Electronic Colloquium on Computational Complexity, Report No. 1 (2005)

Near optimality of the priority sampling procedure

Mario Szegedy*
Rutgers University,
email: szegedy@cs.rutgers.edu;

December 31, 2004

Abstract

Based on experimental results N. Duffield, C. Lund and M. Thorup [DLT2] conjectured that the variance of their highly successful priority sampling procedure is not larger than the variance of the threshold sampling procedure with sample size one smaller. The conjecture's significance is that the latter procedure is provably optimal among all *off-line* sampling procedures. Here we prove this conjecture. In particular, our result gives an affirmative answer to the conjecture of N. Alon, N. Duffield, C. Lund and M. Thorup [ADLT], which states that the standard deviation for the subset sum estimator obtained from k priority samples is upper bounded by $W/\sqrt{k-1}$. (W is the actual subset sum.)

1 Notations

Let $w_1 \geq w_2 \geq \ldots \leq w_n$ be nonnegative weights. Let $w_1 > 0$. We define the polynomial

$$P(\beta) = (1 - w_1\beta)(1 - w_2\beta)\cdots(1 - w_n\beta) = \prod_{i=1}^{n}(1 - w_i\beta).$$

For every integer $0 \leq l \leq n$ we also define the polynomial

$$P_l(\beta) = (1 - w_{l+1}\beta)(1 - w_{l+2}\beta)\cdots(1 - w_n\beta) = \prod_{i=l+1}^{n}(1 - w_i\beta).$$

Thus $P_0 = P$. For a function f and a non-negative integer t we denote the t^{th} derivative of f by $f^{(t)}$. For $t = 1$ we also use the usual f' notation. Let

$$N_l = \{l + 1, l + 2, \ldots, n\}.$$

We have:

$$P_l^{(t)} = (-1)^t t! \sum_{S \subseteq N_l;\, |S|=t} \prod_{i \in S} w_i \prod_{i \in N_l \setminus S} (1 - w_i\beta_i) \tag{1}$$

A sum with higher upper index than lower index, such as $\sum_1^0 exp$ is considered an empty sum with value 0.

*This work was supported by NSF grant 0105692.

ISSN 1433-8092

2 The Priority Sampling

Fix a sample size $k < n$. The priority sampling scheme of N. Duffield, C. Lund and M. Thorup is as follows: Select $\alpha_1, \ldots, \alpha_n$ independently, randomly and uniformly from $(0, 1]$. For $1 \leq i \leq n$ we define priority $q_i = w_i \alpha_i^{-1}$. Let q be the $(k+1)^{\text{st}}$ largest priority (a probability variable, depending on the α_is) and define probability variables $\hat{w}_i = \max\{w_i, q\}\chi_{w_i\alpha_i > q}$, where $\chi_{w_i\alpha_i>q}$ is the indicator function of the event that $w_i\alpha_i > q$. This scheme finds its ancestors in [C, AGPR, AFT, DLT1, GM, HHW, OR, PKC]. For a complete hystory see [DLT2]. N. Duffield, C. Lund and M. Thorup have proved that

1. $E(\hat{w}_i) = w_i$;
2. $\hat{w}_i$ and $\hat{w}_j$ are independent.

N. Alon, N. Duffield, C. Lund and M. Thorup in [ADLT] show that the variance of $\sum_{i=1}^n \hat{w}_i$ is $O(W^2/k)$ and state as a major challenge to replace the big order with $W^2/(k-1)$. Our first theorem verifies this conjecture:

Theorem 1. *$Var\left(\sum_{i=1}^n \hat{w}_i\right) \leq W^2/(k-1)$, where $W = \sum_{i=1}^n w_i$.*

3 The threshold sampling

Threshold sampling is a simple sampling procedure that estimates $\sum_{i=1}^n w_i$ and has the least variance among all sampling procedures for which the expected sample size is the same.

Fix a threshold, τ and define $\tilde{w}_i = w_i$ if $w_i > \tau$. Otherwise select $\alpha_i \in (0, 1]$ randomly and uniformly and independently from other α_is, and set $\tilde{w}_i = \tau$ if $w_i\alpha_i^{-1} > \tau$, else set $\tilde{w}_i = 0$. Like in the case of the priority sampling w_i is considered to be in the sample if $\tilde{w}_i > 0$. Since $\tilde{w}_i$s are independent random variables with

$$\tilde{w}_i = \begin{cases} \tau & \text{with probability } w_i/\tau; \\ 0 & \text{with probability } 1 - w_i/\tau; \end{cases} \qquad (\text{for } i: \quad w_i \leq \tau),$$

it follows that the expected sample size is

$$|\{i | w_i > \tau\}| + \frac{1}{\tau} \sum_{i: w_i \leq \tau} w_i.$$

In order to compare threshold sampling with priority sampling we need to set the expected sample size (nearly) the same. First observe that for a fixed $k > 0$ there is a unique threshold τ_k such that the expected sample size, when the threshold is set to τ is exactly k. Indeed, it is not hard to show that as τ descends from infinity to w_n the expected sample size monotonly and continuously increases from zero to n, and that the increase is strict. In particular, we have:

Lemma 1. *For $1 \leq l \leq n$ let $s_l = \frac{\sum_{i=l}^n w_i}{w_l} + l - 1$. Then for $1 \leq l \leq n-1$ it holds that $s_l \leq s_{l+1}$.*

This lemma can be shown by direct calculation or by simply alluding to the fact that s_l is the expected sample size when the threshold is set to w_l.

