrm{E}[X] > an\} < e^{-nf(a)}$, the error in the exponent is always $o(n)$ (c.f. [SW-92]). Unfortunately, even the Chernoff-Hoeffding estimate (2) is rather inconvenient to use, and the standard approach in the Computer Science literature (c.f. [AV-79], [ASE-92]) often uses approximations that, while based on the Chernoff-Hoeffding estimation procedure, are exponentially worse but considerably more expressive. In summary, it is fair to say that the contributions of Chernoff-Hoeffding bounds are likely endure, and the value of Chernoff-tight bounds is considerable.

Moreover, exponential-based generating functions have turned out to be fundamental to the theory of martingales (c.f. [Do-53], [KT-81], [Wi-91]), and have had significant application to specific processes

such as Brownian motion. The relationship between Chernoff-Hoeffding bounds and martingales is briefly discussed in Section 6.

In this paper, we explore the issue of attaining sharper expressive estimates. More importantly, we expose techniques to export the Chernoff-Hoeffding method to many cases where dependencies and conditionings prohibit the direct application of this method, and appear to complicate considerably the problem of estimating large deviations. We also explore general cases where the random variables are not identically distributed, and attain bounds ranging from Chernoff-tight to reasonably sharp and expressive, for the parameters at hand. An application in Section 7, for example, analyses a sum of dependent heterogeneous Bernoulli trials where the heterogeneity ensures that a deviation of $c\sqrt{n \log n}$ occurs with polynomially small probability, as opposed to the $\Omega(1)$ chance that corresponds to identical collections of comparable (on average) dependent random variables. The exposition is intended to reveal the underlying proof techniques, so that the reader in need can not only access the relevant lemmata, but can also extend these developments as needed.

Despite the value of Chernoff-Hoeffding approximations and the precision of general asymptotic estimation procedures, other methods for attaining probabilistic inequalities have also flourished. In particular, the notion of strong stochastic domination, as exemplified by the Hoeffding bound (1), has been significantly strengthened through the elegant concept of majorization, which imposes, for many distributions, a partial ordering on vectors of independent random variables that is complete with minimal elements. The point of the ordering is that it is preserved under the application of functionals that exhibit Schur-convexity, such as the probability distribution function for suitable classes of random variables x_i and deviations a. In this sense, many bounds originally obtained as Chernoff-Hoeffding estimates have been materially strengthened and refined over the last two decades.

Although Schur-convexity methods appear to be inappropriate for attaining inequalities of this note, the results are quite interesting and worth summarizing. Let $\vec{P}$ and $\vec{Q}$ be the ordered vectors $\vec{P} = (p_1, p_2, \ldots, p_n)$, $\vec{Q} = (q_1, q_2, \ldots, q_n)$, with $p_1 \geq p_2 \geq \cdots \geq p_n$, and $q_1 \geq q_2 \geq \cdots \geq q_n$. Following [MO-79], we write

$$\vec{P} \prec \vec{Q} \quad \text{if} \quad \begin{cases} \sum_{1 \leq i \leq m} p_i \leq \sum_{1 \leq i \leq m} q_i, & \text{for } m = 1, 2, \ldots, n-1; \\ \sum_{1 \leq i \leq n} p_i = \sum_{1 \leq i \leq n} q_i. \end{cases}$$

In words, $\vec{P}$ is *majorized* by $\vec{Q}$. This partial order is extended to unordered vectors by applying it to their coordinate values in sorted order.

The intuition is that $\vec{P}$ has a flatter (or more egalitarian) distribution of values, and the constant distribution of values is the flattest, i.e. it is minimum. In this setting, the real valued functions of interest, which are defined on vectors, are precisely those that are order preserving (or reversing), in terms of the partial order $\prec$. A function f defined on a subset $S \subset R^n$ is defined to be *Schur-convex* if $\forall \vec{P} \prec \vec{Q} \in S:\ f(\vec{P}) \leq f(\vec{Q})$. If equality holds only when $\vec{P}$ and $\vec{Q}$ comprise the same set of values, then f is strictly Schur-convex. Schur-concavity and strict Schur-concavity are defined analogously.

From [MO-79], we have the following characterization, among others:

Theorem(Schur, 1923; Ostrowski, 1952). Let $I \subset R$ be an open interval, and let $f : I^n \rightarrow R$ be continuously differentiable. Necessary and sufficient conditions for f to be Schur-convex are:

1) f is symmetric on I^n;

2) $(x_1 - x_2)\left(\frac{d}{dx_1}f(x) - \frac{d}{dx_2}f(x)\right) \geq 0$ for $x = (x_1, x_2, \ldots, x_n) \in I^n$. □

Thus, for example, the elementary symmetric functions are Schur-concave for $I \subset [0, \infty]$ (Schur, 1923), and numerous generalizations can be found in [MO-79]. The relationship between Schur-convexity and probability is illustrated by the following theorem of Gleser (c.f. [MO-79]).

Theorem(Gleser, 1975). Let $X = x_1 + x_2 + \cdots + x_n$ and $Y = y_1 + y_2 + \cdots + y_n$ each be the sum of n independent Bernoulli trials, where $\mathrm{E}[X] = \mathrm{E}[Y]$, and $(\mathrm{E}[x_1], \ldots, \mathrm{E}[x_n]) \prec (\mathrm{E}[y_1], \ldots, \mathrm{E}[y_n])$. Then for $a \geq 3$, $Pr\{X - \mathrm{E}[X] \geq a\} \geq Pr\{Y - \mathrm{E}[Y] \geq a\}$. If $\mathrm{E}[X]$ is an integer, then the inequality also holds down to $a = 2$. □

In other words, the sum of Bernoulli trials with flatter distributions have greater tails for deviations exceeding 3. A wealth of other random variables and mathematical structures have been put into this powerful framework, and an excellent presentation of this material can be found in [MO-79].

We will exploit methods based upon moment generating functions to attain estimates that do not follow from Schur-convexity, since the constraints, for example, might restrict the admissible probability sequences to a subregion that will not contain minimal members. Moreover, some of our

bounding estimates will turn out to be satisfied by probability sequences that cannot be optimal for all deviation values a.

1.1 Formal definitions and preliminary inequalities

To simplify the notation, we define, for $a \geq 0$, the *deviation* $||\ \cdot\ ||_a^{Dv}$ to be†

$$||X||_a^{Dv} = Pr\{X - \mathrm{E}[X] \geq a\}.$$

It will be very convenient to define the *CH-estimate* $||\ \cdot\ ||_a^{CH}$ to be

$$||X||_a^{CH} = \min_{\lambda \geq 0} e^{-\lambda(a+\mathrm{E}[X])}\mathrm{E}[e^{\lambda X}].$$

It is also convenient to define $||\ \cdot\ ||_{\lambda,a}^{L}$ to be, for $a,\ \lambda \geq 0$,

$$||X||_{\lambda,a}^{L} = e^{-\lambda(a+\mathrm{E}[X])}\mathrm{E}[e^{\lambda X}],$$

so that $||X||_a^{CH} = \min_{\lambda \geq 0} ||X||_{\lambda,a}^{L}$.

Many large deviation estimates begin with the following inequality.

Lemma A(Chebyshev, Markov et al). Let Y be a nonnegative random variable. Then for any $b \geq 0$,

$$Pr\{Y \geq b\} \leq \frac{1}{b}\mathrm{E}[Y].$$

Proof:

$$Pr\{Y \geq b\} = \int_b^\infty 1 Pr\{Y \in [t, t+\mathrm{d}t)\} \leq \int_b^\infty \frac{t}{b} Pr\{Y \in [t, t+\mathrm{d}t)\} \leq \int_0^\infty \frac{t}{b} Pr\{Y \in [t, t+\mathrm{d}t)\} = \frac{1}{b}\mathrm{E}[Y]. \quad \blacksquare$$

Lemma A illustrates the simplest instance of the method of moments, which states that for $a > 0$, $Pr\{|X| > a\} = Pr\{|X|^k > a^k\} \leq \frac{\mathrm{E}[|x|^k]}{a^k}$. This family of estimates, which is a basis for Chernoff-Hoeffding estimates, is often given the name Chebyshev's inequality (c.f [Ch-67]), although some forms are attributable to Markov† and to Bienaymé, among others (c.f. [Lo-77]).

†Unfortunately, the $||\ \cdot\ ||$ estimates and bounds are not norms; they do not even satisfy the triangle inequality. Nevertheless, the expressive power provided by this overloaded notation, and their implicit connotations of size and throw-weight would seem to outweigh any formal disadvantages.

The approach can also be used for one-sided estimates: $Pr\{x > a\} \leq Pr\{|x - \lambda| > a - \lambda\}$, whence the method of moments might be applied and $\lambda < a$ optimized to attain the best result.

Chernoff observed that $Pr\{X - \mathrm{E}[X] \geq a\} = Pr\{e^{\lambda(X-\mathrm{E}[X])} \geq e^{\lambda a}\}$, for $\lambda > 0$, whence Lemma A, applied to the random variable $Y = e^{\lambda(X-\mathrm{E}[X])}$ and deviation $b = e^{\lambda a}$, shows that

$$||X||_a^{Dv} \leq ||X||_a^{CH}.$$

Chernoff's prescription of computing the moment generating function $\mathrm{E}[e^{\lambda X}]$ and optimizing with respect to λ has turned out to provide excellent estimates for many problems. The method embeds the estimation problem in a rich analytic domain, and imparts extra structure to a problem that (lacking Schur-convexity results) appears to be combinatorial, in the case of discrete random variables.

We will also need to use Jensen's inequality, which is an analytic quantification of a little geometry.

Lemma B(Jensen, 1906). Let f be convex. Suppose the coefficients a_i are nonnegative with $a_1 + a_2 + \cdots + a_n = 1$. Then

$$f\Big(\sum_i a_i x_i\Big) \leq \sum_i a_i f(x_i).$$

Proof: See Appendix 1. ∎

A few simple facts about moment generating functions are also worth noting.

Lemma 0. Let X be a random variable with moment generating function $G(\lambda) = \mathrm{E}[e^{\lambda X}]$. Then

0) If X and Y are independent with respective moment generating functions $G(\lambda)$ and $H(\lambda)$, then $X + Y$ has the moment generating function $G(\lambda)H(\lambda)$.

---

†The bound for $k = 1$ given in Lemma A was first published as a lemma in Markov's probability textbook [Ma-13, p. 86]. For $k = 2$, the bound can be traced as far back as Bienaymé [Bi-53], who in 1853 gave the inequality in an effort to convince Cauchy of the value of least squares methods [On-81]. Chebyshev attained the bound independently in 1867 [Ch-67]. Later, he learned of Bienaymé's work and subsequently attributed the method of moments to Bienaymé in a Liouville's Journal article of 1874 [On-81]. Markov attributed the method to his mentor Chebyshev, and observed that it was Chebyshev who recognized the significance of these inequalities and, for example, endeavored to use them to prove the central limit theorem ([Ma-14]). As for Bienaymé, he had the chance, Markov observed, to reference himself in a translation of Chebyshev's work, but did not choose to do so.

1) If $G(\lambda)$ is bounded for $\lambda \in (\alpha, \beta)$ then G is analytic for $\alpha < Re(\lambda) < \beta$, and we may differentiate as often as we please to attain $\frac{\mathrm{d}^k G}{\mathrm{d}\lambda^k} = \mathrm{E}[X^k e^{\lambda X}]$. Moreover, limit arguments can be applied to extend this formulation to the cases where $\lambda = \alpha, \beta$.

2) $G(\lambda) = \mathrm{E}[e^{\lambda X}]$ equals one at $\lambda = 0$, and is strictly increasing (or infinite) for $\lambda > 0$, provided $\mathrm{E}[X] \geq 0$ and $X \not\equiv 0$.

3) $G(\lambda)$ is strictly log-convex where it exists, provided X is not a constant.

Proof: See Appendix 2.1 ∎

In terms of the CH-estimate, Bernoulli trials turn out to be extreme points, and this property is exploited in the Hoeffding bound below. It will, therefore, be convenient to retain the definition of $B(n,p)$ as the the sum of n independent Bernoulli trials, where each has a probability of success equal to p. We shall, in the following theorem and upon other occasions, explicitly rescale a so that X has deviation na. The more general Hoeffding bound of [Ho-63] is as follows.

Theorem(Hoeffding, 1963). Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables where $0 \leq x_i \leq 1$, $p_i = \mathrm{E}[x_i]$, and $\bar{p} = \mathrm{E}[X/n]$. Then for $0 \leq a \leq 1 - \bar{p}$,

$$||X||_{na}^{CH} \leq ||B(n,\bar{p})||_{na}^{CH} = (\frac{\bar{p}}{\bar{p}+a})^{(\bar{p}+a)n}(\frac{1-\bar{p}}{1-\bar{p}-a})^{(1-\bar{p}-a)n}. \tag{2}$$

Proof: The proof is based on three facts. First, since the random variables are independent, we have:

$$\mathrm{E}[e^{\lambda X}] = \prod_i \mathrm{E}[e^{\lambda x_i}].$$

Second, for $\lambda \geq 0$,

$$\mathrm{E}[e^{\lambda x_i}] \leq p_i e^{\lambda \cdot 1} + (1 - p_i)e^{\lambda \cdot 0} \equiv \mathrm{E}[e^{\lambda B(1,p_i)}]; \tag{3}$$

this follows from the inequality

$$e^{\lambda t} \leq 1 + (e^{\lambda} - 1)t, \qquad t \in [0,1], \tag{4}$$

which says that on the interval $(0,1)$, the curve $e^{\lambda t}$ lies below the straight line that interpolates the function between $t = 0$ and $t = 1$. Inequality (4), in turn, follows from the convexity of $e^{\lambda t}$.

Multiplying each side of (4) by x_i's probability measure $Pr\{x_i \in [t, t+dt)\}$, and integrating on $[0,1]$ gives $\int_0^1 e^{\lambda t} Pr\{x_i \in [t,t+dt)\} = \mathrm{E}[e^{\lambda x_i}] \leq 1 + (e^{\lambda} - 1)\mathrm{E}[x_i]$, which, upon rearrangement, gives (3).†

The third fact is that $f(p_i) = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i}$ turns out to be log-concave as a function of p_i: $\frac{\mathrm{d}^2}{\mathrm{d}p^2} \log f(p) \leq 0$, where $f(p) = pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}$. As a consequence of the log-concavity, we may use Jensen's inequality to deduce that $\prod_i^n f(p_i) \leq (f(\sum_i^n p_i/n))^n$.

Combining these observation gives:

$$H(X, an) \leq e^{-\lambda an} \prod_i \mathrm{E}[e^{\lambda(x_i - p_i)}] \leq e^{-\lambda an} \prod_i \left(p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} \right) \leq e^{-\lambda an} \left(\bar{p} e^{\lambda(1-\bar{p})} + (1-\bar{p})e^{-\lambda \bar{p}} \right)^n.$$

Setting $e^{\lambda} = \frac{\bar{p}(1-\bar{p}) + a(1-\bar{p})}{\bar{p}(1-\bar{p}) - a\bar{p}}$ minimizes the last expression and gives (2). ∎

Of course $||X||_{na}^{CH} = 0$, for $a > 1 - \bar{p}$; we shall occasionally interpret $(\frac{\bar{p}}{\bar{p}+a})^{(\bar{p}+a)n} (\frac{1-\bar{p}}{1-\bar{p}-a})^{(1-\bar{p}-a)n}$ as being 0 for such large deviations, and avoid writing explicit case statements.

For completeness, we also state Bernshtein's deviation bound. It should be noted that for this estimate, the deviation a must be bounded by, approximately, the standard deviation $\sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}$, which limits its applicability to modest deviations where the central limit theorem is applicable.

Theorem(Bernshtein, 1924). Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables that have means $\mathrm{E}[x_i] = 0$, and bounded higher moments satisfying $|\mathrm{E}[x_i^k]| \leq \frac{k!}{4!} \left(\frac{L}{5}\right)^{k-4} \mathrm{E}[x_i^4]$, for $i = 1, 2, \ldots, n$, $k \geq 5$. Let $M_k = \sum_{i=1}^n \mathrm{E}[x_i^k]$. Then for $a \leq \frac{5}{4L}\sqrt{2M_2}$,

$$Pr\left\{ |X - \tfrac{a^2}{3}\tfrac{M_3}{M_2}| \geq a\sqrt{2M_2} \left(1 + \tfrac{M_4 a^2}{6M_2^2}\right) \right\} \leq 2e^{-a^2}. \quad \square$$

The recentering of the deviation interval from the nominal mean 0 to the center $\frac{a^2}{3}\frac{M_3}{M_2}$ is designed to give symmetric bounds of e^{-a^2} for the corresponding one-sided estimates [Be-24].

We have seen that the CH-estimate provides fairly strong estimates and simplifies many calculations. Moreover, it solves an implicit optimization question very fairly, as stated in Lemma 1.

---

†The convexity argument actually shows that, among all random variables having mean p, the largest moment generating function results from the distribution with the most weight at the largest possible value (which, according to the constraints of this theorem, must be from the Bernoulli trial $B(1,p)$). For the true deviation $||\cdot||_a^{Dv}$, two point distributions turn out to be maximal, although they may not be scaled Bernoulli trials. See [KS-66].

