THE LIGHT BULB PROBLEM [1]

Ramamohan Paturi[2] **Sanguthevar Rajasekaran**[3] **John Reif**[3]
Univ. of California, San Diego Univ. of Pennsylvania Duke University

Running Title: The Light Bulb Problem

Corresponding Author:

Ramamohan Paturi
Department of Computer Science
Mail Code 0114
University of California, San Diego
La Jolla, CA 92093

ABSTRACT

In this paper, we consider the problem of correlational learning and present algorithms to determine correlated objects.

1. INTRODUCTION

Correlational learning, a subclass of unsupervised learning, aims to identify statistically correlated groups of attributes. In this paper, we consider the following correlational learning problem due to L. G. Valiant, 1985 and 1988: We have a sequence of n random light bulbs each of which is either on or off with equal probability at each time step. Further, we know that a certain pair of bulbs is positively correlated. The problem is to find efficient algorithms for recognizing the unique pair of light bulbs with the maximum correlation. Some preliminary results in this direction are reported in Paturi, 1988.

In this paper, we consider a more general version of the basic light bulb problem. In the general version, we assume that the behavior of the bulbs is governed by some unknown probability distribution except that the pair with the largest pairwise correlation is unique. Our goal would be to find this unique pair. We also consider the extension of the problem to k–way correlations.

Mathematically, we can regard each light bulb l_i at time step t as a random variable X_i^t which takes the values ± 1. We call $(X_1^t, X_2^t, \ldots, X_n^t)$ the t–th *sample*. We also assume that the behavior of the light bulbs is independent of their past behavior. In other words, the samples are independent of each other. We would like to find the desired object (k–tuples with the maximum correlation) *with high probability.* The complexity measures of interest are sample size and the number of operations.

Before we proceed further, we introduce some definitions and facts from probability theory.

We define the *correlation* of a (unordered) pair $\{i, j\}$ of light bulbs l_i and l_j as $\mathbf{P}[X_i = X_j]$. In general, for any $k \geq 2$, the correlation of the unordered k–tuple $\{i_1, i_2, \ldots, i_k\}$ of light bulbs is defined as $\mathbf{P}[X_{i_1} = X_{i_2} = \cdots = X_{i_k}]$. Observe that these definitions differ somewhat from the standard definition of the correlation. However, in the special case of each bulb having the ± 1 value with equal probability, if p is the correlation according to the definition in this paper, then $2p - 1$ is the correlation according to the standard definition.

In the following, we present estimates of the probability of the sum of independent random variables deviating from its mean. Although such large deviation probability estimates are known before, for the sake of completeness, we derive these results using moment generating functions. More information can be found in Alon, Spencer and Erdös, 1992.

Let $z_1, z_2, \ldots, z_m$ be $0, 1$ valued independent random variables such that $\mathbf{P}[z_j = 1] = p_j$ for $1 \leq j \leq m$.

Let $S^m = \sum_{j=1}^m z_j$ and the expectation of S^m be $\mu = \mathbf{E}S^m = \sum_{j=1}^m p_j$. We are interested in the probability that S^m is above or below its expectation. The following lemma bounds the probability that S^m is below its mean. Let $\mu' \leq \mu$ be any lower bound on μ.

Lemma 1 *For $0 \leq T < \mu' \leq \mu$, $\mathbf{P}[S^m < T] \leq e^{-(\mu'-T)^2/(2\mu')}$.*

Proof: We use the moment generating function $\mathbf{E}e^{S^m t}$. Since the random variables z_j are independent, we get

$$\begin{aligned}
\mathbf{P}[S^m < T] &= \mathbf{P}[e^{Tt - S^m t} > 1] \\
&\leq \mathbf{E}e^{Tt - S^m t} \\
&= e^{Tt} \prod_{j=1}^{m} \mathbf{E}e^{-z_j t}
\end{aligned}$$

for any $t > 0$.

