1 OPTIMAL BOUNDS ON TAIL PROBABILITIES: A STUDY OF AN APPROACH

Aviad Cohen

Israel Software Lab, Intel, Haifa, Israel.
aviadc@iil.intel.com

Yuri Rabinovich

Dept. of Math. and Computer Science, Ben-Gurion University, Beer-Sheva 84105, Israel.
uri@cs.bgu.ac.il

Assaf Schuster, Hadas Shachnai

Dept. of Computer Science, Technion IIT, Haifa 32000, Israel.
assaf@cs.technion.ac.il, hadas@cs.technion.ac.il

Abstract: In Computer Science and Statistics it is often desirable to obtain tight bounds on the decay rate of probabilities of the type $\Pr\{S_n - E[S_n] \geq na\}$, where S_n is a sum of independent random variables $\{X_i\}_{i=1}^n$. This is usually done by means of Chernoff inequality, or the more general Hoeffding inequality. The latter inequality is asymptotically optimal as far as the expectations of X_i-s go, but ceases to be so when the variances are also given. The variances are taken into account in the stronger Bennett inequality, which despite its potential usefulness is virtually unknown in CS community.

In this paper we provide a systematic account of the general method (based on Laplace transform) underlying most asymptotically tight estimations of tail probabilities, and show how it can be used in various situations. In particular, we provide new and simple proofs of the Hoeffding and the Bennet bounds, and obtain their natural generalization, which takes into account the first k moments of X_i-s. We discuss also a typical application of the general method to a concrete problem from Computer Science, and obtain estimations superior to those previously known.

The main goal of this work is to give a clear, coherent exposition of the general method and its various aspects, in the belief that a better acquaintance with this powerful tool might prove beneficial in studies involving estimations of tail probabilities, e.g., in analysis of performances of randomized algorithms.

1.1 INTRODUCTION

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent random variables assuming values in a bounded interval (which, without loss of generality, will be assumed to be $[0,1]$), and let $S_n = \sum_{i=1}^n X_i$. Suppose we know something about the distribution of each X_i, e.g., we know the exact distribution, or we know its first k moments $m_k = E[X_i^k]$. In Computer Science, Discrete Mathematics and Statistics, the need often arises to give upper bounds on tail probabilities of the form: [1]

$$\Pr\left[S_n - E[S_n] \geq na\right] \quad .$$

In particular it is often desirable to obtain upper bounds which decrease exponentially fast in n (for a fixed a).

The best known inequalities of this sort are the Chernoff and the Hoeffding inequalities. The Hoeffding inequality [6] asserts that for X_i-s which assume values in the interval $[0,1]$, and all have the same mean μ,

$$\Pr\left[S_n - E[S_n] \geq na\right] \leq \left\{ \left(\frac{\alpha}{\beta}\right)^{\beta} \left(\frac{1-\alpha}{1-\beta}\right)^{1-\beta} \right\}^n \quad , \tag{1.1}$$

[1] Probabilities of the form $\Pr\left[S_n - E[S_n] \leq -na\right]$ can be handled by defining $Y_i = C - X_i$ (C a constant), $S'_n = \sum_{i=1}^n Y_i$ and using

$$\Pr\left[S_n - E[S_n] \leq -na\right] = \Pr\left[S'_n - E[S'_n] \geq na\right]$$

where

$$\alpha = \mu \quad ; \quad \beta = \mu + a \ .$$

The Hoeffding inequality generalizes the Chernoff inequality [3, 9], which makes the same statement about $\{0,1\}$ random variables. The sum of random variables supported on $[0,1]$ turns out to be at least as well concentrated around its mean as the sum of $\{0,1\}$ random variables.

Most research papers employing upper bounds on tails probabilities do not go beyond the Hoeffding inequality. Notice, however, that (1.1) uses only the mean μ, and ignores the variance σ^2 of the X_i-s. It is reasonable to expect that when we know both the mean μ and the variance σ^2, a sharper result should exist. It is indeed so; Bennett has shown in [4][2] that in this case (1.1) holds with α and β defined as

$$\alpha = \frac{\sigma^2}{\sigma^2 + (1-\mu)^2} \quad ; \quad \beta = \frac{\sigma^2 + a(1-\mu)}{\sigma^2 + (1-\mu)^2} \tag{1.2}$$

When the variance assumes its maximal possible value ($\sigma^2 = \mu - \mu^2$), the Bennett inequality (1.2) reduces to the Hoeffding inequality. In all other cases it is strictly stronger. In particular, when the variance is zero, the Bennett inequality asserts that tail probabilities equal to zero, as should be.

Despite the fact that the Bennett inequality should have had numerous applications, it is rarely if ever used in Computer Science literature. We believe that one reason for this is the unappealing form it was given in [4], another is its rather involved proof. This applies to an even larger degree to the natural generalization of both the Hoeffding and Bennett inequalities, when the first k moments of all X_i-s are known and equal, or to the direct computation of the Laplace transform (underlying the proofs of both inequalities) when all X_i-s have the same distribution.

It is precisely this situation the present paper attempts to amend. We describe in detail the generic method for obtaining the tail probabilities (Section 1.2), then use it to derive the Hoeffding and the Bennett inequalities (Sections 1.3 and 1.4, resp.). The latter is furnished with a new and simple proof, and is given an attractive new form. Then we discuss the natural generalization of the two inequalities, which leads to a mathematically rich theory (Section 1.5). Finally, we demonstrate the usefulness of the original generic method by applying it directly to an interesting problem arising in Computer Science (Section 1.6).

Throughout the paper a special effort has been made to simplify and clarify the existing proofs. By gathering the various aspects of the method (which are usually scattered in different advanced textbooks and rarely appear under the same roof), we hope to give a deeper and more complete picture of it. If the detailed exposition of the method and the example of its application to a typical

[2]In fact the Bennett bound holds also for variables that are unbounded from one side [4]. Here we restrict the discussion to bounded variables.

CS problem described in this paper will lead someone to a more sophisticated use of the method than just the Hoeffding inequality, we will have achieved our goal in writing this paper.

1.2 BOUNDING TAIL PROBABILITIES WITH THE LAPLACE TRANSFORM.

Let Ψ be the set of random variables distributed according to some class of probability distributions. The goal is to give upper bounds on tail probabilities, bounds that will hold for *any* sequence of independent random variables $X_i \in \Psi$. The method we shall describe here was apparently presented for the first time in a paper by Bernstein [2]. It uses the Laplace transform, and consists of the following basic steps.

(1) Let χ be the indicator function of the event $S_n - E[S_n] \geq na$. Observe that for any $t > 0$,

$$\chi \leq e^{(S_n - E[S_n] - na)t} \ . \tag{1.3}$$

The latter function will be used as an approximation to χ. For any $t > 0$,

$$\begin{aligned} \Pr[S_n - E[S_n] \geq na] &= E[\chi] \leq E[e^{(S_n - E[S_n] - na)t}] \\ &= e^{-nt(\mu+a)} \prod_{i=1}^{n} E[e^{X_i t}] \\ &= \left(e^{-t(\mu+a)} \phi_n(t)\right)^n \ , \end{aligned} \tag{1.4}$$

where

$$\phi_n(t) \ = \ \left(\prod_{i=1}^{n} E[e^{X_i t}]\right)^{\frac{1}{n}} \ .$$

The function $f(t) = E[e^{Xt}]$ is called the *Laplace transform* of X.

(2) Define $Z(t)$ by:

$$Z(t) = \sup_{Y \in \Psi} E[e^{Yt}]$$

By (1.3), for any $t > 0$,

$$\Pr[S_n - E[S_n] \geq na] \leq \left(e^{-t(\mu+a)} Z(t)\right)^n \ . \tag{1.5}$$

This is the fundamental inequality of the entire method.

(3) The next step is to attempt to express $Z(t)$ as an explicit function of t and the parameters defining the class Ψ, and to plug this function into inequality (1.5). If the explicit form for $Z(t)$ is hard to get, one should attempt to find some other handy representation of it, i.e., an algorithm to compute it, or a nice function which majorizes it.

(4) The final step is to minimize with respect to t the right hand side of inequality (1.5), as obtained in step **(3)**.