N. Duffield, C. Lund and M. Thorup have conjectured that the variance of the k-sample priority sampling procedure is smaller or equal than the variance of the threshold sampling procedure with expected sample size $k-1$. Our main result is to prove this conjecture:

Theorem 2. *$Var(\sum_{i=1}^n \hat{w}_i) \leq Var(\sum_{i=1}^n \tilde{w}_i)$, where sample size for the priority sampling is k, and the threshold for threshold sampling is τ_{k-1}.*

As a result the priority sampling procedure is almost optimal for sample size k while having many advantageous properties the threshold sampling does not have.

Definition 1. *For the rest of the paper let ℓ be the smallest index for which $\tau_{k-1} \geq w_\ell$. Alternatively, ℓ is the smallest integer for which s_ℓ of Lemma 1 is greater or equal than $k-1$.*

From now on we shall be concerned with the case when $\tau = \tau_{k-1}$. Let us denote $\sum_{i=\ell}^n w_i$ by W'. From

$$k - 1 = \text{ expected sample size } = \ell - 1 + \sum_{i=\ell}^{n} w_i/\tau_{k-1} = \ell - 1 + W'/\tau_{k-1},$$

we obtain:

$$\tau_{k-1} = \frac{W'}{k-\ell}, \tag{2}$$

$$Var\left(\sum_{i=1}^n \tilde{w}_i\right) = Var\left(\sum_{i=\ell}^n \tilde{w}_i\right) = \sum_{i=\ell}^n w_i(\tau_{k-1} - w_i) = \frac{(W')^2}{k-\ell} - \sum_{i=\ell}^n w_i^2. \tag{3}$$

4 The Variance of the priority sampling

Because of 1. and 2. of section 2 we have that

$$Var\left(\sum_{i=1}^n \hat{w}_i\right) = \sum_{i=1}^n E\left(\hat{w}_i{}^2\right) - \sum_{i=1}^n w_i^2. \tag{4}$$

Our efforts will mainly go to express $\sum_{i=1}^n E\left(\hat{w}_i{}^2\right)$. For the sake of convenience we introduce the convention $w_0 = \infty$, $w_0^{-1} = 0$. For $0 \leq l \leq k$ let χ_l be the indicator function of the event that $w_l > q \geq w_{l+1}$. Since q is always at least w_{k+1}, we have $1 = \sum_{l=0}^k \chi_l$ (here 1 means the characteristic function of the entire event space, i.e. the identically 1 function). For $0 \leq l \leq k$ define:

$$A_l = E\left(\chi_l \sum_{i=1}^n \hat{w}_i{}^2\right).$$

The first term in (4) is exactly $\sum_{l=0}^k A_l$. To express A_l we further decompose the event $w_l > q \geq w_{l+1}$. Let $l < j \leq n$ be an index and let $S \subseteq N_l$ be a set of size $k-l$ that does not contain j. Define $\chi_{l,j,S}$ as the characteristic function of the event

$$(q = w_j\alpha_j^{-1}) \ \wedge \ (w_l > q \geq w_{l+1}) \ \wedge \bigwedge_{i \in S} (w_i\alpha_i^{-1} \geq q).$$

For the above event the sampled elements are exactly those that have indices from $S \cup \{1, \ldots, l\}$. For $i \in S$ we have $\hat{w}_i = q$. For $i \in \{1, \ldots, l\}$ we have $\hat{w}_i = w_i$. For the other elements $\hat{w}_i = 0$, hence they are not in the sample. The probability that $w_i\alpha_i^{-1} \geq w_j\alpha_j^{-1}$ (for fixed α_j) is $w_i w_j^{-1}\alpha_j$. The probability for $w_i\alpha_i^{-1} < w_j\alpha_j^{-1}$ is $1 - w_i w_j^{-1}\alpha_j$. Therefore

$$E\left(\chi_{l,j,S}\sum_{i=1}^{n}\hat{w}_i{}^2\right) =$$

$$\int_{w_j w_l^{-1}\le\alpha_j<w_j w_{l+1}^{-1}}\left(\sum_{1\le s\le l}w_s^2+(k-l)(w_j\alpha_j^{-1})^2\right)\prod_{i\in S}(w_i w_j^{-1}\alpha_j)\prod_{i\in N_l\setminus S;i\neq j}(1-w_i w_j^{-1}\alpha_j)\,d\alpha_j$$

