The Complexity of Controlled Three Dimensional Rounding

by

Kirk Pruhs

Computer Sciences Technical Report #715
September 1987

The Complexity of Controlled Three Dimensional Rounding

Kirk Pruhs*
Computer Science Department
University of Wisconsin
Madison, Wisconsin
September 4, 1987

Abstract: *It is often necessary to round multidimensional statistical tables in such a way as to preserve row sums. This is called zero restricted rounding. It is also often desirable that the rounding procedure be unbiased. A probabilistic rounding procedure is unbiased if the expected value of the rounding of each table entry is equal to the value of that table entry. A rounding procedure is controlled if it is both zero restricted and unbiased. There is a practical algorithm that generates a controlled rounding for any two dimensional table [Cox]. It is also known that there are three dimensional tables that do not have a zero restricted rounding [CCE]. We prove that the problem of deciding whether a three dimensional table has a zero restricted rounding is NP-complete. We also prove that the problem of deciding whether a controlled rounding is possible for a three dimensional table is NP-complete. The existence of polynomial time algorithm, that is an algorithm that runs in time bounded by a polynomial, for one of these rounding problems would imply that all NP-complete problems have polynomial time solutions. Some of the more famous NP-complete problems are the traveling salesman problem, integer programming, the bin packing problem, the graph coloring problem and the knapsack problem. Despite the fact that NP-complete problems have been widely and extensively studied, there is no known polynomial time algorithm for any NP-complete problem. It is almost unanimously believed that there is no polynomial time algorithm for any NP-complete problem. Therefore it is very unlikely that there is a polynomial time algorithm for either of these three dimensional rounding problems. This implies that it also very unlikely that there is a polynomial time algorithm that generates a zero restricted rounding for those three dimensional tables that have zero restricted roundings, since such an algorithm could be used to decide whether a three dimensional table had a zero restricted rounding.*

* This research was supported in part by the National Science Foundation under grant DCR-8451397.

1. Introduction

Given a table of positive rational numbers it is sometimes necessary to round these numbers to an integer multiple of some rounding base. To round a number x with respect to a base B means x should remain fixed if it is an integer multiple of B and rounded to one of the two adjacent integer multiples of B, $B\lfloor x/B \rfloor$ or $B\lfloor x/B \rfloor + 1$, otherwise. The expression $\lfloor x \rfloor$ denotes the greatest integer less than or equal x. A rounding of a table T is a rounding of every entry in T. One application of rounding is the prevention of statistical disclosure in tabular data [CFGH , New , NS].

It is often necessary that a rounding of a table satisfy some additional conditions. Before we can introduce these conditions we will need the following definitions. If T is an n-dimensional table, we will denote an entry in T by $T(x_1, x_2, \ldots, x_n)$. A *hyperplane* of a table T is obtained from T by fixing some subset of the indices and is denoted by placing $*$'s in the unfixed positions. For example, $T(x_1, *, *, x_4)$ is a two dimensional table derived from T by looking only at those entries whose the first coordinate is x_1 and whose fourth coordinate is x_4. A one dimensional hyperplane is called a *row* and and a two dimensional hyperplane is called a *sheet*. One basic condition that a rounding may be required to satisfy is that the rounding be zero restricted. A rounding is *zero restricted* if all the hyperplane sums in the rounded table are roundings of the corresponding hyperplane sums in the original table. Alternatively without loss of generality we may assume that the hyperplane sums are integer multiples of the rounding base and define a zero restricted rounding as a rounding that preserves hyperplane sums [Cox]. A rounding is *unbiased* if for every entry x the expected value of the rounding of x is equal to the value of x. A second condition that a rounding may be required to satisfy is that the rounding be controlled. A rounding is *controlled* if it is zero restricted and unbiased.

There is a procedure that generates a controlled rounding for any two dimensional table T [Cox]. If T is a m by n table this procedure runs in time bounded by $c \cdot z \cdot min(m, n)$, where z is the number of nonzero entries in T and c is a constant. This is assuming the table is represented by an adjacency list data structure [AHU]. This algorithm runs in linear time on a one dimensional table.

In contrast it is not true that every three dimensional table has a zero restricted rounding [CCE]. Therefore we would like to examine the complexity of deciding whether a particular three dimensional table has a zero restricted rounding and the complexity of deciding whether a particular three dimensional table has an controlled rounding. We will give very strong evidence that there are no efficient algorithms for either of these problems. Before we start we need some preliminary definitions.

