Exponentially Small Bounds on the Expected Optimum of the Partition and Subset Sum Problems*

George S. Lueker
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425

1. Introduction

Define the *partition problem* as follows. Given n numbers $x_1, x_2, \ldots, x_n$, find values for $\gamma_i \in \{-1, 1\}$ so as to minimize

$$\Bigl|\sum_{i=1}^{n} \gamma_i x_i\Bigr|. \tag{1.1}$$

Also define a related problem called the *subset sum* problem; here we are given a target value t and asked to choose $\delta_i \in \{0, 1\}$ to minimize

$$\Bigl|t - \sum_{i=1}^{n} \delta_i x_i\Bigr|. \tag{1.2}$$

Determining whether the minimum achievable in (1.1), or in (1.2), is 0 is NP-complete, and thus either minimization problem is NP-hard (see [GJ79, Karp72]).

In this paper we are interested in behavior of this problem when the x_i are i.i.d. random variables. Under fairly general conditions, the median of the solution for the subset sum problem has been shown to be exponentially small when t is near $\mathsf{E}\left[\sum_{i=1}^{n} x_i\right]$ [Luek82]; this result has found application in the probabilistic analysis of approximation algorithms for the 0-1 Knapsack problem [Luek82, GMS84]. The median solution to the partition problem is known to be exponentially small [KKLO86] under fairly general conditions; this paper commented "a significant question which our results leave open is the *expected* value of the difference for the best partition" [KKLO86, p. 643].

Under fairly general conditions on the distribution of the X_i, we show that the expected value of the solution to these problems is also exponentially small, i.e., of the form $O(e^{-cn})$, though we make no claim that we have the best value for the constant c. The proof method is in some ways similar to the argument in [PIA78]: we model the problem by a sequence of random variables and then apply a nonlinear transformation to make the sequence amenable to analysis by martingale theory.

We note that while the bounds developed in [KKLO86, Luek82] on the median are much more precise than those we show here on the expectation, the bounds in [KKLO86,

* The journal version of this has now appeared in *Random Structures and Algorithms* 12 (1998), pp. 51–62.

Luek82] are not strong enough to show that the expectation is exponentially small; see the first two paragraphs in [KKLO86, Section 4]. Moreover, the results of the present paper show that it is likely that for *every* value z in some interval, some partition difference or subset sum comes close to z (see Corollary 2.5 and the Corollaries of Section 3 for more precise statements).

The result in this paper is simply a statement of the behavior of the optimum; we do not know whether it can be achieved by a polynomial-time algorithm. We note that algorithms for the partition problem have achieved considerable attention; see [CL91] for details. In [KK82] the notion of differencing two variables is used. *Differencing* x and y means replacing them by their difference $|x-y|$; this simply corresponds to placing x and y on opposite sides of the partition. [KK82] showed that a fairly complicated algorithm based on this idea tended to achieve a difference of only $n^{-\Omega(\log n)}$. In [Yaki96] this same result was proven for a much simpler and more natural implementation of the differencing method.

2. The Expected Subset-Sum Solution

Assume that the X_i are uniformly distributed over $[-1,1]$. Also assume that some $\eta \in (0, \frac{1}{2})$ is specified. If A is some event, the *indicator* for A, written 1_A, is the random variable which is 1 if A holds and 0 otherwise. Let $\hat{f}_{k,\eta}(z)$, or more briefly $\hat{f}_k(z)$, be the indicator for the event

$$\exists \delta_i \in \{0,1\} \text{ such that } \left|\sum_{i=1}^{k} \delta_i X_i - z\right| \leq \eta. \tag{2.1}$$