Step **(3)** is the crucial step of the strategy. It requires an explicit (or at least convenient) expression of $Z(t)$ as a function of t and the parameters defining the class Ψ. In all cases discussed in this paper, there will be a single member $X \in \Psi$, whose Laplace transform $E[e^{Xt}]$ simultaneously majorizes the Laplace transforms of all other $Y \in \Psi$, for all $t > 0$. In this case $Z(t) = E[e^{Xt}]$; expressing $E[e^{Xt}]$ explicitly, one gets an optimal bound, as far as the above strategy is concerned.

1.2.1 The Optimality of the Method: Cramér's Theorem

It is natural to ask how good the upper bound given by inequality (1.5) is. It is easy to see that unless $\mu + a = \max X$, the inequality (1.3) is strict on a set of positive probability, and therefore the bound of (1.5) is not optimal. However, it turns out to be optimal in a certain asymptotical sense.

The following theorem is a special case of Cramér's Theorem, one of the cornerstones of the Large Deviations Theory (see [10] for more details). For the sake of simplicity, we consider the case when all Y_i-s have the same distribution.

Theorem 1 *Let $\{Y_i\}_{i=1}^{\infty}$ be independent equi-distributed random variables taking values in the interval $[0,1]$ and having the mean μ, and let $S_n = \sum_{i=1}^{n} Y_i$. Then, for any $\delta > \mu$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \Pr\left[\frac{S_n}{n} > \delta\right] \;\geq\; \inf_{t>0} \left(-t\delta + \log E[e^{Yt}]\right) \;.$$

Proof: The theorem is obviously true when $\delta \geq \max Y$: in this case the right-hand side tends to 0 as t tends to ∞. In what follows we assume $\delta < \max Y$.

Let z, where $\delta < z < \max Y$, be a number, and let us choose $t > 0$ such that $E[(Y-z)e^{Yt}] = 0$. Such a t exists: viewing $E[(Y-z)e^{Yt}] = 0$ as a function of t, we see that this function is continuous, negative for $t = 0$ (since $z > \mu$), and positive for $t = \infty$ (since $z < \max Y$).

Let F denote the probability distribution on $[0,1]$ corresponding to the random variable Y. Define a new random variable $Y^{(t)}$ on $[0,1]$, distributed according to $F^{(t)}$, defined by:

$$dF^{(t)}(y) \;=\; E(e^{Yt})^{-1}\, e^{ty}\, dF(y) \;.$$

Our first observation is that the mean of $Y^{(t)}$ is z. Indeed,

$$\begin{aligned} \int_0^1 y\, dF^{(t)}(y) &= E(e^{Yt})^{-1} \int_0^1 y\, e^{yt}\, dF(y)\, dy \\ &= E(e^{Yt})^{-1}\, E(Y\, e^{Yt}) = E(e^{Yt})^{-1}\, E(z\, e^{Yt}) = z \quad , \end{aligned}$$

where the last inequality is due to our choice of t: $E[(Y-z)e^{Yt}] = 0$.

Let also D_n and $D_n^{(t)}$ denote the probability distributions corresponding to the random variables S_n/n and $S_n^{(t)}/n$, respectively. It can be readily checked that

$$dD_n^{(t)}(y) \;=\; E(e^{Yt})^{-n}\, e^{ynt} dD_n(y) \;\;.$$

Let $\epsilon > 0$ be small enough such that $z - \epsilon > \delta$. Then,

$$\begin{aligned} Pr\left[\frac{S_n}{n} > \delta\right] \;\; &\geq \;\; Pr\left[\frac{S_n}{n} \in [z-\epsilon, z+\epsilon]\right] \\ &= \;\; \int_{z-\epsilon}^{z+\epsilon} dD_n(y) = E(e^{Yt})^n \int_{z-\epsilon}^{z+\epsilon} e^{-ynt}\, dD_n^{(t)}(y) \;\;. \end{aligned}$$

Since for $y \in [z-\epsilon, z+\epsilon]$ it holds $e^{-ynt} \geq e^{-znt}\, e^{-\epsilon nt}$, we may conclude

$$Pr\left[\frac{S_n}{n} > \delta\right] \;\; \geq \;\; \left[E(e^{Yt})\, e^{-zt} \;\cdot\; e^{-\epsilon t}\right]^n \int_{z-\epsilon}^{z+\epsilon} dD_n^{(t)}(y) \;. \qquad (1.6)$$

Now we need two key observations. First, since the mean of $Y^{(t)}$ is z, the Low of Large Numbers implies that

$$Pr\left[\frac{S_n^{(t)}}{n} \;\in\; [z-\epsilon, z+\epsilon]\right] \;\; \rightarrow \;\; 1 \;,$$

as $n \rightarrow \infty$. Thus, the rightmost multiplier of (1.6) tends to 1, as n tends to infinity.

Second, consider the function $c_z(t) \;=\; E(e^{Yt})\, e^{-zt} \;=\; E(e^{(Y-z)t})$. The function is concave: $c_z^{''}(t) \;=\; E((Y-z)^2 e^{(Y-z)t}) \;\geq\; 0$. Therefore, it has at most one minimum. The necessary and sufficient condition for this minimum is $c_z^{'}(t) \;=\; E((Y-z)e^{(Y-z)t}) \;=\; 0$. But this is precisely how t was chosen in the first place! Thus, the global minimum of $c_z(x)$ is achieved at our t.

Keeping the two observations in mind, and taking logarithms in (1.6), we conclude:

$$\liminf_{n\rightarrow\infty} \frac{1}{n} \log \Pr\left[\frac{S_n}{n} > \delta\right] \;\; \geq \;\; \inf_{t>0}\{-zt + \log E(e^{Yt})\} \;-\; t\epsilon \;.$$

Observing that $\inf_{t>0}\{-zt + \log E(e^{Yt})\}$ is monotone non-increasing in z, and letting z tend to δ and ϵ tend to 0, we conclude the proof of the theorem. ∎

Since by inequality (1.5), for every n we have

$$\frac{1}{n} \log \Pr\left[\frac{S_n}{n} > \delta\right] \leq \;\; \inf_{t>0}\{-t\delta + \log E(e^{Yt})\} \;,$$

we conclude that

$$\lim_{n\rightarrow\infty} \frac{1}{n} \log \Pr\left[\frac{S_n}{n} > \delta\right] \;\; = \;\; \inf_{t>0}\{-t\delta + \log E(e^{Yt})\} \;.$$

Thus, the bound of inequality (1.5) is asymptotically optimal.

1.3 WHEN ONLY THE MEAN IS GIVEN: THE HOEFFDING BOUND

In this section we apply the method described in Section 1.2 to the case in which all we know about (or all we want to use of) the X_i-s is that they have the same mean μ. The Hoeffding inequality [6] claims that in this case, for any $a \leq 1 - \mu$,

$$\Pr\left[S_n - E[S_n] \geq na\right] \leq \left\{ \left(\frac{\alpha}{\beta}\right)^{\beta} \left(\frac{1-\alpha}{1-\beta}\right)^{1-\beta} \right\}^n \tag{1.7}$$

where

$$\alpha = \mu \quad ; \quad \beta = \mu + a \ .$$

Let $\Psi(\mu)$ be a class of random variables on $[0, 1]$ having the mean μ. Following the strategy of Section 1.2, we show:

1. There exists a unique member $X \in \Psi(\mu)$, which simultaneously attains the maximum of all $E(e^Y t)$, $Y \in \Psi(\mu)$, for all $t > 0$.

2. We plug the Laplace transform of X into the rightmost part of (1.3) and optimize the resulting expression with respect to t.

1.3.1 Maximizing the Laplace transform

Let $\rho(\mu)$ be a random variable defined as

$$\rho(\mu) \ = \ \begin{cases} 1 \text{ with probability } \mu \\ 0 \text{ with probability } 1 - \mu \end{cases} \quad .$$

Lemma 1.3.1 *$\rho(\mu)$ has the maximal possible moments of any order among the members of $\Psi(\mu)$.*

Proof: This is obvious: all the moments m_k of $\rho(\mu)$ are equal to $m_1 = \mu$, while for any $Y \in \Psi(\mu)$,

$$m_k(Y) \ = \ \int_0^1 t^k \, d\sigma(t) \ \leq \ \int_0^1 t \, d\sigma(t) \ = \ \mu \ . \qquad \blacksquare$$

Corollary 1.3.1 *For any $t \geq 0$, $E[e^{\rho(\mu)t}] \ \geq \ E[e^{Yt}]$ for any $Y \in \Psi(\mu)$.*

Proof: This is an immediate consequence of the expansion

$$E[e^{tX}] \ = \ \sum_{i=0}^{\infty} \frac{t^i E[X^i]}{i!} \ = \ \sum_{i=0}^{\infty} \frac{t^i \, m_i(X)}{i!},$$

and the fact that each m_i is maximized by $\rho(\mu)$. $\blacksquare$

Notice that $E[e^{\rho(\mu)t}] \ = \ (1 - \mu) + \mu e^t$.