We select $t = (\mu' - T)/\mu' \leq 1$. We obtain $\mathbf{E}e^{-z_j t} = 1 - p_j + p_j e^{-t} \leq 1 - p_j((\mu' - T)(\mu' + T)/(2(\mu')^2)) \leq e^{-p_j(\mu'-T)(\mu'+T)/(2(\mu')^2)}$. Using this upper bound, we get

$$\begin{aligned}
\mathbf{P}[S^m < T] &\leq e^{(\mu'-T)T/\mu'} e^{-(\mu'-T)(\mu'+T)\sum_{j=1}^m p_j/(2(\mu')^2)} \\
&\leq e^{-(\mu'-T)^2/(2\mu')}.
\end{aligned}$$

which is the desired inequality. 2

The following lemma bounds the probability that S^m is above its mean. Let $\mu' \geq \mu$ be any upper bound on μ.

Lemma 2 *For $\mu \leq \mu' < T \leq 2\mu'$, $\mathbf{P}[S^m \geq T] \leq e^{-(T-\mu')^2/(3\mu')}$.*

Proof: We use the moment generating function $\mathbf{E}e^{S^m t}$. Due to the independence of the variables involved, we get

$$\begin{aligned}\mathbf{P}[S^m \geq T] &= \mathbf{P}[e^{S^m t - Tt} \geq 1]\\ &\leq \mathbf{E}e^{S^m t - Tt}\\ &= e^{-Tt}\prod_{j=1}^{m}\mathbf{E}e^{z_j t}\end{aligned}$$

for any $t > 0$.

Following Raghavan, 1986, we select $t = \ln(T/\mu')$. We now obtain

$$\mathbf{E}e^{z_j t} = 1 - p_j + p_j e^t = 1 + p_j(T-\mu')/\mu' \leq e^{p_j(T-\mu')/\mu'}.$$

Using this upper bound, we get

$$\begin{aligned}\mathbf{P}[S^m \geq T] &\leq e^{-T\ln(T/\mu')}e^{(T-\mu')\sum_{j=1}^{m}p_j/\mu'}\\ &\leq e^{T-\mu'-T\ln(T/\mu')}\\ &\leq e^{-(T-\mu')^2/(3\mu')}.\end{aligned}$$

The last inequality follows since $(T-\mu')/\mu' \leq 1$ and $(T-\mu'-T\ln(T/\mu'))/\mu' = ((T-\mu')/\mu' - T\ln(1+(T-\mu')/\mu')/\mu')$ is less than $-1/3$ for $0 \leq (T-\mu')/\mu' \leq 1$. 2

We say that a statement holds *with high probability* , if it holds with probability $1 - n^{-\alpha}$ for some $\alpha \geq 1$.

2. A QUADRATIC–TIME ALGORITHM

We now present an algorithm called algorithm Q which samples each pair of bulbs $O(\ln n)$ times to determine the pair with the largest correlation. We first develop some definitions that will be used in our algorithms.

Let $S_{ij}^t = |\{1 \leq u \leq t | X_i^u = X_j^u\}|$. In other words, S_{ij}^t is the number of times the bulbs l_i and l_j have identical output when t samples are considered. Let p_1 be the largest pairwise correlation and p_2 be the second largest correlation. If p_1 and p_2 are very close to each other, one would expect to look at a larger number of samples to isolate the pair with the largest correlation. However, we do not know the separation between p_1 and p_2. Hence we introduce the *accuracy* parameter $\gamma > 1$ as one of the inputs to our algorithms. The parameter γ is an estimate of $p_1/(p_1 - p_2)$. Our algorithms, with high probability, output a pair whose correlation is greater than $p_1(1 - 1/\gamma)$. If $p_2 \leq p_1(1-1/\gamma)$, our algorithms, with high probability, output the pair with the largest correlation. All our algorithms also have a certainty parameter $\alpha \geq 1$ as their input. Our algorithms succeed with probability at least $1 - n^{-\alpha}$.