Informally, $\hat{f}_k(z)$ tells us whether z can be approximated to within η by summing some subset of the first k variables. Note that $\hat{f}_0(z) = 1_{|z|\leq\eta}$, i.e., $\hat{f}_0(z)$ is simply 1 if $|z| \leq \eta$ and 0 otherwise. Also note that, letting $\vee$ denote the operator **or** as usually defined for 0-1 variables, we have for $0 < k < n$

$$\hat{f}_{k+1}(z) = \hat{f}_k(z) \vee \hat{f}_k(z - X_{k+1}) = \hat{f}_k(z) + \big(1 - \hat{f}_k(z)\big)\hat{f}_k(z - X_{k+1}). \tag{2.2}$$

For our analysis it will be useful to restrict the choices for the δ_i in (2.1). Say that a choice of values for $\delta_1, \delta_2, \ldots, \delta_k$ is *admissible* (for a given $z \in [-\frac{1}{2}, \frac{1}{2}]$) if

$$\forall k' \in \{1, \ldots, k\}\,,\; z - \sum_{i=k'}^{k} \delta_i X_i \in \left[-\tfrac{1}{2}, \tfrac{1}{2}\right].$$

Say that z has an *admissible η-approximation* if (2.1) holds even when we are only allowed to consider admissible choices for the δ_i. Define $f_k(z)$ to be the indicator for the event that z has an admissible η-approximation. Then as before we have

$$f_0(z) = 1_{|z|\leq\eta}, \tag{2.3}$$

and the recurrence (2.2) must be modified to

$$f_{k+1}(z) = f_k(z) + \big(1 - f_k(z)\big) 1_{z - X_{k+1} \in [-1/2, 1/2]} f_k(z - X_{k+1}). \tag{2.4}$$

Next define p_k to be the random variable (depending on $X_1, \dots, X_k$)

$$p_k = \int_{-1/2}^{1/2} f_k(z)\, dz. \tag{2.5}$$

Informally, this tells us the fraction of the interval $[-\frac{1}{2}, \frac{1}{2}]$ which has an admissible η-approximation; the essence of the proof is to study how this fraction grows as k increases. From (2.3) we have $p_0 = 2\eta$. Note also that we must have

$$p_{k+1} \le 2p_k, \tag{2.6}$$

since the fraction of the interval which is covered can at most double. Also, if $p_k < 1$ and we fix the value of $X_1, \dots, X_k$, then from (2.4), (2.5), and the fact that the density of X_{k+1} is $\frac{1}{2}$ over $[-1, 1]$, we can compute the following recurrence for the expected value of p_{k+1}:

$$\begin{aligned}
\mathsf{E}[p_{k+1}] &= \mathsf{E}\Big[\int_{-1/2}^{1/2} f_{k+1}(z)\, dz\Big] \\
&= \int_{-1/2}^{1/2} f_k(z)\, dz + \int_{-1/2}^{1/2} \big(1 - f_k(z)\big) \int_{-1}^{1} \tfrac{1}{2} 1_{z-x\in[-1/2,1/2]} f_k(z-x)\, dx\, dz \\
&= p_k + \int_{-1/2}^{1/2} \big(1 - f_k(z)\big) \int_{-1/2}^{1/2} \tfrac{1}{2} f_k(u)\, du\, dz \\
&= p_k + \tfrac{1}{2} \int_{-1/2}^{1/2} \big(1 - f_k(z)\big)\, dz \int_{-1/2}^{1/2} f_k(u)\, du \\
&= p_k + \tfrac{1}{2}(1 - p_k)p_k.
\end{aligned} \tag{2.7}$$

(Here $X_1, X_2, \dots, X_k$ are considered fixed, and the expectation is taken with respect to X_{k+1}.) Now for $k + 1 \in \{1, \dots, n\}$ let Z_{k+1} be the random variable defined by

$$Z_{k+1} = \begin{cases} \dfrac{p_{k+1} - p_k}{p_k(1 - p_k)} & \text{if } p_k < 1, \text{ and} \\ \frac{1}{2} & \text{if } p_k = 1. \end{cases} \tag{2.8}$$