1.3.2 Obtaining the Inequality

Let $\delta = a + \mu$. Combining Corollary 1.3.1 with the inequality (1.3), we get for any $t > 0$:

$$\Pr[S_n - n\mu \geq na] \;\leq\; \left(e^{-t\delta} E[e^{\rho(\mu)}]\right)^n \;=\; \left((1-\mu)e^{-t\delta} \;+\; \mu e^{t(1-\delta)}\right)^n \;. \quad (1.8)$$

What value of $t > 0$ makes this bound the best? We need to find the minimum of $B(t) = (1-\mu)e^{-t\delta} + \mu e^{t(1-\delta)}$ over all $t > 0$. Differentiating $B(t)$ with respect to t, we conclude that the minimum is achieved at $t = \tau$, where

$$e^{\tau} = \frac{1-\mu}{\mu} \cdot \frac{\delta}{1-\delta} \quad .$$

Substituting $t = \tau$ into (1.8), and making the cancelations, we obtain:

$$\Pr[S_n - n\mu \geq na] \;\leq\; \left\{ \left(\frac{\mu}{\delta}\right)^{\delta} \left(\frac{1-\mu}{1-\delta}\right)^{1-\delta} \right\}^n \quad , \qquad (1.9)$$

as claimed.

1.3.3 A Remark on the Form of the Upper Bound

The upper bound of equation (1.9) contains an expression of the form

$$F(\alpha, \beta) = \left(\frac{\alpha}{\beta}\right)^{\beta} \left(\frac{1-\alpha}{1-\beta}\right)^{1-\beta} \quad .$$

We shall meet the same kind of expression later, in the Bennett Inequality. It is not hard to check that it will appear whenever the extremal distribution is concentrated on two points. Here we would like to analyze this expression, and give it a potentially useful alternative form.

Consider the Entropy function $H(x), 0 \leq x \leq 1$, and its three first derivatives:

$$H(x) = -x \log x - (1-x)\log(1-x)$$

$$H'(x) = \log \frac{1-x}{x} \quad ; \quad H''(x) = -\frac{1}{x(1-x)} \quad ; \quad H'''(x) = \frac{1}{x^2} - \frac{1}{(1-x)^2}$$

Taking the logarithm of $F(\alpha, \beta)$ and inserting $\Delta = \beta - \alpha$ we get

$$\begin{aligned} \log F(\alpha,\beta) &= -\beta\log\beta - (1-\beta)\log(1-\beta) + \beta\log\alpha + (1-\beta)\log(1-\alpha) \\ &= H(\beta) + (\alpha+\Delta)\log\alpha + ((1-\alpha)-\Delta)\log(1-\alpha) \\ &= H(\beta) - H(\alpha) - \Delta H'(\alpha) \quad . \end{aligned}$$

The Taylor expansion of $H(\beta)$ at α gives:

$$H(\beta) = H(\alpha) + \Delta H'(\alpha) + \frac{\Delta^2}{2} H''(\alpha) + R_2 \,,$$

where R_2 is the remainder term of order 2; its exact form is

$$R_2 = \frac{1}{2}\int_\alpha^\beta (\beta - t)^2 H'''(t)dt = \frac{1}{2}\int_\alpha^\beta (\beta - t)^2 \left(\frac{1}{t^2} - \frac{1}{(1-t)^2}\right) dt \ .$$

Thus,

$$\log F(\alpha, \beta) \ = \ \frac{\Delta^2}{2} H''(\alpha) + R_2 \ = \ \frac{(\beta - \alpha)^2}{2\alpha(1-\alpha)} + R_2 \ .$$

In the case of the Hoeffding inequality we have $\alpha = \mu$, $\beta = \mu + a$. Thus, (1.9) can be alternatively presented as follows:

Let $\{X_i\}_{i=1}^\infty$ be a sequence of independent random variables on $[0, 1]$, all having the same mean μ. Then,

$$\Pr[S_n - E[S_n] \geq na] \leq \exp\left(-\frac{na^2}{2\mu(1-\mu)} + nR_2\right)$$

The last equation becomes particularly useful when $\mu \geq \frac{1}{2}$: in this case R_2 is negative, and it holds

$$\Pr[S_n - E[S_n] \geq na] \leq \exp\left(-\frac{na^2}{2\mu(1-\mu)}\right) \ . \tag{1.10}$$

1.4 WHEN THE MEAN AND THE VARIANCE ARE GIVEN: A SIMPLE PROOF OF THE BENNETT BOUND

In this section we apply the method described in Section 1.2 to the case in which we know that all X_i-s have the same mean μ, and the same variance σ^2. Equivalently, they have second moment $m_2 = \nu = \mu^2 + \sigma^2$. The Hoeffding inequality can be sharpened in this case; the sharper version is called the Bennett inequality [4]:

For any $a \leq 1 - \mu$,

$$\Pr\left[S_n - E[S_n] \geq na\right] \leq \left\{\left(\frac{\alpha}{\beta}\right)^\beta \left(\frac{1-\alpha}{1-\beta}\right)^{1-\beta}\right\}^n \tag{1.11}$$

where

$$\alpha = \frac{\sigma^2}{\sigma^2 + (1-\mu)^2} \quad ; \quad \beta = \frac{\sigma^2 + a(1-\mu)}{\sigma^2 + (1-\mu)^2} \quad .$$

We establish the Bennett inequality in essentially the same way as the Hoeffding inequality was established. Let $\Psi(\mu, \nu)$ be the class of random variables on $[0, 1]$ whose mean is μ, and whose second moment is ν. Following the same strategy as in Section 1.2, we show:

1. There exists a unique member $X \in \Psi(\mu, \nu)$, which simultaneously attains the maximum of all $E(e^Y t)$, $Y \in \Psi(\mu)$, for all $t > 0$.

2. We plug the Laplace transform of this X into the rightmost part of (1.3) and optimize the resulting expression with respect to t.

1.4.1 Maximizing the Laplace transform

Define a random variable $\rho(\mu,\nu) \in \Psi(\mu,\nu)$ by the following conditions: its second moment is precisely ν, and it is supported on two points, one of which is 1. Such $\rho(\mu,\nu)$ is well defined: if $\rho(\mu,\nu)$ assumes the value λ with probability p, and the value 1 with probability q, then p, q, λ are determined by

$$p+q=1 \quad ; \quad p\lambda+q=\mu \quad ; \quad p\lambda^2+q=\nu \ .$$

The solution of these equations can be expressed in the form:

$$\lambda=\frac{\mu-\nu}{1-\mu} \quad ; \quad p=\frac{1-\mu}{1-\lambda} \ ; \quad q=\frac{\mu-\lambda}{1-\lambda} \ . \tag{1.12}$$

Lemma 1.4.1 *$\rho(\mu,\nu)$ has the maximal possible moments of any order among the members of $\Psi(\mu,\nu)$.*

Proof:let $\{d_l\}_{l=1}^{\infty}$ be the sequence of moments of $\rho(\mu,\nu)$. It suffices to show that for any $X \in \Psi(\mu,\nu)$ with moments $\{m_i\}_{i=1}^{\infty}$, and for all $i \geq 0$:

$$m_i - m_{i+1} \geq d_i - d_{i+1}$$

or

$$\frac{m_i-m_{i+1}}{1-\mu} \ \geq \ \frac{d_i-d_{i+1}}{1-\mu} \ = \ \frac{(p\lambda^i+q)-(p\lambda^{i+1}+q)}{1-\mu} \ = \ \frac{p\lambda^i(1-\lambda)}{1-\mu} \ = \ \lambda^i$$

Let F be the distribution of X. Let Y be a random variable on $[0,1]$ with a distribution function G defined by

$$dG(x)=\frac{1}{1-\mu}(1-x)\,dF(x) \ .$$

It is easy to check that Y is well defined. It holds that

$$E[Y^i]=\frac{1}{1-\mu}\int_0^1 x^i(1-x)dF(x)=\frac{m_i-m_{i+1}}{1-\mu}$$

Since $E[Y^i]^{1/i}$ is a nondecreasing function of i (see e.g., [12]), one has

$$\frac{m_i-m_{i+1}}{1-\mu} \ = \ E[Y^i] \ \geq \ (E[Y])^i \ = \ \left(\frac{m_1-m_2}{1-\mu}\right)^i \ = \ \left(\frac{d_1-d_2}{1-\mu}\right)^i \ = \ \lambda^i$$

as desired. ∎

Arguing as in the proof of Corollary 1.3.1, we conclude that

Corollary 1.4.1 *For any $t \geq 0$, $E[e^{\rho(\mu,\nu)t}] \ \geq \ E[e^{Yt}]$ for any $Y \in \Psi(\mu,\nu)$.*

Notice that $E[e^{\rho(\mu,\nu)t}] \ = \ pe^{t\lambda}+qe^t$.