The algorithm Q relies on the basic fact that the value of S_{ij}^t tends to be larger if the pair $\{i, j\}$ is more correlated. Hence, if we take a sufficiently large t, we can guarantee with a high probability that the pair $\{i, j\}$ with the largest S_{ij}^t has the maximum correlation provided p_1 and p_2 are sufficiently separated. Let $T' = O(\gamma^2 \ln n)$. Each step of the algorithm Q consists of obtaining a sample and updating S_{ij}'s. At step t of the algorithm, for each pair $\{i, j\}$, we check if $S_{ij}^t \geq T'$. If there exists a pair $\{i, j\}$ such that $S_{ij}^t \geq T'$, we output $\{i, j\}$ and stop. Otherwise, we look at the random variables X_i^{t+1} at step $t+1$ and update S_{ij}^{t+1} and repeat the above computation until we succeed. Each step of this computation takes $O(n^2)$ operations since we examine each of the $O(n^2)$ pairs. Note that this algorithm does not depend on prior knowledge of the probabilities p_1 and p_2. We show that the algorithm with high probability outputs a pair whose correlation is greater than $p_1(1 - 1/\gamma)$ and terminates in $O((\gamma^2 \ln n)/p_1)$ steps. Hence if $p_2 \leq p_1(1 - 1/\gamma)$, with high probability, the algorithm outputs the pair with the maximum correlation using $O((\gamma^2 \ln n)/p_1)$ samples and its time complexity is $O((\gamma^2 n^2 \ln n)/p_1)$.

In the following, we give the algorithm and its analysis.

Algorithm Q:

```
    Inputs: $\alpha \geq 1$, the certainty parameter and $\gamma > 1$, the accuracy
    parameter;

    1.  $t = 1$;
```

```
2. for each pair $\{i, j\}$, $S_{ij} \leftarrow \begin{cases} +1, & \text{if } X_i^t = X_j^t \\ 0, & \text{otherwise} \end{cases}$

3. $T' \leftarrow 12(2+\alpha)\gamma^2 \ln n$;

4. while (TRUE ) do
       if there exists a pair $\{i, j\}$ such that $S_{ij} \geq T'$ then
       output the pair $\{i, j\}$ and stop;
       else $t \leftarrow t + 1$ and, for each pair $\{i, j\}$,
       $S_{ij} \leftarrow \begin{cases} S_{ij} + 1, & \text{if } X_i^t = X_j^t \\ S_{ij}, & \text{otherwise} \end{cases}$
```

The following theorem gives the performance of the algorithm. It should be noted that we used simpler constants rather than striving for the optimal constants.

Theorem 1 *For all $n \geq 3$, Algorithm Q terminates with $t \leq \lfloor (18(2+\alpha)\gamma^2 \ln n)/p_1 \rfloor$ and outputs a pair whose correlation is greater than $p_1(1 - 1/\gamma)$ with probability at least $1 - n^{-\alpha}$.*

Proof: Let $T = \lfloor (18(2+\alpha)\gamma^2 \ln n)/p_1 \rfloor = \lfloor 3T'/(2p_1) \rfloor$ where $T' = 12(2+\alpha)\gamma^2 \ln n$. We show that the algorithm Q terminates with $t \leq T$ outputting a pair whose correlation is greater than $p_1(1-1/\gamma)$ with probability at least $1 - n^{-\alpha}$.

Let $t = \lfloor (1 + \frac{1}{2\gamma})T'/p_1 \rfloor \leq T$. If the pair $\{i, j\}$ is the pair with the maximum correlation p_1, we then have $\mathbf{E}[S_{ij}^t] = p_1 t \geq (1 + \frac{1}{2\gamma})T' - 1$. Hence, from lemma 1, we get $\mathbf{P}[S_{ij}^t < T'] < n^{-\alpha}/2$.

On the other hand, we show that, for any pair $\{i, j\}$ whose correlation is less than or equal to $p_1(1 - 1/\gamma)$, the probability $S_{ij}^{t'} \geq T'$ for any $1 \leq t' \leq t$ is small. Let $\{i, j\}$ be a pair whose correlation is at most $p_1(1 - 1/\gamma)$. Since $S_{ij}^t < T'$ implies $S_{ij}^{t'} < T'$ for any $1 \leq t' \leq t$, we need only consider the probability that $S_{ij}^t \geq T'$ and show that it is small. We have $\mathbf{E}[S_{ij}^t] \leq tp_1(1 - 1/\gamma) \leq (1 + \frac{1}{2\gamma})(1 - \frac{1}{\gamma})T' < T'$. To apply lemma 2, we use the upper bound $\max((1 - \frac{1}{\gamma})(1 + \frac{1}{2\gamma})T', T'/2)$ for $\mathbf{E}[S_{ij}^t]$. We then get $\mathbf{P}[S_{ij}^t \geq T'] \leq n^{-2-\alpha}$. Since the number of the events $[S_{ij}^t \geq T']$ is at most $n^2/2$, we get that the probability that $S_{ij}^{t'} \geq T'$ for some $1 \leq t' \leq t$ and a pair $\{i, j\}$ with correlation less than or equal to $p_1(1 - 1/\gamma)$ is bounded by $n^2 n^{-2-\alpha}/2 < n^{-\alpha}/2$.

