CS3150 - Homework — Week 2 (2.25, 3.22, 4.9)*

Dan Li, Xiaohui Kong †, Hammad Ibqal and Ihsan A. Qazi
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260
† Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260

January 20, 2006

Contents

*This was written by Dan Li

1 Problem 2.25

A blood test is being performed on n individuals. Each person can be tested separately, but this is expensive. Pooling can decrease the cost. The blood sample of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the group of k in test suffices for the group of k individuals. If the test is positive, then each of the k person must be tested separately and thus $k+1$ total tests are required for the k people.

Suppose that we create n/k disjoint groups of k people(where k divides n) and use the pooling method. Assume that each person has a positive result on the test independently with probability p.

(a) What is the probability that the test for a pooled sample of k people will be positive?

Answer: The result of the pooled sample is positive means that at least one of the k tested samples has positive result, which probability is:

1 - P(all of the k people have negative sample).

Since we assume that each person has a positive result on the test independently with probability p, so that the probability that each person has negative result is $1-p$, and the probability that all of the k persons have negative results is $(1-p)^k$.

Finally we have the probability that the test for the pooled sample of k people is positive is:

$$1-(1-p)^k \tag{1}$$

(b) What is the expected number of tests necessary?

Answer: At least one test is needed no matter what the test result of the pooled sample is. When the result is positive, k extra tests are needed. The probability that k extra tests are needed is:

$$1-(1-p)^k \tag{2}$$

so that the expected number of tests for each group of k people is

$$1+k\cdot[1-(1-p)^k] = 1+k-k\cdot(1-p)^k \tag{3}$$

And there are n/k groups, so the total number of tests is:

$$N(n,k) = \frac{n}{k}\cdot[1+k-k\cdot(1-p)^k] = n\cdot(1+\frac{1}{k}-(1-p)^k) \tag{4}$$

(c) Describe how to find the best value of k.

Answer: In order to find the best value of k, we need to find such a value of k that the number from (b) reach its minimum value. This is done by the following:

$$\begin{aligned} \frac{\partial N(n,k)}{\partial k} &= \frac{\partial}{\partial k} n\cdot[1+\frac{1}{k}-(1-p)^k] \\ &= n\cdot[-\frac{1}{k^2}-(1-p)^k\cdot ln(1-p)] \\ &= 0 \end{aligned} \tag{5}$$

which gives:

$$k^2 \cdot (1-p)^k = -\frac{1}{ln(1-p)} \tag{6}$$

The value of k can not be solved in close form.

(d) Give an inequality that shows for what values of p pooling is better than just testing every individual.

Answer: When the number of tests using pooled sample is less that the number of tests for testing every individual, the pooling method is better. This is obtained by having:

$$\begin{aligned}
n[1+\frac{1}{k}-(1-p)^k] &< n \\
1+\frac{1}{k}-(1-p)^k &< 1 \\
\frac{1}{k} &< (1-p)^k \\
k &> (\frac{1}{1-p})^k \\
k^{\frac{1}{k}} &> \frac{1}{1-p} \\
1-p &> (\frac{1}{k})^{\frac{1}{k}} \\
p &< 1-(\frac{1}{k})^{\frac{1}{k}}
\end{aligned} \tag{7}$$

2 Problem 3.22

Suppose that we flip coin n times to obtain n random bits. Consider all $m = \binom{n}{2}$ pairs of these bits in some order. Let Y_i be the exclusive-or of the ith pair of bits, and let $Y = \sum_{i=1}^{m} Y_i$ be the number of Y_i that equal 1.

(a) Show that each Y_i is 0 with probability 1/2 and 1 with probability 1/2.

Answer: Each Y_i is the exclusive-or of two bits. Assume $Y_i = x_j \oplus x_k$, then

$$\begin{aligned} P(Y_i = 1) &= P((x_j = 0 \cap x_k = 1) \cup (x_j = 1 \cap x_k = 0)) \\ &= P(x_j = 0 \cap x_k = 1) + P(x_j = 1 \cap x_k = 0) \\ &= P(x_j = 0) \cdot P(x_k = 1) + P(x_j = 1) \cdot P(x_k = 0) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{2} \end{aligned} \tag{8}$$

and

$$\begin{aligned} P(Y_i = 0) &= P((x_j = 0 \cap x_k = 0) \cap (x_j = 1 \cap x_k = 1)) \\ &= P(x_j = 0 \cap x_k = 0) + P(x_j = 1 \cap x_k = 1) \\ &= P(x_j = 0) \cdot P(x_k = 0) + P(x_j = 1) \cdot P(x_k = 1) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{2} \end{aligned} \tag{9}$$

(b) Show that the Y_i are not mutually independent.
Answer: Mutually independent means for every subset, the probability

$$Pr(Y_i \cap Y_j \cap \cdots \cap Y_r) = Pr(Y_i) \cdot P(Y_j) \cdots P(Y_r)$$

If we choose such a subset that those $Y_i's$ have factors in common, for example, we choose $Y_i = x_a \oplus x_b$, $Y_j = x_a \oplus x_c$ and $Y_k = x_b \oplus x_c$, then

$$P(Y_i = 1 \cap Y_j = 1 \cap Y_k = 1) = 0$$

but

$$P(Y_i = 1)P(Y_j = 1)P(Y_k = 1) = \frac{1}{8}$$

They are not equal. So the $Y_i's$ are not mutually independent.