From (2.7) we conclude that, regardless of the value of $X_1 \dots, X_k$, we have

$$\mathsf{E}[Z_{k+1}] = \tfrac{1}{2}. \tag{2.9}$$

Moreover, since using (2.6) we have $p_k \le p_{k+1} \le \min(2p_k, 1)$, one easily computes that

$$0 \le Z_{k+1} \le 2. \tag{2.10}$$

Thus the sequence $-k/2 + \sum_{i=1}^{k} Z_i$, for $k = 0, 1, \dots, n$ is a martingale so a standard application of a Hoeffding bound [Hoef63] yields

Lemma 2.1. *For $\alpha \le n/2$,*

$$\Pr\Big\{\sum_{i=1}^{n} Z_i \le \alpha\Big\} \le \exp\left(-\frac{(n/2-\alpha)^2}{2n}\right).$$

In order to monitor the evolution of the sequence p_k, it is useful to consider the function

$$\psi(p) = \lg p - \ln(1-p) + p/2, \tag{2.11}$$

so

$$\psi'(p) = \frac{\lg e}{p} + \frac{1}{1-p} + \frac{1}{2}, \tag{2.12}$$

(To avoid having to deal with special cases when the argument of ψ is 1, in the following we assume the following conventions: $\psi(1) = \infty$, $\infty \ge r$ and $\infty + r = \infty$ for all real r, and $\infty \ge \infty$. Also, we assume that division has precedence lower than multiplication, so that we can write, for example, $e^{n/2C}$ instead of the more cumbersome $e^{n/(2C)}$.)

Lemma 2.2. *For $p_k \in (0,1]$, we have*

$$\psi(p_{k+1}) \ge \psi(p_k) + Z_{k+1}. \tag{2.13}$$

Proof. If $p_{k+1} = 1$, then (2.13) holds since the left side is ∞. Also, if $Z_{k+1} = 0$, then $p_k = p_{k+1}$ and (2.13) holds trivially. Otherwise we need to show that

$$1 \le \frac{\psi(p_{k+1}) - \psi(p_k)}{Z_{k+1}} = \frac{\psi\big(p_k + Z_{k+1}p_k(1-p_k)\big) - \psi(p_k)}{Z_{k+1}}. \tag{2.14}$$

Consider several cases.

Case 1. $p_k \in [0, 1/4]$. Then since $p_k + Z_{k+1}p_k(1-p_k) = p_{k+1} \le 2p_k$ we have $Z_{k+1} \le 1/(1-p_k)$. Since ψ' is decreasing over $(0, \frac{1}{2})$, the right hand side of (2.14) is bounded below (see Appendix) by

$$\begin{aligned}\frac{\psi(2p_k) - \psi(p_k)}{1/(1-p_k)} &= (1-p_k)\left(1 + \ln\frac{1-p_k}{1-2p_k} + \frac{p_k}{2}\right) \\ &\ge (1-p_k)\left(1 + \frac{p_k}{1-p_k}\right) \\ &= 1.\end{aligned}$$

Case 2. $p_k \in (1/4, 1/2]$. Straightforward computation shows that ψ' has a minimum, over $(0,1)$, of

$$\big(1 + (\lg e)^{1/2}\big)^2 + \tfrac{1}{2} \ge \frac{16}{3}.$$

Hence the right-hand side of (2.14) is at least $\frac{16}{3}p_k(1-p_k)$, which is at least 1 for any $p_k \in (1/4, 1/2]$.

Case 3. $p_k \in (1/2, 1)$. Then the right-hand side of (2.14) is at least 1 since one easily sees that ψ' is bounded below by $1/\big(p_k(1-p_k)\big)$ over the interval $(p_k, 1)$. ■

Lemma 2.3. *If*

$$\sum_{i=1}^{n} Z_i \geq (1 + \lg e) \ln \eta^{-1} - \tfrac{1}{2}, \tag{2.15}$$

then every number $z \in [-\frac{1}{2}, \frac{1}{2}]$ has an admissible 2η-approximation.