Hence, with probability at least $1 - n^{-\alpha}$, the algorithm Q will output a pair whose correlation is greater than $p_1(1 - 1/\gamma)$ and terminates with $t \leq T$. 2

Corollary 1 *Algorithm Q outputs the pair with the largest correlation using $O((\alpha\gamma^2 \ln n)/p_1)$ samples in time $O((\alpha\gamma^2 n^2 \ln n)/p_1)$ with probability at least $1 - n^{-\alpha}$ if $p_2 \leq p_1(1 - 1/\gamma)$. More generally, given m pairs of random variables, the algorithm Q outputs the pair with the largest correlation in time $O((\alpha\gamma^2 m \ln n)/p_1)$ if $p_2 \leq p_1(1 - 1/\gamma)$.*

Can we do better? In the rest of the paper, we present an algorithm which takes *subquadratic* time and extend it to k–way correlations.

3. ALGORITHM B

To understand this algorithm, consider the special case with $p_1 = 1$. In this case, the problem is reduced to sorting. We want to determine the pair that produced identical outputs. This can be done by sorting the strings $s_i^t = X_i^1 X_i^2 \ldots X_i^t$. If we consider $O(\gamma \ln n)$ samples, with high probability, we can find a pair whose correlation is at least $p_1(1-1/\gamma)$. The total number of operations in this special case is $O(\gamma n \ln n)$.

Even in the more general case, we can use the above idea to reduce the number of pairs to be considered. We classify the random variables based on their $s_i^t = X_i^1 \cdots X_i^t$. We say that two bulbs i and j fall into the same bucket if $s_i^t = s_j^t$. We consider all the pairs $\{i, j\}$ for i and j in the same bucket. We select t such that the number of pairs $\{i, j\}$ of bulbs with $s_i^t = s_j^t$ is on the average no more than cn for a suitably chosen constant c. On the other hand, if t is not too large, the maximally correlated pair falls into the same bucket with a sufficiently large probability. If this process is repeated for a sufficient number of times, the maximally correlated pair occurs more frequently than all the others. We use this frequency count to isolate the pair with the maximum correlation. In addition to the inputs γ and α, the algorithm B requires an upper bound q_2 on the second largest pair–wise correlation. With high probability, it outputs the pair whose correlation is greater than $p_1(1 - 1/\gamma)$.

In the following, we give the algorithm and its analysis. By a family of buckets of bulbs, we mean a partition of the set of bulbs.

Algorithm B:

Inputs: $\alpha \geq 1$, the certainty parameter, $\gamma > 1$, the accuracy parameter, an upper bound q_2 on the second largest pair–wise correlation;

```
0. Let $T' = 12(2+\alpha)\gamma^2 \ln n$

1. Let $PAIRS$ be the empty list of pairs and their counts;
2. while (TRUE) do
    2.1. Let $t \leftarrow 0$ and $F$ be the family of buckets of
         bulbs obtained by placing all the bulbs in one bucket.
    2.2. while ($t < \lfloor \frac{\ln n}{\ln(1/q_2)} \rfloor$) do
        2.2.1. Obtain the sample vector $(X_1^{t+1}, X_2^{t+1}, \ldots, X_n^{t+1})$. Split
               the buckets if necessary so that all bulbs in
               the same bucket agree on all the $t+1$ sample
               vectors seen so far. Let $F$ denote the family
               of buckets so obtained.
        2.2.2. $t \leftarrow t + 1$
    2.3. For each bucket of bulbs in $F$ consider all possible
         pairs of bulbs from the bucket. Add any new pairs to
         the list $PAIRS$ with count initialized to one. Increase
         the count of the other pairs by one. If the count of
         any pair is at least $T'$, output the pair and exit.
```