1.4.2 Obtaining the Inequality

Let $\delta = a + \mu$. Combining Corollary 1.4.1 with the inequality (1.3), we get for any $t > 0$:

$$\Pr[S_n - n\mu \geq na] \;\leq\; \left(e^{-t\delta} E[e^{\rho(\mu,\nu)}]\right)^n \;=\; \left(e^{-t\delta} \cdot (pe^{t\lambda} + qe^t)\right)^n \quad , \quad (1.13)$$

where p, q, λ are as in (1.12). Let $B(t) = e^{-t\delta} \cdot (pe^{t\lambda} + qe^t)$. Differentiating $B(t)$ with respect to t, we find that this expression is minimized for $t = \tau$, satisfying

$$e^{\tau(\lambda-1)} = \frac{q}{p} \cdot \frac{1-\delta}{\delta - \lambda} \; .$$

Notice that $\delta > \lambda$, and thus $\tau > 0$. Substituting $t = \tau$, we get:

$$\begin{aligned}
\Pr[S_n - n\mu \geq na] &\leq \left\{ \left(\frac{q}{p} \cdot \frac{1-\delta}{\delta-\lambda}\right)^{\frac{1-\delta}{\lambda-1}} \frac{q(1-\lambda)}{\delta-\lambda} \right\}^n \\
&= \left\{ \left(\frac{\mu-\lambda}{\delta-\lambda}\right)^{1-\frac{1-\delta}{1-\lambda}} \left(\frac{1-\mu}{1-\delta}\right)^{\frac{1-\delta}{1-\lambda}} \right\}^n \\
&= \left\{ \left(\frac{\alpha}{\beta}\right)^{\beta} \left(\frac{1-\alpha}{1-\beta}\right)^{1-\beta} \right\}^n \quad ,
\end{aligned}$$

with

$$\alpha = \frac{\mu-\lambda}{1-\lambda} = \frac{\sigma^2}{\sigma^2 + (1-\mu)^2} \quad ; \quad \beta = \frac{\delta-\lambda}{1-\lambda} = \frac{\sigma^2 + a(1-\mu)}{\sigma^2 + (1-\mu)^2} \; .$$

This is exactly what we wanted to show.
The alternative form presented in Section 1.3.3 applies here as well. For Bennett's bound it says:

$$\Pr[S_n - E[S_n] \geq na] \leq \exp\left(-\frac{na^2}{2\sigma^2} + nR_2\right) \quad (1.14)$$

with

$$R_2 \;=\; \frac{1}{2}\int_{\alpha}^{\beta} (\beta - t)^2 \left(\frac{1}{t^2} - \frac{1}{(1-t)^2}\right) dt \; .$$

1.5 WHEN THE FIRST N MOMENTS ARE GIVEN: A GLIMPSE OF THE GENERAL THEORY

Generalizing the results of Sections 1.3 and 1.4, we consider now the case when all X_i-s have the same first n moments $m_k = E(X_i^k)$, $k = 0, 1, .., n$. Although the situation becomes considerably more involved, it can still be satisfactorily analyzed, and the main results can still be stated in a clear way.

Let $\Psi(\mathbf{m})$ be class of random variables on $[0, 1]$ whose first n moments are given by $\mathbf{m} = (m_0, m_1, ..., m_n)$. In order to estimate the $\Pr[S_n - E[S_n] \geq na]$ using the strategy of Section 1.2, we need to

1. Find an easy-to-handle expression for the function $Z(t)$,

$$Z(t) = \sup_{Y \in \Psi(\mathbf{m})} E[e^{Yt}].$$

2. Minimize with respect to t the expression

$$e^{-t\delta)} Z(t) \ ,$$

where $\delta = \mu + a$.

The first question falls into the circle of problems related to the the so-called Markov Moment Problem. The underlying general theory is elaborated in the excellent (both in its scope and and conceptual clarity) book [11]. Our presentation will be close to that of [11].

The answer to the first question is:

(a) There exists a unique member $\rho(\mathbf{m}) \in \Psi(\mu)$, which simultaneously attains the maximum of all $E(e^Y t)$, $Y \in \Psi(\mu)$, for all $t > 0$.

(b) $\rho(\mathbf{m})$ is discrete, and is supported on at most n points. That is, there exist at most n points $\Xi = \{\xi_i\}_{i=0}^r$ such that $\Pr[\rho(\mathbf{m}) = \xi_i] > 0$.

(c) The points $\Xi = \{\xi_i\}_{i=0}^r$ can be efficiently computed; they are roots of some explicitly constructed polynomial.

(d) Once the set $\{\xi_i\}_{i=0}^r$ is determined, the corresponding weights $w_i = \Pr[\rho(\mathbf{m}) = \xi_i]$ can be obtained by solving the (nonsingular) system of equations

$$\sum_{i=0}^{r} w_i \, \xi_i^j \ = \ m_j, \qquad j = 0, ..., r \ . \tag{1.15}$$

(e) Finally,

$$Z(t) \ = \ \sum_{i=0}^{r} w_i \, e^{\xi_i t} \ .$$

It is easy to perceive a similarity between the previously studied cases $n = 1$ and $n = 2$, and the current general situation.

Consider now the second question. Although in general there is no closed-form solution, it can still be solved reasonably well. The function we wish to minimize,

$$e^{-t\delta} Z(t) \ = \ \sum_{i=0}^{r} w_i \, e^{(\xi_i - \delta)t},$$

is concave, and thus has a unique minimum. Differentiating, we conclude that this minimum is achieved at $t = \tau > 0$ such that

$$\sum_{i=0}^{r} w_i (\xi_i - \delta) \, e^{\xi_i \tau} = 0 \ .$$

While the form of the present solution is more complex than the one corresponding to $n \leq 2$, it is still not too hard to work with, both numerically and theoretically.

It remains to give a justification to facts **(a)**, **(b)**, **(c)**. Our main goal in the rest of this section is to give the reader an intuitively clear outline of the relevant parts of the general theory. While we shall make a keen attempt to make the proofs mathematically sound, plausible but somewhat technically involved details will occasionally be omitted. For full details and the full account of the beautiful general theory, the reader is referred to [11].

1.5.1 Preliminaries: The Geometry of the Moment Space

Define the *moment space* $\mathcal{M}_n \subseteq R^{n+1}$ as

$$\mathcal{M}_n = \left\{ \mathbf{m} \;=\; (m_0, .., m_n) \mid m_i = \int_0^1 t^i \, d\sigma(t), \;\; i = 0, 1, .., n \right\} \quad ,$$

where σ ranges over all probability distributions on $[0, 1]$. Observe that $\mathcal{M}_n$ is precisely the convex hull of the moment curve

$$\mathcal{C}_n = \left\{ (t^0, t^1, .., t^n) \mid t \in [0, 1] \right\} \subseteq R^{n+1} \quad .$$

Indeed, $\mathcal{M}_n$ contains the moment curve: the vector $(t^0, t^1, .., t^n)$ corresponds to σ, which has all its weight on the point t. Since $\mathcal{M}_n$ is obviously convex, this implies $conv\,(\mathcal{C}_n) \subseteq M_n$. On the other hand, since an integral of a function with respect to a probability measure can be viewed as a convex combination of the function's values, for any $\mathbf{m}$ and corresponding σ we have

$$\mathbf{m} = (m_0, .., m_n) = \int_0^1 (t^0, t^1, ..., t^n) \, d\sigma(t) \in \mathrm{conv}(\mathcal{C}_n) \quad .$$

Notice also that $\mathcal{M}_n$, being a convex hull of a compact set $\mathcal{C}_n$, is compact.