Replace parameter α_j in the integral with $\beta = w_j^{-1}\alpha_j$ and rewrite the above integral as

$$w_j\int_{w_l^{-1}\le\beta<w_{l+1}^{-1}}\left(\sum_{1\le s\le l}w_s^2+(k-l)\beta^{-2}\right)\prod_{i\in S}w_i\beta\prod_{i\in N_l\setminus(S\cup\{j\})}(1-w_i\beta)\,d\beta \;=$$

$$\int_{w_l^{-1}\le\beta<w_{l+1}^{-1}}\left(\sum_{1\le s\le l}w_s^2+(k-l)\beta^{-2}\right)\beta^{k-l}\prod_{i\in S\cup\{j\}}w_i\prod_{i\in N_l\setminus(S\cup\{j\})}(1-w_i\beta)\,d\beta$$

Since $\chi_l = \sum_{j=l+1}^{n}\sum_{S\subseteq N_l;\,|S|=l;\,j\notin S}\chi_{l,j,S}$, we can write:

$$\begin{aligned}
A_l &= \sum_{S\subseteq N_l;\,|S|=k-l;\,j\notin S} E\left(\chi_{l,j,S}\sum_{i=1}^{n}\hat{w}_i{}^2\right)\\
&= \sum_{S\subseteq N_l;\,|S|=k-l;\,j\notin S}\int_{w_l^{-1}}^{w_{l+1}^{-1}}\left(\sum_{1\le s\le l}w_s^2+(k-l)\beta^{-2}\right)\beta^{k-l}\prod_{i\in S\cup\{j\}}w_i\prod_{i\in N_l\setminus(S\cup\{j\})}(1-w_i\beta)\,d\beta\\
&= \frac{(-1)^{k-l+1}}{(k-l)!}\int_{w_l^{-1}}^{w_{l+1}^{-1}}\left(\sum_{1\le s\le l}w_s^2+(k-l)\beta^{-2}\right)\beta^{k-l}P_l^{(k-l+1)}d\beta.
\end{aligned}$$

For $0\le l\le k$ we can write $A_l = B_l + C_l$, where $B_0 = C_k = 0$ and

$$\begin{aligned}
B_l &= \frac{(-1)^{k-l+1}}{(k-l)!}\sum_{1\le s\le l}w_s^2\int_{w_l^{-1}}^{w_{l+1}^{-1}}\beta^{k-l}P_l^{(k-l+1)}d\beta \quad (l\neq 0);\\
C_l &= \frac{(-1)^{k-l+1}}{(k-l-1)!}\int_{w_l^{-1}}^{w_{l+1}^{-1}}\beta^{k-l-2}P_l^{(k-l+1)}d\beta. \quad (l\neq k).
\end{aligned}$$

5 Integration

In this section we compute the primitive functions of the integrals of the previous section. By repeated integration by part we get:

Lemma 2. *Let t and s be non-negative integers, $s\ge t+1$. Let f be a function such that $f^{(1)} = f', f^{(2)},\ldots,f^{(s)}$ exist. Then*

$$\int x^t f^{(s)} = \sum_{r=0}^{t}(-1)^{t+r}\frac{t!}{r!}\beta^r f^{(s-1-t+r)}.$$

From Lemma 2 we obtain:

$$\int \beta^{k-l} P_l^{(k-l+1)} d\beta = \sum_{r=0}^{k-l} (-1)^{k+l+r} \frac{(k-l)!}{r!} \beta^r P_l^{(r)}; \tag{5}$$

$$\int \beta^{k-l-2} P_l^{(k-l+1)} d\beta = \sum_{r=0}^{k-l-2} (-1)^{k+l+r} \frac{(k-l-2)!}{r!} \beta^r P_l^{(r+2)} \qquad (\text{for } l \leq k-2). \tag{6}$$

We are left to compute the integral associated with C_{k-1}. Notice that this is the only integral among the C_ls in which β has a negative exponent. We use the estimate:

Lemma 3.

$$C_{k-1} \leq w_{k-1} \int_{w_{k-1}^{-1}}^{w_k^{-1}} P_{k-1}^{(2)} d\beta = w_{k-1} \left(P'_{k-1}(w_k^{-1}) - P'_{k-1}(w_{k-1}^{-1}) \right).$$

6 Properties of P_l and its derivatives

Let $0 \leq l \leq k$, $1 \leq r$ fixed.