The following theorem gives the time and the sample complexity of algorithm B. In the analysis, we make use of the fact that p_2, the second largest pairwise correlation is at least $\frac{1}{2} - \frac{3}{2n}$. To see this, observe that given any sequence $X_1, \ldots, X_n$ of binary values, for a random pair $\{i, j\}$ other than the pair with the largest correlation, the probability that $X_i = X_j$ is at least $\frac{1}{2} - \frac{3}{2n}$ for all $n \geq 3$. Hence, for any probability distribution, there is a pair other than the pair with the largest correlation whose correlation is at least $\frac{1}{2} - \frac{3}{2n}$ for all $n \geq 3$.

Theorem 2 *Algorithm B outputs a pair whose correlation is greater than $p_1(1 - 1/\gamma)$ with probability at least $1 - n^{-\alpha}$ and terminates in expected time $O((2+\alpha)\gamma^2 n^{1+\frac{\ln p_1}{\ln q_2}} \ln^2 n)$ with $O((2+\alpha)\gamma^2 n^{\frac{\ln p_1}{\ln q_2}} \ln^2 n)$ expected number of samples. In particular, if $p_2 \leq p_1(1 - 1/\gamma)$, then the algorithm B outputs the pair with the maximum correlation with probability at least $1 - n^{-\alpha}$.*

Proof: Consider any invocation of the inner loop (step 2.2). Let F be the family of buckets obtained at the end of this invocation. We have the probability that the pair with the largest correlation falls in the same bucket of this family is $p_1^{\ln n/\ln(1/q_2)} = n^{-\frac{\ln p_1}{\ln q_2}}$. On the other hand, for any other pair, the probability that it falls in the same bucket of F is at most $p_2^{\ln n/\ln(1/q_2)} = n^{-\frac{\ln p_2}{\ln q_2}} \leq 1/n$. Hence, the expected number of pairs obtained from F is at most n.

We will show that the count of the maximally correlated pair with high probability exceeds T' before the count of any other pair whose correlation is at most $p_1(1-1/\gamma)$ does. Let f_1^j denote the count of the maximally correlated pair at the end of the first j iterations of the outer loop. Let k be the smallest integer such that $\mathbf{E}[f_1^k] \geq (1 + \frac{1}{2\gamma})T'$ where $T' = 12(2+\alpha)\gamma^2 \ln n$. Since in any iteration of the outer loop the pair with the largest correlation falls in the same bucket with probability $n^{-\frac{\ln p_1}{\ln q_2}}$, we have $k \leq 1 + (1 + \frac{1}{2\gamma})T' n^{\frac{\ln p_1}{\ln q_2}}$. Since the iterations of the outer loop are independent, we get, applying lemma 1, $\mathbf{P}[f_1^k \geq T'] \geq 1 - n^{-\alpha}/2$.

Let p_3 be the correlation of any pair with $p_3 \leq p_1(1 - \frac{1}{\gamma})$. Let f_3^j be the maximum count for this pair at the end of the first j iterations of the outer loop. Clearly, $\mathbf{E}[f_3^k] \leq k n^{-\frac{\ln p_3}{\ln q_2}} \leq [1 + (1 + \frac{1}{2\gamma})T' n^{\frac{\ln p_1}{\ln q_2}}] n^{-\frac{\ln p_3}{\ln q_2}}$. But we have $n^{\frac{\ln p_1}{\ln q_2}} \times n^{-\frac{\ln p_3}{\ln q_2}} \leq n^{-\frac{\ln(1-1/\gamma)}{\ln q_2}}$. Since $q_2 \geq p_2 \geq \frac{1}{5}$ for all $n \geq 5$, we have that $n^{-\frac{\ln(1-1/\gamma)}{\ln q_2}} \leq (1 - \frac{1}{\gamma})$ for all $n \geq 5$. This implies that $\mathbf{E}[f_3^k] \leq$

$(1-\frac{1}{\gamma})(1+\frac{1}{2\gamma})T'+o(1)$. To apply lemma 2, we use the upper bound $\max((1-\frac{1}{\gamma})(1+\frac{1}{2\gamma})T'+o(1), T'/2)$ for $\mathbf{E}[f_3^k]$. We then get $\mathbf{P}[f_3^k \geq T'] \leq n^{-2-\alpha}$. Since there are at the most $\frac{n^2}{2}$ pairs, we get the probability that for some pair whose correlation is at most $p_1(1-1/\gamma)$ the count will be be at least T' at the end of the first k iterations of the outer loop is at most $n^{-\alpha}/2$.