Proof. First note that

$$\psi(p_0) = \psi(2\eta) = \lg(2\eta) - \ln(1 - 2\eta) + 2\eta/2 \geq 1 + \lg(\eta), \tag{2.16}$$

and

$$\psi(p_n) = \lg(p_n) - \ln(1 - p_n) + p_n/2 \leq -\ln(1 - p_n) + \tfrac{1}{2}$$

i.e.,

$$-\ln(1 - p_n) + \tfrac{1}{2} \geq \psi(p_n). \tag{2.17}$$

Using Lemma 2.2 and the assumption of this lemma we have

$$\psi(p_n) \geq \psi(p_0) + \sum_{i=1}^{n} Z_i \geq \psi(p_0) + (1 + \lg e) \ln \eta^{-1} - \tfrac{1}{2}. \tag{2.18}$$

Adding the left and right sides of (2.16), (2.17), and (2.18) gives

$$\psi(p_0) - \ln(1 - p_n) + \tfrac{1}{2} + \psi(p_n) \geq 1 + \lg(\eta) + \psi(p_n) + \psi(p_0) + (1 + \lg e) \ln \eta^{-1} - \tfrac{1}{2},$$

which simplifies to

$$-\ln(1 - p_n) \geq \lg(\eta) + (1 + \lg e) \ln \eta^{-1} = -\ln \eta,$$

implying $1 - p_n \leq \eta$. Thus the measure of the portion of $[-\frac{1}{2}, \frac{1}{2}]$ over which f_n is 0 is at most η. Hence each point z of the interval $[-\frac{1}{2}, \frac{1}{2}]$ either has $f_n(z) = 1$ or is within η of a point z' for which $f_n(z') = 1$. From the definition of f_n, this implies that each point in $[-\frac{1}{2}, \frac{1}{2}]$ has an admissible 2η-approximation. ■

Since we will frequently use the constant $1 + \lg e$, we will henceforth let C denote this constant. (The numerical value of C is approximately 2.442695.)

Theorem 2.4. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. uniform over $[-1, 1]$, and let $0 < \eta < \frac{1}{2}$. Suppose that $n/2 \geq C \ln \eta^{-1}$. Then, except with probability bounded by*

$$\exp\left(-\frac{\big(n/2 - C \ln \eta^{-1}\big)^2}{2n}\right),$$

all values in $[-\frac{1}{2}, \frac{1}{2}]$ have admissible 2η-approximations.

Proof. This follows immediately from Lemma 2.1 and Lemma 2.3. ■

By omitting the condition about admissibility, and noting that the theorem is trivial for $\eta > \frac{1}{2}$, we have

Corollary 2.5. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. uniform over $[-1, 1]$, and let $\eta \geq e^{-n/2C}$ be given. Then, except with probability bounded by*

$$\exp\left(-\frac{\left(n/2 - C \ln \eta^{-1}\right)^2}{2n}\right),$$

we have

$$\forall z \in [-\tfrac{1}{2}, \tfrac{1}{2}],\ \exists S \subseteq \{1, 2, \ldots, n\} \text{ such that } \left|z - \sum_{i \in S} X_i\right| \leq 2\eta.$$

Now, define the $[a, b]$-*subset-sum gap* of $X_1, X_2, \ldots, X_n$ to be the smallest value of 2η such that each $z \in [a, b]$ can be approximated to within 2η by summing some sublist of the X_i.

Theorem 2.6. *The expected value of the $[-\frac{1}{2}, \frac{1}{2}]$-subset-sum gap for n variables $X_1, X_2, \ldots, X_n$ distributed uniformly over $[-1, 1]$ is at most*

$$2e^{-n/2C}\left(1 + (2\pi n)^{1/2}\, C^{-1} e^{n/2C^2}\right) = \exp\left(-\tfrac{1}{2}\left(\frac{1}{C} - \frac{1}{C^2}\right)n + o(n)\right).$$