2. Definitions

The class P is defined as the class of problems that have polynomial time solutions. An algorithm solves a problem in polynomial time if there is some polynomial $p(x)$ such that on any input α, of size n, the algorithm solves the problem for α in no more than $p(n)$ steps. The class P is generally regarded as the class of problems that have practical solutions [Edm, GJ]. The class NP is defined as the the class of problems where it is possible, in polynomial time, to guess a candidate solution and then verify whether the candidate is indeed a solution. As an example of a problem in NP, consider the Hamiltonian cycle problem. The input to this problem is a set of cities and a listing of the cities that have direct plane flights between them. The problem is then to decide whether there is a scheduling of flights that starts and finishes in the same city and visits each other city exactly once. While finding a Hamiltonian cycle is apparently very difficult, it is easy to guess at some subset of flights and then decide if they form a Hamiltonian cycle.

A problem $\mathcal{C}$ is *NP-hard* if a polynomial time algorithm for $\mathcal{C}$ implies a polynomial time algorithm for every problem in NP. More technically a problem $\mathcal{C}$ is NP-hard if for every problem $\mathcal{A}$ in NP there is a polynomially computable function f such that an instance x of $\mathcal{A}$ has a solution if and only if the instance $f(x)$ of $\mathcal{C}$ has a solution. A problem is *NP-complete* if it is in NP and it is NP-hard. The above mentioned Hamiltonian cycle problem is NP-complete. Some examples of NP-complete problems are the traveling salesman problem, satisfiability of boolean formulas, the graph coloring problem, the bin packing problem, the knapsack problem, and integer programming. See Garey and Johnson [GJ] for more information about NP-completeness, including a 100 page collection of some of the more significant NP-complete problems. The class of NP-complete problems contains a rich variety of problems that have been extensively studied and yet there are no known subexponential time algorithms for any NP-complete problem.

As a result it is almost unanimously believed that proving a problem NP-hard implies the problem has no polynomial time solution.

An NP-complete problem we will use later is the 2-*in*-4 SAT problem [Sch]. An instance of 2-*in*-4 SAT is a set of boolean variables $X = \{x_1, x_2, ..., x_n\}$ and a set of clauses $C = \{C_1, C_2, ..., C_m\}$. Each clause is a set of four variables from X. The 2-*in*-4 SAT problem is determining whether there is a truth assignment which assigns true to exactly two variables per clause.

3. NP-completeness Results

We prove that the problem of deciding whether a three dimensional table has a zero restricted rounding is NP-complete. We also prove that the problem of deciding whether a three dimensional table has a controlled rounding is NP-complete. This trivially implies NP-completeness for tables of dimension greater than three. In fact these rounding problems remain NP-complete even if sheet sums instead of row sums are required to be zero restricted.

To prove that the zero restricted three dimensional rounding problem is NP-hard we will exhibit a polynomial time transformation that takes an instance A of 2-*in*-4 SAT and creates a three dimensional table T. The table T has the property that it has a zero restricted rounding if and only if the instance A, of the 2-*in*-4 SAT problem, has a satisfying assignment.

Theorem 1 The zero restricted three dimensional rounding problem is NP-complete.

proof: The problem is in NP because it is trivial to verify that a potential rounding is zero restricted. To show that the problem is NP-hard, let $X = \{x_1, x_2, ..., x_n\}$ be a set of variables and $C = \{C_1, C_2, ...C_m\}$ be a set of clauses of an arbitrary instance of 2-in-4 SAT. We will denote the jth occurrence of variable x_i by x_i^{j-1} and the number of times x_i appears by $\#x_i$. Let $V = \{x_i^j | 1 \leq i \leq n, 0 \leq j \leq \#x_i - 1\}$ be a set with a distinct element for each variable occurrence. V has $4m$ elements. We are now going to construct a $4m$ by $4m$ by $5m$ table T. Let $T(\alpha, \beta, \gamma)$ denote a single entry in this table where α and β are elements of V and γ is an element of $V \cup C$. That is the first two indices are variable occurrences and the third index is either a variable occurrence or a clause. There will be a total of $8m$ nonzero entries in T, 2 for each variable occurrence in the 2-in-4 SAT instance. All the entries will be 0 or $\frac{1}{2}$ and the rounding base will be 1. If variable x_i^j occurs in clause C_k then set $T(x_i^j, x_i^j, C_k)$ to be $\frac{1}{2}$. Rounding an entry $T(x_i^j, x_i^j, C_k)$ to 1 corresponds to assigning x_i true and rounding an entry $T(x_i^j, x_i^j, C_k)$ to 0 corresponds to assigning x_i false. Therefore each sheet $T(*, *, C_k)$ has four entries each with value $\frac{1}{2}$ and the sheet sum of $T(*, *, C_k)$ is 2. The restriction that exactly two literals for clause C_k must be assigned true is analogous to the restriction that exactly two out of the four entries on sheet $T(*, *, C_k)$ must be rounded to 1.