Lemma 4. *P_l and its higher derivatives are continuous. Moreover, $(-1)^r P_l(\beta)^{(r)} > 0$ on $(0, w_{l+1}^{-1})$.*

The lemma easily follows from (1). From (1) we also obtain:

Lemma 5. *For $0 \leq l \leq k$ and $1 \leq r$ we have:*

$$P_l^{(0)}(w_{l+1}^{-1}) = 0 \tag{7}$$

$$P_l^{(r)}(w_{l+1}^{-1}) = -r w_{l+1} P_{l+1}^{(r-1)}(w_{l+1}^{-1}) \tag{8}$$

Definition 2. *For $0 \leq l \leq k$ and $0 \leq r$ we define the constants:*

$$p_{l,r} = \frac{(-1)^r}{r!} w_l^{-r} P_l^{(r)}(w_l^{-1}).$$

By Lemma 5 we have:

$$\frac{(-1)^r}{r!} w_{l+1}^{-r} P_l^{(r)}(w_{l+1}^{-1}) = p_{l+1,r-1}. \tag{9}$$

7 An interpretation of $p_{l,r}$

Fix l. In this section we reinterpret $p_{l,0}, p_{l,1}, p_{l,2}, \ldots$ as a probability distribution related to an independent sequence of Bernouli trials with different biases. Indeed, let $X_{l+1}, \ldots, X_n$ be independent zero-one valued random variables such that $Prob(X_i = 1) = w_i/w_l$ and $Prob(X_i = 0) = 1 - w_i/w_l$. Then

$$Prob(\sum_{i=l+1}^{n} X_i = r) = \sum_{|S|=r} \prod_{i \in S} \frac{w_i}{w_l} \prod_{i \notin S} (1 - w_i w_l^{-1}) = \frac{(-1)^r}{r!} w_l^{-r} P_l^{(r)}(w_l^{-1}).$$

The generator function for $\sum_{i=l+1}^{n} X_i$ is

$$G(\lambda) = \prod_{i=l+1}^{n} (1 - \frac{w_i}{w_l} + \lambda \frac{w_i}{w_l}) \ \ (= P_l(\frac{1-\lambda}{w_l})).$$

Also, $G(\lambda) = \sum_{r=0}^{\infty} p(l,r)\lambda^r$. Thus we have:

Lemma 6.

$$\begin{aligned} \sum_{r=1}^{\infty} r p_{l,r} &= G'(1) = \sum_{i=l+1}^{n} \frac{w_i}{w_l}; \\ \sum_{r=2}^{\infty} r(r-1) p_{l,r} &= G''(1) = \sum_{l+1 \le i \ne j \le n} \frac{w_i w_j}{w_l^2}; \end{aligned}$$

8 Summing it up

Recall that for $0 \le l \le k-1$:

$$C_l = \frac{(-1)^{k-l+1}}{(k-l-1)!} \int_{w_l^{-1}}^{w_{l+1}^{-1}} \beta^{k-l-2} P_l^{(k-l+1)} d\beta. \quad (l \ne k).$$

and $C_k = 0$. Let $l \le k-2$. From (2) we can express

$$C_l = \frac{(-1)^{k-l+1}}{(k-l-1)!} \left[\sum_{r=0}^{k-l-2} (-1)^{r+1} \frac{(k-l-2)!}{r!} \beta^r P_l^{(r+2)} \right]_{w_l^{-1}}^{w_{l+1}^{-1}}$$

We simplify the above expression as

$$C_l = -\left[\frac{1}{(k-l-1)} \sum_{r=0}^{k-l-2} \frac{(-1)^r}{r!} \beta^r P_l^{(r+2)} \right]_{w_l^{-1}}^{w_{l+1}^{-1}}.$$

Thus $C_l = D_l - E_l$, where

$$\begin{aligned} D_l &= \frac{1}{k-l-1} \sum_{r=0}^{k-l-2} \frac{(-1)^r}{r!} w_l^{-r} P_l^{(r+2)}(w_l^{-1}) = w_l^2 \sum_{r=0}^{k-l-2} \frac{(r+1)(r+2)}{k-l-1} p_{l,r+2}; \\ E_l &= \frac{1}{k-l-1} \sum_{r=0}^{k-l-2} \frac{(-1)^r}{r!} w_{l+1}^{-r} P_l^{(r+2)}(w_{l+1}^{-1}) = w_{l+1}^2 \sum_{r=0}^{k-l-2} \frac{(r+1)(r+2)}{k-l-1} p_{l+1,r+1}. \end{aligned}$$

By renaming the running indices in the above expressions we can write:

$$D_l = w_l^2 \sum_{r=2}^{k-l} \frac{(r-1)r}{k-l-1} p_{l,r} = w_l^2 \sum_{r=0}^{k-l} \frac{(r-1)r}{k-l-1} p_{l,r}; \quad (10)$$

$$E_l = w_{l+1}^2 \sum_{r=1}^{k-l-1} \frac{r(r+1)}{k-l-1} p_{l+1,r} = w_{l+1}^2 \sum_{r=0}^{k-(l+1)} \frac{r(r+1)}{k-(l+1)} p_{l+1,r} \quad (11)$$