In what follows, given a vector of coefficients $\mathbf{a} = (a_0, a_1, .., a_n)$, we define the polynomial $P_{\mathbf{a}}(x)$ as $P_{\mathbf{a}}(x) = \sum_{i=0}^n a_i x^i$.

The following important structure theorem can be viewed as a dual characterization of $\mathcal{M}_n$:

Theorem 2 *A sequence of real numbers* $\mathbf{s} = \{(s_0, s_1, ..., s_n)\}$ *represents the first* $n+1$ *moments (counting the zero-moment) of some probability distribution on* $[0, 1]$ *if and only if* $s_0 = 1$, *and for any polynomial* $P(x) = \sum_{i=0}^n a_i x^i$ *nonnegative on* $[0, 1]$, *the value of* $\sum_{i=0}^n a_i s_i$ *is also nonnegative.*

Proof: If $\mathbf{m}$ is a moment sequence of some σ, and $P(x) = \sum_{i=0}^n a_i x^i$ is a nonnegative polynomial on $[0, 1]$, we get:

$$\sum_{i=0}^n a_i m_i \;=\; \sum_{i=0}^n a_i \int_0^1 t^i \, d\sigma(t) \;=\; \int_0^1 P(t) \, d\sigma(t) \;\geq\; 0 \quad .$$

This, together with the fact that

$$m_0 = \int_0^1 t^0 \, d\sigma(t) = 1 \quad ,$$

establishes the "only if" part of the theorem.

Recall that any closed convex set is the intersection of the (closed) half-spaces defined by its supporting hyperplanes. What are the supporting hyperplanes of $\mathcal{M}_n$? It is easy to visualize that $\mathcal{M}_n$ is the section of the cone $Cone_n = \{c\mathbf{m} \mid c \geq 0;\ \mathbf{m} \in \mathcal{M}_n\}$ by the hyperplane $s_0 = 0$. Thus, to show the "if" part, it suffices to show that any supporting hyperplane of $Cone_n$ is of the form $\mathbf{a} \cdot \mathbf{s} = 0$, where $P_{\mathbf{a}}(x)$ is nonnegative on $[0, 1]$. Recall that from the "only if" part we already know that for every such $\mathbf{a}$, $\mathcal{M}_n$ and consequently $Cone_n$, are contained in the half-space $\mathbf{a} \cdot \mathbf{s} \geq 0$.

Observe that cones with an apex at 0 have only supporting hyperplanes of the form $\mathbf{a} \cdot \mathbf{s} = 0$. Consider one such supporting hyperplane. Assume for contradiction that the corresponding polynomial $P_{\mathbf{a}}(x)$ is not nonnegative on $[0, 1]$, i.e., there exists $\xi \in [0, 1]$ such that $P_{\mathbf{a}}(\xi) < 0$. Consider a distribution whose entire mass is concentrated at ξ. Let $\mathbf{m} = (\xi^0, \xi^1, ..., \xi^n)$ be the moment sequence of this distribution. Then, $\mathbf{a} \cdot \mathbf{m} = P_{\mathbf{a}}(\xi) < 0$, contrary to our assumption that $\mathbf{a} \cdot \mathbf{s} = 0$ is a supporting hyperplane of $Cone_n$. ∎

1.5.2 On Finite Representations of $\mathbf{m}$

Given a sequence of moments $\mathbf{m} = (m_0, ..., m_n)$ a *finite representation* (or just *representation*) σ of $\mathbf{m} = (m_0, ..., m_n)$ is a probability distribution on $[0, 1]$ with moments given by $\mathbf{m}$, whose support is a discrete set of (distinct) points $\xi_0, ..., \xi_r \in [0, 1]$. In what follows, the points $\{\xi_i\}_{i=0}^r$ will be called the *roots* of the representation σ. The weights $\{w_i\}_{i=0}^r$ associated these points will be called the *weights* of σ.

Given a representation σ, define the *index* of a root ξ_i to be $ind(\xi_i) = 1$ if ξ_i is 0 or 1, and $ind(\xi_i) = 2$ if $0 < \xi_i < 1$. The index of a non-root will be defined as 0. Define the index of the representation σ as the sum of the indices of its roots.

Call a sequence of moments $\mathbf{m} = (m_0, ..., m_n)$ *singular* if it has a representation of index $\leq n$.

Theorem 3 *A sequence of moments* $\mathbf{m}$ *is singular if and only if there exists* $\mathbf{a} = (a_0, a_1, ..., a_n)$ *such that* $\mathbf{a} \cdot \mathbf{m} = \sum_{i=0}^n a_i m_i = 0$, *while* $P_{\mathbf{a}}(x) = \sum_{i=0}^n a_i x^i$ *is nonnegative on* $[0, 1]$.
Moreover, for a singular $\mathbf{m}$, *there exists a unique (up to subsets of measure zero) probability distribution* σ *whose moment sequence coincides with* $\mathbf{m}$. *It is precisely the least degree representation of* $\mathbf{m}$.

Proof: Let $\{\xi_i\}_{i=0}^r$ be the roots of a representation of index d, $d \leq n$, of $\mathbf{m}$. Define a polynomial

$$P(x) = (1 - x)^{ind(1)} \prod_{\xi \neq 1} (x - \xi)^{ind(\xi)} \ .$$

It is a simple matter to check that the degree of $P(x)$ is exactly d, and that $P(x)$ is nonnegative on $[0,1]$. Let $\mathbf{a} = (a_0, a_1, ..., a_n)$ be the vector of coefficients of $P(x) = \sum_{j=0}^{n} a_j x^j$; if $d < n$, we take $a_{d+1} = 0, ..., a_n = 0$. It holds that

$$\mathbf{a} \cdot \mathbf{m} = \sum_{j=0}^{n} a_j \left(\sum_{i=0}^{r} w_i \xi_i^j \right) = \sum_{i=0}^{r} w_i P(\xi_i) = 0 \ ,$$

establishing the "if" part of the theorem.

Assume now there exists an $\mathbf{a}$ such that $\mathbf{a} \cdot \mathbf{m} = 0$ while $P_{\mathbf{a}}(x)$ is nonnegative on $[0,1]$. Let $\{\xi_i\}_{i=0}^{r}$, $r \leq n$, be the roots of $P_{\mathbf{a}}(x)$ lying in the interval $[0,1]$. Then, for any probability distribution σ with moments specified by $\mathbf{m}$,

$$0 = \mathbf{a} \cdot \mathbf{m} = \mathbf{a} \cdot \int_0^1 (t^0, t^1, ..., t^n) \, d\sigma(t) = \int_0^1 P_{\mathbf{a}}(t) \, d\sigma(t) \quad .$$

Clearly, this is possible only when σ assigns measure 0 to the set $[0,1] - \{\xi_i\}_{i=0}^{r}$. Thus, σ must be supported on the set $\{\xi_i\}_{i=0}^{r}$ of index $d \leq n$. Thus, $\mathbf{m}$ has a representation of index $\leq n$. This establishes the "if" part.

Putting together the observations we have made so far, we conclude that any probability distribution σ on $[0,1]$ with moments specified by $\mathbf{m}$, must be supported on the zeroes of any representation of index $\leq n$. Consider the least-index representation of $\mathbf{m}$. Its roots $\{\xi_i\}_{i=0}^{r}$ are uniquely defined, since they form a subset of roots of any representation of $\mathbf{m}$ of index $\leq n$, and in particular of the one of the smallest possible index. The corresponding weights $\{w_i\}_{i=0}^{r}$ are also uniquely defined: since $r \leq n$,

$$\begin{pmatrix} \xi_0^0 & \xi_1^0 & ... & \xi_r^0 \\ \xi_0^1 & \xi_1^1 & ... & \xi_r^1 \\ \vdots & \vdots & \vdots & \vdots \\ \xi_0^r & \xi_1^r & ... & \xi_r^r \end{pmatrix}^{-1} \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_r \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_r \end{pmatrix} \quad .$$

Notice that by choosing the least-degree representation we have ensured that all w_i-s are strictly positive. ∎

Consider now a non-singular $\mathbf{m}$. A representation of $\mathbf{m}$ of index $n+1$ will be called *principal*. Observe that there can be only two kinds of principal representations:
For n even, it is either

$$0 = \xi_0 < \xi_1 < ... < \xi_{\frac{n}{2}} < 1, \quad \text{or} \quad 0 < \xi_0 < \xi_1 < ... < \xi_{\frac{n}{2}} = 1 \ . \tag{1.16}$$

We call the two kinds the *lower* and the *upper* principal representations, respectively.
For n odd, the lower and the upper principal representation will be, respectively,

$$0 < \xi_0 < \xi_1 < ... < \xi_{\lceil \frac{n}{2} \rceil} < 1, \quad \text{and} \quad 0 = \xi_0 < \xi_1 < ... < \xi_{\lceil \frac{n}{2} \rceil} = 1 \ . \tag{1.17}$$

It will be shown that for a nonsingular $\mathbf{m}$ the upper and the lower principal representations always exist, and, moreover, are uniquely defined.