Lemma 1. Let $x_1, x_2, \ldots, x_n$ be n independent random variables. Then for $a > 0$,

1) $$||\sum_{i=1}^{n} x_i||_a^{CH} = \max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^{n} ||x_i||_{a_i}^{CH}, \quad a \geq 0.$$

As a very weak consequence,

2) $$||X+Y||_a^{CH} \geq ||X||_a^{CH}, \quad \text{for independent } X \text{ and } Y.$$

Proof: See Appendix 2.2 ∎

Thus the CH-estimate distributes the total deviation a quite judiciously among the x_i. Of course

$$\max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^{n} ||x_i||_{a_i}^{Dv} \leq ||\sum_i x_i||_a^{Dv} \leq \max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^{n} ||x_i||_{a_i}^{CH},$$

so the Chernoff-Hoeffding approximation provides a simple vehicle to go from a multiplicative underestimate to a multiplicative upper bound. This formulation is quite expressive, and can sometimes enable simplifying estimates to be made quite easily.

Lemma 1.1 and even the simple fact stated in Lemma 1.2 have application in some of the inequalities established in subsequent sections. Yet it should be noted that the analogous formulation of Lemma 1.2 in terms of pure probabilities $||\cdot||_a^{Dv}$ is false. The Chernoff-Hoeffding formulation saves us from the difficulty of quantifying and proving a theorem about something that is almost always true and always almost true.

When estimating deviations for dependent random variables, a slightly different optimization problem occurs. The basic question, of course, is the same: how to distribute a large aggregate deviation to a pool of random variables.

It will become evident that Lemma 0 could have listed a few more facts. For example, the function $||X||_a^{CH}$ will turn out to be log-concave a function of a. As a function of a, the value of λ where $||X||_a^{CH} = ||X||_{\lambda,a}^{L}$ will turn out to be strictly monotone in a. The proof of Lemma 1.1 shows that the maximizing a_i's have corresponding λ's (which minimize $||x_i||_{\lambda,a_i}^{L}$) that must all be the same, as long as the product is not zero. This is more than enough information to deduce that when $X = x_1+x_2+\cdots+x_n$ is the sum of n independent identically distributed random variables,

$$||X||_{na}^{CH} = \left(||x_1||_a^{CH}\right)^n.$$

This fact is usually proved more directly by observing that the independence ensures multiplicativity of moment generating functions: $\mathrm{E}[e^{\lambda \sum x_i}] = \prod_i \mathrm{E}[e^{\lambda x_i}]$. Indeed, there is a fundamental duality between CH-estimates and moment generating functions that can simplify probabilistic calculations. This duality, which is a special case of Legendre† transformations, is captured by the following lemma.

Lemma C. Let $h(\lambda)$ be a convex function of λ, and define $f(c) = \min_\lambda(-\lambda c + h(\lambda))$. Then f is concave and for $\lambda \in \mathrm{Domain}(h)$, $h(\lambda) = \max_c(\lambda c + f(c))$.

Proof: See Appendix 2.3 ∎

The transformation $e^{h(\lambda)} = \mathrm{E}[e^{\lambda X}]$ defines a convex function h (Lemma 0.3), and the Chernoff-Hoeffding probability estimate $e^{f(a)} = \min_\lambda e^{-\lambda a + h(\lambda)}$ defines f as the Legendre transform of h. Obviously, an upper bound for h translates into an upper bound for f and vice versa. In this sense, $f(a)$ and $h(\lambda)$ are equivalent. From a computational perspective, we may take more liberties in computations with h, since the evaluation of $e^{-\lambda a + h(\lambda)}$ for any value of λ is certain to give an overestimate. For sums of random variables, this difference is more significant. For independent X and Y, we see from Lemma 1 that $||X+Y||_a^{CH} = \max_{a_1+a_2=a}(||X||_{a_1}^{CH} ||Y||_{a_2}^{CH})$. The Legendre transform presents the more familiar formulation $\mathrm{E}[e^{\lambda(Y+Y)}] = \mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\lambda Y}]$ as stated in Lemma 0. In the analysis of martingales, where the random variables are not necessarily independent, both formulations produce corresponding inequalities, when quantified with proper conditioning, but the latter representation seems to be preferable.

This paper is organized as follows. Section 1 gives a generalization of the Chernoff-Hoeffding bound that is specific to Bernoulli trials as a function of mean and variance (question 4). The subsequent sections concern more general families of random variables. Section 3 presents some methods to estimate the probability of large deviations in the presence of side conditioning on the number of certain kinds of outcomes (question 3). Section 4 considers some dependencies that are graph or set based (question 1). Section 5 presents approximations to the traditional Chernoff-Hoeffding bounds, which combine sharpness with expressiveness (subquestion 2). More importantly, some of these bounds

†Legendre first investigated the transformation in 1789; prior discovery, however, can be credited to Euler in 1776 [Vi-90]. More about Legendre transformations and convex analysis can be found in [Se-89] and [ET-76].

are shown to yield fairly strong expressive bounds when quantifiable set-based dependencies are known (question 2). Section 6 illustrates further computational tradeoffs for probabilistic inequalities, and their implications for martingales. Section 7 presents an example to illustrate how a number of these estimates may be combined to solve complicated problems. Section 8 contains the conclusions.

2. Heterogeneous Bernoulli trials

Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli trials, where $Pr\{x_i = 1\} = p_i$, and $Pr\{x_i = 0\} = 1 - p_i$, and $\bar{p} = \mathrm{E}[X]/n$. Under this restriction, Hoeffding [Ho-56] established the following strong stochastic domination bound for $an \geq 1$:

$$||X||_{na}^{Dv} \leq ||B(n, \bar{p})||_{na}^{Dv}, \quad na \geq 1.$$

If $n\bar{p}$ is an integer, then the inequality extends down to $a \geq 0$. If only n and $\bar{p}$ are known, we cannot estimate $||X||_{na}^{Dv}$ any better, and even $||B(n, \bar{p})||_{na}^{CH}$ is a fairly sharp estimate for $||B(n, \bar{p})||_{na}^{Dv}$.

On the other hand, it is reasonable to expect applications where the variance of X is also known, and perhaps p_i or $\bar{p}$ satisfy some a priori bounds. Hoeffding addressed some of these problems as well [Ho-63], but did not establish tight bounds for Bernoulli trials in these cases.

For example, suppose we have $n/2$ coins stuck at 0, and another $n/2$ stuck at 1. Then the probability that the number of successes is at least $n/2 + 1$ is, of course zero, and an abstract measure of the absence of randomness in this problem is that the variance is zero. Thus we can expect improved deviation bounds if both the mean and variance are used. Moreover, the improvement can be much more than a constant factor in the exponent of an exponential rate of decay, if the probabilities are sufficiently biased to be predominantly near zero or one. Just what role a diminished variance plays in large deviation bounds is the subject of this section. It should be emphasized that Theorem 1 and Corollary 1 are specific to Bernoulli trials, and not arbitrary random variables confined to, say, $[0, 1]$.

Define σ^2 to be the average variance of the x_i, so that $\sigma^2 = (\mathrm{E}[X^2] - \mathrm{E}[X]^2)/n = \sum_i p_i(1 - p_i)/n$. As a function of n, a, $\bar{p}$ and σ^2, we have the following estimate for $||X||_{na}^{CH}$,

Theorem 1. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli trails. Let $\bar{p} = \mathrm{E}[X]/n$

and $\sigma^2 = (\mathrm{E}[X^2] - \mathrm{E}^2[X])/n$. Then

$$||X||_{na}^{CH} \leq ||B(n\frac{(1-\bar{p})^2}{1-\bar{p}-\sigma^2}, \frac{\sigma^2}{1-\bar{p}})||_{na\frac{1-\bar{p}-\sigma^2}{(1-\bar{p})^2}}^{CH}.$$

Proof: See Appendix 2.4 ∎

The estimate is sharp in the sense that a sequence of n Bernoulli trials comprising $n\frac{(1-\bar{p})^2}{1-\bar{p}-\sigma^2}$ trials with mean $\frac{\sigma^2}{1-\bar{p}}$ and $n\frac{\bar{p}-\bar{p}^2-\sigma^2}{1-\bar{p}-\sigma^2}$ (constant) trials with mean 1 has mean $\bar{p}$, variance σ^2 and the $||\cdot||_{na}^{CH}$-estimate stated above. Consequently the tightness is the same as for the subsequence where the constant trials are omitted. It should also be noted that this formulation admits the more expressive approximations of Section 5, since the formulation is in terms of a standard Hoeffding bound, for suitably adjusted parameters.

A little algebra can recast the formula implicit in Theorem 1 into a simpler form.

Corollary 1. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli trails. Let $\bar{p} = \mathrm{E}[X]/n$ and $\sigma^2 = (\mathrm{E}[X^2] - \mathrm{E}^2[X])/n$. Let $\hat{p} = \dfrac{\sigma^2}{1 - \frac{\sigma^2}{1-\bar{p}}}$. Then

$$||X||_{na}^{CH} \leq (\frac{\hat{p}}{\hat{p}+a})^{(a+\hat{p})n}(\frac{1-\bar{p}}{1-\bar{p}-a})^{(1-\bar{p}-a)n}.$$

Corollary 1 follows from simplifying Theorem 1. ∎

For completeness, we note that if only the average variance σ^2 is specified for a sum of Bernoulli trials, then the Hoeffding bound derived for $p \leq \frac{1}{2}$ with $p(1-p) = \sigma^2$ turns out to be maximal as well as tight. This can be established by observing

1) For $p \leq 1/2$ and $\lambda > 0$, $\mathrm{E}[e^{\lambda B(1,p)}] \geq \mathrm{E}[e^{\lambda B(1,1-p)}]$ where $B(1,p)$ is a Bernoulli trial with probability of success p.

2) For $p \leq 1/2$, $\mathrm{E}[e^{\lambda B(1,p)}]$ is log-concave as a function of $p(1-p)$.

Finally, it is worth noting that Hoeffding [Ho-63] established deviation bounds for $\sum_i x_i$ where the x_i have a common mean $\mathrm{E}[x_i] = p$ and upper bound b, but differing variances.

Theorem(Hoeffding, 1963). Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables,

where $x_i \leq b$, and $\mathrm{E}[x_i] = 0$, for $i = 1, 2, \ldots, n$. Let $\sigma^2 = \mathrm{E}[X^2]/n$. Then for $0 \leq a \leq b$,

$$||X||_{na}^{CH} \leq \left(\frac{\sigma^2}{\sigma^2 + ba}\right)^{n\frac{\sigma^2+ba}{\sigma^2+b^2}} \left(\frac{b}{b-a}\right)^{n\frac{b^2-ba}{\sigma^2+b^2}}. \quad \square$$

3.0 Simple approximations to eliminate conditioning

Sometimes probabilistic algorithms use processes that are similar to Bernoulli trials with extra constraints. The most obvious example is that from a set of n random variables, exactly k are sampled, and a large deviation bound is needed for some function of the sample. We shall use a slightly restricted but reasonably general formulation of this problem.

Definition.

Let $x_1, x_2, \ldots, x_n$ be a sequence of n random variables, with no assumptions about independence. Let $b_1, b_2, \ldots, b_n$ be a sequence of n mutually independent Boolean trials, which are also independent of the x_i-s. Let $f(y_1, y_2, \ldots, y_n)$ be an n-ary real function. Let $H[X; a]$ be a real valued probabilistic functional of the random variable X and the scalar a. Define the *Bernoulli-sampler* $F(a)$ to be $F(a) = H[f(b_1x_1, b_2, x_2, \ldots, b_nx_n); a]$, and define the *k-sampler* $F_k(a)$ as the conditional evaluation $F_k(a) = \mathrm{E}[H[f(b_1x_1, b_2, x_2, \ldots, b_nx_n); a] \mid \sum_i b_i = k]$.

For example, f could be the sum of the n random variables, and $H[X; a]$ might be $Pr\{X > a\}$, so that $F_k(a) = Pr\{b_1x_1 + b_2 + x_2 + \cdots + b_nx_n > a \mid \sum_i b_i = k\}$. Alternatively, $H[X; a]$ might be the estimate $\min_{\lambda \geq 0} \mathrm{E}[e^{\lambda(X-a)}]$.

The technical difficulty in bounding deviations for the k-sampler F_k is that its selection events are not quite independent, and consequently the results provided by Chernoff-Hoeffding are formally inapplicable. Yet the dependencies are quite mild; surely they cannot cause much difference, for a well behaved F.

Definition.

Let the n-ary function f, the random variables $x_1, \ldots, x_n$, the Bernoulli trials $b_1, \ldots, b_n$, and the functional H define the Bernoulli-sampler F and the k-samplers $F_1, F_2, \ldots, F_n$. We say that the F_i are *increasing* if for all $a > 0$,

$$F_k(a) \leq F_l(a), \text{ for } 0 \leq k \leq l \leq n.$$

This increasing property for samplers can occur in practice. For example, it can be readily deduced from Lemma 1.2 that if $F(a) = \min_{\lambda \geq 0} \mathrm{E}[e^{\lambda(x_1+x_2+\ldots+x_n-a)}]$, then the F_i are increasing provided $\mathrm{E}[x_i] \geq 0$, for $i = 1, 2, \ldots, n$. This sections shows that Chernoff-Hoeffding estimates for suitable k-samplers may be derived from the corresponding Bernoulli sampler, where the sampling has no conditioning.

We shall, for the moment, assume that each k–set of events is equally likely to occur, although more general assumptions will soon be accommodated.

Theorem 2. Let $x_1, x_2, \ldots, x_n$ be a sequence of n random variables, and $b_1, \ldots, b_n$ be fully independent Bernoulli trials, each with probability of success p. Let $B = \sum_i b_i$. Let F be a Bernoulli sampler with an increasing family $F_1, F_2, \ldots, F_n$ of k-samplers. Then for $k \leq np$, $a \geq 0$,

$$F_k(a) \leq 2F(a).$$

Proof: We use a remarkable fact about Bernoulli trials, which is due to Jogdeo and Samuels [JS-68], and based, in part, on the strong Hoeffding bound (1). In particular, [JS-68] shows that if B is the sum of n independent Bernoulli trials with integer mean $\mathrm{E}[B] = k$, then $Pr\{B \geq k\} > \frac{1}{2}$; the mean is the median. The individual trials need not have the same probability of success. If the expectation is not an integer, then the median is one of the two adjacent integers. Accordingly, we have:

$$\begin{aligned}
F_k(a) &= \mathrm{E}[F(a) |\ B = k], \\
&\leq \mathrm{E}[F(a) |\ B \geq k], \quad \text{because the } F_i \text{ are increasing}, \\
&\leq \frac{\mathrm{E}[F(a) \bigwedge B \geq k]}{Pr\{B \geq k\}}, \\
&\leq \frac{F(a)}{Pr\{B \geq k\}}, \\
&\leq 2F(a). \quad \blacksquare
\end{aligned}$$

Interestingly, we may even allow the probabilities of success to be different, provided the expected number of successes is k (or more, or perhaps between $k-1$ and k, in which case the factor of 2 should be replaced by an estimate of $\frac{1}{\frac{1}{2} - Pr\{B=k-1\}}$.) When the number of events selected is conditioned to

be k, straightforward conditioning shows that the effective probability of an event r_i with Bernoulli selection probability p_i to be $\hat{p}_i = \frac{p_i}{S_k^{1/k}(1-p_i)}$, where $S_k = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} \frac{p_{i_1} p_{i_2} \cdots p_{i_k}}{(1-p_{i_1})(1-p_{i_2})\cdots(1-p_{i_k})}$. With these values, direct calculation verifies that a k-way selection of the r_i's has a conditional probability that is just the product of the corresponding $\hat{p}$'s.

Lemma 2. Let $x_1, x_2, \ldots, x_n$ be a collection of possibly dependent random variables, and $b_1, b_2, \ldots, b_n$ be n independent Bernoulli trials with $\mathrm{E}[b_i] = p_i$. Let $R = \cup_{i:b_i=1}\{x_i\}$, so that R depends on outcome of the random variables b_i. Let f be a real valued increasing function on subsets of $\{x_1, \ldots, x_n\}$: if $A \subset B$ then $\mathrm{E}[f(A)] \le \mathrm{E}[f(B)]$. Then $\mathrm{E}[f(R)|b_1 + \cdots + b_n = k]$ is increasing in k.

Proof: The proof is by induction on (n, k). The base cases are $n = k - 1$, and $k = 0$. In either instance, it is trivially seen that $\mathrm{E}[f(R)|b_1 + b_2 + \cdots + b_n = k+1] \ge \mathrm{E}[f(R)|b_1 + b_2 + \cdots + b_n = k]$.

So we may suppose that the inequality holds for $(n-1, k+1)$ and $(n-1, k)$.

Lemma 3. Let $\xi_k = Pr\{b_1 = 1 | b_1 + b_2 + \cdots + b_n = k\}$. Then ξ_k is monotone increasing in k.

Proof: See Appendix 2.5 ∎

There follows:

$$\begin{aligned}
\mathrm{E}[f(R)|b_1 + b_2 + \cdots + b_n = k+1] &= \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k]\xi_{k+1} \\
&\quad + \mathrm{E}[f(R)|b_1 = 0, b_2 + \cdots + b_n = k+1](1 - \xi_{k+1}) \\
&= \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k]\xi_k \\
&\quad + \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k](\xi_{k+1} - \xi_k) \\
&\quad + \mathrm{E}[f(R)|b_1 = 0, b_2 + \cdots + b_n = k+1](1 - \xi_{k+1}) \\
&\ge \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k-1]\xi_k \\
&\quad + \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k](\xi_{k+1} - \xi_k) \\
&\quad + \mathrm{E}[f(R)|b_1 = 0, b_2 + \cdots + b_n = k](1 - \xi_{k+1}) \\
&\ge \mathrm{E}[f(R)|b_1 = 1, b_2 + \cdots + b_n = k-1]\xi_k \\
&\quad + \mathrm{E}[f(R)|b_1 = 0, b_2 + \cdots + b_n = k](1 - \xi_k) \\
&\ge \mathrm{E}[f(R)|b_1 + b_2 + \cdots + b_n = k]. \qquad \blacksquare
\end{aligned}$$

Applying Theorem 2, for $k \le \sum p_i$, gives $\mathrm{E}[f(R)|b_1 + b_2 + \cdots + b_n = k] \le 2\mathrm{E}[f(R)]$.