Proof. Let 2η be the random variable (depending on $X_1, X_2, \ldots, X_n$) giving the value of the $[-\frac{1}{2}, \frac{1}{2}]$-subset-sum gap, and define $\eta_0 = e^{-n/2C}$, i.e.,

$$\frac{n}{2} = C \ln \eta_0^{-1}. \tag{2.19}$$

Now using Corollary 2.5 we can write

$$\begin{aligned} \mathsf{E}[\eta] &= \int_0^\infty \Pr\{\eta \geq z\}\, dz \\ &\leq \eta_0 + \int_{\eta_0}^\infty \Pr\{\eta \geq z\}\, dz \\ &\leq \eta_0 + \int_{\eta_0}^\infty e^{-(n/2 - C \ln z^{-1})^2/2n}\, dz. \end{aligned} \tag{2.20}$$

To evaluate the integral on the right side we make the substitution $z = \eta_0 u$ to obtain

$$\begin{aligned} \int_{\eta_0}^\infty e^{-(n/2 - C \ln z^{-1})^2/2n}\, dz &= \int_1^\infty e^{-\left(n/2 - C \ln(\eta_0 u)^{-1}\right)^2/2n} \eta_0\, du \\ &= \eta_0 \int_1^\infty e^{-(n/2 - C \ln \eta_0^{-1} - C \ln u^{-1})^2/2n}\, du \\ &= \eta_0 \int_1^\infty e^{-(C \ln u)^2/2n}\, du \\ &\qquad \text{by (2.19)} \\ &\leq \eta_0 \,(2\pi n)^{1/2}\, C^{-1} e^{n/2C^2}. \end{aligned} \tag{2.21}$$

(See Appendix.) Substituting (2.21) into (2.20) results in the bound on $\mathsf{E}[2\eta]$ appearing in the Theorem. ■

3. Generalizations

Note that the results of the previous section say not only that a particular $z \in [-\frac{1}{2}, \frac{1}{2}]$ is likely to be near some subset sum of $X_1, X_2, \ldots, X_n$, but in fact that it is likely that for *all* $z \in [-\frac{1}{2}, \frac{1}{2}]$ some subset sum of $X_1, X_2, \ldots, X_n$ is near z. This makes it easy to prove a variety of corollaries showing that related quantities have exponentially small expectation.

First we note that we can easily expand the range of values having good approximations to an interval much larger than $[-\frac{1}{2}, \frac{1}{2}]$.

Corollary 3.1. *Given any $\xi > 0$, there exists a $c > 0$ such that the expected value of the $\big[-(1-\xi)n/4, (1-\xi)n/4\big]$-subset-sum gap for n variables $X_1, X_2, \ldots, X_n$ distributed uniformly over $[-1, 1]$ is $O\big(e^{-cn}\big)$.*

Proof. Let $\xi' = \xi/2$ and consider two subsets of the random variables:

$$A = \left\{X_1, X_2, \ldots, X_{\lceil \xi' n \rceil}\right\} \quad \text{and} \quad B = \left\{X_{\lceil \xi' n \rceil + 1}, X_{\lceil \xi' n \rceil + 2}, \ldots, X_n\right\}.$$

Let ϵ be the $[-\frac{1}{2}, \frac{1}{2}]$-subset-sum gap of A; by Theorem 2.6 we know that $\mathsf{E}[\epsilon]$ is exponentially small. By a straightforward application of a Hoeffding bound, we can establish that, except with exponentially small probability, the lowest subset sum achievable from B is less than $-(1-\xi)n/4$ and the highest subset sum achievable from B is at least $(1-\xi)n/4$. But since the range of the X_i is $[-1, 1]$, if we look at all subset sums achievable from B in sorted order, they cannot be more than a distance of 1 apart. Thus, except with exponentially small probability, we can approximate any $z \in [-(1-\xi)n/4, (1-\xi)n/4]$ to within $\frac{1}{2}$ from B, and then to within ϵ by fine-tuning the approximation using elements of A. ■

Note that the constant c may become quite small as ξ approaches 0. Also note that one could not hope to improve the range of approximable numbers substantially, since the expected sum of all of the positive (resp. negative) X_i is $n/4$ (resp. $-n/4$).