1.5.3 The Extremal Properties of Principal Representations

Without risk of running into confusion, let us identify for the rest of this section random variables and their underlying distributions. Thus, $\Psi(\mathbf{m})$ will denote the set of all probability distribution on $[0,1]$ whose first n moments (not counting m_0) are given by $\mathbf{m}$. The relevance of the principal representations to the main theme of this paper comes forward in the following theorem:

Theorem 4 *Let* $\mathbf{m} = (m_0, m_1, ..., m_n)$ *be a sequence of moments of some probability distribution on* $[0,1]$. *Assume further that* $\mathbf{m}$ *is non-singular. Then, among all* $\sigma \in \Psi(\mathbf{m})$, *the maximal* k*-th moment* $m_k(\sigma)$, $k > n$, *is achieved on an upper principal representation of* $\mathbf{m}$. *Similarly, the minimal* $m_k(\sigma)$ *is achieved on a lower principal representation.*

Proof: We will prove only the part concerning the upper principal representation; the part concerning of the lower one has a very similar proof, and is omitted.

Using arguments similar to those used in establishing the structure of the moment space (its convexity and its dual characterization), we arrive at similar conclusions:

1. The set

$$\left\{ (m_0, .., m_n, m_k) \mid m_i = \int_0^1 t^i \, d\sigma(t), \ i = 0, 1, .., n, k \right\} \quad ,$$

   where σ ranges over all probability distributions on $[0,1]$, is precisely the convex hull of the curve

$$\left\{ (t^0, t^1, .., t^n, t^k) \mid t \in [0,1] \right\} \ .$$

   Thus, it is compact, and the maximum of m_k on $\Psi(\mathbf{m})$ is achieved.

2. Given a sequence of moments $\mathbf{m}$ and a real number s_k, there exists a probability distribution $\sigma \in \Psi(\mathbf{m})$ whose k-th moment is equal to s_k, if and only if the linear form $a_k s_k + \sum_{i=0}^n a_i m_i$ is nonnegative, whenever the corresponding polynomial $a_k x^k + \sum_{i=0}^n a_i x^i$ is nonnegative on $[0,1]$.

This implies that the maximal value of m_k is defined by the linear programming problem

$$\text{Min} \sum_{i=0}^n a_i m_i \ \text{ subject to } \ -x^k + \sum_{i=0}^n a_i x^i \ \geq \ 0 \quad \text{for any } x \in [0,1] \ . \quad (1.18)$$

Notice that (1.18) is nothing but the dual of the primal program

$$\text{Max}_\sigma \int_0^1 t^k \, d\sigma(t) \text{ subject to } \int_0^1 t^i \, d\sigma(t) = m_i, \ i = 0, 1, .., n \ . \quad (1.19)$$

What can be said about the vector $\mathbf{a}$ achieving the optimal value in (1.18)? We need first a preliminary fact:

Fact: *A polynomial of the form* $P(x) = a_k x^k + \sum_{i=0}^{n} a_i x^i \neq a_k x^k$, $k > n$, *can have at most* $n+1$ *nonnegative real roots (counting their cardinalities).*

This fact can be proven by induction on n, using a simple corollary to Rolle's Theorem, implying that the number of nonnegative roots of $P'(x)$ is at least that of $P(x)$, minus one.

Claim 1.5.1 *If a polynomial* $-x^k + \sum_{i=0}^{n} a_i x^i$, *nonnegative on* $[0,1]$, *has less than* $n+1$ *zeroes (counting their cardinalities) in* $[0,1]$, *the value of* $\sum_{i=0}^{n} a_i m_i$ *is not optimal for the linear program (1.18).*

Proof: Indeed, let $P(x)$ be such a polynomial, and let $\{\xi_i\}_{i=0}^{r}$ be the set of its distinct roots in $[0,1]$. Let $d(\xi_i)$ be the cardinality of the root ξ_i; if ξ is not a root, let $d(\xi)$ be 0. Let $z(P) = \sum_{i=0}^{r} d(\xi_i)$ be number of roots of $P(x)$ (counting their cardinalities) in $[0,1]$. By our assumption, $z(P) \leq n$.

Let us represent $P(x)$ as $P(x) = R(x)Q(x)$, where $R(x)$ has no zeroes in the interval $[0,1]$, and $Q(x)$ is of the form

$$Q(x) = (1-x)^{d(1)} \prod_{\xi_i \neq 1} (x - \xi_i)^{d(\xi_i)} = \sum_{i=0}^{n} b_i x^i \quad .$$

Notice that the degree of $Q(x)$ is $z(P) \leq n$. If it is strictly less than n, the leading b_i-s will be 0. If $R(x)$ has no zeroes in $[0,1]$, let $Q(x) \equiv 1$.

Clearly, all $d(\xi_i)$-s (with a possible exclusion of $d(0)$ and $d(1)$) must be even – otherwise $P(x)$ changes its sign on $[0,1]$. Thus, $Q(x)$ is nonnegative on $[0,1]$. Since $P(x)$ is also nonnegative there, and $P(x) = R(x)Q(x)$, the polynomial $R(x)$ must be nonnegative on $[0,1]$ as well. In fact, since $R(x)$ has no zeroes in this interval, it must be strictly positive there. Let $\alpha > 0$ be the minimum of $R(x)$ on $[0,1]$. Consider now a new polynomial,

$$P_1(x) = \sum_{i=0}^{n} (a_i - \alpha b_i) x^i = P(x) - \alpha Q(x) = (\, R(x) - \alpha \,)\, Q(x) \quad .$$

By the choice of α, $P_1(x)$ is also nonnegative on $[0,1]$. To conclude the proof of the claim, it remains to show that the value of the linear form (as in (1.18)) corresponding to $P_1(x)$ is smaller than the one corresponding to $P(x)$. Indeed,

$$\sum_{i=0}^{n} (a_i - \alpha b_i) m_i \;=\; \sum_{i=0}^{n} a_i m_i \;-\; \alpha \sum_{i=0}^{n} b_i m_i \;.$$

Keeping in mind that $\mathbf{m}$ is nonsingular, and that $\mathbf{b} = (b_0, ..., b_n)$ is the sequence of coefficients of a nonnegative polynomial $Q(x)$, we conclude that $\mathbf{b} \cdot \mathbf{m} > 0$. ∎

We return to the proof of our theorem. By Claim 1.5.1, the maximal m_k is equal to some $\mathbf{a} \cdot \mathbf{m}$, such that the corresponding polynomial $-x^k + P_{\mathbf{a}}(x)$ is nonnegative on $[0,1]$, and has $n+1$ roots (i.e., all its roots) there. Consider a distribution σ whose moments are $m_0, ..., m_n$ and $m_k = \mathbf{a} \cdot \mathbf{m}$. Using the

argument we have used many times by now, σ has all its weight on the roots of $-x^k + P_{\mathbf{a}}(x)$. Since all the inner roots are of even cardinality, we conclude that σ is a representation of index $n+1$ or less. But it cannot be less: by our assumptions $\mathbf{m}$ is nonsingular. Thus, σ is a principal representation. Moreover, since $-x^k + P_{\mathbf{a}}(x)$ has all its roots in the interval $[0,1]$, and its value at $-\infty$ is $-\infty$, σ must have 1 among its roots. We conclude that σ is an upper principal representation. ∎