While the x_i need not be independent, they are likely to be so in practice. The contribution of Theorem 2 and Lemma 2 is due to the fact that the unconditioned random variable $f(R)$ can usually be bounded much more easily than $(f(R) \mid k = |R|)$ This might be especially true for a complicated f, which depends on more than the sum of the selected x_i. The Lueker and Molodowitch analysis of double hashing [LM-88] illustrates this kind of problem. The main calculation is a large deviation estimate for the number of probe possibilities, in double hashing, that would cause an item to be placed in a fixed vacant slot of a hash table that has n slots and contains αn items that are uniformly distributed (i.e. not inserted by double hashing). The estimate is based on an approximation where each location is occupied according to a Bernoulli trial with probability of success α. For this application, an item is more likely to hash into the given location if additional items belong to the table, so the increasing property holds. Section 7 presents this analysis in greater detail.

For completeness, it is worth noting that when $f(R) = f(\sum_i b_i x_i)$ for convex f, and each k-set of the r's is equally likely to be selected, then $\mathrm{E}[f(R)|B = k]$ turns out to be convex in k. Jensen's inequality establishes the equivalent of Theorem 2 without the factor of 2. For this case, a related approach was given by Hoeffding [Ho-63].

Theorem(Hoeffding, 1963). Let $x_1, x_2, \ldots, x_n$ be a sequence of possibly dependent random variables. Let $X_1, X_2, \ldots, X_k$ be k samples of the x_i without replacement: each random variable produces at most one random sample. Let $Y_1, Y_2, \ldots, Y_k$ be k samples of the x_i with replacement. Let f be a convex real valued function. Then

$$\mathrm{E}[f(\sum_1^k X_i)] \leq \mathrm{E}[f(\sum_1^k Y_i)]. \quad \square$$

Choosing $f(x) = e^{\lambda x}$ gives a useful method to estimate the deviation, since the Y_i are identically distributed and $\mathrm{E}[e^{\lambda Y_i}] = \frac{1}{n}\sum_1^n \mathrm{E}[e^{\lambda X_i}]$. Moreover, the probability community has derived an astonishingly rich diversity of generalizations for symmetric sampling schemes (c.f. [MO-79]). For example, Karlin identifies a large class of symmetric Schur-convex-like f where $\mathrm{E}_{\vec{P}}[f(X_1, \ldots, X_n)] \leq \mathrm{E}_{\vec{Q}}[f(X_1, \ldots, X_n)]$ when $P \prec Q$, for a suitably defined partial order $\prec$ [MO-79]. Here the expectation $\mathrm{E}_{\vec{P}}$ is the average (among $n!$ possibilities) of f applied to samples of the X_i with repetition P_i, where each of the $n!$ samples uses repetitions from a different permutation of the n integer components of $\vec{P}$. The maximal

vector has all of its weight concentrated at one coordinate. These inequalities use sampling schemes that have the same total weight; they do not capture Theorem 2 because the factor of 2 is essential for increasing F_k. On the other hand, the use of Theorem 2 for Chernoff-Hoeffding estimates will require an upper bound for $\mathrm{E}[F] - \mathrm{E}[F_k]$, which, as illustrated in Section 7, will often be suitably small.

4. Generic Dependencies

Sometimes a set of Bernoulli trials or other real random variables has dependencies that might be very difficult to quantify. This might happen, for example, because of a probabilistic decomposition of a problem with parameter n into k subproblems with positive parameters $n_1, n_2, \ldots, n_k$, where $n_1 + \cdots + n_k = n$, but where the actual values of the n_i are unknown. Suppose that the i-th subproblem depends on n_i independent Bernoulli trials with probability of success p, but there are unknown dependencies among the subproblems. Let X_i denote the sum of the Bernoulli trials in the i-th subproblem. Finally, suppose we need a deviation bound for $X = X_1 + \cdots + X_k$. We seek a tractable quantification for the worst case deviation. Our intuition would suggest that the worst case ought to occur when the X_i are the same random variable. We show that although the intuition is false for general random variables and exact probabilities, it is always true for Chernoff-Hoeffding estimates.

Let, for positive a_i, $a_1 + a_2 + \cdots + a_k = a$. Then $||X||_a^{Dv} \leq \sum_{i=1}^{a} ||X_i||_{a_i}^{Dv}$, and one question is how to partition the a's in an optimal manner. It will become evident, in this section, that this formulation has already forfeited opportunities to exploit the full benefits of convexity. Hoeffding [Ho-63], for example, uses convexity to attain the following deviation estimate.

Theorem(Hoeffding, 1963). Suppose $X = p_1X_1 + p_2X_2 + \cdots + p_kX_k$ where the X_i are not necessarily independent, the p_i are positive and $\sum_i p_i = 1$. Then for $\lambda \geq 0$,

$$||X||_a^{CH} \leq \sum_{i=1}^{k} p_i \mathrm{E}[e^{\lambda(x_i - \mathrm{E}[x_i] - a)}]. \quad \square$$

Unfortunately, this bound does not appear to have a more natural (or closed) formulation. It also forces the rescaling parameter λ to be the same for all k random variables. By exploiting convexity to apportion the deviation a and by optimizing both the a apportionments and implicit p_i weightings, we get sharper and more tractable formulations with individually optimized λ's.

Theorem 3. Let $X = \sum_{i=1}^{k} X_i$, where X_i, $i = 1, 2, \ldots, k$, are arbitrary possibly dependent real valued random variables. Let $a = a_1 + a_2 + \cdots + a_k$, and suppose that $||X_i||_{a_i}^{CH} = C$, for $i = 1, 2, \ldots, k$. Then

$$||X||_a^{CH} \leq C.$$

Proof: We may normalize the X_i, so that they have zero means. Let p_i be positive with $p_1 + p_2 + \cdots + p_k = 1$. Evidently,

$$||\sum_i X_i||_a^{CH} \leq \mathrm{E}[e^{\sum_{j=1}^{k} p_j \lambda (X_j - a_j)/p_j}] \leq \mathrm{E}[\sum_{j=1}^{k} p_j e^{\lambda (X_j - a_j)/p_j}], \tag{5}$$

due to the convexity of e^x and Jensen's inequality. We seek the best λ, p_j and a_j. It turns out that a straightforward derivation suffices to solve this optimization problem, but a verification argument is simpler still. Let the a_i partition a as prescribed by the Theorem. Let λ_i be the optimum λ defined by $||X_i||_{a_i}^{CH} = \mathrm{E}[e^{(\lambda_i (X_i - \mathrm{E}[X_i] - a_i)}]$. Put $\lambda = 1/(\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_k})$, and set $p_i = \lambda/\lambda_i$. By construction, $\sum_i p_i = 1$. Substituting these values into equation (5) shows that $||\sum_i X_i||_a^{CH} \leq \sum_i p_i C = C$. ∎

The bound is clearly Chernoff-tight.

There are, however, circumstances where the assumptions of Theorem 3 cannot be fulfilled. Suppose, for example, that X_1 comprises a single Bernoulli trial $B(1, \frac{1}{2})$, and X_2 comprises 9 such trials, which are independent. Let X_1 and X_2 be mutually dependent. Now $||X_1||_{\frac{1}{2}}^{CH} = \frac{1}{2}$, and $||X_1||_a^{CH} = 0$ for $a > \frac{1}{2}$, which means that the C of Theorem 3 cannot have values in $(0, \frac{1}{2})$. Suppose we wish to estimate $||X||_5^{CH}$; the probability of such a deviation is clearly below $\frac{1}{2}$.

The problem is easily resolved in terms of the discontinuities of the Chernoff-Hoeffding estimates $||X_i||_{a_i}^{CH}$; the following definition is intended to serve this purpose.

Definition.

Let the value a^+ denote a plus an infinitesimal amount, and define the real evaluation $f(a^+) = \lim_{b \downarrow a} f(b)$.

Corollary 2. Let $X = \sum_{i=1}^{k} X_i$, where X_i, $i = 1, 2, \ldots, k$, are arbitrary possibly dependent real valued random variables. Let $a = a_1 + a_2 + \cdots + a_k$. Suppose that $\min_i ||X_i||_{a_i}^{CH} = C$, and for $i = 1, \ldots, k$, either $||X_i||_{a_i}^{CH} = C$ or $||X_i||_{a_i^+}^{CH} = 0$. Then

$$||X||_a^{CH} \leq C.$$

Proof: Let $a = a_1 + a_2 + \cdots + a_k$, and suppose that $a_i > 0$, for $i = 1, 2, \ldots, k$. We must show that for any l,

$$||X||_a^{CH} \leq \max(||X_l||_{a_l}^{CH}, \max_i ||X_i||_{a_i^+}^{CH}). \tag{6}$$

The key point is that the CH-estimate is continuous from the left: $\lim_{a \uparrow b} ||Y||_a^{CH} = ||Y||_b^{CH}$ for $a > 0$. In fact, it is also continuous from the right unless $b = \max(Y - \mathrm{E}[Y])$. This because $||Y||_a^{CH} = 0$ for $a > b$, while $||Y||_b^{CH}$ is the discrete mass probability $Pr\{Y = b + \mathrm{E}[Y]\}$, which may be positive. For $a \geq b$, the optimizing λ will be infinite: $||Y||_a^{CH} = ||Y||_{\infty,a}^{L} = 0$, $a \geq \max(Y - \mathrm{E}[Y])$.

Let $||X_l||_{a_l}^{CH} = \min_i(||X_i||_{a_i}^{CH})$, and consider the partition $\tilde{a}_l = a_l - (k-1)\epsilon$, and $\tilde{a}_i = a_i + \epsilon$, for $i \neq l$. This is a permissible partition of a for small ϵ, and we may apply the convexity argument of Theorem 3. Those X_i where $||X_i||_{a_i+\epsilon}^{CH} = 0$, for all positive ϵ, will have corresponding $p_i = 0$, and contribute nothing to the estimate. The resulting expression, as $\epsilon \downarrow 0$, is $\sum_i p_i ||X_i||_{a_i}^{CH}$, with the aforementioned p_i set to zero. This averaging is bounded by the maximum Chernoff-Hoeffding estimate having a non-zero weight, which establishes (6). ∎

The general applicability of Corollary 2 comes from the fact that we can always optimize the a_i, with the result that either $||X_i||_{a_i}^{CH} = \min_j(||X_j||_{a_j}^{CH})$, or the value is larger but cannot be diminished because $||X_i||_{a_i^+}^{CH} = 0$. More abstractly,

$$||X||_a^{CH} \leq \min_{\substack{a_1 + \cdots + a_k = a \\ a_i \geq 0}} \max(\min_i ||X_i||_{a_i}^{CH},\ \max_i ||X_i||_{a_i^+}^{CH}) = \inf_{\substack{a_1 + \cdots + a_k = a \\ a_1, a_2, \ldots, a_k \geq 0}} \max_i ||X_i||_{a_i}^{CH},$$

The first formulation has a partition of a where the minimum must be achieved, but the second might not achieve its minimum value; the use of the infimum, while denotationally equivalent to minimum, is intended to emphasize this distinction.

As for our motivating problem, we see that X_1 should be assigned the deviation $a_1 = \frac{1}{2}^+$, and the Chernoff-Hoeffding estimate becomes $||X_2||_{4.5}^{CH}$.

Section 5.1 examines the problem of deriving strong and expressive estimates from Theorem 3. Meanwhile, it should be noted that probabilistic behavior sometimes ensures, with high probability, that a dependent collection of random variables can be partitioned into only a few sets, which each comprise mutually independent random variables. When the decomposition happens to be probabilistic, the detailed information required for Theorem 3 and Section 5.1 may not be available. The

following weak consequence of Theorem 3 may be adequate for some of these circumstances. The deviation bound is still Chernoff-tight for a worst-case decomposition.

Corollary 3. Let $X = \sum_{i=1}^{k} X_i$, where X_i, $i = 1, 2, \ldots, k$, are arbitrary possibly dependent real valued random variables, and let $Y = \sum_{i=1}^{k} Y_i$, where the Y_i are mutually independent with Y_i equal to X_i in distribution: $Pr\{Y_i \leq a\} = Pr\{X_i \leq a\}$, for $i = 1, 2, \ldots, k$. Then

$$||X||_a^{CH} \leq \left(||Y||_a^{CH}\right)^{1/k}.$$

Proof: Let $a_1, a_2, \ldots, a_k$ be the partition of a yielding the estimate C of Corollary 2, so that $||X_i||_{a_i}^{CH} \geq C$ and $||X||_a^{CH} \leq C$. From Lemma 1.1, $||Y||_a^{CH} \geq \prod_{i=1}^{k} ||X_i||_{a_i}^{CH}$, whence $||Y||_a^{CH} \geq C^k$, and hence $||X||_a^{CH} \leq \left(||Y||_a^{CH}\right)^{1/k}$. ∎

Corollary 3 also results if the optimization in Theorem 3 is performed over free λ and a_i with $p_1 = p_2 = \cdots = p_k = 1/k$.

Natural applications for Corollary 3 would use estimates for Y. The X_i might themselves be sums of independent random variables, some of which might belong to more than one X_i.

The technical difficulty suppressed by the Chernoff-Hoeffding estimate is that deviation probabilities are not concave functions, which is why the rescaling of deviations was necessary. Of course $k||Y||_{\frac{na}{k}}^{Dv}$ is a correct upper bound for $||X||_a^{Dv}$, and it is easy to see that this expression is optimal among products of a scalar and a probability. On the other hand, the natural analog to Corollary 3, $\left(||Y||_a^{Dv}\right)^{1/k}$, turns out to be incorrect.

A counterexample to this Dv formulation of $\left(||Y||_a^{Dv}\right)^{1/k}$ can be constructed as follows. Suppose that each X_i equals $\pm a$ with probability $\frac{1}{k+1}$ and zero with probability $\frac{k-1}{k+1}$. By correlating the k random variables so that they are all equal to $-a$ simultaneously, while only one is positive at any other time, we see that $||\sum_i^k X_i||_a^{Dv} = \frac{k}{k+1}$. If we take the X_i's to be independent and have k even, then symmetry allows us to infer that the sum will be at least as large as a half the time it is unequal to zero. The probability that the sum is zero is

$$\sum_{j=0}^{k} \binom{k}{j, j, k-2j} \left(\frac{1}{k+1}\right)^{2j} \left(\frac{k-1}{k+1}\right)^{k-2j}.$$

Taking the first two terms gives a value exceeding $2\left(\frac{k-1}{k+1}\right)^k$. Our counterexample will be established if we show that $\left(\frac{1}{2} - \left(\frac{k-1}{k+1}\right)^k\right)^{1/k} < \frac{k}{k+1}$. Simplifying gives the requirement $\frac{1}{2} < \left(\frac{k-1}{k+1}\right)^k + \left(\frac{k}{k+1}\right)^k$. Letting $k \to \infty$ gives, in the limit: $\frac{1}{2} < e^{-2} + e^{-1}$. Upon multiplying by e^2, we see that it suffices to show that $x^2/2 - x - 1 \mid_{x=e} < 0$. But the roots of this quadratic are $1 \pm \sqrt{3}$, and it suffices to show that $e < 1 + \sqrt{3}$, which follows since $\sqrt{3} > 1.73$, and $e < 2.72$. Consequently, the inequality holds for sufficiently large k. In fact, it is not difficult to show that $\left(\frac{k-1}{k+1}\right)^k + \left(\frac{k}{k+1}\right)^k > \frac{1}{2}$ for $k > 1$.

5.0 Expressiveness and dependencies

While the Hoeffding bound (2) for the sum of n random variables with range in $[0, 1]$ is quite strong, and even better estimates can be attained based on the variance or a sharper analysis of binomial distributions, most applications do not need such precision. Rather, proof techniques benefit more from simpler formulations that express adequate bounds in a tractable form. This section introduces simple approximations that are applied to instances of dependent random variables in Section 5.1, and are sharpened in Section 5.2

Let $X = \sum_{i=1}^{n} X_i$ be the sum of n independent random variables, $0 \le X_i \le 1$ with mean probability of success $p = \sum \mathrm{E}[X_i]/n$. While we give estimates below for $Pr\{X - \mathrm{E}[X] \ge an\}$, for $a > 0$, comparable estimates for $Pr\{X - \mathrm{E}[X] \le -an\}$ can be written by replacing p with $1 - p$.

The most direct means for deriving deviation estimates is by approximating the (fairly) sharp bound (2) stated explicitly by Hoeffding [Ho-63]. We can prove that $(f(a))^n \le (g(a))^n$ by showing that $\log(f)$ and its first $k-1$ derivatives, say, agree with those of $\log(g)$ at $a = 0$. If $\frac{\mathrm{d}^k}{da^k}\log(f(a)) \le \frac{\mathrm{d}^k}{da^k}\log(g(a))$ for $a \ge 0$, then integrating the inequality from $a = 0$ k times establishes the desired bound. This approach works with $k = 2$ for each of the approximations listed below. In each case, $(f)^n$ is the Hoeffding bound, and $\log f(0) = (\log f)_a(0) = 0 = \log g(0) = (\log g)_a(0)$. We omit the simple calculations for $\log g(0)$ and $(\log g)_a(0)$, and just list the straightforward $(\log g)_{aa}(a)$. The last entry is the Hoeffding bound f.

Theorem. For $0 \le a \le 1-p$, $Pr\{X - \mathrm{E}[X] \ge na\} < (g(a))^n$.