Similarly, for B_l $(1 \le l \le k)$ we have:

$$B_l = \frac{(-1)^{k-l+1}}{(k-l)!} \sum_{1\le s\le l} w_s^2 \int_{w_l^{-1}}^{w_{l+1}^{-1}} \beta^{k-l} P_l^{(k-l+1)} d\beta; \;\; =$$

$$-\left[\sum_{1\le s\le l} w_s^2 \sum_{r=0}^{k-l} \frac{(-1)^r}{r!} \beta^r P_l^{(r)}\right]_{w_l^{-1}}^{w_{l+1}^{-1}} \;\; =$$

$$\sum_{1\le s\le l} w_s^2 \sum_{r=0}^{k-l} \frac{(-1)^r}{r!} w_l^{-r} P_l^{(r)}(w_l^{-1}) - \sum_{1\le s\le l} w_s^2 \sum_{r=0}^{k-l} \frac{(-1)^r}{r!} w_{l+1}^{-r} P_l^{(r)}(w_{l+1}^{-1}) \;\; =$$

$$\sum_{1\le s\le l} w_s^2 \sum_{r=0}^{k-l} p_{l,r} - \sum_{1\le s\le l} w_s^2 \sum_{r=0}^{k-l-1} p_{l+1,r}$$

How about C_{k-1}? From Lemma 3 we estimate

$$C_{k-1} \;\le\; w_{k-1}\left(P'_{k-1}(w_k^{-1}) - P'_{k-1}(w_{k-1}^{-1})\right) \;\; =$$

$$-w_{k-1}P'_{k-1}(w_{k-1}^{-1}) - w_{k-1}w_k P_k(w_k^{-1}) \;\; \le$$

$$-w_{k-1}P'_{k-1}(w_{k-1}^{-1}) - w_k^2 P_k(w_k^{-1}) \;=\; w_{k-1}^2 p_{k-1,1} - w_k^2 p_{k,0}. \tag{12}$$

We denote the right hand side by C'_{k-1}. Now we start to compute the total. A comparison of the negative terms of B_l and the positive terms of B_{l+1} shows that

$$\sum_{l=0}^{k} B_l = \sum_{l=1}^{k} B_l = \sum_{l=1}^{k} w_l^2 \sum_{r=0}^{k-l} p_{l,r}. \tag{13}$$

Let us pay attention to the form of our expressions for D_l, E_l, C'_{k-1} and $\sum_{l=0}^{k} B_l$. We find that

$$\sum_{l=0}^{k}(B_l + C_l) \le \sum_{l=0}^{k} B_l + \sum_{l=0}^{k-2} D_l - \sum_{l=0}^{k-2} E_l + C'_{k-1} = \sum_{l=0}^{k} w_l^2 \sum_{r=0}^{k-l} \theta_l(r) p_{l,r}$$

for some coefficients $\theta_l(r)$ $(0 \le l \le k,\, 0 \le r \le k-l)$, where $\theta_l(r)$ does not depend on the weights, only on k, l and r. First we compute these coefficients in the typical case, i.e. when l does not equal to 0, $k-1$ or k. For $1 \le l \le k-2$ from (10), (11) and (13) we get:

$$\theta_l(r) = 1 - \frac{r(r+1)}{k-l} + \frac{r(r-1)}{k-l-1}. \tag{14}$$

The above is a second degree polynomial in r, which is equal to 1, when $r = 0$ and has roots at $k-l$ and $k-l-1$. Hence for $1 \le l \le k-1$ we have:

$$\theta_l(r) = \frac{(k-l-r)(k-l-r-1)}{(k-l)(k-l-1)}. \tag{15}$$

Next we compute D_0 outright instead of computing the $\theta_0(r)$s. Keeping in mind that in our

shorthand notation w_0^{-1} stands for 0, we have:

$$D_0 = \frac{1}{k-1}\sum_{r=0}^{k-2}\frac{(-1)^r}{r!}w_0^{-r}P_0^{(r+2)}(w_0^{-1}) = \frac{1}{k-1}P_0^{(2)}(0) = \frac{1}{k-1}\sum_{1\le i,i'\le n} w_i w_{i'} = \frac{W^2-\sum_{i=1}^n w_i^2}{k-1}.$$

Finally, we compute the missing special cases:

$$\begin{aligned}\theta_{k-1}(0) &= 1-0+0=1;\\ \theta_{k-1}(1) &= 1-\frac{1\times 2}{1}+1=0;\\ \theta_k(0) &= 1-1=0.\end{aligned}$$

Note that in (10), (11), (12) and (13) r never exceeds $k-l$. Summing up everything we get:

$$Var(\sum_{i=1}^n \hat{w}_i) \le \frac{W^2-\sum_{i=1}^n w_i^2}{k-1} + w_{k-1}^2 p_{k-1,0} + \sum_{l=1}^{k-2} w_l^2 \sum_{r=0}^{k-l}\theta_l(r)p_{l,r} - \sum_{i=1}^n w_i^2, \qquad (16)$$

where $\theta_l(r)$ is as in (14). From this expression we can easily prove:

Theorem 3.

$$Var\left(\sum_{i=1}^n \hat{w}_i\right) \le \frac{W^2-\sum_{i=1}^n w_i^2}{k-1} - \sum_{i=k}^n w_i^2.$$

Proof: From (15) it follows that $\theta_l(r) \le 1$ for $1 \le l \le k-2$, $0 \le r \le k-l$. Since $p_{l,r}$ $(r \ge 0)$ is a probability distribution for a fixed l, it follows that $\sum_{r=0}^{k-l}\theta_l(r)p_{l,r} \le 1$. Also, $p_{k-1,0} \le 1$. Thus the sum of the second and third terms of (16) is upper bounded by $\sum_{i=1}^{k-1} w_i^2$, and the theorem follows. As a consequence of Theorem 3 we get Theorem 1.

9 The ϕ and ψ functions

Although for general $0 \le l \le k$ we defined $\theta_l(r)$ only for $0 \le r \le k-l$, by Expression (14) in the special case when $1 \le l \le k-2$, we can extend $\theta_l(r)$ to the case, when r is an arbitrary non-negative integer. Towards proving Theorem 2 from 16 we define:

$$\begin{aligned}\varphi(w_l,\ldots,w_n) &= w_l^2\sum_{r=0}^{k-l}\theta_l(r)p_{l,r};\\ \psi(w_l,\ldots,w_n) &= w_l^2\sum_{r=0}^{\infty}\theta_l(r)p_{l,r}.\end{aligned}$$

Remark 1. *We did not index φ and ψ with l. Syntactically, l can be deduced from the number of arguments, since k and n are considered fixed. In our view φ and ψ are functions of a variable length argument list of positive numbers $x, x_1, x_2, \ldots, x_m$ with the requirement that $x \geq x_i$ for $1 \leq i \leq m$. We define p_r as the probability of having r ones in the sum of m independent Bernouli trials, where the i^{th} trial has bias x_i/x towards 1. Then*

$$\begin{aligned} \varphi(x, x_1 \ldots, x_m) &= x^2 \sum_{r=0}^{K} \frac{(K-r)(K-r-1)}{K(K-1)} p_r, \\ \psi(x, x_1 \ldots, x_m) &= x^2 \sum_{r=0}^{\infty} \frac{(K-r)(K-r-1)}{K(K-1)} p_r \end{aligned}$$

for some fixed K given in advance. In our case $K = k - l$.

In order to show that ψ is an upper bound on ϕ, it is sufficient:

Lemma 7. *Let $1 \leq l \leq k-2$. For every integer $r \geq 0$ it holds that $\theta_l(r) \geq 0$*

Indeed, this is straightforward from (15).

Consequence 1. *For $1 \leq l \leq k-2$ we have $\varphi(w_l, \ldots, w_n) \leq \psi(w_l, \ldots, w_n)$.*

While we can only estimate φ, we can compute ψ exactly:

$$\psi(w_l, \ldots, w_n) = w_l^2 + \frac{\sum_{l+1 \leq i \neq j \leq n} w_i w_j}{(k-l)(k-l-1)} - \frac{2w_l \sum_{i=l+1}^{n} w_i}{k-l}. \tag{17}$$

Indeed, using (14) we get:

$$\theta_l(r) = 1 - \frac{r(r+1)}{k-l} + \frac{r(r-1)}{k-l-1} = 1 - \frac{r(r-1)}{k-l} + \frac{r(r-1)}{k-l-1} - \frac{2r}{k-l} \quad = \\ 1 + \frac{r(r-1)}{(k-l)(k-l-1)} - \frac{2r}{k-l}.$$

Then Equation (17) follows from Lemma 6.

10 The monotonicity of φ

Before making our variance estimates we need one more lemma that concerns the monotone decreasing property of φ in its second, third, etc. variables.

Lemma 8. *Let $1 \leq l \leq k-2$ and let $i > l$. As w_i decreases (but stays greater or equal than 0), $\varphi(w_l, \ldots, w_n)$ does not decrease. Above we assume all other arguments stay the same.*

Proof: Recall how we got $p_{l,r}$ from $n-l$ independent Bernouli trials. Consider the set of $n-l-1$ Bernouli trials, which leaves out the one associated with w_i, and denote the probability that in this set of trials we obtain r ones by p_r^-. Define $p = w_i/w_l$ and $q = 1 - w_i/w_n$. As w_i decreases, p decreases and q increases while their sum remains 1. Clearly,

$$p_{l,r} = p \times p_{r-1}^- + q \times p_r^- \tag{18}$$

for every $0 \leq r$. Above we have set p^-_{-1} to 0. Define

$$\vartheta_l(r) = \begin{cases} \theta_l(r) & \text{if } r \leq k-l; \\ 0 & \text{if } r > k-l. \end{cases}$$