Now we must guarantee that every occurrence of x_i is assigned the same truth value. For each variable occurrence x_i^j set $T(x_i^j, x_i^{(j+1) \bmod \# x_i}, x_i^j)$ to $\frac{1}{2}$. All other entries in T are zero. Then each sheet of the form $T(x_i^j, *, *)$ or $T(*, x_i^j, *)$ has two entries and a sheet sum of 1. This creates a cycle which guarantees that for each x_i all the nonzero entries of the form $T(x_i^j, x_i^j, C_k)$ are rounded in the same direction. To see this let x^j be an occurrence of some variable x, we will ignore subscripts for the moment. Without loss of generality we will assume that $(j+1) \bmod \# x$ equals $j+1$. Also assume that x^j occurs in clause C_k and x^{j+1} occurs in clause C_l. Then by the zero restrictedness condition on sheet $T(x^j, *, *)$ we have that $T(x^j, x^j, C_k)$ and $T(x^j, x^{j+1}, x^j)$ must be rounded in opposite directions. Also by the zero restrictedness condition on sheet $T(*, x^{j+1}, *)$ we have that $T(x^{j+1}, x^{j+1}, C_l)$ and $T(x^j, x^{j+1}, x^j)$ must be rounded in opposite directions. Therefore $T(x^j, x^j, C_k)$ and $T(x^{j+1}, x^{j+1}, C_l)$ must be rounded in the same direction. By repeatedly applying this argument all entries of the form $T(x^j, x^j, C_m)$ must be rounded in the same direction.

This construction gives us a one to one correspondence between zero restricted roundings of T and satisfying assignments of the 2-*in*-4 SAT instance. Assigning a variable x_i true corresponds to rounding to 1 all nonzero table entries of the form $T(x_i^j, x_i^j, C_k)$ and rounding to 0 all table entries of the form $T(x_i^j, x_i^{(j+1) \bmod \# x_i}, x_i^j)$. Assigning x_i false corresponds to rounding to 1 all table entries of the form $T(x_i^j, x_i^{(j+1) \bmod \# x_i}, x_i^j)$ and to 0 all table entries of the form $T(x_i^j, x_i^j, C_k)$. ♠

We now turn our attention to the controlled three dimensional rounding problem. To show the controlled rounding problem is in NP we need the following standard lemma [PS].

Lemma 1 Let $Ax = b, x_i \geq 0$ be a linear program, where $x = (x_1, \ldots, x_n)$ is a column vector of indeterminants. If this linear program has a solution then it has a solution $\hat{x}$ where the columns of A, which correspond to the nonzero entries of $\hat{x}$, form a linearly independent set.

Theorem 2 The controlled three dimensional rounding problem is NP-complete.

proof : To prove membership in NP let T be a three dimensional table. For this problem we may assume without loss of generality that each dimension is of size n and the rounding base is 1 [Cox]. For convenience we will consider T as one dimensional vector of size n^3. A base 1 rounding s of T can then be thought of as a 0-1 vector of length n^3. Let $\{s_1, s_2, ..., s_m\}$ be the set of zero restricted roundings for T and let S be a n^3 by m matrix whose ith column is s_i. Let $w = (w_1, \ldots, w_m)$ be a vector of indeterminants of size m. T has a controlled rounding if and only if there is a solution to the following linear program:

$$\begin{aligned} \sum_{i=1}^{m} w_i &= 1 \\ Sw &= T \\ w_i &\geq 0, \qquad i \in \{1 \ldots m\} \end{aligned}$$

In these formulas the value of w_i represents the probability that solution s_i is generated. The second equation then expresses the restriction that the rounding must be unbiased. We cannot construct the matrix S in polynomial time since m may be exponentially larger than n. Yet we can nondeterministically determine whether this linear program has a solution without explicitly constructing S. This follows from the lemma 1. Since the rank of S is at most n^3, if the above linear program has a solution, then it has one where at most $n^3 + 1$ of the w_i's are nonzero.

To solve the controlled rounding problem nondeterministically guess a number k, $1 \geq k \geq n^3 + 1$, and k 0-1 vectors $a_1, a_2, ..., a_k$ each of size n^3. Verify in polynomial time that each a_i is a zero restricted rounding of T. Let A be the matrix whose ith column is a_i and $v = (v_1, \ldots, v_k)$ be a k-dimensional vector of indeterminants. Then check whether the following polynomial sized linear program has a solution:

$$\begin{aligned} \sum_{i=1}^{k} v_i &= 1 \\ Av &= T \\ v_i &\geq 0, \qquad i \in \{1 \ldots k\} \end{aligned}$$

Using the ellipsoid method of Khachian a linear program can be solved in polynomial time [PS].