1.5.4 The Uniqueness of the Principal Representations, and their Efficient Computation

Given the sequence of moments $\mathbf{m} = (m_0, m_1, .., m_n)$, let us define the two polynomials $\overline{P}(x)$ and $\underline{P}(x)$ in the following manner. For n even define

$$\overline{P}(t) = (1-x)\det\begin{pmatrix} m_0 - m_1 & m_1 - m_2 & ... & m_{\frac{n}{2}-1} - m_{\frac{n}{2}} & t^0 \\ m_1 - m_2 & m_2 - m_3 & ... & m_{\frac{n}{2}} - m_{\frac{n}{2}+1} & t^1 \\ \vdots & \vdots & & \vdots & \vdots \\ m_{\frac{n}{2}} - m_{\frac{n}{2}+1} & m_{\frac{n}{2}+1} - m_{\frac{n}{2}+2} & ... & m_{n-1} - m_n & t^{\frac{n}{2}} \end{pmatrix}$$

and

$$\underline{P}(t) = x\det\begin{pmatrix} m_1 & m_2 & ... & m_{\frac{n}{2}} & t^0 \\ m_2 & m_3 & ... & m_{\frac{n}{2}+1} & t^1 \\ \vdots & \vdots & & \vdots & \vdots \\ m_{\frac{n}{2}+1} & m_{\frac{n}{2}+2} & ... & m_n & t^{\frac{n}{2}} \end{pmatrix} .$$

For n odd define

$$\overline{P}(t) = x(1-x)\det\begin{pmatrix} m_1 - m_2 & m_2 - m_3 & ... & m_{\lfloor\frac{n}{2}\rfloor} - m_{\lfloor\frac{n}{2}\rfloor+1} & t^0 \\ m_2 - m_3 & m_3 - m_4 & ... & m_{\lfloor\frac{n}{2}\rfloor+1} - m_{\lfloor\frac{n}{2}\rfloor+2} & t^1 \\ \vdots & \vdots & & \vdots & \vdots \\ m_{\lfloor\frac{n}{2}\rfloor+1} - m_{\lfloor\frac{n}{2}\rfloor+2} & m_{\lfloor\frac{n}{2}\rfloor+2} - m_{\lfloor\frac{n}{3}\rfloor+2} & ... & m_{n-1} - m_n & t^{\lfloor\frac{n}{2}\rfloor} \end{pmatrix}$$

and

$$\underline{P}(t) = \det\begin{pmatrix} m_0 & m_1 & ... & m_{\lfloor\frac{n}{2}\rfloor} & t^0 \\ m_1 & m_2 & ... & m_{\lfloor\frac{n}{2}\rfloor+1} & t^1 \\ \vdots & \vdots & & \vdots & \vdots \\ m_{\lfloor\frac{n}{2}\rfloor+1} & m_{\lfloor\frac{n}{2}\rfloor+2} & ... & m_n & t^{\lfloor\frac{n}{2}\rfloor+1} \end{pmatrix} .$$

Theorem 5 *Let* $\mathbf{m}$ *be a nonsingular sequence of moments. Then the roots of* $\underline{P}(x)$ *are precisely the roots of the lower principal representation of* $\mathbf{m}$*, while the roots of* $\overline{P}(x)$ *are precisely the roots of the upper principal representation of* $\mathbf{m}$*. Since these roots uniquely define the weights, the upper and the lower principal representations are uniquely defined.*

Proof: We already know that $\mathbf{m}$ admits an upper and a lower principal representation. Let

$$0 < \xi_1 < \xi_2 < ... < \xi_r < 1 ,$$

be the set of the inner roots of such a presentation. Notice that the corresponding (i.e., depending on whether n is even or odd, and the representation is lower or upper) $P(x)$ is always of the form

$$P(x) = (1-x)^{ind(1)}\, x^{ind(0)} \det\left(M_{r+1\,\times\, r+1}(t)\right) \quad .$$

Thus, it suffices to prove that $\{\xi_i\}_{i=1}^{r}$ are exactly the roots of $M(t)$. Recall that

$$\begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ m_n \end{pmatrix} = ind(0)\, w(0) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \sum_{i=1}^{r} w_i \begin{pmatrix} 1 \\ \xi_i^1 \\ \vdots \\ \xi_i^n \end{pmatrix} + ind(1)\, w(1) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad .$$

It is easy to check that in each case $M(t)$ was constructed in such a manner that the contribution of the moment vectors corresponding to 0 and 1 is nil. Moreover, the first r columns of $M(t)$ are linear combinations of vectors $\{(\xi_i^0, \xi_i^1, ..., \xi_i^r)\}_{i=1}^{r}$, and the explicit computation of their coefficients shows that these rows are independent. We leave the verification of this fact to the reader; it follows easily from the non-singularity of the Vandermonde matrices.

But then, we are done. The matrix $M(t)$ becomes singular if and only if $(t^0, t^1, ..., t^r)$ belongs to the span of these vectors. This happens precisely when $t = \xi_i$ for some $1 \le i \le r$. ∎

1.5.5 The Computation of the Minimum Index Representation of a Singular $\mathbf{m}$

It remains to take care of the case when $\mathbf{m}$ is singular. In what follows let $\overline{P}_k(x)$ and $\underline{P}_k(x)$, $k < n$, be defined as the corresponding polynomials for the moment sequence $(m_0, m_1, .., m_k)$.

Theorem 6 *Assume that* $\mathbf{m}$ *has a representation* σ *of index* $k+1 < n+1$*, and no representations of a lesser index. Assume further that* σ *has the form of an upper (lower) principal representation of index* k*. Then it is indeed the upper (lower) representation of* $(m_0, m_1, .., m_k)$*, and its roots are the roots of* $\overline{P}_k(x)$ *(of* $\underline{P}_k(x)$*,respectively). Furthermore, the polynomial* $\overline{P}_{k+1}(x)$ *(respectively,* $\underline{P}_{k+1}(x)$*) is identically equal to zero, while* $\underline{P}_{k+1}(x)$ *(respectively,* $\overline{P}_{k+1}(x)$*) is not.*

Proof: The fact that σ is a principal representation of $(m_0, m_1, .., m_k)$ follows directly from the definition of the principal representation. Thus, by Theorem 5, its roots are the roots of $\overline{P}_k(x)$ (or of $\overline{P}_k(x)$, respectively). The second part of the theorem can be verified using an argument similar to that used in the proof of Theorem 5, and performing case analysis. We omit the technical details. ∎

1.5.6 The Conclusion: Finding $\rho(\mathbf{m})$

Consider first the case when $\mathbf{m}$ is nonsingular. By Theorem4, the maximal value of moments of all orders for $X \in \Psi(\mathbf{m})$ is attained on the upper principal representation of $\mathbf{m}$. By Corollary 1.3.1, the maximal value of all Laplace transforms $E[e^t X]$, $t \geq 0$, must be also attained there. Finally, by Theorem 5 the upper principal representation of $\mathbf{m}$ is unique. This is exactly the $\rho(\mathbf{m})$ we are looking for!

To find it, one needs to explicitly compute the polynomial $\overline{P}(x)$. By Theorem 5, the roots of $\rho(\mathbf{m})$ are precisely the roots of $\overline{P}(x)$. The weights of $\rho(\mathbf{m})$ can be computed by solving the system of linear equations 1.15.

Consider now the case when $\mathbf{m}$ is singular. The situation becomes simpler: by Theorem 3 $\Psi(\mathbf{m})$ consists of a single probability distribution σ (up to subsets of measure 0). The corresponding random variable X will be our $\rho(\mathbf{m})$.

How shall we find the roots of σ? By Theorem 6, it suffices to find the minimum $k+1$ such that exactly one of $\overline{P}_{k+1}(x)$, $\underline{P}_{k+1}(x)$ is identically 0. Say, it is $\underline{P}_{k+1}(x)$. Then the roots of $\rho(\mathbf{m})$ are exactly the roots of $\underline{P}_k(x)$. The weights are found as before.

1.6 AN APPLICATION: IMPROVED BOUNDS FOR THE LIST UPDATE PROBLEM

In this section we apply the general strategy introduced in Section 1.2 to a concrete distribution arising from a well-studied problem, and obtain better results in simulations than those obtained by other methods.

The problem is the List Update Problem (see,e.g., [8, 7]), in which a set of n items held as a linear list is accessed randomly, according to some fixed probability distribution. Each request involves a search for a specific item (identified uniquely by its key). The probability of accessing the i-th element R_i, $1 \leq i \leq n$, is p_i. The p_i's are fixed but initially *unknown*. The list is dynamically reorganized along a reference sequence, so as to improve the relative ordering of the items. Each request is implemented as a sequential search starting at the header. Clearly, the optimal static arrangement of the items for this implementation is by decreasing order of the access probabilities. The Counter Scheme (*CS*), which maintains a reference count for each element, and rearranges the list in decreasing order of the counters, can be shown to converge to the optimal ordering [7]. The goal is to estimate in advance the number m of samples (equivalently, to find the *stopping point*) so that the arrangement produced by the Counter Scheme after m rounds will be not much worse than the optimal arrangement.