Now define the $[a, b]$-*partition gap* of $X_1, X_2, \ldots, X_n$ to be the smallest value of 2η such that each $z \in [a, b]$ can be approximated to within 2η by a sum of the form

$$\sum_{i=1}^{n} \gamma_i X_i \quad \text{for} \quad \gamma_i \in \{-1, 1\}. \tag{3.1}$$

Corollary 3.2. *Given any $\xi > 0$, there exists a $c > 0$ such that the expected value of the $[-(1-\xi)n/2, (1-\xi)n/2]$-partition gap for n variables $X_1, X_2, \ldots, X_n$ distributed uniformly over $[-1, 1]$ is at most $O(e^{-cn})$.*

Proof. Let $\xi' = \xi/3$ and consider two subsets of the random variables:

$$A = \left\{X_1, X_2, \ldots, X_{\lceil \xi' n \rceil}\right\} \quad \text{and} \quad B = \left\{X_{\lceil \xi' n \rceil + 1}, X_{\lceil \xi' n \rceil + 2}, \ldots, X_n\right\}.$$

Let ϵ be the $[-\frac{1}{2}, \frac{1}{2}]$-subset-sum gap of A; by Theorem 2.6 we know that $\mathsf{E}[\epsilon]$ is exponentially small. By setting

$$\gamma_i = \begin{cases} 1 & \text{for } X_i \geq 0 \\ -1 & \text{for } X_i < 0 \end{cases}$$

and using a Hoeffding bound, we can establish that, except with exponentially small probability, the highest partition difference achievable from B is at least $(1-2\xi')n/2$; similarly, except with exponentially small probability the lowest (signed) partition difference achievable from B is less than $-(1-2\xi')n/2$. But since the range of the X_i is $[-1,1]$, if we look at all partition differences achievable from B in sorted order, they cannot be more than a distance of 2 apart. Thus, except with exponentially small probability, we can approximate any $z \in [-(1-2\xi')n/2, (1-2\xi')n/2]$ to within 1 from B. Except with exponentially small probability we also have

$$\left|\sum_{i\in A} X_i\right| \leq \xi' n/2,$$

in which case we can also approximate any $z \in [-(1-3\xi')n/2, (1-3\xi')n/2] = [-(1-\xi)n/2, (1-\xi)n/2]$ to within 1 by selecting values for γ_i (for $i \in B$) by a sum of the form

$$\sum_{i\in B} \gamma_i X_i - \sum_{i\in A} X_i.$$

Assume that we now fix z and the corresponding values of γ_i for $i \in B$, and let

$$z' = z - \sum_{i\in B} \gamma_i X_i + \sum_{i\in A} X_i \in [-1,1]. \tag{3.2}$$

Since A has a $[-\frac{1}{2}, \frac{1}{2}]$-subset-sum gap of ϵ, and $|z'| \leq 1$, we can choose values for $\delta_i \in \{0,1\}$ (for $i \in A$) so that

$$\left|z' - 2\sum_{i\in A} \delta_i X_i\right| \leq 2\epsilon.$$

Letting $\gamma_i = 2\delta_i - 1$ (for $i \in A$), this means there are $\gamma_i \in \{-1,1\}$ (for $i \in A$) such that

$$\left|z' - \sum_{i\in A} (\gamma_i + 1) X_i\right| \leq 2\epsilon. \tag{3.3}$$

Substituting in (3.2) into (3.3) gives

$$2\epsilon \geq \left|z - \sum_{i\in B} \gamma_i X_i + \sum_{i\in A} X_i - \sum_{i\in A} (\gamma_i+1) X_i\right| = \left|z - \sum_{i\in B} \gamma_i X_i - \sum_{i\in A} \gamma_i X_i\right| = \left|z - \sum_{i=1}^{n} \gamma_i X_i\right|,$$

giving us the desired approximation for z. ■

These results can easily be generalized to a much larger class of distributions. Let $U(a,b)$ denote the uniform distribution over $[a,b]$. Say that a distribution G *contains some uniform distribution* if there exists a distribution G_1 and constants $\alpha \in (0,1]$, c, and $h > 0$ such that

$$G = (1-\alpha)G_1 + \alpha U(c-h, c+h).$$

If in particular $c = 0$, say the distribution *contains some uniform distribution centered at 0*.