Hence, after at most $1+(1+\frac{1}{2\gamma})T'n^{\frac{\ln p_1}{\ln q_2}}$ iterations of the outer loop, with probability at least $1-n^{-\alpha}$, the count for the maximally correlated pair exceeds T' before the count of any other pair whose correlation is at most $p_1(1-1/\gamma)$ does.

In order to obtain the expected run time of the algorithm, observe that the expected number of iterations of the outer loop is $O(T'n^{\frac{\ln p_1}{\ln q_2}})$ since using lemma 1 we get that $\mathbf{P}[f_1^l < T']$ decreases exponentially in j where $l = jk$. In each iteration of the outer loop, the expected number of pairs considered is $O(n)$. Each iteration of the outer loop on the average requires $O(n \ln n)$ time steps and $O(\ln n)$ samples since $q_2 \geq p_2 \geq \frac{1}{2} - \frac{3}{2n}$. Combining these estimates, we get that the Algorithm B terminates in expected time $O((2+\alpha)\gamma^2 n^{1+\frac{\ln p_1}{\ln q_2}} \ln^2 n)$ and the expected number of samples used is $O((2+\alpha)\gamma^2 n^{\frac{\ln p_1}{\ln q_2}} \ln^2 n)$. 2

The previous theorem gives a bound on the expected run time and the number of samples of the algorithm. With an additional log factor, we can show that the time and the sample bounds also hold with high probability. We use the following fact for this purpose. See also Karp, 1991 for a generalization of this technique.

Lemma 3 *For any $c \geq 0$, if a randomized algorithm $\mathcal{A}$ produces the correct answer with probability at least $1-n^{-\alpha}$ and runs in expected time T, then there is a randomized algorithm that with probability at least $1-(cn^{-\alpha}\log_2 n + n^{-c})$ runs in time $O(cT\log_2 n)$ and produces the correct answer.*

Proof: We use the following algorithm to achieve the time bound mentioned in the theorem. Run $c\log_2 n$ copies of the algorithm independently in parallel. Stop as soon as one of the copies outputs a pair and this pair will be the output of the algorithm. (If two or more copies produce their outputs simultaneously, select one of them arbitrarily as the output of the algorithm.)

The probability that all the $c\log_2 n$ copies produce the correct answer is at least $1-cn^{-\alpha}\log_2 n$. Moreover, using Markov's inequality, we get that the probability that $\mathcal{A}$ does not output a pair in time $2T$ is at most $1/2$. Thus, the probability that none of the $c\log_2 n$ copies outputs a pair in time $2T$ is at most $(1/2)^{c\log_2 n} = n^{-c}$. Therefore, the modified algorithm computes the correct result and runs in time $O(cT\log_2 n)$ with probability at least $1-(cn^{-\alpha}\log_2 n + n^{-c})$. 2

Using the above lemma, we get the following theorem.

Theorem 3 *Algorithm B, with probability at least* $(1-n^{-\alpha})$*, outputs a pair whose correlation is greater than* $p_1(1-1/\gamma)$ *and terminates in time* $O((2+\alpha)\alpha\gamma^2 n^{1+\frac{\ln p_1}{\ln q_2}}\ln^3 n)$ *with* $O((2+\alpha)\alpha\gamma^2 n^{\frac{\ln p_1}{\ln q_2}}\ln^3 n)$ *samples. In particular, if* $p_2 \le p_1(1-1/\gamma)$*, then the algorithm B outputs the pair with the maximum correlation with probability at least* $1-n^{-\alpha}$*.*

4. BOOTSTRAP TECHNIQUE

Algorithm B uses a large number $(O((2+\alpha)\alpha\gamma^2 n^{\frac{\ln p_1}{\ln q_2}}\ln^3 n))$ of samples. We can reduce the sample size to $O(\ln n)$ using the bootstrap technique (Diaconis and Efron, 1980, Efron, 1979 and 1982).