We have

$$\varphi(w_l, \ldots, w_n) = w_l^2 \sum_{r=0}^{\infty} \vartheta_l(r) p_{l,r}.$$

We also have that $\vartheta_l(r) \geq \vartheta_l(r+1)$ for every integer $r \geq 0$. From (18) we obtain:

$$\varphi(w_l, \ldots, w_n) = w_l^2 \sum_{r=0}^{\infty} \vartheta_l(r)(p \times p^-_{r-1} + q \times p^-_r) \quad = \\ w_l^2 \sum_{r=0}^{\infty} (\vartheta_l(r+1)p + \vartheta_l(r)q) p^-_r \,.$$

The lemma is now implied by the monotone non-decreasing property of $\vartheta_l(r+1)p + \vartheta_l(r)q$ in q.

11 Proof of Theorem 2

First we outline the proof. Define $\varphi_l = \varphi(w_l, \ldots, w_n)$, $\psi_l = \psi(w_l, \ldots, w_n)$. Thus the right hand side of (16) is

$$\frac{W^2 - \sum_{i=1}^n w_i^2}{k-1} \; + \; w_{k-1}^2 p_{k-1,0} \; + \; \sum_{l=1}^{k-2} \varphi_l - \sum_{i=1}^n w_i^2,$$

For $1 \leq m \leq k-1$ define

$$V_m = \frac{\left(\sum_{i=m}^n w_i\right)^2 - \sum_{i=m}^n w_i^2}{k-m} \; + \; w_{k-1}^2 p_{k-1,0} \; + \; \sum_{l=m}^{k-2} \varphi_l - \sum_{i=m}^n w_i^2.$$

We shall prove:

Lemma 9. *$V_{m+1} = V_m - \varphi_m + \psi_m$ for $1 \leq m \leq k-2$.*

Consequence 2. *V_m is non-decreasing in m.*

Lemma 10. *If for some $1 \leq m \leq k-1$ we have $\frac{\sum_{i=m}^n w_i}{k-m} \geq w_m$ then it holds that*

$$V_m \leq \frac{\left(\sum_{i=m}^n w_i\right)^2}{k-m} - \sum_{i=m}^n w_i^2.$$

These lemmas now imply Theorem 2, since if m is the first index for which $\frac{\sum_{i=m}^n w_i}{k-m} \geq w_m$ then, using Lemma 1 we have $m = \ell \leq k-1$. Expressions (3), (16) and the above lemmas then imply

$$Var\left(\sum_{i=1}^n \tilde{w}_i\right) = \frac{(W')^2}{k-\ell} - \sum_{i=\ell}^n w_i^2 \geq V_\ell = V_m \geq V_1 \geq Var\left(\sum_{i=1}^n \hat{w}_i\right).$$

Proof of Lemma 9: Using (17):

$$\begin{aligned} V_m - \varphi_m + \psi_m &= \\ \frac{\sum_{m\le i\neq j\le n} w_i w_j}{k-m} + w_{k-1}^2 p_{k-1,0} + \sum_{l=m+1}^{k-2} \varphi_l - \sum_{i=m}^{n} w_i^2 + \psi_m &= \\ \frac{\sum_{m+1\le i\neq j\le n} w_i w_j + 2w_m \sum_{i=m+1}^{n} w_i}{k-m} + w_{k-1}^2 p_{k-1,0} + \sum_{l=m+1}^{k-2} \varphi_l - \sum_{i=m+1}^{n} w_i^2 + & \\ + \frac{\sum_{m+1\le i\neq j\le n} w_i w_j}{(k-m)(k-m-1)} - \frac{2w_m \sum_{i=m+1}^{n} w_i}{k-m} &= \\ \frac{\sum_{m+1\le i\neq j\le n} w_i w_j}{k-m-1} + w_{k-1}^2 p_{k-1,0} + \sum_{l=m+1}^{k-2} \varphi_l - \sum_{i=m+1}^{n} w_i^2 = V_{m+1}. & \end{aligned}$$

Proof of Lemma 10: Let m be the (first) index for which $\frac{\sum_{i=m}^{n} w_i}{k-m} \ge w_m$. For the rest of the section we fix this m. By Lemma 1 we have $\frac{\sum_{i=t}^{n} w_i}{k-t} \ge w_t$ for $m \le t \le k-1$. For $m \le t \le k-1$ define

$$v_t = \frac{\left(\sum_{i=m}^{n} w_i\right)^2}{k-m} - \sum_{i=m}^{n} w_i^2 + w_{k-1}^2 p_{k-1,0} - \frac{\sum_{i=t}^{k-1} w_i^2}{k-t} + \sum_{l=t}^{k-2} \varphi_l.$$

Clearly,

$$\begin{aligned} V_m &\le v_m; \\ v_{k-1} &\le \frac{\left(\sum_{i=m}^{n} w_i\right)^2}{k-m} - \sum_{i=m}^{n} w_i^2. \end{aligned}$$

Hence we are done if we can prove that $v_m \le v_{m+1} \le \ldots \le v_{k-1}$.