Corollary 3.3. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. bounded random variables. Suppose that the distribution of X_1 contains some uniform distribution. Let*

$$\mu_- = \mathsf{E}\left[1_{X\leq 0}X\right], \quad \mu_+ = \mathsf{E}\left[1_{X>0}X\right], \quad \text{and} \quad \mu_{\text{abs}} = \mathsf{E}\left[|X|\right] = \mu_+ - \mu_-.$$

(Note that $\mu_- \leq 0$.) Finally, choose any $\xi > 0$. Then both the expected value of the $\left[(\mu_- + \xi)n, (\mu_+ - \xi)n\right]$-subset-sum gap and the expected value of the $\left[(-\mu_{\text{abs}} + \xi)n, (\mu_{\text{abs}} - \xi)n\right]$-partition gap for $X_1, X_2, \ldots, X_n$ are exponentially small.

Proof. First consider the partition gap. Let the support of X_1 be contained in $[-d, d]$, and let $\xi' = \xi/2d$. Partition the variables into two sets

$$A = \left\{X_1, X_2, \ldots, X_{2\lceil\xi' n/2\rceil}\right\} \quad \text{and} \quad B = \left\{X_{2\lceil\xi' n/2\rceil+1}, X_{2\lceil\xi' n/2\rceil+2}, \ldots, X_n\right\}.$$

First consider the variables in A. Recalling that the distribution of these variables contains some uniform distribution, by definition we can find constants $\alpha > 0$, c, and $h > 0$, and a distribution G_1 such that the variables in A can be considered to have been generated as follows: flip a biased coin which comes up heads with probability α. If it comes up heads, return a uniform draw from $[c - h, c + h]$; if it comes up tails, return a value chosen according to the distribution G_1. Partition A as $A_u \cup A_G$, where the variables in A_u correspond to heads and those in A_G correspond to tails. Then by a Hoeffding bound, except with exponentially small probability, we have

$$|A_u| \geq \alpha\xi' n/2 + 1. \tag{3.4}$$

If $|A_u|$ is odd, move the last variable from A_u to A_G, so that $|A_u|$ becomes even.

Finally consider the variables in A_u, which by (3.4) we may index as $X_1, X_2, \ldots, X_{2k}$ with $2k \geq \alpha\xi' n/2$. As in [Tsai92], we first perform a preprocessing step in which we difference these in pairs to obtain

$$X_1 - X_2,\ X_3 - X_4,\ X_5 - X_6,\ \ldots,\ X_{2k-1} - X_{2k}. \tag{3.5}$$

This corresponds to deciding that the differenced variables in each pair will appear on opposite sides of the partition. Note that each of these differences has a triangular distribution centered at 0. By a resampling argument like that in [KK82], we can partition these differences into two sets D_u and D_o, such that the variables in D_u have a uniform distribution, and (except with exponentially small probability, by a Hoeffding bound) $|D_u| \geq k/3 = \Theta(n)$. By Corollary 3.2, the $[-2d, 2d]$-partition gap of the values in D_u, say ϵ, has an exponentially small expectation.

By another application of the Hoeffding bound, we can conclude that, except with exponentially small probability, the sum of the absolute values of the variables in set B is at least $(\mu_{\text{abs}} - \xi)n$. If so, then since all values in $B \cup A_G \cup D_o$ lie in $[-2d, 2d]$, the $[-(\mu_{\text{abs}} - \xi)n, (\mu_{\text{abs}} - \xi)n]$-partition gap of $B \cup A_G \cup D_o$ must be at most $2d$.