To show that the controlled rounding problem is NP-hard we will exhibit a polynomial time transformation from the 2-*in*-4 SAT problem to the controlled three dimensional rounding problem. This transformation is identical to the one used in theorem 1. Let $\{x_1, x_2, ..., x_n\}$ be a set of variables and $\{C_1, C_2, ...C_m\}$ a set of clauses of an instance of the 2-*in*-4 SAT problem. Notice that if δ is any satisfying assignment for this instance, another satisfying assignment $\bar{\delta}$ can be constructed from δ by reversing the truth assignment of each variable. A identical phenomenon occurs in the table T constructed by the transformation. The table T has $8m$ nonzero entries. Every zero restricted rounding σ of T rounds $4m$ of the nonzero entries to 1 and $4m$ of the nonzero entries to 0. If σ is a zero restricted rounding of T then another zero restricted rounding $\bar{\sigma}$ can be constructed from σ by rounding each nonzero entry of T in the opposite direction. If T has a zero restricted rounding σ then T has a controlled rounding. The controlled rounding can be obtained by picking σ with probability $\frac{1}{2}$ and $\bar{\sigma}$ with probability $\frac{1}{2}$. Therefore T has a zero restricted rounding if and only if it has a controlled rounding. We have already shown in theorem 1 that that T has a zero restricted rounding if and only if the original 2-*in*-4 SAT instance was solvable. Therefore we may conclude that the controlled three dimensional rounding problem is NP-hard. ♠

It is interesting to note that the table T constructed in the above theorems has only one nonzero entry in each row, and hence only the sheet additivity conditions were necessary to prove the problems NP-hard. This implies the following two corollaries.

Corollary 1 Given a three dimensional table T and a rounding base B it is NP-complete to decide whether T has a rounding that preserves sheet sums.

Corollary 2 Given a three dimensional table T and a rounding base B it is NP-complete to decide whether T has an unbiased rounding that preserves sheet sums.

4. Conclusions and Open Questions

It has been shown that some three dimensional tables do not have zero restricted roundings [CCE]. We have proved that it is very unlikely that there is an efficient algorithm that can decide which tables have zero restricted roundings. This implies that it is also very unlikely that there is an efficient algorithm that generates a zero restricted rounding for those three dimensional tables that have zero restricted roundings. For such an algorithm could be used to decide which tables had zero restricted roundings.

Our results indicate that any attempt to develop an efficient algorithm for the general three dimensional zero restricted rounding problem will probably be unsuccessful, since such an algorithm would imply that all NP-complete problems are efficiently solvable. This does not rule out the possibility that important special cases of the three dimensional zero restricted rounding problem are efficiently solvable. One special case that is of interest is when the size of one or two of the dimensions is small.

Another avenue for further investigation is approximation algorithms. For many applications we would be willing to settle for a rounding that was in some sense almost zero restricted or almost controlled. One possibility for loosening the restrictions would be by relaxing the rounding constraints. For example, we may not want to require that multiples of the rounding base remain fixed. We could allow these entries to be rounded to one of the two adjacent integer multiples of the rounding base. With these constraints it is not known if all three dimensional tables possess zero restricted roundings [Cox]. We conjecture that there are tables of higher dimension that do not possess zero restricted roundings under these constraints. We further conjecture that it is NP-complete to decide which of these tables possess zero restricted roundings under these constraints.

5. References

[AHU] Aho, Hopcroft and Ullman, *The Design and Analysis of Computer Algorithms*, Addison - Wesley(1974).

[CFGH] Cox, Fagan, Greenberg and Hemmig, "Research at the Census Bureau into Disclosure

Avoidance Techniques for Tabular Data," *Proceedings of the Survey Research Methods Section, American Statistical Association*, to appear.

[CMN] Cox, McDonald and Nelson, "Confidentiality Issues at the U.S. Bureau of Census," *Journal of Official Statistics*, 2, 2, 135-160, (1986).

[Coc] W.G. Cochran, *Sampling Techniques*, (*3rd* addition), John Wiley, New York, (1977).

[Cox] Lawrence Cox, "A Constructive Procedure for Unbiased Controlled Rounding," to appear in JASA.

[Edm] J. Edmonds,"Paths, Flowers, and Trees," *Canadian Journal of Mathematics*, 17, 449-467, (1965).

[GJ] Gary and Johnson, *Computers and Intractability; A Guide to the Theory of NP-completeness*, W.H. Freeman and Company, (1979).

[New] Newman, "Techniques for Ensuring the Confidentiality of Census Information in Great Britain," *Proceedings of the 40th Session, International Statistical Institute*, (1975).

[NS] Nargundkar and Saveland, "Random Rounding to Prevent Statistical Disclosures," *Proceedings of the Social Statistics Section, American Statistical Association*, 382-385, (1972).

[PS] Papadimitriou and Steiglitz, *Combinatorial Optimization; Algorithms and Complexity*, Prentice - Hall, (1982).

[Sch] T.J. Schaefer, "The Complexity of Satisfiability Problems," *Proceedings of the 10th Annual ACM Symposium on Theory of Computing*, 216-226, (1978).