(c) Show that the Y_i satisfy the property that $E[Y_iY_j] = E[Y_i]E[Y_j]$.

Answer: $E[Y_iY_j] = Pr(Y_iY_j = 1) = Pr(Y_i = 1 \cap Y_j = 1)$.

If Y_i and Y_j do not have factor in common, i.e. $Y_i = x_a \oplus x_b$ and $Y_j = x_c \oplus x_d$, then

$$\begin{aligned}
Pr(Y_i = 1 \cap Y_j = 1) &= Pr((x_a \oplus x_b = 1) \cap (x_c \oplus x_d = 1)) \\
&= Pr(x_a = 0)Pr(x_b = 1) \cdot P(x_c \oplus x_d = 1) \\
&\quad + Pr(x_a = 1)Pr(x_b = 0) \cdot P(x_c \oplus x_d = 1) \\
&= \frac{1}{4}Pr(x_c \oplus x_d = 1) + \frac{1}{4}Pr(x_c \oplus x_d = 1) \\
&= \frac{1}{2}Pr(x_c \oplus x_d = 1) \\
&= \frac{1}{4} \qquad (10)
\end{aligned}$$

If Y_i and Y_j have one factor in common, i.e. $Y_i = x_a \oplus x_b$ and $Y_j = x_b \oplus x_c$, then

$$\begin{aligned}
Pr(Y_i = 1 \cap Y_j = 1) &= Pr((x_a \oplus x_b = 1) \cap (x_b \oplus x_c = 1)) \\
&= Pr(x_a = 0)Pr(x_b = 1)P(x_c = 0) \\
&\quad + Pr(x_a = 1)Pr(x_b = 0)P(x_c = 1) \\
&= \frac{1}{8} + \frac{1}{8} \\
&= \frac{1}{4} \qquad (11)
\end{aligned}$$

While, for any i,

$$\begin{aligned}
E[Y_i] &= Pr(Y_i = 1) \\
&= Pr(x_a = 0 \cap x_b = 1) + Pr(x_a = 1 \cap x_b = 0) \\
&= \frac{1}{4} + \frac{1}{4} \\
&= \frac{1}{2} \qquad (12)
\end{aligned}$$

So that $E[Y_i]E[Y_j] = \frac{1}{4}$.
In any case, the equality $E[Y_iY_j] = E[Y_i]E[Y_j]$ holds.

(d) Using Exercise 3.15, find $Var[Y]$.
Answer: Using Exercise 3.15, since the above euality holds, and $Y = \sum_{i=1}^{m} Y_i$,

$$Var[Y] = \sum_{i=1}^{m} Var[Y_i]$$

$$\begin{aligned}
Var[Y_i] &= \overline{Y_i^2} - (\overline{Yi})^2 \\
&= Pr(Y_i = 1) - (Pr(Y_i = 1))^2 \\
&= \frac{1}{2} - (\frac{1}{2})^2 \\
&= \frac{1}{4} \qquad (13)
\end{aligned}$$

So, $Var[Y] = m/4$.

(e) Using Chebyshev's inequality, prove a bound on $Pr(|Y - E[Y]| \geq n)$.

Answer: Using Chebyshev's inequality,

$$
\begin{aligned}
Pr(|Y - E[Y]| \geq n) &\leq \frac{Var[Y]}{n^2} \\
&= \frac{m/4}{n^2} \\
&= \frac{n-1}{8n} \\
&= \frac{1}{8}(1 - \frac{1}{n})
\end{aligned} \tag{14}
$$

3 Problem 4.9

Suppose that we can obtain independent samples $X_1, X_2, \cdots$ of a random variable X and that we want to use these samples to estimate $E[X]$. Using t samples, we use $(\sum_{i=1}^{t} X_i)/t$ for estimate of $E[X]$. We want the estimate to be within $\varepsilon E[X]$ from the true value of $E[X]$ with probability at least 1-δ. We may not be able to use Chernoff's bound directly to bound how good our estimate is if X is not a 0-1 random variable, and we do not know its moment generating function. We develop an alternative approach that requires only having a bound on the variance of X. Let $r = \sqrt{Var[X]}/E(X)$.