Proof: $(\log g)_{aa}(a) \ge (\log f)_{aa}(a)$.		
$g(a)$	$(\log g)_{aa}$	References
$\frac{e^a}{(1+\frac{a}{p})^{a+p}}$, $e^{\frac{-a^2}{2(1-p)}}$; $\frac{e^a}{(1+\frac{a}{1-p})^{a+1-p}}$	$\frac{-1}{p+a}$, $\frac{-1}{1-p}$; $\frac{-1}{1-p+a}$	[ASE-91]; [Ra-88]
$\frac{e^{\frac{a}{1-p}}}{(1+\frac{a}{p})^{\frac{a}{1-p}+\frac{p}{1-p}}}$	$\frac{-1}{p(1-p)+a(1-p)}$	[Be-62]-[Ho-63]†
$\frac{e^{\frac{a}{(1-p)^2}}}{(1+\frac{a(1-p)}{p})^{\frac{a}{(1-p)^2}+\frac{p}{(1-p)^3}}}$	$\frac{-1}{p(1-p)+a(1-p)^2}$	
$\frac{e^{\frac{a}{1-2p}}}{(1+\frac{a(1-2p)}{p(1-p)})^{\frac{a}{1-2p}+\frac{p(1-p)}{(1-2p)^2}}}$	$\frac{-1}{p(1-p)+a(1-2p)}$	
$f = \left(\frac{p}{p+a}\right)^{p+a}\left(\frac{1-p}{1-p-a}\right)^{1-p-a}$	$\frac{-1}{p(1-p)+a(1-2p)-a^2}$	[Ho-63]

Table 1

The bound $\frac{e^a}{(1+\frac{a}{1-p})^{a+1-p}}$ and its stronger counterpart $e^{\frac{-a^2}{2(1-p)}}$ are intended to be used with small p for large deviations below the mean, which means with p and $1-p$ interchanged. Each of the expressions in the above table is sharper than its predecessors, but all, apart from $e^{\frac{-a^2}{2(1-p)}}$ involve ratios of increasing values. In this sense, these bounds are less expressive than the following preliminary bound that, although incomparable but often weaker than the weaker bounds above, is never much weaker; it has an exponent that is off by a modest factor. We emphasize its utility by stating it as a theorem, and mentioning a few consequences; stronger formulations are given in Section 5.2.

Theorem 4. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables, where $0 \le x_i \le 1$, and $p = \mathrm{E}[X]/n$. Then for $a \ge 0$,

$$||X||_{na}^{CH} \le \left(1 + \frac{2a(1-p)}{3p}\right)^{-\frac{3an}{4(1-p)^2}} \le \left(1 + \frac{2a}{3p}\right)^{-\frac{3an}{4}}.$$

†The estimate originates with Bennett [Be-62] for somewhat different bounds on the X_i, and is discussed by Hoeffding [Ho-63] within this modified context, but Hoeffding's function inequalities are independent of the X_i's.

Proof: It is not difficult to check that for $g(a) = \left(1 + \frac{2a(1-p)}{3p}\right)^{-\frac{3an}{4(1-p)^2}}$, $\log g(0) = 0 = (\log g)_a(0)$. Finally, straightforward calculations verify that

$$(\log g)_{aa}(a) = \frac{-1}{p(1-p) + \frac{2}{3}a(1-p)^2} + \frac{3a}{(3p + 2a(1-p))^2} \geq \frac{-1}{p(1-p) + a(1-2p-a)}. \quad \blacksquare$$

Hoeffding [Ho-63] also gives an approximation for his bound:

$$||X||_{na}^{CH} \leq e^{-na^2 g(p)},$$

where

$$g(p) = \begin{cases} \frac{1}{1-2p} \log \frac{1-p}{p}, & \text{for } 0 < p < \frac{1}{2}; \\ \frac{1}{2p(1-p)}, & \text{for } \frac{1}{2} \leq p < 1. \end{cases}$$

The drawback with this estimate is that it is weak in cases where p is very small, say, $p \approx \frac{1}{\sqrt{n}}$, and a is a moderate multiple of the expectation np. Such circumstances are of interest in many computer science applications. The bound also has cases, which may make its application more difficult in circumstances where the p is not explicitly known.

The expressiveness of Theorem 4 is sufficient to permit a trivial weakening that gives the following formulation.

Corollary 4. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables where $0 \leq x_i \leq 1$ and $p = \mathrm{E}[X]/n$. Then for $a \geq 0$,

$$||X||_{na}^{CH} \leq e^{-\frac{a^2 n}{2p(1-p) + 2a(1-p)^2/3}} < e^{\frac{-an}{2(1-p)^2}} + e^{\frac{-a^2 n}{3p(1-p)}}.$$

Proof: We use Theorem 4. The first inequality follows from the fact that $\log(1+x) \geq \frac{x}{1+x/2}$, for $x \geq 0$ (and the inequality reverses when $x < 0$): substituting $\frac{2a(1-p)}{3p}$ for x, multiplying by $\frac{-3an}{4(1-p)^2}$ and exponentiating completes the derivation.

The second inequality follows from noting that $\frac{a^2 n}{2p(1-p) + 2a(1-p)^2/3} > a^2 n \min(\frac{1}{2a(1-p)^2}, \frac{1}{3p(1-p)})$, where one of the two terms in the denominator is individually taken as dominant, according to whether or not $p(1-p) < 2a(1-p)^2/3$. $\blacksquare$

Other separation points are also quite natural. For example, the separation point $2p(1-p) = a(1-p)^2/3$ gives $||X||_{na}^{CH} < e^{\frac{-an}{(1-p)^2}} + e^{\frac{-a^2 n}{6p(1-p)}}$.

The price for this simplification to pure exponentials is that the first term has an exponent that is too small by a factor of $\log(1+a/p)$, approximately, and the second term has an exponent that is deficient by a factor of $3/2$, or so.

The bounds compare fairly well with the estimate

$$||X||_{na}^{CH} \le \begin{cases} e^{-a^2n/2p+a^3n/2p^2}, & \text{if } a \le 2p/3; \\ e^{-2np/27}, & \text{if } a > 2p/3 \end{cases}$$

listed by Alon and Spencer [ASE-91]. A direct expansion, as in [ASE-91], when applied to Corollary 4 gives

$$||X||_{na}^{CH} \le \begin{cases} e^{-a^2n/2p(1-p)+a^3n/6p^2}, & \text{if } a \le 2p/(1-p); \\ e^{-3an/5(1-p)^2}, & \text{if } a > 2p/(1-p). \end{cases}$$

Modest improvements could be derived from Section 5.2. Of course any approximation that is an exponential of a rational function will lose the factor of $\log(1+a/p)$ in the exponent.

The use of expressive Chernoff-Hoeffding approximations originates with Angluin and Valiant [AV-79], who presented the inequality $||X||_{\epsilon\mathrm{E}[X]}^{CH} \le e^{-\epsilon^2\mathrm{E}[X]/3}$, for $0 \le \epsilon \le 1$. Comparable datings can be attained for the approximation $||X||_{\epsilon\mathrm{E}[X]}^{CH} \le e^{-\epsilon\mathrm{E}[X]\frac{\log(1+\epsilon)}{2}}$, for $\epsilon \ge 1$. Recently, similar bounds have been established for limited independence; the strongest to date appear in [SSS-93], where the first bound is shown to hold when every subset of $\epsilon^2\mathrm{E}[X]$ Bernoulli trials is guaranteed to be mutually independent, for $\epsilon < 1$, and the CH-estimate bound $||X||_{\epsilon\mathrm{E}[X]}^{CH}$, as defined for full independence is shown to hold provided the mutual independence occurs for any set of $\epsilon\mathrm{E}[X]$ Bernoulli trials, for $\epsilon \ge 1$.

5.1 Good bounds from heterogeneous dependencies

Other kinds of dependence also occur quite naturally in probabilistic algorithms, including some that are based upon sets of mutually independent random variables.

Theorem 5. Let $X = X_1 + X_2 + \cdots + X_k$ be the sum of k possibly dependent random variables. Suppose that X_i, for $i = 1, 2, \ldots, k$, is the sum of n_i mutually independent random variables having values in the interval $[0,1]$. Let $\mathrm{E}[X_i] = n_i\bar{p}_i$. Then

$$||X||_a^{CH} < e^{-\frac{a^2}{8\left(\sum_i \sqrt{\bar{p}_i(1-\bar{p}_i)n_i}\right)^2}} + e^{-\frac{3a}{4\sum_i(1-\bar{p}_i)^2}}.$$

Proof: If we estimate $||X_i||_{a_i}^{CH}$ as $e^{-\frac{a_i^2}{2\bar{p}(1-\bar{p})n_i+2a_i(1-2\bar{p}_i)/3}}$, then selecting the a_i to make each term have the same value, as prescribed by Theorem 3, should give a strong bound. Executing this objective

gives an overestimate of e^{-c} where

$$c = \frac{a_i^2}{2\bar{p}_i(1-\bar{p}_i) + 2a_i(1-\bar{p}_i)^2/3}, \quad i = 1, 2, \ldots, k.$$

Solving for a_i gives

$$a_i = c(1-\bar{p}_i)^2/3 + \sqrt{(c(1-\bar{p}_i)^2/3)^2 + 2\bar{p}_i(1-\bar{p}_i)n_i c}.$$

Summing over i yields:

$$a = \sum_i c(1-\bar{p}_i)^2/3 + \sum_i \sqrt{(c(1-\bar{p}_i)^2/3)^2 + 2\bar{p}_i(1-\bar{p}_i)n_i c},$$

whence

$$a < \sum_i c(1-\bar{p}_i)^2/3 + \sum_i \left(c(1-\bar{p}_i)^2/3 + \sqrt{2\bar{p}_i(1-\bar{p}_i)n_i c} \right).$$

It follows that

$$\sqrt{c} > \frac{a}{2\sqrt{c}\sum_i (1-\bar{p}_i)^2/3 + \sum_i \sqrt{2\bar{p}_i(1-\bar{p}_i)n_i}}.$$

Separating the cases according to which term in the denominator is larger gives

$$c > \min\left(\frac{a}{4\sum_i (1-\bar{p}_i)^2/3}, \frac{a^2}{8(\sum_i \sqrt{p_i(1-p_i)n_i})^2}\right),$$

whence

$$||X||_a^{CH} < e^{-\frac{a^2}{8\left(\sum_i \sqrt{\bar{p}_i(1-\bar{p}_i)n_i}\right)^2}} + e^{-\frac{3a}{4\sum_i (1-\bar{p}_i)^2}}. \quad \blacksquare$$

As far as exponential estimates go, these inequalities are only off by small constant factors in the exponent. The second term, of course, should capture a decay that has an additional factor of, approximately, $\log(1 + a/p)$ in the exponent, but these aggregate bounds compare rather favorably with the exponential estimates for the fully independent case. This estimate is shown to be useful in Section 7.

5.2 Further tradeoffs between expressiveness and sharpness

The term $e^{\frac{-an}{2(1-p)^2}}$, in the estimate of Corollary 4, is unnecessary when $p \geq \frac{1}{2}$. This snap action becomes evident when, as suggested by rows 3 and 4 of Table 1, the factors $(1-p)^2$ are replaced by the more precise $1-2p$, and such a strengthening can be performed, for example, to Theorem 4.

Theorem 6. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables, where $0 \le x_i \le 1$, and $\bar{p} = \mathrm{E}[X]/n$. Then

$$||X||_{na}^{CH} \le \left(1 + \frac{2a(1-2p)}{3p(1-p)}\right)^{-\frac{3an}{4(1-2p)}}.$$

Proof: It is not difficult to check that for $g(a) = \left(1 + \frac{2a(1-p)}{3p}\right)^{-\frac{3an}{4(1-p)^2}}$, $\log g(0) = 0 = (\log g)_a(0)$. Finally, straightforward calculations verify that

$$(\log g)_{aa}(a) = \frac{-1}{p(1-p) + \frac{2}{3}a(1-2p)} + \frac{3a(1-2p)}{(3p(1-p) + 2a(1-2p))^2} \ge \frac{-1}{p(1-p) + a(1-2p) - a^2}. \quad \blacksquare$$

While sharper than the bound of Theorem 4, this expression may be a little less natural to use, due to its peculiar, albeit well-defined, behavior (removable singularity) at $p = 1/2$. A straightforward weakening of this inequality gives the more natural formulation $||X||_{na}^{CH} \le e^{-\frac{a^2 n}{2p(1-p)+2a(1-2p)/3}}$, which can sustain a slight improvement as given below.

Theorem 7. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables, where $0 \le x_i \le 1$, and $\bar{p} = \mathrm{E}[X]/n$. Then

$$||X||_{na}^{CH} \le e^{-\frac{a^2 n}{2p(1-p)+2a(1-2p)/3-2a^2/9}}.$$

Proof: It is not difficult to check that for $g(a) = -\frac{a^2 n}{2p(1-p)+2a(1-2p)/3-2a^2/9}$, $g(0) = 0 = g_a(0)$. Finally,

$$\begin{aligned} g_{aa}(a) &= \frac{-p^2}{(p + a/3)^3} + \frac{-(1-p)^2}{(1-p-a/3)^3} = \frac{-1}{p + a + \frac{a^2}{3p}(1 + \frac{a}{3p})} + \frac{-1}{1-p-a + \frac{a^2}{3(1-p)}(1 - \frac{a}{3(1-p)})} \\ &\ge \frac{-1}{p+a} + \frac{-1}{1-p-a} = \frac{-1}{p(1-p) + a(1-2p) - a^2}. \quad \blacksquare \end{aligned}$$

It is worth noting that the factor $\frac{-1}{2p(1-p)+2a(1-2p)/3-2a^2/9}$ achieves the maximum value -2 when $a = 3(1-2p)/2$. This demonstrates that, as originally observed by Hoeffding [Ho-63],

$$||X||_{na}^{CH} \le e^{-2a^2 n}. \tag{7}$$

(This inequality is improved, slightly, in Section 6.1.) Similar reasoning verifies the estimates $Pr\{X \ge (1+\epsilon)\mathrm{E}[X]\} \le e^{-\frac{3}{8}\epsilon^2 \mathrm{E}[X]}$, for $\epsilon \le 1$ [AV-79]; $Pr\{X \le (1-\epsilon)\mathrm{E}[X]\} \le e^{-\epsilon^2 \mathrm{E}[X]/2}$ [ASE-91]; and $Pr\{X \ge$

$(1+\epsilon)\mathrm{E}[X]\} \leq e^{-\frac{3}{8}\epsilon \mathrm{E}[X]}$, for $\epsilon \geq 1$. Theorem 4 improves this last estimate to $Pr\{X \geq (1+\epsilon)\mathrm{E}[X]\} \leq e^{\frac{-3\mathrm{E}[X]\log(1+2\epsilon/3)}{4}} \leq e^{\frac{-\mathrm{E}[X]\log(1+\epsilon)}{2}}$.

Interestingly, a moment's thought shows that the three bounds containing the expression $1-2p$, and Hoeffding's bound (2) all hold as written for $a < 0$, with the understanding that for $a < 0$, $||X||_a^{CH} = Pr\{X - \mathrm{E}[X] \leq a\}$:

$$\forall a: \ Pr\left\{\frac{X-\mathrm{E}[X]}{an} \geq 1\right\} \leq \left(\frac{p}{p+a}\right)^{np+na}\left(\frac{1-p}{1-p-a}\right)^{n-np-na} \leq \begin{cases} e^{-\frac{a^2 n}{2p(1-p)+2a(1-2p)/3-2a^2/9}} \\ \left(1+\frac{2a(1-2p)}{3p(1-p)}\right)^{-\frac{3an}{4(1-2p)}} \\ \frac{e^{\frac{na}{1-2p}}}{(1+\frac{a(1-2p)}{p(1-p)})^{\frac{na}{1-2p}+\frac{np(1-p)}{(1-2p)^2}}}. \end{cases}$$

6.0 Martingales and sums of bounded heterogeneous random variables

For completeness and generality, we include a few standard results that have been shown to be of use in the analysis of probabilistic algorithms. The derivations in this section also show, en passant, the computational expressiveness of Theorem 7. These results are sharpened, very slightly, in Section 6.1.

Suppose $X = x_1 + x_2 + \cdots + x_n$ is the sum of n independent random variables that have differing ranges: $c_i \leq x_i \leq d_i$, and $\mathrm{E}[x_i] = m_i$. The estimate of Theorem 7 enables a direct estimate of the tail distribution of X.

It is convenient to normalize x_i, for the moment, by setting $c_i = 0$; we shall use δ_i to represent $d_i - c_i$, and μ_i to represent $m_i - c_i$. Integrating (as before) the convexity statement: for $t \in [0,\delta]$, $\lambda \geq 0$, $e^{\lambda t} \leq 1 + (e^{\lambda\delta}-1)\frac{t}{\delta}$ with respect to the probability distribution function for x_i, and multiplying by $e^{-\mu_i\lambda}$ gives the following.

$$\mathrm{E}[e^{(x_i-\mu_i)\lambda}] \leq (1-\frac{\mu_i}{\delta_i})e^{-\mu_i\lambda} + \frac{\mu_i}{\delta_i}e^{(\delta_i-\mu_i)\lambda} = (1-\frac{\mu_i}{\delta_i})e^{-\frac{\mu_i}{\delta_i}\lambda\delta_i} + \frac{\mu_i}{\delta_i}e^{\frac{(\delta_i-\mu_i)}{\delta_i}\lambda\delta_i}.$$

It follows, from the change of variables $\lambda\delta_i \to \lambda$, that $||x_i||_a^{CH} \leq ||B(1, \frac{m_i-c_i}{d_i-c_i})||_{a/(d_i-c_i)}^{CH}$.