Lemma 11. Proof: *For any $m \le t \le k-2$ we have $v_t \le v_{t+1}$.*

For $m \le t \le k-2$ we have:

$$v_{t+1} - v_t = \frac{\sum_{i=t}^{k-1} w_i^2}{k-t} - \frac{\sum_{i=t+1}^{k-1} w_i^2}{k-t-1} - \varphi_t = \frac{w_t}{k-t} - \frac{\sum_{i=t+1}^{k-1} w_i^2}{(k-t)(k-t-1)} - \varphi_t. \tag{19}$$

In order to upper bound φ_t we shall use the monotonicity property from Lemma 8. Recall that $w_t \le \frac{\sum_{i=t}^{n} w_i}{k-t}$ holds. Hence:

$$w_t + w_{t+1} + w_{t+2} + \cdots w_{k-1} \le (k-t) w_t \le \sum_{i=t}^{n} w_i.$$

Introduce $0 \le w_i' \le w_i$ for $k \le i \le n$ in any manner such that $\sum_{i=t}^{k-1} w_i + \sum_{i=k}^{n} w'(i) = (k-t) w_t$

holds. Then from Lemma 8 and (16):

$$\begin{aligned}
\varphi_l \le \varphi(w_t,\dots,w_{k-1},w'_k,\dots,w_n) \le \psi(w_t,\dots,w_{k-1},w'_k,\dots,w_n) &= \\
w_t^2 + \frac{\left(\sum_{i=t+1}^{k-1} w_i + \sum_{i=k}^{n} w'_i\right)^2 - \sum_{i=t+1}^{k-1} w_i^2 + \sum_{i=k}^{n}(w'_i)^2}{(k-t)(k-t-1)} - \frac{2w_t\left(\sum_{i=t+1}^{k-1} w_i + \sum_{i=k}^{n} w'_i\right)}{k-t} &= \\
w_t^2 + \frac{(k-t-1)^2 w_t^2 - \sum_{i=t+1}^{k-1} w_i^2 + \sum_{i=k}^{n}(w'_i)^2}{(k-t)(k-t-1)} - \frac{2(k-t-1)w_t^2}{k-t} &\le \\
w_t^2 + \frac{(k-t-1)^2 w_t^2 - \sum_{i=t+1}^{k-1} w_i^2}{(k-t)(k-t-1)} - \frac{2(k-t-1)w_t^2}{k-t} &= \\
w_t^2 + \frac{-\sum_{i=t+1}^{k-1} w_i^2}{(k-t)(k-t-1)} - \frac{(k-t-1)w_t^2}{k-t} &= \\
\frac{w_t^2}{k-t} - \frac{\sum_{i=t+1}^{k-1} w_i^2}{(k-t)(k-t-1)}.
\end{aligned}$$

Replacing the above into (19) the right hand side cancels to 0.

References

[AGPR] Swarup Acharya, Phillip B. Gibbons, Viswanath Poosala, Sridhar Ramaswamy: The Aqua Approximate Query Answering System. SIGMOD Conference 1999: 574-576

[AFT] Adler, RJ, RE Feldman, MS Taqqu. 1998 . A Practical Guide to Heavy TailsStatistical Techniques and Applications. Birkhauser, Boston, MA

[ADLT] N. Alon, N. Duffield, C. Lund and M. Thorup: Estimating arbitrary subset sums with few probes, submitted.

[C] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. J. Comput. Syst. Sci., 55(3):441–453, 1997.

[DLT1] Nick G. Duffield, Carsten Lund, Mikkel Thorup: Estimating flow distributions from sampled flow statistics. SIGCOMM 2003: 325-336

[DLT2] Nick G. Duffield, Carsten Lund, Mikkel Thorup: Priority sampling estimating arbitrary subset sums, submitted

[GM] Phillip B. Gibbons, Yossi Matias: New Sampling-Based Summary Statistics for Improving Approximate Query Answers. SIGMOD Conference 1998: 331-342

[HHW] Joseph M. Hellerstein, Peter J. Haas, Helen J. Wang: Online Aggregation. SIGMOD Conference 1997: 171-182

[OR] Frank Olken, Doron Rotem: Random sampling from databases - A survey. `http://pueblo.lbl.gov/ olken`

[PKC] Kihong Park, Gitae Kim, Mark Crovella: On the relationship between file sizes, transport protocols, and self-similar network traffic. ICNP 1996: 171-180

ECCC ISSN 1433-8092
http://www.eccc.uni-trier.de/eccc
ftp://ftp.eccc.uni-trier.de/pub/eccc
ftpmail@ftp.eccc.uni-trier.de, subject 'help eccc'