Hofri and Shachnai present in [7] a stopping point for this reorganization process in the case in which the vector of access probabilities $\bar{p} = (p_1, \ldots, p_n)$ is known, but the permutation assigning these probabilities to the elements $\{R_i\}_{i=1}^n$ is not known. In what follows we assume $p_1 \geq p_2 \geq ... \geq p_n$. Denote by $C_m(CS|\bar{p})$ the expected average access cost to the list after m references,

and by $C(OPT \mid \bar{p})$ the actual optimal average access cost. Notice that

$$C(OPT \mid \bar{p}) \;=\; \sum_{i=1}^{n} i\, p_i \;.$$

Let σ_m denote the CS order of the list elements after the mth reference, and let $\Pr_{CS}[\sigma_m(j) < \sigma_m(i)]$ be the probability that R_j precedes R_i in σ_m. In [7] it is shown using the additivity of expectation that

$$C_m(CS|\bar{p}) \;=\; C(OPT \mid \bar{p}) \;+\; \sum_{1 \le i < j \le n} (p_i - p_j) \cdot \Pr[\sigma_m(j) < \sigma_m(i)] \;. \quad (1.20)$$

The usual approach to estimating the gap $C_m(CS|\bar{p}) - C(OPT \mid \bar{p})$ is by providing good tail estimations on probabilities $\Pr[\sigma_m(j) < \sigma_m(i)]$. Here we shall use a superior (in particular, asymptotically optimal) method for estimating these probabilities; the numerical simulations indeed show that the bounds obtained by our method significantly outperform those based on Chebyshev and Hoeffding inequalities.

Lemma 1.6.1 *Assume that $j > i$, or, equivalently, $p_j < p_i$. Then,*

$$\Pr[\sigma_m(j) < \sigma_m(i)] \;\le\; \left(1 - (\sqrt{p_i} - \sqrt{p_j})^2\right)^m \;.$$

Proof: Let Y_k, $k = 1, .., m$, be a random variable, which takes value -1 if R_i was referred to in the k-th stage, 1 if R_j was referred in the k-th stage, and 0 otherwise. Clearly, the Y_k-s are independent, and they all have the same distribution[3]

$$Y_k = \begin{cases} -1 & p_i \\ 0 & 1 - p_i - p_j \\ 1 & p_j \end{cases} \;. \quad (1.21)$$

Let $S_m = \sum_{k=1}^{m} Y_k$. The expectation of Y_k is $\mu = p_j - p_i$. Evidently,

$$\Pr[\sigma_m(j) < \sigma_m(i)] \;=\; \Pr[S_m \ge 0] \;=\; \Pr[S_m - n\mu \ge -n\mu] \;.$$

The Laplace transform of Y_k is

$$Z(t) \;=\; E[e^{tY_k}] = p_i e^{-t} + p_j e^{t} + (1 - p_i - p_j) \quad ;$$

thus, by (1.5),

$$\Pr[S_m - n\mu \ge n\mu] \;\le\; \min_{t>0} \left(Z(t)\right)^m \;=\; \min_{t>0} \left(p_i e^{-t} + p_j e^{t} + (1 - p_i - p_j)\right)^m \;.$$

It is easy to check that the minimum of $p_i e^{-t} + p_j e^{t} + (1 - p_i - p_j)$ is attained at $\tau = \ln \sqrt{p_i / p_j} > 0$, and its value is

$$2\sqrt{p_i p_j} + 1 - p_i - p_j \;=\; 1 - (\sqrt{p_i} - \sqrt{p_j})^2 \;.$$

[3]Indeed, since we know the distribution of the Y_k's, $Z(t)$ is chosen as the Laplace transform of Y_k. Hence, our derivation here follows the steps of the Chernoff bound technique [3].

This completes the proof of the lemma. ∎

As an immediate consequence of Lemma 1.6.1 and inequality (1.20) we obtain the most interesting result of this section:

Theorem 7 (A Stopping Point for the CS) *For a list of n items with the probability vector $(p_1, \ldots, p_n)$, and any $0 < \varepsilon < 1$,*

$$C_m(CS|\bar{p}) \;\leq\; (1+\epsilon)C(OPT\,|\,\bar{p})\;,$$

provided that the number of references m satisfies

$$\sum_{i<j}(p_i - p_j)(1-(\sqrt{p_i}-\sqrt{p_j})^2)^m \;\leq\; \epsilon C(OPT\,|\,\bar{p}) \;=\; \epsilon\sum_{i=1}^{n} i\,p_i\;. \qquad \blacksquare \quad (1.22)$$

We provide Tables 1.1 and 1.2, which compare the bound on m given by Theorem 7 to the bounds obtained by estimating the probabilities $\Pr[\sigma_m(j) < \sigma_m(i)]$ by means of Chebyshev and Hoeffding inequalities.

We use for our test the family of so-called Zipf distributions Z_n, where $p_i = 1/(H_n i)$, $H_n = \sum_{i=1}^{n} \frac{1}{n}$. Numerical simulations show that the reference process is often governed by a Zipf distribution, especially when keys are drawn from a text file (as in Lisp [1]). Thus, they are close to the distributions met in practice, and can serve as good test distributions. In comparison with Chebyshev-based bounds (Table 1.1), which also take the variance into account, the improvement dramatically increases as ϵ becomes smaller. The comparison with Hoeffding-based bounds is even more favorable.

n\ ϵ	0.0001	0.001	0.01	0.05	0.1	0.15	0.2
10	22.06	5.77	3.25	2.77	2.41	2.36	2.24
20	11.66	4.63	3.61	2.96	2.6	2.47	2.38
25	9.85	4.48	3.78	3.00	2.66	2.51	2.46
50	6.78	4.45	4.13	3.04	2.72	2.52	2.52
100	5.56	4.78	4.14	3.10	2.73	2.71	2.58

Table 1.1 The required number of references for the reorganization process under CS to approach the optimum within $1+\epsilon$: The ratio between a Chebyshev-based bound and the stopping point in Theorem 7.

Acknowledgments

We would like to thank Jim Fill, Micha Hofri, Kurt Mehlhorn and Ofer Zeitouni for helpful comments and suggestions.

n \ ϵ	0.05	0.1	0.15	0.2
10	7.73	6.53	6.06	5.37
20	18.21	15.75	14.25	13.0
25	23.84	20.81	18.84	17.48
50	52.13	48.24	34.58	34.28
100	87.03	79.83	65.24	52.33

Table 1.2 The required number of references for the reorganization process under CS to approach the optimum within $1 + \epsilon$: The ratio between a Hoeffding-based bound and the stopping point in (1.22).

References

[1] Bentley J. L., McGeogh, *Amortized Analyses of Self-Organizing Sequential Search Heuristics*, *Comm. ACM*, **28**, pp. 404-411, 1985.

[2] S. Bernstein, *Sur une modification de l'inequalite de Tchebichef*, Annals Science Institute Sav. Ukraine, Sect. Math. I, 1924 (Russian, French summary).

[3] H. Chernoff, *A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations*, Annals of Math. Stat., **23**, 1952, 493–507.

[4] G. Bennett, *Probability Inequalities for the Sum of Independent Random Variables*, J. Am. Stat. Ass., **57**, 1962, 33–45.

[5] Hester J. H., Hirschberg D. S., *Self-organizing Linear Search*, *ACM Comput. Surveys*, **17**, pp. 295-312, 1985.

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

[7] Hofri M., Shachnai H., *Self-Organizing Lists and Independent References - a Statistical Synergy*, *Jour. of Alg.*, **12**, pp. 533-555, 1991.

[8] McCabe J., *On Serial Files with Relocatable Records*, *Operations Res.*, **13**, pp. 609-618, 1965.

[9] T. Hagerup and C. Rüb, *A Guided Tour of Chernoff Bounds*, Inf. Proc. Lett., **33**, 1990, 305–308.

[10] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer, April 1998 (to appear).

[11] M.G. Krein and A.A. Nudelman, *The Markov Moment Problem and Extremal Problems*, Translations of Math. Monographs (From Russian), Vol. 50, 1977, American Mathematical Society.

[12] W. Feller, *An Introduction to Probability Theory and its Applications*, John Wiley and Sons, 1957.