From this fact, Lemma 1, and the application of Theorem 7 to $||B(1, \frac{m_i-c_i}{d_i-c_i})||_{a/(d_i-c_i)}^{CH}$, for each i, we conclude that

$$||X||_a^{CH} \leq \max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^{n} e^{\frac{-a_i^2}{2(m_i-c_i)(d_i-m_i)-2a_i(d_i+c_i-2m_i)-2a_i^2/9}}.$$

Now, this optimization problem is rather messy, but we can simplify it considerably by setting the parameters m_i on the right, to maximize each factor. This is done by maximizing each denominator with the selection $m_i = (c_i + d_i)/2 - a_i/3$, which yields,

$$||X||_a^{CH} \leq \max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^{n} e^{\frac{-2a_i^2}{(d_i-c_i)^2}}.$$

Equivalently, this product results from rescaling the Hoeffding bound (7). In any case, this product is easily maximized by setting the a_i so that the partial derivatives of the product with respect to a_i are all the same. This occurs when $a_i = a(d_i - c_i)^2 / \sum_i (d_i - c_i)^2$. Substituting and simplifying gives the Hoeffding bound [Ho-63] below.

Theorem(Hoeffding, 1963). Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent random variables where $c_i \leq x_i \leq d_i$.

Then

$$||X||_a^{CH} \leq e^{\frac{-2a^2}{\sum_i (d_i-c_i)^2}}. \quad \blacksquare \tag{8}$$

This inequality will be reestablished via standard, more general methods in the next section. But first, a few observations are worth making. Since the bound (8) is independent of the underlying probability distributions, we might expect the inequality to hold even if the distributions are modified in an adversarial manner. That is, suppose the random variables produce sample values sequentially; after the values $x_1, \ldots, x_k$ are known, an adversary can select the worst distribution for x_{k+1}, as long as its range is restricted to the interval $[c_{k+1}, d_{k+1}]$ and the (conditional) expectation is still m_{i+1}. The deviation bound (8) will still turn out to be valid. Then the bound must also hold for $\max_{k \leq n} \sum_1^k x_i$, since a deviation of a or more for a current partial sum can be preserved simply by fixing all subsequent random variables at their expected values.

6.1 Martingales

As noted in the discussion of Legendre transforms, many probabilistic inequalities can be attained a little more easily by applying bounds directly to the resulting generating functions, and this difference is especially useful for martingales.

A martingale is a special kind of stochastic process, which, in turn, is a family of random variables that might be viewed as a random variable evolving over time. Given n random variables $x_1, \ldots, x_n$, we can view, for example, the partial sums $Y_k = x_1 + x_2 + \cdots + x_k$ as a discrete time evolution process. Other families may be indexed over a continuous parameter space.

Informally, a sequence of random variables $Y_1, Y_2 \ldots$ is a martingale if they represent a fair game: $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_{k-1}] = Y_{k-1}$. Thus in the previous example, the Y's would be a martingale if the x_i are independent and have mean zero. In fact, they need not be independent; the probability distribution for x_i can depend on the outcome of the previous x_i's, provided the mean is always zero. A rigorous theory rests upon a foundation of measure theory; see [Do-53], [KT-81], [Lo-77] or [Wi-91].

Among the relevant developments is the following special case of Doob's submartingale inequality.

Theorem(Doob). Let $Y_1, Y_2, \ldots, Y_n$ be a martingale. Then for $a \geq 0$,

$$Pr\{\max_{1 \leq k \leq n} Y_k \geq a\} \leq \min_{\lambda > 0} e^{-\lambda a} \mathrm{E}[e^{\lambda Y_n}].$$

Proof: Actually, the theorem is more general than this statement. See Appendix 3 for the details and proof. ∎

Now, $\mathrm{E}[e^{\lambda(Y_n)}]$ might be rather difficult to compute, especially when the increments $Y_{i+1} - Y_i$ are not independent. On the other hand, the increments might satisfy a priori bounds. To be specific, let $Y_k = x_1 + x_2 + \cdots + x_k$ be a martingale, and suppose that $c_i \leq x_i \leq d_i$, for constants c_i, d_i, $i = 1, 2, \ldots, n$, and, let $\gamma_i = d_i - c_i$.

Let $h(\lambda)$ be a function that bounds, for $\lambda > 0$ the moment generating functions of any Bernoulli trial: $\max_{p \in [0,1]} \mathrm{E}[e^{\lambda(B(1,p)-p)}] \leq h(\lambda)$. Then $\max_{p \in [0,1]} \mathrm{E}[e^{\lambda\gamma(B(1,p)-p)}] \leq h(\gamma\lambda)$, and $h(\lambda\gamma_i)$ is an upper bound for the function $\mathrm{E}[e^{\lambda x_i} \mid Y_1, Y_2, \ldots, Y_{i-1}]$, since the moment generating functions of the rescaled Bernoulli trials comprise the extreme points of the convex region defined by such expectations.

A bound for $\mathrm{E}[e^{\lambda Y_n}]$ is established as follows.

$$\begin{aligned}
\mathrm{E}[e^{\lambda Y_n}] &= \mathrm{E}[\mathrm{E}[e^{\lambda Y_{n-1}} e^{\lambda x_n} \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&= \mathrm{E}[e^{\lambda Y_{n-1}} \mathrm{E}[e^{\lambda x_n} \mid Y_1, Y_2, \ldots, Y_{n-1}]] \\
&\leq \mathrm{E}[e^{\lambda Y_{n-1}} h(\lambda \gamma_n)] \\
&\leq \mathrm{E}[e^{\lambda Y_{n-1}}] h(\lambda \gamma_n) \\
&\leq \quad \cdots \\
&\leq \prod_i h(\lambda \gamma_i),
\end{aligned}$$

whence

$$Pr\{\max_{1 \leq k \leq n} Y_k \geq a\} \leq \min_{\lambda > 0} e^{-\lambda a} \prod_i h(\lambda \gamma_i). \tag{9}$$

Hoeffding [Ho-63] uses the (slightly bemusing) inequality

$$\mathrm{E}[e^{\lambda(B(1,p)-p)}] \leq e^{\lambda^2/8} = h(\lambda), \tag{10}$$

for h in (9), and selects the minimizing $\lambda = 4a/\sum_i (d_i - c_i)^2$, which establishes (8) in the more general context of martingales: $Pr\{\max_{1 \leq k \leq n} Y_k \geq a\} \leq e^{\frac{-2a^2}{\sum_i (d_i - c_i)^2}}$. The bound (10) is usually proved via some cleverness and sophistication (c.f. [Ho-63], [ASE-91]). An alternative proof is to apply (7): if (10) were false at some λ_0, for some $B(1,p)$, pick a so that $||B(1,p)||^L_{\lambda,a}$ is optimized at λ_0. The resulting value would then be greater than $\min_{\lambda > 0} e^{-a\lambda + \lambda^2/8} = e^{-2a^2}$, which would contradict (7): $||B(1,p)||^{CH}_a \leq e^{-2a^2}$.

More abstractly, $\log(e^{-2a^2})$ is the dual (under the Legendre transform) of $\log e^{\lambda^2/8}$; the inequalities are equivalent, in the respective spaces of Chernoff-Hoeffding estimates and moment generating functions. Actually, it is worth observing that both the e^{-2a^2} estimate (7) and dual $e^{\lambda^2/8}$ bound (10) are easily established by applying the logarithm and differentiating twice, in the respective variables a and λ. More generally, when these simple steps suffice, as in Table 1, to confirm that, say $e^f < e^g$, for convex f and g, then, as is easily verified, the same approach must also succeed for their Legendre transforms.

Now, the underlying reason the inequality $||X||^{CH}_a \leq e^{-2a^2}$, for $X:\ 0 \leq X \leq 1$, and its dual are not Chernoff-tight is that solving for the exact bound, $\max_p ||B(1,p)||^{CH}_a = \max_p \left(\frac{p}{p+a}\right)^{a+p} \left(\frac{1-p}{1-p-a}\right)^{1-a-p}$

leads to a transcendental equation. On the other hand, we can attain a slight improvement via the following observation. Let $f(a) = \left(\frac{p}{p+a}\right)^{a+p}\left(\frac{1-p}{1-p-a}\right)^{1-a-p}$. We have shown that $f(a) \leq e^{-2a^2}$ by observing that $\log f(0) = (\log f)_a(0) = 0$, and calculating that $(\log f)_{aa}(a) = -\frac{1}{(p+a)(1-p-a)} \leq -4 = (-2a^2)_{aa}$. Since $(\log f)_{tt}(t) = -\frac{1}{(p+t)(1-p-t)}$, for $t \in [0,a]$, we may fix $p = \frac{1}{2} - \frac{a}{2}$, and observe that these (negative) values as a weighted set $\{(-\frac{1}{(p+t)(1-p-t)}, \mathrm{d}t)\}$ are as large as possible, since the denominator is as large as possible (and the numerator is negative.) Unfortunately, integrating this $(\log f)_{tt}(t)$ twice does not give the maximal $\log f$, but if we rearrange the collection of values to be monotone decreasing, then we will get an overestimate for f. Informally, the greatest acceleration should come first, if the maximum distance is to be traversed, and this fact is readily formalized[†] via integration by parts. In any case, the consequence is

$$\log(f(a)) \leq \int_0^a \int_0^s \frac{-1}{(1/2+t/2)(1/2-t/2)} dt ds = 4\int_0^{a/2}\int_0^x \frac{-1}{(1/2+y)(1/2-y)} dy dx.$$

This last expression is just $\log\left(||B(1,\frac{1}{2})||^{CH}_{a/2}\right)^4 = \log\left(||B(4,\frac{1}{2})||^{CH}_{2a}\right)$. If $X_i = x_1, x_2, \ldots, x_i$ is the sum of i independent random variables, with $0 \leq x_k \leq 1$, for each k, then

$$Pr\{\max_{1\leq i\leq n}(X_i - \mathrm{E}[X_i]) \geq an\} \leq ||B(4n,\frac{1}{2})||^{CH}_{2an} = \left(\frac{1}{1+a}\right)^{2n+2na}\left(\frac{1}{1-a}\right)^{2n-2na}.$$

Equivalently, the dual formulation reads, $\mathrm{E}[e^{\lambda(X_n - \mathrm{E}[X_n])}] \leq (\cosh(\frac{\lambda}{4}))^{4n}$.

This expression has the virtue of being a Chernoff-Hoeffding estimate for an actual probability distribution, although it is only slightly sharper than the estimate of $e^{\lambda^2 n/8}$, since the optimal λ will be small. Moreover, this improvement only complicates the resulting optimization problem. On the other hand, we may always use the most convenient formulation to calculate or estimate λ, and safely substitute the value elsewhere. This simple rearrangement approach can be readily sharpened when the x_i are, say, Bernoulli trials having expectations confined to some fixed subinterval of $[0,1]$. Such estimates might occur from upper and lower bounds for the optimal p that yields the maximum Chernoff-Hoeffding estimate.

[†]Also needed is the following inequality of Chebyshev: Suppose $f(x)$ and $g(x)$ are nonnegative and monotone increasing. Then $\int f(x)g(x)\mathrm{d}x \geq \int f(x)g(\pi(x))\mathrm{d}x$, where π is a measure preserving transformation.

If $Y_i = x_1 + x_2 + \cdots + x_i$ happens to be a martingale with bounded increments, $x_i \in [-b_i, c - b_i]$, then rescaling shows that

$$Pr\{\frac{1}{c}\max_{i\le n} Y_i > an\} \le ||B(4n,\frac{1}{2})||^{CH}_{2an} = \left(\frac{1}{1+a}\right)^{2n+2an}\left(\frac{1}{1-a}\right)^{2n-2an}.$$

For completeness, we note that the (mean zero) increments x_i often belong to a symmetric interval: $x_i \in [-c/2, c/2]$. For this case, Azuma [AZ-76] presented a lemma establishing the following Chernoff-tight estimate.

$$Pr\{\frac{1}{c}\max_{i\le n} Y_i > an\} \le ||B(n,\frac{1}{2})||^{CH}_{an} = \left(\frac{1}{1+2a}\right)^{n/2+an}\left(\frac{1}{1-2a}\right)^{n/2-an}.$$

The bound follows from the inequality $\mathrm{E}[e^{\lambda x_i}] \le \mathrm{E}[e^{c\lambda(B(1,\frac{1}{2})-\frac{1}{2})}]$, which, in turn, is derived from a convexity argument similar to that given for Hoeffding's bound (2). The lemma is sometimes called the Azuma-Hoeffding inequality (c.f. [Wi-91]).

6.2 An easy extension

Hoeffding pointed out [Ho-63] that in the case of fully independent random variables, Doob's Theorem strengthens the inequalities to bound the probability that any (of the linearly ordered) partial sums exceed the deviation. It is straightforward to observe that this strengthening also applies to the Chernoff-Hoeffding bounds with conditioning in Section 3.

6.3 A very simple stochastic domination bound

In some computer science applications, items disappear according to a complicated Bernoulli process, and estimates are needed to bound the likelihood that an unusually large number have survived, at specific points of the process. Here the complication is that item i may have a probability of survival p_i that depends on the survival properties of the other items. On the other hand, it is often the case that the p_i can be easily bounded by readily attained ρ_i.

Theorem(Folklore). Let $X_k = \sum_{i=1}^{k} x_i$ and $Y_k = \sum_{i=1}^{k} y_i$, where x_i and y_i are Bernoulli trials. Let the y_i be mutually independent, and suppose that $\mathrm{E}[x_i \mid X_{i-1}] \le \mathrm{E}[y_i]$. Then $\forall a$:

$$Pr\{X_n \ge a\} \le Pr\{Y_n \ge a\}.$$

Proof: Given X_k, we may compute x_{k+1} and y_{k+1} jointly via the following simplified accept-reject method. Let $r_{k+1} = \mathrm{E}[x_{k+1} \mid X_k]$, and compute a random y_{k+1}, which is 1 with probability $\rho_{k+1} = \mathrm{E}[y_{k+1}]$. If $y_{k+1} = 0$, we take $x_{k+1} = 0$ as well. Otherwise, we compute a second random Bernoulli trial that has outcome 1 with probability $r_{k+1}/\rho_{k+1} < 1$. If this second outcome is also 1, we set $x_{k+1} = 1$ as well; otherwise it is taken to be 0. It is easy to see that this method computes X_{k+1} according to its correct distribution. Moreover, the x's will always have a summation that is bounded by the y's. The inequality follows. ∎

This elementary inequality illustrates two points. First, deviation bounds (from the mean) are more sophisticated than some simple survival probabilities. Second, the proof shows how resampling can be applied to use one process to attain bounds on another. The next section gives a more complicated example of this resampling approach, and shows how many of the inequalities of the previous sections may be useful en route to a complicated estimation result.

7. Applications

Hashing comprises a variety of methods for storing and retrieving data in an array A, based on computed index probes. Formally, let S be a of αn items, where $\alpha \leq 1$. The array A will have n locations and will be used to store the elements of S, or perhaps pointers to the elements. The elements are referenced by their names, which are called hash keys. The idea behind hashing is to have a simple function h that maps each hash key to an array location j, $1 \leq j \leq n$. Functions where the keys are mapped 1 to 1 into the range of array indices are called perfect hash functions (c.f [FKS-84]).

Most classical hashing schemes use random functions to compute array indices. When two items hash to the same location, we say that a collision has occurred. The issue of how collisions are resolved is what distinguishes the bulk of the classical hashing schemes.

Closed hashing (also called hashing with open addressing) uses additional probes (randomly computed array locations) into A to place a colliding item in the first vacant slot that is found. Double hashing, for example, uses the key of an element x to compute an arithmetic progression of array indices and places x in the first vacant location found in the progression. The analysis of double hashing is the subject of Section 7.2.

Open hashing links colliding items together in a linked list. Search proceeds to the hash index and then along the linked list without subsequent hashing probes. The reason that these schemes have variations is that the colliding items may be stored within the hash table A, or outside in some auxiliary storage area. Alternatively, a mix of these two possibilities is feasible, and suitable hybrid schemes turn out to have the best performance among this set of hash algorithms ([VC-87]). Such schemes are called coalesced hashing because the linked lists can coalesce when list elements are located within the hash table.

The analysis of coalesced hashing leads to the need to compute stopping times and the probability of large deviations for such times.

7.1 A stopping time calculation

The basic stopping time problem for coalesced hashing can be formulated as follows.

At time $T = 0$, an array has n vacant slots. A cellar count C is initially zero. At each time $T = j > 0$, a slot is selected at random and, if it is vacant, marked full. If the the slot is already full, C is increased by one. In either case, the slot is kept available for subsequent selection. The stopping time S is defined to be the time when the cellar count C first becomes αn, for a fixed value α.

Part of the analysis of optimal coalesced hashing (c.f. [VC-87]) requires an estimate of the probability that S has a large deviation from its mean. Let x_j be one if, at time $T = j$, the probe into the array selects a full location, and let x_j be zero if the location is vacant. Then we may model C as the sum of the Bernoulli trials: $C(t) = \sum_{j=1}^{t} x_j$. Since the number of full locations, at the beginning of time $t+1$, is $t - C(t)$,

$$x_{t+1} = \begin{cases} 1 & \text{with probability } \frac{t-C(t)}{n}, \\ 0 & \text{with probability } 1 - \frac{t-C(t)}{n}. \end{cases}$$

We may transform $C(t)$ into a martingale by noting that $\mathrm{E}[C(t+1) \mid C(t)] = \frac{n-1}{n}C(t) + \frac{t}{n}$. In particular, set $Z(1) = 0$, $Z(t) = (\frac{n}{n-1})^{t-1}(C(t) - t + n) - n + 1$. Direct substitution shows that $\mathrm{E}[Z(t) \mid Z(t-1)] = Z[t-1]$, so Z is a martingale.