Thus, much as before, we can approximate any desired value in $[-(\mu_{\text{abs}} - \xi)n, (\mu_{\text{abs}} - \xi)n]$ to within $2d$ using variables in $B \cup A_G \cup D_o$, and then to within ϵ using the variables in D_u.

A fairly similar proof holds for the case of the subset-sum gap. This time, for the variables in A_u, for each i we include X_{2i-1} in the sum and then decide whether or not to include $X_{2i} - X_{2i-1}$; this is equivalent to deciding whether to include X_{2i-1} or X_{2i} in the sum. We omit the details. ■

Acknowledgements

It is a pleasure to thank Ed Coffman, Brad Hutchings, David Johnson, Andy Odlyzko, and the referees for comments on an earlier manuscript and/or comments on the related literature.

References

[CL91] Jr. Coffman, E. G. and George S. Lueker. *Probabilistic Analysis of Packing and Partitioning Algorithms.* Wiley Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, New York, NY, 1991.

[GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, New York, 1979.

[GMS84] A. V. Goldberg and A. Marchetti-Spaccamela. On finding the exact solution of a zero-one knapsack problem. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, pages 359–368, May 1984.

[Hoef63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[Karp72] Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, New York, 1972.

[KK82] Narendra Karmarkar and Richard M. Karp. The differencing method of set partitioning. Technical Report UCB/CSD 82/113, Computer Science Division (EECS), University of California, Berkeley, December 1982.

[KKLO86] Narendra Karmarkar, Richard M. Karp, George S. Lueker, and Andrew M. Odlyzko. Probabilistic analysis of optimum partitioning. *Journal of Applied Probability*, 23(3):626–645, 1986.

[Luek82] G. S. Lueker. On the average difference between the solutions to linear and integer knapsack problems. In Ralph L. Disney and Teunis J. Ott, editors, *Applied Probability—Computer Science, The Interface*, volume I, pages 489–504. Birkhäuser, Boston, 1982.

[PIA78] Yehoshua Perl, Alon Itai, and Haim Avni. Interpolation search—a $\log\log n$ search. *Communications of the ACM*, 21(7):550–553, July 1978.

[Tsai92] Li-Hui Tsai. Asymptotic analysis of an algorithm for balanced parallel processor scheduling. *SIAM Journal on Computing*, 21(1):59–64, February 1992.

[Yaki96] Benjamin Yakir. The differencing algorithm LDM for partitioning: A proof of Karp's conjecture. *Mathematics of Operations Research*, 21(1):85–99, February 1996.

A. Appendix

This appendix gives the details of a few omitted computations, in the hope that this may save the interested reader time.

For verifying Case 1 of Lemma 2.2, we use the following simple observation, letting $f = \psi$, $x = p_k$, $u = Z_{k+1}(1 - p_k)p_k$, and $u_0 = 2p_k$.

Observation A.1. *Suppose that f'' exists and is negative over $[x, x + u_0]$. Then for any u with $0 < u \leq u_0$, we have*

$$\frac{f(x+u) - f(x)}{u} \geq \frac{f(x+u_0) - f(x)}{u_0}.$$

Then we use the fact that for $0 \leq x < 1/2$, we have

$$\ln \frac{1-x}{1-2x} = \int_{1-2x}^{1-x} \frac{1}{z}\, dz \geq \int_{1-2x}^{1-x} \frac{1}{1-x}\, dz = \frac{x}{1-x},$$

letting $x = p_k$.

For verifying (2.21), note that if we let $x = e^u$, so $dx = e^u du$, then

$$\int_1^\infty e^{-a \ln^2 x} dx = \int_0^\infty e^{-au^2} e^u du \leq \int_{-\infty}^\infty e^{-au^2+u} du = (\pi/a)^{1/2}\ e^{1/4a}.$$