Assume that we are given a data set (i.e., a random sample of size d) $D = \{x_1, x_2, \ldots, x_d\}$ from an unknown distribution, and we want to estimate some statistic, say θ. The idea of bootstrap is to generate a large number of new data sets from D and estimate θ on each one of the generated data sets to obtain a better estimate of θ. A data set is generated by drawing samples independently with replacement from D with each element in D being equally likely.

In the following, we explain how one can reduce the number of samples to $O(\ln n)$ by increasing the time complexity of the algorithm B. We first make $d\ln n$ observations for some constant $d \ge 1$. Our new probability distribution D' is obtained by drawing the data sets with uniform probability from among the $d \ln n$ samples obtained.

If d is chosen to be large, we show that, with high probability, the separation between the largest correlation and the second largest correlation in the sample data does not decrease significantly in the distribution D'. Let ${p'}_1$ and ${p'}_2$ be the largest and the second largest correlations in D'. For $n \geq 5$ and $d \geq 240\gamma^2(2+\alpha)$, we will show that with probability at least $1 - n^{-\alpha}$, ${p'}_1$ is at least $(1 - \frac{1}{2\gamma})p_1$ and ${p'}_2$ is at most $(1 + \frac{1}{4(\gamma-1)})p_2$.

Let $\{i, j\}$ be the largest correlation pair in the original probability distribution. Let f_1 be the number of samples among the $d \ln n$ samples for which i and j have the same value. Then, $\mathbf{E}[f_1] = p_1 d \ln n$. Hence $\mathbf{P}[f_1 \leq (1 - \frac{1}{2\gamma})p_1 d \ln n] \leq n^{-\alpha}/2$ using lemma 1.

Let $\{i, j\}$ be any other pair. Let f_2 be the number of samples among the $d \ln n$ samples for which i and j have the same value. We have $\mathbf{E}[f_2] = p_2 d \ln n$. Hence for $n \geq 5$, using lemma 2 with $\mathbf{E}[f_2]$ upper bounded by $\max(p_2 d \ln n, (1 + \frac{1}{4(\gamma-1)})p_2 d \ln n/2)$, we get $\mathbf{P}[f_2 \geq (1 + \frac{1}{4(\gamma-1)})p_2 d \ln n] \leq n^{-2-\alpha}/2$ since $p_2 \geq \frac{1}{5}$. Since there are at most $n^2/2$ pairs, we have that the probability that one of the pairs other than the one with the largest correlation agrees on more than $(1+\frac{1}{4(\gamma-1)})p_2 d \ln n$ samples is at most $n^{-\alpha}/2$.

Finally, we run the algorithm B with the inputs $\alpha' = \alpha$, $\gamma' = 4\gamma$ and ${q'}_2 = (1 + \frac{1}{4(\gamma-1)})q_2$ using the distribution D'. Using the bounds on ${p'}_1$ and ${p'}_2$, we conclude that algorithm B, with probability at least $1 - n^{-\alpha}$, outputs a pair whose correlation is greater than $p_1(1 - \frac{1}{2\gamma})(1 - \frac{1}{4\gamma})$ and terminates in time $O((2+\alpha)\alpha\gamma^2 n^{1+\frac{\ln(p_1(1-\frac{1}{2\gamma}))}{\ln(q_2(1+\frac{1}{4(\gamma-1)}))}} \ln^3 n)$ using $d \ln n$ samples for $d \geq 240\gamma^2(2+\alpha)$.

5. k–WAY CORRELATION

We can modify the algorithm B to detect a k–tuple with the largest k–way correlation. The run time of the algorithm would then be $O((k+\alpha)(k-1)\alpha\gamma^2 n^{1+(k-1)\frac{\ln p_1}{\ln q_2}} \ln^3 n)$. In the following, we present the algorithm B$_k$ to detect the k–tuple with the largest k–way correlation.