Furthermore, the increment $Z(t+1) - Z(t) = (\frac{n}{n-1})^{t-1}\frac{C(t)+nx_{t+1}-t}{n-1}$, which is easily seen to lie in the interval $[-(\frac{n}{n-1})^{t-1}, (\frac{n}{n-1})^{t-1}]$, since $x_{t+1} = 1$ if all n slots are already full. Thus the increments are bounded by $(\frac{n}{n-1})^{t-1}$, whence (8) is applicable to give $Pr\{Z_t - \mathrm{E}[Z_t] > a(\frac{n}{n-1})^{t-1}\} \leq e^{\frac{-a^2}{2n}}$. In terms of

C, we see that $Pr\{C(t) - \mathrm{E}[C(t)] > a\} \leq e^{\frac{-a^2}{2n}}$. If t is taken as $\mathrm{E}[S]$, then the cellar count will only have a polynomially small probability of varying by more than $c\sqrt{n \log n}$. We may infer that the probability that the cellar first becomes full at a time outside of $[\mathrm{E}[S] - c\sqrt{n \log n}, \mathrm{E}[S] + c\sqrt{n \log n}]$ is polynomially small.

Evidently, the number of filled slots, at this stopping time S will also fluctuate by a modest $O(\sqrt{n \log n})$, with high probability. Similarly, it would not be difficult to normalize the random variable counting the number of filled table slots, to attain, thereby, a martingale with bounded increments.

To estimate $\beta n \equiv \mathrm{E}[S]$, we may reason as follows. If we stop the process at the exact time βn, then number of items in the cellar should be, approximately, αn. $\mathrm{E}[C(\beta n)] \approx \alpha n$ Substituting these values into the formulation for our Martingale at time $s = \beta n$ gives $0 = \mathrm{E}[Z(s)] \approx (\frac{n}{n-1})^{s-1}(C(s) - s + n) - n + 1$, or $e^{\beta}(\alpha - \beta + 1) = 1$, as a formulation for a sharp estimate. Even more can be deduced about these stopping times by saddle point methods. See, for example, [Si-94].

As this section illustrates, martingale formulations give sharp and direct procedures for capturing larger deviation estimates in some processes such as stopping times, despite the presence of dependencies within the underlying Bernoulli trials.

7.2 Double hashing

Double hashing is a method for storing and retrieving data in an array A based on computed index probes. The insertion procedure places an item x in a table of size n, for prime n, by inserting the item in the first vacant location in the table slots $A[h(x) - (k)f(x) \bmod n]$, $k = 0, 1, 2, \ldots, n-1$. Here $h(x)$ is assumed to be a random function uniformly distributed over $[0, n-1]$ and the stride f is independent and uniformly distributed over $[1, n-1]$. Lookup works by testing the elements found according to the same probe sequence.

The performance analysis of this method is complicated by the mild dependencies induced by restricting the probe sequences to arithmetic progressions. Lueker and Molodowitch, in an elegant analysis of double hashing [LM-88], show that the expected number of probes to insert the $(\alpha n + 1)$-st item, in double hashing, is bounded by the expected number of probes to insert the $(\alpha n +$

$O(\sqrt{n}\log^{5/2} n))$-th item into a table where the items are distributed randomly† with each configuration equally likely. It follows, therefore, that the expected number of probes to insert this item is $\frac{1}{1+\alpha} + O(\log^{5/2} n/\sqrt{n})$.

The proof is inductive, and is based on a scheme that keeps the table configuration completely random. They achieve this by selecting the next vacant slot for insertion according to a formal accept-reject probabilistic calculation of the double hashing probability that is embedded within a uniform selection procedure.

In particular, let A be currently filled with βn items, and let p_i be the probability that the i-th vacant slot would get the next item were it to be inserted according to double hashing. Let $H = \max_i(p_i)$, and consider a bar graph of the p_i in a box of dimension $H \times (1-\beta)n$. Imagine tossing a dart randomly, with uniform distribution, into the box. If the dart lands on the j-th bar, then the item is inserted in the j-th vacant slot, and the insertion conforms to the double hashing distribution. In the (moderately improbable) event that the dart lands above some bar, a dummy item is inserted in the corresponding slot, and the (double hashing) insertion attempt is repeated for the datum at hand with an updated bar graph. Thus vacant slots receive items according to the uniform distribution, while the actual data is double hashed into the table, which happens to have endured the occasional insertion of a few extra items by some other scheme.

Lueker and Molodowitch exploit monotonicity in the sense that the insertion of additional items into a hash table can only increase the number of probes needed for subsequent insertions. This is a form of probabilistic resampling, which replaces one algorithm by a similar one that is easier to analyze and has a guaranteed performance that is worse (better) for upper (lower) bounds.

The technical portion of the proof is to show that for a random table containing βn items uniformly distributed among all $\binom{n}{\beta n}$ possibilities, with overwhelming probability (over the table configurations),

---

†Moreover, the hashing model with uniform distribution and random probe selection is the optimum for this genre of hashing [Ya-85].

the box height H satisfies:

$$H = \left(1 + O(\frac{\log^{5/2} n}{\sqrt{n}})\right) \frac{1}{(1-\beta)n}.$$

This bound guarantees that the next insertion will be a dummy with probability $O(\frac{\log^{5/2} n}{\sqrt{n}})$. Then the (uninteresting) probability estimate of Section 6.3 can be used to show that with overwhelming probability, only $n \times O(\frac{\log^{5/2} n}{\sqrt{n}})$ dummy items will be inserted into the table.

We now show that the Chernoff-Hoeffding bounds of the previous sections combine quite naturally to simplify the calculations in the Lueker-Molodowitch proof strategy, and give the marginally better error of $O(\sqrt{\frac{\log n}{n}})$, as opposed to $O(\frac{\log^{5/2} n}{\sqrt{n}})$.

Lueker and Molodowitch bound H by computing (1) a large deviation bound for the number of different double hashing probe sequences $(f, h) \in [0, n-1] \times [1, n-1]$ that hit vacant slot l in a table containing βn randomly placed items. This estimate is based on (2) a corresponding bound where each location is occupied with independent Bernoulli probability $\delta = \beta + c_2\sqrt{\frac{\log n}{n}}$. Each probe sequence (f, h) is represented as an arithmetic progression (s, k) having strides of length s and requiring k steps to reach l. A k-step sequence has a probability δ^k that all k locations will be occupied, which would cause location l to be hit by the implicit double hashing probe sequence. But the Bernoulli trials represented by the sequences are not independent. A specific k-step sequence will have locations in common with at most $k(k-1)$ other k-step sequences. Lueker and Molodowitch use (3) a complicated partitioning argument to partition the set of all $n-1$ k-step sequences into subsets where each location (other than l) belongs to at most one sequence within a given subset, so that the implicitly represented Bernoulli events within each partition will be independent. They then use (4) a Chernoff-Hoeffding estimate to bound the probability that any of these subsets has a deviation that is in excess of some share of the $O(\sqrt{n}\log^{5/2} n)$ aggregate deviation.

Theorem 2 gives a generic method that in this application transforms the bound achieved for step (2) to the model in step (1) very conveniently and with less cost than the approach in [LM-88]. Corollary 3, plus the fact that a degree $k(k-1)$ graph is k^2 colorable gives a simplified and stronger transition from (3) to (4). Now a trivial Chebyshev bound can used to bound k. Finally, the expressive Chernoff-Hoeffding bound stated in Theorem 5 allows, as outlined below, a simple deviation

of $a = c\sqrt{n \log n}$ to be allotted implicitly in a way which results in large deviation probabilities that decay quite rapidly.

The net result is a fully systematic and completely simple emulation proof that inserts, with overwhelming probability, $O(\sqrt{n \log n})$ extra items instead of $O(\sqrt{n} \log^{5/2} n)$. The streamlined calculations are as follows. Given a table containing βn items distributed at random, let slot l be selected at random, without regard, say, to its vacancy. In the case of random probing (without replacement), the probability that a probe sequence will reach l before reaching a vacant slot is $p_U = \frac{1}{n} + \frac{\beta}{(1-\beta)n+1}$, where the first term is the probability for the case that l is vacant, and the latter for the event that l is occupied. In double hashing, the probability for an individual slot depends on the distribution of the occupied slots, but the expected probability for a randomly selected slot must be the same, since the occupancy distribution is random. If each slot is occupied according to an independent Bernoulli trial with probability β, then the expected probability, in both cases, is $p_B = \frac{1}{n}(1 + \beta + \beta^2 + \cdots + \beta^{n-1}) = \frac{1-\beta^n}{(1-\beta)n}$. Evidently, $p_B - p_U \approx \frac{\beta}{(n-\beta n)^2}$, and the expected number of double hashing probes that reach l differ by less than $\frac{\beta}{(1-\beta)^2}$, for the two models.

Let

$$P \equiv Pr\{ \text{ at least } n(n-1)p_U + \frac{\beta}{(1-\beta)^2} + c\sqrt{n \log n} \text{ probe sequences hit slot } l \}.$$

According to Theorem 2, $P \leq 2Pr_B\{$ at least $n(n-1)p_B + c\sqrt{n \log n}$ probe sequences hit slot $l\}$, where Pr_B uses a table that has each slot independently occupied with probability β. Consequently, we have that $P \leq 2Pr_B\{Z \geq 1\} + 2||\sum_{k=0}^{c_1 \log n} X_k||_{c\sqrt{n \log n}}^{CH}$, where Z and X_k are defined as follows. X_k is the number of arithmetic progressions comprising k occupied slots that terminate at l, in the Bernoulli trial model. Z is the number of arithmetic progressions comprising $c_1 \log n$ filled slots that terminate at location l. $Pr_B\{Z \geq 1\}$ is an overestimate of the probability any probe progression to l has length $c_1 \log n$ or more, since the existence of a longer sequence implies that a length $c_1 \log n$ one also exists. Its presence enables the summation for the X_k to terminate at $k = c_1 \log n - 1$. Now the random variable X_0 is a constant, it cannot contribute to the variance. Similarly, X_1 is the same constant in double hashing and random probing without replacement, when the βn items are distributed at random, and its mean turns out to be unchanged for the Bernoulli model. Consequently, we may take

the index k in the summation to satisfy $k \geq 2$.

Altogether, $\sum_{k=2}^{c_1 \log n} X_k$ comprises $(c_1 \log n - 1)(n-1)$ partially dependent Bernoulli trials, which we shall count as $c_1 n \log n$. X_k is the sum of $n-1$ dependent Bernoulli trials with individual probability of success β^k, and a k^2 colorable conflict graph. Corollary 3 says that the worst case Chernoff-Hoeffding estimate occurs if we have k^2 copies of the same n/k^2 independent trials. Estimating $Pr\{Z \geq 1\}$ as E$[Z]$ via Chebyshev's inequality, and applying Theorem 5 (along with Corollary 3) to the remaining term gives

$$P < 2n\beta^{c_1 \log n} + 2e^{-\frac{c^2 n \log n}{8\left(\sum_k^{c_1 \log n} k^2 \sqrt{\beta^k (1-\beta^k) n/k^2}\right)^2}} + 2e^{-\frac{3c\sqrt{n \log n}}{4\sum_k^{c_1 \log n} k^2 (1-\beta^k)^2}}.$$

The estimate is polynomially small because $\left(\sum_k k\sqrt{\beta^k(1-\beta^k)n}\right)^2$ is clearly linear in n, due to the exponential decay of β^k. The other term is negligible because the summation stops at $k = c_1 \log n$.

Note that Corollary 3 is not strong enough to be applied directly to the collection of all $c_1 n \log n$ trials because it is a worst case bound that knows nothing about how the sets are partitioned. Indeed, we are saved by the fact that the less probable events exhibit the greater dependence, and not vice versa. Restated, the heterogeneity of the problem and Theorem 5 combine to guarantee that a $c\sqrt{n \log n}$ deviation will occur with a polynomially small probability, as opposed to an estimate of $e^{\frac{-c^2}{\log n}}$, which would result from Corollary 3, and which would indeed be Chernoff-tight in the homogeneous case, where all set sizes and probabilities are the same as their average.

8. Conclusions

We have shown how to achieve deconditioning by translating selection problems into comparable inequalities for Bernoulli problems, for random variables that exhibit a form of monotonic conditioning. Furthermore, methods to handle complicated dependencies have been presented and shown to be sharp and useful. Similarly, probability inequalities have been derived that exploit heterogeneity within the underlying probability distributions. Moreover, these techniques can be combined to simplify complicated problems while yielding improved deviation estimates.

The expressive Chernoff-Hoeffding estimate stated in Theorem 7 has the advantages of being case free and readily applicable to decompositions with different probabilities. These properties may be of especial benefit when probabilistic decompositions give subevents of different sizes and probabilities,

and where global statistics are readily available. This is also true of Theorem 5, and the more general estimation formulations in Theorem 3 and its corollaries.

Acknowledgements

It is a pleasure to thank Joel Spencer for helpful advice and suggestions, which in no small way motivated this work. Don Coppersmith and Jeanette Schmidt provided insightful and stimulation discussion, which is gratefully acknowledged. Steve Samuels and S. R. S. Varadhan were wonderful sources of published results, as were Samuel Karlin, Larry Shepp, and Alan Weiss.

References

[AZ-67] K. Azuma. Weighted sums of certain dependent random variables. *Tôhoko Mathematics Journal*, **19**(3), 1967. 357–367.

[ASE-91] N. Alon, J. H. Spencer and P. Erdos. *The Probabilistic Method with appendix of open problems by P. Erdos.* Wiley-Interscience, 1991.

[AV-79] D. Angluin and L. G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *JCSS* **18**, 1979, 155–193.

[Ba-60] R. R. Bahadur. Some approximations to the binomial distribution function. *Ann. Math. Stat.*, 31(1960), 43–54.

[Be-24] S. Bernshtein. Sur une Modification de l'inégalité de Tchebichef. *Ann. Sci. Instit. Sav. Ukraine, Sect. Math.*, **I**(Russian, French Summary), 38–48, 1924. See also *S. Bernshtein Collected Works*, (translit.) *Sobranie sochineniĭ*, Vol. IV, #5, The Academy of Sciences of U.S.S.R., Moscow, CCCP, 1964, 71–79, (Russian).

[Be-62] G. Bennett. Probability inequalities for the sum of independent random variables. *J. Am. Stat. Ass.*, 57(1962), 33–45.

[Bi-53] I. J. Bienaymé. Considérations a l'appui de la découverte de Laplace sur la loi des probabilitès dans le mèthod des moindres carrès. *Comptes Rendus Acad. Sci.*, Paris, Vol. 37, 1853.

[Ch-67] P. Chebyshev (also spelled Tchebichef, Tchebycheff). Des valuers moyennes. *Journal de Mathématics Pures et Appliquées*, 12(1867).

[Ch-52] H. Chernoff. A measure of asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations. *Ann. Math. Statist.*, 23(1952), 493–507.

[Do-53] J. L. Doob. *Stochastic Processes.* Wiley & Sons, Inc., 1953.

[ET-76] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems.* North-Holland, 1976.

[FKS-84] M.L. Fredman, J. Komlós and E. Szemerédi. Storing a Sparse Table with $O(1)$ Worst Case Access Time. *Journal of the Association for Computing Machinery*, Vol 31, No. 3, July 1984, pp. 538–544.

[Ho-56] W. Hoeffding. On the distribution of the number of successes in independent trials. *Ann. Math. Stat.*, 27(1956), 713–721.

[Ho-63] W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *J. Am. Stat. Ass.*, 58(1963), 13–30.

[Ho-87] M. Hofri. *Probabilistic Analysis of Algorithms.* Springer-Verlag, 1987.

[Je-06] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math.* **30**(1906), 175–193.

[JS-68] K. Jogdeo and S. M. Samuels. Monotone Convergence of Binomial Probabilities and a Generalization of Ramanujan's Equation. *Ann. Math. Stat.*, 39(1968), 1191–1195.

[KS-66] S. Karlin and W. J. Studden. *Tchebycheff Systems: With Applications in Analysis and Statistics.* Wiley-Interscience, 1966.

[KT-81] S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, 1981.

[Ko-28] A. N. Kolmogorov. Ueber die Summen durch den Zufall bestimmter unabhängiger Grössen. *Math. Ann.*, 99(1928), 309–319.

[Lo-77] M. Loève. *Probability Theory I*, Springer-Verlag, 1977.

[LM-88] G. Lueker and M. Molodowitch. More Analysis of Double Hashing. *Proceedings of the Twentieth Annual Symposium on Theory of Computing*, 1988, 354–359.

[Ma-13] A. A. Markov. *Ischislenie Veroiatnosteĭ* (Trans. The Calculus of Probabilities), 3rd ed., Gosizdat, Moscow, 1913, (Russian).

[Ma-14] A. A. Markov. The Bicentennial of the Law of Large Numbers, an address to the Russian Academy of Sciences, Dec. 1, 1913, *Journal (Vestnik) of Experimental Physics and Elementary Mathematics*, 2nd series, Semester 1, No. 3, Odessa, 1914, 59–64, translated [On-83] 158–163.

[MO-79] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, 1979.

[Mo-70] W. Molenaar. *Approximations to the Poisson, Binomial and Hypergeometric Distribution Functions.* Mathematical Centre Tracts, 31, Mathematisch Centrum, Amsterdam, 1970.

[Ne-83] P. Ney. Dominating points and the asymptotics of large deviations for random walk on $I\!R^d$. *Ann. Prob.*, 11(1983), 158–167.