(a) Show using Chebyshev's inequality that $O(r^2/\varepsilon^2\delta)$ samples are sufficient to solve the problem.
Answer:

$$\begin{aligned} Pr(\sum_{i=1}^{t} X_i/t \le (1+\varepsilon)E[X]) &= 1 - Pr(\sum_{i=1}^{t} X_i/t > (1+\varepsilon)E[X]) \\ &= 1 - Pr(\sum_{i=1}^{t} X_i > t(1+\varepsilon)E[X]) \end{aligned} \tag{15}$$

$$E(\sum_{i=1}^{t} X_i) = t \cdot E(X_i) = t \cdot E(X) \tag{16}$$

and

$$Var(\sum_{i=1}^{t} X_i) = t \cdot Var(X_i) = t \cdot Var(X) \tag{17}$$

Using Chebyshev's Inequality, and write $Y = \sum_{i=1}^{t} X_i$,

$$\begin{aligned} Pr(\sum_{i=1}^{t} X_i > t(1+\varepsilon)E[X]) &= Pr(Y > E(Y) + \varepsilon E(Y)) \\ &= Pr(Y - E(Y) > \varepsilon E(Y)) \\ &\le Pr(|Y - E(Y)| > \varepsilon E(Y)) \\ &\le \frac{Var(Y)}{(\varepsilon E(Y))^2} \\ &= \frac{Var(\sum_{i=1}^{t} X_i)}{(\varepsilon t E(X))^2} \\ &= \frac{t \cdot Var(X)}{t^2 \varepsilon^2 E(X)^2} \\ &= \frac{r^2}{t \cdot \varepsilon^2} \end{aligned} \tag{18}$$

As long as $t \geq r^2/(\varepsilon^2\delta)$, we have

$$\begin{aligned} Pr(\sum_{i=1}^{t} X_i/t \leq (1+\varepsilon)E[X]) &= 1 - Pr(\sum_{i=1}^{t} X_i > t(1+\varepsilon)E[X]) \\ &= 1 - \frac{r^2}{t \cdot \varepsilon^2} \\ &\geq 1 - \delta \end{aligned} \tag{19}$$

So the number of estimates needed is: $r^2/(\varepsilon^2\delta) = O(r^2/\varepsilon^2\delta)$.
But, if we have $O(r^2/\varepsilon^2\delta)$ samples, it does not guarantee the probability of $1-\delta$.

(b) Suppose that we need only a weak estimate that is within $\varepsilon E[X]$ of $E[X]$ with probability at least $3/4$. Argue that $O(r^2/\varepsilon^2)$ samples are enough for this weak estimate.

Answer: Probability of 3/4 means $\delta = 1/4$. By setting $\delta = 1/4$ in $O(r^2/\varepsilon^2\delta)$, we have $O(4r^2/\varepsilon^2) = O(r^2/\varepsilon^2)$.

(c) Show that, by taking the median of $O(log(1/\delta))$ weak estimates, we can obtain an estimate within $\varepsilon E[X]$ of $E[X]$ with probability at least 1-δ. Conclude that we need only $O((r^2 log(1/\delta))/\varepsilon^2)$ samples.

Answer: If the median of the weak estimates satisfies the condition, it means less than half of the weak estimates are not within $\varepsilon E[X]$ of the true value of $E[X]$. Let's use a new random variable X_i:

$$X_i = \begin{cases} 1 & \text{if the } i th \text{ weak estimate fall above } \varepsilon E(X) \text{ of } E(X) \\ 0 & \text{if the } i th \text{ weak estimate fall below } \varepsilon E(X) \text{ of } E(X) \end{cases}$$

X_i follows binomial distribution with probability of 1/4 or more to be 1 and 3/4 or less to be 0. For simplicity, we use 1/4 in this problem. Lower probability will need lower number of estimates.

If we use $X = \sum_{i=1}^{m} X_i$ to represent how many weak estimates fall above $(1+\varepsilon)E(X)$, we will be able to use Chernoff bound to for the value of m so that $Pr(X >= m/2) < \delta$.

Chernoff bound gives:

$$Pr(X \geq (1+\delta')E(X)) \leq e^{-E(X)\delta'^2/3}$$

where E(X) = m/4. Use $\delta' = 1$, we have

$$Pr(X \geq m/2) \leq e^{-m/12}$$

By using $m = 12 \cdot log(1/\delta)$, we have $Pr(X \geq m/2) \leq \delta$, so that the probability that the median of weak estimates gives result within $\varepsilon E(X)$ is at least $1-\delta$.

Each weak estimate uses $O(r^2/\varepsilon^2)$ samples, and there are $O(log(1/\delta))$ weak estimates so that the total number of samples is $O(r^2 log(1/\delta)/\varepsilon^2)$.