Algorithm B$_k$:

Inputs: $\alpha > 0$, the certainty parameter, $\gamma > 1$, the accuracy parameter, an upper bound q_2 on the second largest k–way correlation;

```
0. Let $T' = 12(k+\alpha)\gamma^2 \ln n$
1. Let $k$–$TUPLES$ be the empty list of $k$–tuples and their counts;
2. while (TRUE) do
    2.1. Let $t \leftarrow 0$ and $F$ be the family of buckets of bulbs obtained by placing all the bulbs in one bucket.
    2.2. while ($t < \lfloor (k-1)\frac{\ln n}{\ln(1/q_2)} \rfloor$) do
        2.2.1. Obtain the sample vector $(X_1^{t+1}, X_2^{t+1}, \ldots, X_n^{t+1})$. Split the buckets if necessary so that all bulbs in the same bucket agree on all the $t+1$ sample vectors seen so far. Let $F$ denote the family of buckets so obtained.
        2.2.2. $t \leftarrow t+1$
    2.3. For each bucket of bulbs in $F$ consider all possible $k$–tuples of bulbs from the bucket. Add any new $k$–tuples to the list $k$–$TUPLES$ with count initialized to one. Increase the count of the other $k$–tuples by one. If the count of any $k$–tuple is at least $T'$, output the $k$–tuple and exit.
```

Theorem 4 *Algorithm B_k, with probability at least $(1 - n^{-\alpha})$, outputs a k–tuple whose correlation is greater than $p_1(1 - 1/\gamma)$ and terminates in time $O((k+\alpha)(k-1)\alpha\gamma^2 n^{1+(k-1)\frac{\ln p_1}{\ln q_2}} \ln^3 n)$ with $O((k+\alpha)(k-1)\alpha\gamma^2 n^{(k-1)\frac{\ln p_1}{\ln q_2}} \ln^3 n)$ samples. In particular, if $p_2 \leq p_1(1 - 1/\gamma)$, then the algorithm B_k outputs the k–tuple with the maximum correlation with probability at least $(1 - n^{-\alpha})$.*

Proof. The analysis of Algorithm B$_k$ is similar to that of Algorithm B and hence is omitted. 2

OPEN PROBLEM: Is there an $O(n \ln n)$ algorithm for determining the pair with the maximum correlation?

Acknowledgments: We would like to thank Prof. Les Valiant for suggesting this problem and for listening to our solutions. We also would like to thank our referees for a careful and helpful review.

References

[1] ALON, N., SPENCER, J., AND ERDÖS, P. (1992), "The Probabilistic Method", Wiley–Interscience Series, John Wiley & Sons Inc.

[2] DIACONIS, P., AND EFRON, B. (1980), Computer-Intensive Methods in Statistics. *Annals of Statistics*, pp 116–130.

[3] EFRON, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics* **7**, pp 1–26.

[4] EFRON, B. (1982), "The Jacknife, the Bootstrap and Other Resampling Plans", SIAM, Philadelphia, Pennsylvania.

[5] KARP, R. (1991), Probabilistic Recurrence Relations, *in* "Proceedings of the 23rd Symposium on Theory of Computing", pp 190–197.

[6] KEARNS, M., AND LI, M. (1988), Learning in the Presence of Malicious Errors. *in* "Proceedings of the 20th Symposium on Theory of Computing", pp 267–280.

[7] PATURI, R. (1988), The Light Bulb Problem. Technical Report CS88–129, University of California, San Diego.

[8] PATURI, R., RAJASEKARAN, S., and REIF, J.H. (1989), The Light Bulb Problem. *in* "Second Work shop on Computational Learning Theory," pp. 261–268.

[9] RAGHAVAN, P., (1986), Probabilistic Construction of Deterministic Algorithms: Approximating Packing Integer Programs, *in* "27th IEEE Symposium on Foundations of Computer Science", pp. 10–18.

[10] VALIANT. L.G. (1984), A Theory of the Learnable. *Communications of the ACM* **27**, pp 1134–1142.

[11] VALIANT, L.G. (1985), Learning Disjunctions of Conjunctions. *in* “Proceedings of the 9th International Joint Conference on Artificial Intelligence”, Los Angeles, pp 560–566.

[12] VALIANT, L.G. (1985), Private Communication.

[13] VALIANT, L.G. (1988), Functionality in Neural Nets. *in* “First Work shop on Computational Learning Theory”, pp 28–39.