[On-83] Kh. O. Ondar, editor. *The Correspondence Between A.A. Markov and A.A. Chuprov on the Theory of Probability and Mathematical Statistics.* Translated by Charles and Margaret Stein, Springer-Verlag, 1981.

[Pr-53] Y. U. Prohorov. Asymptotic behavior of the binomial distribution. English version appears in *Selected Translations in Math. Stat. and Prob.*, vol. 1, A.M.S., 1961, 87–96.

[Ra-88] P. Raghavan. Probabilistic Construction of Deterministic Algorithms: Approximating Packing Integer Programs. *JCSS*, 37(1988), 130–143.

[SSS-93] J. P. Schmidt, A. Siegel and A. Srinivasan. Chernoff-Hoeffding Bounds for Applications with Limited Independence. *Proc. 4th Ann. ACM-SIAM Symp. on Discrete Algorithms*, 1993, 331–340.

[SW-92] A. Schwartz and A. Weiss. *Large Deviations and their Applications to Computer Science and Communication Systems.* Draft, 1992.

[Se-87] M. J. Sewell. *Maximum and minimum principles.* Cambridge University Press, 1987.

[Si-94] A. Siegel. Coalesced hashing is randomizable and computable via universal hash functions. In progress.

[Vi-90] I.M. Vinogradov, ed. *Encyclopedia of Mathematics*, Kluwer Academic Publishing, Vol. 5, 1990, 374, (trans. *Soviet Mathematical Encyclopedia*, 1977–85).

[VC-78] J. S. Vitter. *Design and Analysis of Coalesced Hashing* Oxford University Press, 1987.

[Wi-91] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[Ya-85] A. C. Yao. Uniform Hashing Is Optimal. *JACM*, Vol. 32, No. 3, 1985, 687–693.

Appendix 1. Convexity and Jensen's Inequality

A function $f(t)$ on $[a,b]$ is defined to be convex if for every $t \in (a,b)$, and every x, y, where $a \le x < t < y \le b$, the point $(t, f(t))$ lies below (or on) the line segment from $(x, f(x))$ to $(y, f(y))$.

Consequences of convexity are:

1) For each point $(t, f(t))$, there is at least one line l passing through it where no point on the curve $(z, f(z))$ lies below l. Such a line is called a support line.

2) The function f is continuous on (a,b), and upper semicontinuous at a and b.

3) If f is convex and l is a secant line through $X_1 = (x, f(x))$ and $X_2 = (y, f(y))$ with $x < y$, then for $t < x$ or $t > y$, $(t, f(t))$ lies above (or possibly on) l.

4) A function f is convex iff $\frac{\mathrm{d}f}{\mathrm{d}x}$ is nondecreasing (i.e. $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$ is a positive distribution.)

The first fact follows from observing that $(t, f(t))$ will have no support line iff there is a line l, which passes through the point, and has points on the curve $(z, f(z))$ that lie below l for $z < t$ and for $t > z$, whence lowering l slightly shows that f is not convex.

The second and third facts follow similarly.

The fourth fact follows from considering a secant line through $X_1 = (x, f(x))$ and $X_2 = (y, f(y))$, for $x < y$. Since f is convex, the line must be rotated (at least zero radians) clockwise about X_1 to become a support line, and counterclockwise about X_2. Hence the slopes of the support lines at X_1 are no larger than the slopes of those at X_2.

It follows that for a convex function f, the convex closure of the points $(x_1, f(x_1))$, $(x_2, f(x_2))$, $\ldots$, $(x_n, f(x_n))$ lie on or above the curve $(t, f(t))$.

Lemma B(Jensen, 1906). Let f be convex. Suppose the coefficients a_i are nonnegative with $a_1 + a_2 + \cdots + a_n = 1$. Then

$$f\left(\sum_i a_i x_i\right) \le \sum_i a_i f(x_i).$$

Proof: The point defined by the convex average, in two dimensional space, $\sum_i a_i(x_i, f(x_i))$ lies on or above the curve $y = f(x)$, which is to say that $f\left(\sum_i a_i x_i\right) \le \sum_i a_i f(x_i)$. ∎

Of course the inequality is reversed for concave f. If $\log f$ is convex, then the lemma applies to products: $f\left(\sum_i a_i x_i\right) \le \prod_i \left(f(x_i)\right)^{a_i}$.

Appendix 2. Five Proofs

2.1. Lemma 0. Let X be a random variable with moment generating function $G(\lambda) = \mathrm{E}[e^{\lambda X}]$. Then

0) If X and Y are independent with respective moment generating functions $G(\lambda)$ and $H(\lambda)$, then $X + Y$ has the moment generating function $G(\lambda)H(\lambda)$.

1) If $G(\lambda)$ is bounded for $\lambda \in (\alpha, \beta)$ then G is analytic for $\alpha < Re(\lambda) < \beta$, and we may differentiate as often as we please to attain $\frac{\mathrm{d}^k G}{\mathrm{d}\lambda^k} = \mathrm{E}[X^k e^{\lambda X}]$. Moreover, limit arguments can be applied to extend this formulation to the cases where $\lambda = \alpha, \beta$.

2) $G(\lambda) = \mathrm{E}[e^{\lambda X}]$ equals one at $\lambda = 0$, and is strictly increasing (or infinite) for $\lambda > 0$, provided $\mathrm{E}[X] \ge 0$ and $X \not\equiv 0$.

3) $G(\lambda)$ is strictly log-convex where it exists, provided X is not a constant.

Proof:

0) X and Y are mutually independent if and only if $Pr\{X \in [t, t+dt) \bigwedge Y \in [s, s+ds)\} = Pr\{X \in [t, t+dt)\} Pr\{Y \in [s, s+ds)\}$. Using this fact gives:

$$\begin{aligned}
\mathrm{E}[e^{\lambda(X+Y)}] &= \int_{t,s} e^{\lambda(t+s)} Pr\{X \in [t, t+dt) \bigwedge Y \in [s, s+ds)\} \\
&= \int_{t,s} e^{\lambda t} e^{\lambda s} Pr\{X \in [t, t+dt)\} Pr\{Y \in [s, s+ds)\} \\
&= \int_t e^{\lambda t} Pr\{X \in [t, t+dt)\} \int_s e^{\lambda s} Pr\{Y \in [s, s+ds)\} \\
&= \mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\lambda Y}].
\end{aligned}$$

1) These facts are based on standard analytic function theory. The gist is that a convergent sum of analytic functions is also analytic provided there is sufficient regularity to permit the differentiation operator to be passed through the summation (integration) operator.

2) By definition, $G(0) = \mathrm{E}[1] = 1$. Differentiating with respect to λ gives $\frac{\mathrm{d}}{\mathrm{d}\lambda}G(0) = \mathrm{E}[X]$, and $\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}G(\lambda) = \mathrm{E}[X^2 e^{\lambda X}] > 0$. Hence G is strictly increasing on $\lambda \geq 0$.

3) Differentiating gives, $G_\lambda = \mathrm{E}[X e^{\lambda X}]$, $G_{\lambda\lambda} = \mathrm{E}[X^2 e^{\lambda X}]$. Now,

$$\begin{aligned}(\log G)_{\lambda\lambda} &\equiv \frac{G_{\lambda\lambda}}{G} - \left(\frac{G_\lambda}{G}\right)^2 \\ &= \frac{\mathrm{E}[X^2 e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]} - \left(\frac{\mathrm{E}[X e^{\lambda X}]}{\mathrm{E}[e^{\lambda X}]}\right)^2 \\ &= \mathrm{E}[Y^2] - \mathrm{E}[Y]^2, \quad \text{where } Pr\{Y \in [t, t+dt)\} = e^{\lambda t}\tfrac{Pr\{X \in [t,t+dt)\}}{\mathrm{E}[e^{\lambda X}]}, \\ &= \mathrm{E}[(Y - \mathrm{E}[Y])^2] > 0. \quad \blacksquare\end{aligned}$$

2.2. Lemma 1. Let $x_1, x_2, \ldots, x_n$ be n independent random variables. Then for $a > 0$,

1) $$||\sum_{i=1}^n x_i||_a^{CH} = \max_{\substack{a_1+a_2+\cdots+a_n=a \\ a_i \geq 0}} \prod_{i=1}^n ||x_i||_{a_i}^{CH}, \quad a \geq 0.$$

As a weak consequence,

2) $$||X + Y||_a^{CH} \geq ||X||_a^{CH}.$$

Proof: 1) It suffices to show that

$$||X + Y||_a^{CH} = \max_{0 \leq b \leq a} ||X||_b^{CH} ||Y||_{a-b}^{CH}.$$

Now, for any $\lambda \geq 0$ and $b < a$, $||X + Y||_{\lambda,a}^L \geq ||X||_b^{CH} ||Y||_{a-b}^{CH}$, since the right hand side is defined as a minimization over a pair of independent parameters. Hence

$$||X + Y||_a^{CH} \geq ||X||_b^{CH} ||Y||_{a-b}^{CH}.$$

We need to establish the reverse inequality for some b. We may, without loss of generality, assume that $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, and suppose that neither X nor Y is identically zero. Define the function $\lambda\{a\}$ so that for $\lambda \geq 0$, $e^{-a\lambda}\mathrm{E}[e^{\lambda X}]$ is minimized at $\lambda = \lambda\{a\}$. Similarly, define $\delta\{a\}$ so that $e^{-a\delta}\mathrm{E}[e^{\delta Y}]$ is minimized at $\delta\{a\}$.

Consider the factors $||X||_b^{CH}$, and $||Y||_{a-b}^{CH}$, and the parameters λ, δ where $||X||_b^{CH} = ||X||_{\lambda,b}^L$, and $||Y||_{a-b}^{CH} = ||Y||_{\delta,a-b}^L$. Either $\lambda = 0$ or $\frac{\mathrm{d}}{\mathrm{d}\lambda}||X||_{\lambda,b}^L \Big|_{\lambda=\lambda\{b\}} = 0$, because $\lambda\{b\}$ gives a minimum value for $||X||_{\lambda\{b\},b}^L = ||X||_b^{CH}$. Assuming the derivative is in fact zero gives

$$e^{-b\lambda\{b\}}(-b\mathrm{E}[e^{\lambda\{b\}X}] + \mathrm{E}[X e^{\lambda\{b\}X}]) = 0. \tag{11}$$

Similarly, either $\delta = 0$ or

$$e^{-(a-b)\delta\{a-b\}}(-(a-b)\mathrm{E}[e^{\delta\{a-b\}Y}] + \mathrm{E}[Ye^{\delta\{a-b\}Y}]) = 0. \tag{12}$$

Now suppose that the product $||X||_c^{CH}||Y||_{a-c}^{CH}$ is maximal at the value $c = b$. If $0 < b < a$, then we must have $\frac{\mathrm{d}}{\mathrm{d}b}||X||_b^{CH}||Y||_{a-b}^{CH} = 0$, since b is a maximum. Evaluating the derivative, in this case, and substituting equations (11) and (12) gives the following.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}b}||X||_b^{CH}||Y||_{a-b}^{CH} &= \frac{\mathrm{d}}{\mathrm{d}b}||X||_{\lambda\{b\},b}^{L}||Y||_{\delta\{a-b\},a-b}^{L} \\
&= e^{-b\lambda-(a-b)\delta}(-\lambda\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\delta Y}] + \delta\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\delta Y}] \\
&\quad + \mathrm{E}[Xe^{\lambda X}]\mathrm{E}[e^{\delta Y}]\frac{\mathrm{d}\lambda}{\mathrm{d}b} + \mathrm{E}[e^{\lambda X}]\mathrm{E}[Ye^{\delta Y}]\frac{\mathrm{d}\delta}{\mathrm{d}b} \\
&\quad - b\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\delta Y}]\frac{\mathrm{d}\lambda}{\mathrm{d}b} - (a-b)\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\delta Y}]\frac{\mathrm{d}\delta}{\mathrm{d}b}); \quad \lambda = \lambda\{b\},\ \delta = \delta\{a-b\} \\
&= e^{-b\lambda-(a-b)\delta}(\delta - \lambda)\mathrm{E}[e^{\lambda X}]\mathrm{E}[e^{\delta Y}].
\end{aligned} \tag{13}$$

Hence $\frac{\mathrm{d}}{\mathrm{d}b}||X||_b^{CH}||Y||_{a-b}^{CH} = 0$ implies that

$$\delta\{a-b\} = \lambda\{b\}.$$

In this case, $||X||_b^{CH}||Y||_{a-b}^{CH} = ||X||_{\lambda\{b\},b}^{L}||Y||_{\delta\{a-b\},a-b}^{L} = ||X+Y||_{\lambda\{b\},a}^{L} \geq ||X+Y||_a^{CH}$ as desired.

The boundary constraints are readily handled. First, it is easy to see that for the optimal b, $\lambda\{b\} = \infty$ precisely when $Pr\{X > b\} = Pr\{Y > a-b\} = 0$, in which case the lemma is trivially true.

For $b > 0$, $\frac{\mathrm{d}}{\mathrm{d}\lambda}e^{-\lambda b}\mathrm{E}[e^{\lambda X}]\ |_{\lambda=0} = -b\mathrm{E}[1] + \mathrm{E}[X] = -b$, which shows that the minimum of $||X||_{\lambda,b}^{L}$ cannot be at $\lambda = 0$. Similarly, $\delta\{a-b\}$ cannot be zero for $b < a$.

Finally, we address the cases $b = 0$, $b = a$. Since $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathrm{E}[e^{\lambda X}]\ |_{\lambda=0} = 0$, and $\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\mathrm{E}[e^{\lambda X}] = \mathrm{E}[X^2e^{\lambda X}]$, which is positive for $X \not\equiv 0$, it follows that $\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathrm{E}[e^{\lambda X}] > 0$ for $\lambda > 0$ and hence $\lambda\{b\}\ |_{b=0} = 0$. But then (13) shows that $\frac{\mathrm{d}}{\mathrm{d}b}||X||_b^{CH}||Y||_{a-b}^{CH}\ |_{b=0}$ is positive, if $\delta\{a\} > 0$. In this case, $b = 0$ cannot be a maximum, while the case $\delta\{a\} = \lambda\{0\} = 0$ again establishes the desired inequality. The exact same reasoning also holds if $b = a$.

2) This follows immediately from 1); in fact, 2) also follows from Lemma 0.2. ∎

2.3. Lemma C. Let $h(\lambda)$ be a convex function of λ, and define $f(c) = \min_\lambda(-\lambda c + h(\lambda))$. Then f is concave[†] and for $\lambda \in \text{Domain}(h)$, $h(\lambda) = \max_c(\lambda c + f(c))$.

Proof: We may suppose that $h(\lambda)$ is continuously differentiable and strictly convex, as otherwise we may construct a sequence of strictly convex continuously differentiable functions h_n, where $h_n \downarrow h$ as $n \to \infty$. This gives a decreasing family f_n. Passing to the limit and exploiting the continuity of limits will establish the claims for convex h.

Let the derivative of h, $h_\lambda(\lambda)$, map λ **onto** the range $[a, b]$, and for $c \in [a, b]$, define $f(c) = \min_\lambda(-\lambda c + h(\lambda))$. Let λ_0 satisfy

$$h_\lambda(\lambda_0) = c; \tag{14}$$

there is such a λ_0, since $c \in [a, b]$. Moreover, the λ_0 must be a local minimum for $-\lambda c + h(\lambda)$, since h is convex, and $\frac{\mathrm{d}}{\mathrm{d}\lambda}(-\lambda c + h(\lambda)) \Big|_{\lambda=\lambda_0} = -c + h_\lambda(\lambda_0) = 0$. The convexity of h ensures that there are no local maxima, whence λ_0 must be a global minimum. Hence $f(c) = -\lambda c + h(\lambda) \Big|_{\lambda=\lambda_0}$.

We may use the chain rule to compute the first derivative of f, with the understanding that λ is a function of c, in the definition of f:

$$f_c(c) = -\lambda + (-c + h_\lambda(\lambda))\lambda_c = -\lambda + (0)\lambda_c = -\lambda,$$

where we have used (14) to attain the zero factor. It follows that $f_{cc} = -\lambda_c$, and we may differentiate (14) to compute λ_c, which gives $h_{\lambda\lambda}(\lambda)\lambda_c \Big|_{\lambda=\lambda_0} = 1$. Hence

$$f_{cc}(c) = -\lambda_c = -1/h_{\lambda\lambda}(\lambda_0).$$

It follows that f is strictly concave.

Moreover, for $-\lambda \in \text{Range}(f_c)$, $\max_c(\lambda c + f(c)) = \lambda c + f(c) \Big|_{c=c_0}$, where c_0 satisfies $\lambda + f_c(c_0) = 0$. To see that the range of f_c is just the negative of the domain of h, we observe that for any $\lambda \in \text{Domain}(h)$, we may use (14) to set $c = h_\lambda(\lambda)$; then $f(c)$ is computed from h at λ, and $f_c = -\lambda$.

[†]Here the notion of concavity (as well as convexity) needs to be extended under pointwise convergence to include functions that are infinite outside of an interval. This accommodates, for example, the consequences of setting $h(x) = ax + b$.

Finally, let $g(\gamma) = \max_c(\gamma c + f(c))$. Then the g is computed at a value c_0 where $f_c(c_0) = -\gamma$. But the computation for $f(c_0)$ is computed from h at λ where $f_c(c_0) = -\lambda$. So $g(\gamma) = \gamma c_0 + (-\gamma c_0 + h(\gamma)) = h(\gamma)$, and h is computed from f as claimed. ∎

2.4. Theorem 1. Let $X = x_1 + x_2 + \cdots + x_n$ be the sum of n independent Bernoulli trails. Let $\bar{p} = \mathrm{E}[X]/n$ and $\sigma^2 = (\mathrm{E}[X^2] - \mathrm{E}^2[X])/n$. Then

$$||X||_{na}^{CH} \leq ||B(n\frac{(1-\bar{p})^2}{1-\bar{p}-\sigma^2}, \frac{\sigma^2}{1-\bar{p}})||_{na\frac{1-\bar{p}-\sigma^2}{(1-\bar{p})^2}}^{CH}.$$

Proof: Form the moment generating function $G(\lambda) = \mathrm{E}[e^{\lambda(X)}] = \prod_i (1 + p_i(e^\lambda - 1))$, and consider the factors of G. We may take fractional powers of them, and pair them together to achieve the balanced "dipoles" $(1 + (\bar{p} - c)(e^\lambda - 1))^\kappa (1 + (\bar{p} + \kappa)(e^\lambda - 1))^c$. These factors correspond to a weighted pair of Bernoulli trials where one trial has weight κ and probability of success $\bar{p} - c$, while the other has weight c and probability of success $\bar{p} + \kappa$. The mean of the pair is $\bar{p}$ with joint weight $c + \kappa$. The weighted variance of the pair is $\kappa(\bar{p} - c)(1 - \bar{p} + c) + c(\bar{p} + \kappa)(1 - \bar{p} - \kappa) = (c + \kappa)\bar{p}(1 - \bar{p}) - (c + \kappa)(c\kappa)$. It is convenient to normalize such a factor to have unit weight; so we will analyze $w(c, \kappa) = (1 + (\bar{p} - c)(e^\lambda - 1))^{\frac{\kappa}{c+\kappa}} (1 + (\bar{p} + \kappa)(e^\lambda - 1))^{\frac{c}{c+\kappa}}$, which has mean $\bar{p}$ with weight 1, and weighted variance $\bar{p}(1 - \bar{p}) - c\kappa$.

Notice that an individual Bernoulli trial with mean $\bar{p}$ can be represented as a balanced dipole with $c = 0$ and $\kappa = 1 - \bar{p}$. Thus any $G(\lambda)$ can be decomposed as a product of (fractional) dipoles $\prod w(c_i, \kappa_i)^{f_i}$ where the f_i are positive, $\sum_i f_i = n$, and $\sum f_i c_i \kappa_i = n\bar{p}(1 - \bar{p}) - n\sigma^2$.

The objective is to dominate $G(\lambda)$ by a maximal product; we must first find the maximum $w(c, \kappa)$, for fixed λ and fixed product $c\kappa = \gamma$. The constraints are that $0 \leq c \leq \bar{p}$ and $0 \leq \kappa \leq 1 - \bar{p}$. Let $z(c, \kappa) = \log(w(c, \kappa))$. We now show that subject to these constraints, $\frac{\mathrm{d}}{\mathrm{d}\kappa} z$ is positive, which ensures that taking κ to be a large as possible maximizes w.

Since $c = \gamma/\kappa$, $\frac{\mathrm{d}}{\mathrm{d}\kappa} c = -\gamma/\kappa^2$. Differentiating z with respect to κ gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\kappa}z &= \frac{\kappa^2}{\gamma+\kappa^2}\frac{\gamma(e^\lambda-1)}{\kappa^2(1+(\bar{p}-c)(e^\lambda-1))} + \frac{2\gamma\kappa}{(\gamma+\kappa^2)^2}\log(1+(\bar{p}-c)(e^\lambda-1)) \\
&\quad + \frac{\gamma}{\gamma+\kappa^2}\frac{\gamma(e^\lambda-1)}{\kappa^2(1+(\bar{p}+\kappa)(e^\lambda-1))} - \frac{2\gamma\kappa}{(\gamma+\kappa^2)^2}\log(1+(\bar{p}+\kappa)(e^\lambda-1)) \\
&= \frac{\gamma}{\gamma+\kappa^2}\left((e^\lambda-1)\left(\frac{1}{1+(\bar{p}-c)(e^\lambda-1)}+\frac{1}{1+(\bar{p}+\kappa)(e^\lambda-1)}\right)\right. \\
&\qquad \left. + \frac{2\kappa}{\gamma+\kappa^2}\log(1-\frac{(\kappa+c)(e^\lambda-1)}{1+(\bar{p}+\kappa)(e^\lambda-1)})\right) \\
&= \beta(\frac{1}{x-\delta}+\frac{1}{x}+\frac{2}{\delta}\log(1-\frac{\delta}{x})),
\end{aligned}$$

where $\beta = \frac{\gamma(e^\lambda-1)}{\gamma+\kappa^2}$, $x = 1+(\bar{p}+\kappa)(e^\lambda-1)$, and $\delta = (\kappa+c)(e^\lambda-1) = \frac{\kappa^2+\gamma}{\kappa}(e^\lambda-1)$.

We need to show that for $0<\delta<x$, $\frac{1}{x-\delta}+\frac{1}{x}+\frac{2}{\delta}\log(1-\frac{\delta}{x})>0$. It suffices to show that $\frac{\delta}{x-\delta}+\frac{\delta}{x}+2\log(1-\frac{\delta}{x})>0$.

If $\delta=0$, the expression is zero. If we can show that its first derivative (with respect to δ) is positive on $(0,x)$, then integrating shows that the expression is positive. Differentiating with respect to δ gives $\frac{\delta}{(x-\delta)^2}+\frac{1}{x}-\frac{1}{x-\delta} = \frac{\delta^2}{x(x-\delta)^2} > 0$ on $(0,x)$. We conclude that for $c\kappa$ fixed at γ, κ should be maximal, which means $\kappa = 1-\bar{p}$ unless a priori bounds on the size of p_i constrain them to be smaller.

We now show that for fixed κ, $w(c_i,\kappa)$ is log-concave, whence Jensen's inequality ensures that $\prod w(c_i,\kappa)^{f_i}$ is maximized by setting all c_i to their mean, which is $\frac{\bar{p}(1-\bar{p})-\sigma^2}{\kappa}$. Let

$$z(c) = \log\left((1+(\bar{p}-c)(e^\lambda-1))^{\frac{\kappa}{c+\kappa}}(1+(\bar{p}+\kappa)(e^\lambda-1))^{\frac{c}{c+\kappa}}\right).$$

Differentiating twice with respect to c gives

$$\begin{aligned}
z_{cc} &= \frac{2\kappa}{(c+\kappa)^3}\log(1+(\bar{p}-c)(e^\lambda-1)) + \frac{2\kappa}{(c+\kappa)^2}\frac{(e^\lambda-1)}{(1+(\bar{p}-c)(e^\lambda-1))} \\
&\quad - \frac{\kappa}{c+\kappa}\frac{(e^\lambda-1)^2}{(1+(\bar{p}-c)(e^\lambda-1))^2} - \frac{2\kappa}{(c+\kappa)^3}\log(1+(\bar{p}+\kappa)(e^\lambda-1)) \\
&= \beta\left(\log(x)+\frac{\delta}{x}-\frac{\delta^2}{2x^2}-\log(x+\delta)\right),
\end{aligned}$$

where $\beta = \frac{2\kappa}{(c+\kappa)^3}$, $x = 1+(\bar{p}-c)(e^\lambda-1)$, and $\delta = (c+\kappa)(e^\lambda-1)$.

We wish to show that $\log(x)+\frac{\delta}{x}-\frac{\delta^2}{2x^2}-\log(x+\delta)$ is nonpositive. When $\delta=0$, the expression is zero. Differentiating with respect to δ gives $\frac{1}{x}-\frac{\delta}{x^2}-\frac{1}{x+\delta} = \frac{\delta}{x(x+\delta)}-\frac{\delta}{x^2} < 0$, for $\delta>0$. It follows that $w(c)$ is log-concave, and X is maximal in the CH-estimate when it comprises $n\frac{(1-\bar{p})^2}{1-\bar{p}-\sigma^2}$ identical Bernoulli trials having mean $\frac{\sigma^2}{1-\bar{p}}$, and $n\frac{\bar{p}-\bar{p}^2-\sigma^2}{1-\bar{p}-\sigma^2}$ (constant) trials with mean 1. Theorem 1 now follows from Hoeffding's bound (2) applied to the nonconstant trials. ∎

2.5. Lemma 3. Let b_i be Bernoulli trials, and set $\xi_k = Pr\{b_1 = 1 | b_1 + b_2 + \cdots + b_n = k\}$. Then ξ_k is monotone increasing in k.

Proof: Let $\zeta_k = Pr\{b_2 + b_3 + \cdots + b_n = k\}$, and recall that $p_i = Pr\{b_i = 1\}$. Now, $\xi_k = \frac{p_1 \zeta_{k-1}}{p_1 \zeta_{k-1} + (1-p_1)\zeta_k}$, and it follows that $\xi_{k+1} \geq \xi_k$ iff $\frac{\zeta_k}{\zeta_{k+1}} \geq \frac{\zeta_{k-1}}{\zeta_k}$. Thus we must show that $\zeta_k^2 \geq \zeta_{k+1}\zeta_{k-1}$.

Put $Q = \prod_1^n (1 - p_i)$, and let $r_i = p_i/(1 - p_i)$. Then $\zeta_k = Q \sum_{i_1 < i_2 < \cdots < i_k} r_{i_1} r_{i_2} \cdots r_{i_k}$, whence

$$\zeta_k^2 = Q^2 \sum_{j=0}^{k} \sum_{i_1 < i_2 < \cdots < i_{2k-j}} \left(\binom{2k-2j}{k-j} r_{i_1} r_{i_2} \cdots r_{i_{2k-j}} \sum_{\{s_1, s_2, \ldots, s_j\} \subset \{i_1, \ldots, i_{2k-j}\}} r_{s_1} r_{s_2} \cdots r_{s_j} \right).$$

Moreover,

$$\zeta_{k-1}\zeta_{k+1} = Q^2 \sum_{j=0}^{k-1} \sum_{i_1 < i_2 < \cdots < i_{2k-j}} \left(\binom{2k-2j}{k-j-1} r_{i_1} r_{i_2} \cdots r_{i_{2k-j}} \sum_{\{s_1, s_2, \ldots, s_j\} \subset \{i_1, \ldots, i_{2k-j}\}} r_{s_1} r_{s_2} \cdots r_{s_j} \right).$$

Thus ζ_k^2 contains a superset of the terms in $\zeta_{k-1}\zeta_{k+1}$, and with corresponding coefficients of $Q^2\binom{2k-2j}{k-j}$ as opposed to $Q^2\binom{2k-2j}{k-j-1}$, which shows that the inequality is in fact strict, unless $\zeta_k = 0$. ∎

Appendix 3. Doob's Submartingale inequality

Informally, a sequence of random variables $Y_1, Y_2, \ldots$ is a submartingale if they represent a game with winning expectations: $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_{k-1}] \geq Y_{k-1}$.

It is critical to understand just what the notion of conditional expectation means. A family of random variables must have a joint probability distribution if the probability of joint events is to be defined. Informally, this means they are defined as random functions on a single probability space S that has a probability measure P. Martingales and submartingales such families. The interesting part concerns conditional expectations. Formally, $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_j]$ is equal to a new random function g that is defined (or indexed) over a j dimensional space. The randomness results from the randomness in assigning values to the Y_i, for $i = 1 = 2 = \ldots, j$. In the discrete case, we may write the deterministic value returned by g, for $Y_i = z_i$, $i = 1, 2, \ldots, j$ as $g(z_1, z_2, \ldots, z_j) = \frac{\sum_{x \in S} x\, Pr\{Y_k = x, Y_1 = z_1, Y_2 = Z_2, \ldots, Y_j = z_j\}}{Pr\{Y_1 = z_1, Y_2 = Z_2, \ldots, Y_j = z_j\}}$. Restated, the constraints $Y_i = z_i$, for $i = 1, 2, \ldots, j$ define a sample region $S_{\bar{z}} \subset S$, and by definition, $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_j]$ is constant on this subset; the value of the constant is just the average value of Y_k on $S_{\bar{z}}$. Of course $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_j]$ is really defined on the probability space S. Given an $s \in S$,

s defines the values for $Y_1, Y_2, \ldots Y_j$. The subset of S where $Y_1, Y_2, \ldots, Y_j$ have these specific values is the averaging region used to compute the value of the conditional expectation $\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_j](s)$. In light of this, we see that the submartingale requirement

$$\mathrm{E}[Y_k \mid Y_1, Y_2, \ldots, Y_{k-1}] \geq Y_{k-1}$$

is an inequality about functions defined on S, and the inequality holds everywhere, except, perhaps, for some set of probability zero.

It is important to note that averaging on subsets conserves the mean on coarser sets. For example, $\mathrm{E}[\mathrm{E}[Y \mid W]] = \mathrm{E}[Y]$. More generally, $\mathrm{E}[\mathrm{E}[Y \mid W, X] \mid X] = \mathrm{E}[Y \mid X]$. The finest partition, in some sense, that is admissible for Y is defined by Y itself, which is the collection sets $Y^{inv}(x)$, for $x \in \Re$: $\mathrm{E}[Y \mid Y] = Y$. Even finer partitions can be used, although they have no computational effect: $\mathrm{E}[Y \mid W, Y] = Y$. The coarsest partition is the whole space S (defined by, say, constant functions on S, which contain no information whatsoever): $\mathrm{E}[Y \mid 1] = \mathrm{E}[Y]$. That is, the average value of Y, when $1 = 1$, is $\mathrm{E}[Y]$.

A key question is the following. When can we assert that, for a subset $R \subset S$, $\int_R \mathrm{E}[Y \mid X] dP = \int_R Y dP$? The answer is that if for every $s \in R$, $\mathrm{E}[Y \mid X](s)$ is computed as an average of Y on some subset $Q_s \subset R$, with no contributions from $S - R$, then the integrals will be the same. Indeed, the definition of conditional expectation dictates that $\int_{Q(s)} \mathrm{E}[Y \mid X] dP = \int_{Q(s)} Y dP$ for the s's averaging region $Q(s)$. This explains the idea of coarseness: R should comprise a union of subsets that have been used in the averaging process. Finally, constants can be factored through expectations. For example, $\mathrm{E}[XY \mid X, W] = X\mathrm{E}[Y \mid X, W]$, since the random variable X is constant on points $(X = a, W = b)$ which comprise the averaging regions used to compute the random variable $\mathrm{E}[Y \mid X, W]$.

A rigorous presentation of conditional expectation requires the notion of measurability, since the individual averaging regions may have measure zero. Nevertheless, this brief digression on conditional expectation, it is hoped, might be sufficient preparation to present Doob's elegant submartingale inequality and its consequences.

Theorem(Doob). Let $Y_1, Y_2, \ldots, Y_n$ be a non-negative submartingale. Then for $a > 0$,

$$Pr\{\max_{i \le n} Y_i \ge a\} \le \mathrm{E}[Y_n]/a.$$

Proof: Let S be the probability space with probability measure $Pr\{\ \cdot\ \}$ for the joint distribution of the Y_i's. Define $S_i \subset S$ to be the subregion where $Y_i \ge a$, but $Y_h < a$, for $h < i$. Then the S_i's are disjoint and

$$Pr\{\max_{i \le n} Y_i \ge a\} = Pr\{\cup_{i=1}^{n} S_i\},$$

by definition of the S_i. There follows:

$$\begin{aligned}
aPr\{\max_{i \le n} Y_i \ge a\} &= aPr\{\cup_{i=1}^{n} S_i\}, \\
&\le \sum_i \int_{S_i} Y_i dP, \\
&\le \sum_i \int_{S_i} \mathrm{E}[Y_n \mid Y_1, Y_2, \ldots, Y_i] dP \quad \text{since } Y \text{ is a submartingale,} \\
&\le \sum_i \int_{S_i} Y_n dP \quad \text{since } S_i \text{ is completely defined by the behavior of } Y_1, \ldots, Y_i, \\
&\le \int_{\cup S_i} Y_n dP \quad \text{since the } S_i \text{ are disjoint,} \\
&\le \int_{S} Y_n dP \quad \text{since } Y_n \text{ is non-negative.} \quad \blacksquare
\end{aligned}$$

Finally, let $X_i = x_1 + x_2 + \cdots + x_i$ be a martingale. This is equivalent to the requirement that $\mathrm{E}[x_i \mid X_1, X_2, \ldots, X_{i-1}] = 0$. We now show that for any real λ, $Y_i = e^{\lambda X_i}$ is a non-negative submartingale. We must verify that $\mathrm{E}[Y_i \mid Y_1, Y_2, \ldots, Y_{i-1}] \ge Y_{i-1}$.

$$\begin{aligned}
\mathrm{E}[Y_i \mid Y_1, Y_2, \ldots, Y_{i-1}] &= \mathrm{E}[Y_{i-1} e^{\lambda x_i} \mid Y_1, Y_2, \ldots, Y_{i-1}] \\
&= Y_{i-1} \mathrm{E}[e^{\lambda x_i} \mid Y_1, Y_2, \ldots, Y_{i-1}] \\
&\ge Y_{i-1} e^{\mathrm{E}[\lambda x_i \mid Y_1, Y_2, \ldots, Y_{i-1}]} = Y_{i-1} e^0 = Y_{i-1},
\end{aligned}$$

where the inequality follows from Jensen's inequality and the convexity of $e^{\lambda x}$.

Consequently, Doob's submartingale theorem can be used to attain Hoeffding bounds for martingales, and

$$Pr\{\max_{i \le n} X_i \ge a\} \le e^{-\lambda a} \mathrm{E}[e^{\lambda X_n}],$$

as claimed in Section 6.1. $\blacksquare$

The partitioning technique via disjoint S_i originates with Kolmogorov, who used it to derive a second moment version of the theorem for the partial sums of independent random variables with mean zero [Ko-28].