

Unbiased Matrix Rounding

Tobias Friedrich

*Max-Planck-Institut für Informatik
Saarbrücken, Germany*

Benjamin Doerr, Christian Klein, and Ralf Oswald

*Max-Planck-Institut für Informatik
Saarbrücken, Germany*

Abstract

We show several ways to round a real matrix to an integer one such that the rounding errors in all rows and columns as well as the whole matrix are less than one. This is a classical problem with applications in many fields, in particular, statistics.

We improve earlier solutions of different authors in two ways. For rounding matrices of size $m \times n$, we reduce the runtime from $O((mn)^2)$ to $O(mn \log(mn))$. Second, our roundings also have a rounding error of less than one in all initial intervals of rows and columns. Consequently, arbitrary intervals have an error of at most two. This is particularly useful in the statistics application of controlled rounding.

The same result can be obtained via (dependent) randomized rounding. This has the additional advantage that the rounding is unbiased, that is, for all entries y_{ij} of our rounding, we have $E(y_{ij}) = x_{ij}$, where x_{ij} is the corresponding entry of the input matrix.

Keywords: Rounding, Controlled rounding, Approximation algorithm, Linear discrepancy.

1 Introduction

In this paper, we analyze a rounding problem with strong connections to statistics, but also to different areas in discrete mathematics, computer science, and operations research. We show how to round a matrix to an integer one such that rounding errors in intervals of rows and columns are small.

Let m, n be positive integers. For some set S , we write $S^{m \times n}$ to denote the set of $m \times n$ matrices with entries in S . For real numbers a, b let $[a..b] := \{z \in \mathbb{Z} \mid a \leq z \leq b\}$. We show the following.

Theorem 1.1 *For all $X \in [0, 1]^{m \times n}$ a rounding $Y \in \{0, 1\}^{m \times n}$ such that*

$$\begin{aligned} \forall b \in [1..n], i \in [1..m] : \left| \sum_{j=1}^b (x_{ij} - y_{ij}) \right| < 1, \\ \forall b \in [1..m], j \in [1..n] : \left| \sum_{i=1}^b (x_{ij} - y_{ij}) \right| < 1, \\ \left| \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - y_{ij}) \right| < 1 \end{aligned}$$

can be computed in time $O(mn \log(mn))$.

This result extends the famous rounding lemma of Baranyai [3] and several results on controlled rounding in statistics by Bacharach [2] and Causey, Cox and Ernst [6].

2 Baranyai's Rounding Lemma and Applications in Statistics

Baranyai [3] used a weaker version of Theorem 1.1 to obtain his well-known results on coloring and partitioning complete uniform hypergraphs. He showed that any matrix can be rounded such that the errors in all rows, all columns and the whole matrix are less than one. He used a formulation as flow problem to prove this statement. This yields an inferior runtime than the bound in Theorem 1.1. However, algorithmic issues were not his focus.

In statistics, Baranyai's result was independently obtained by Bacharach [2] (in a slightly weaker form) and again independently by Causey, Cox and Ernst [6]. There are two statistical applications for such rounding results. Note first that instead of rounding to integers, our result also applies to rounding to multiples of any other base (e.g., multiples of 10). Such a rounding can

be used to improve the readability of data tables.

The main reason, however, to apply such a rounding procedure is confidentiality protection. Frequency counts that directly or indirectly disclose small counts may permit the identification of individual respondents. There are various methods to prevent this [21], one of which is *controlled rounding* [8]. Here, one tries to round an $(m + 1) \times (n + 1)$ -table \tilde{X} given by

$$\begin{array}{c|c} (x_{ij})_{\substack{i=1\dots m \\ j=1\dots n}} & \left(\sum_{j=1}^n x_{ij}\right)_{i=1\dots m} \\ \hline \left(\sum_{i=1}^m x_{ij}\right)_{j=1\dots n} & \sum_{i=1}^m \sum_{j=1}^n x_{ij} \end{array}$$

to an $(m + 1) \times (n + 1)$ -table \tilde{Y} such that additivity is preserved, i.e., the last row and column of \tilde{Y} contain the associated totals of \tilde{Y} . In our setting we round the $m \times n$ -matrix X defined by the mn inner cells of the table \tilde{X} to obtain a controlled rounding.

The additivity in the rounded table allows to derive information on the row and column totals of the original table. In contrast to other rounding algorithms, our result also permits to retrieve further reliable information from the rounded matrix, namely on the sums of consecutive elements in rows or columns. Such queries may occur if there is a linear ordering on statistical attributes. Here an example. Let x_{ij} be the number of people in country i that are j years old. Say Y is such that $\frac{1}{1000}Y$ is a rounding of $\frac{1}{1000}X$ as in Theorem 1.1. Now $\sum_{j=20}^{40} y_{ij}$ is the number of people in country i that are between 20 and 40 years old, apart from an error of less than 2000. Note that such guarantees are not provided by the results of Baranyai [3], Bacharach [2], and Causey, Cox and Ernst [6].

3 Unbiased Rounding

We present a randomized algorithm computing roundings as in Theorem 1.1. It has the additional property that each matrix entry is rounded up with probability equal to its fractional value. This is known as randomized rounding [16] in computer science and as unbiased controlled rounding [7,12] in statistics. Here, a controlled rounding is computed such that the expected values of each table entry (including the totals) equals its fractional value in the original table.

To state our result more precisely, we introduce the following notation. For $x \in \mathbb{R}$ write $\lfloor x \rfloor := \max\{z \in \mathbb{Z} \mid z \leq x\}$, $\lceil x \rceil := \min\{z \in \mathbb{Z} \mid z \geq x\}$ and $\{x\} := x - \lfloor x \rfloor$.

Definition 3.1 Let $x \in \mathbb{R}$. A random variable y is called *randomized rounding* of x , denoted $y \approx x$, if $\Pr(y = \lfloor x \rfloor + 1) = \{x\}$ and $\Pr(y = \lfloor x \rfloor) = 1 - \{x\}$. For a matrix $X \in \mathbb{R}^{m \times n}$, we call an $m \times n$ matrix-valued random variable Y randomized rounding of X if $y_{ij} \approx x_{ij}$ for all $i \in [1..m], j \in [1..n]$.

We then get the following randomized version of Theorem 1.1.

Theorem 3.2 Let $X \in [0, 1]^{m \times n}$ be a matrix having entries of binary length at most ℓ . Then a randomized rounding Y fulfilling the additional constraints that

$$\begin{aligned} \forall b \in [1..n], i \in [1..m]: \sum_{j=1}^b x_{ij} &\approx \sum_{j=1}^b y_{ij}, \\ \forall b \in [1..m], j \in [1..n]: \sum_{i=1}^b x_{ij} &\approx \sum_{i=1}^b y_{ij}, \\ \sum_{i=1}^m \sum_{j=1}^n x_{ij} &\approx \sum_{i=1}^m \sum_{j=1}^n y_{ij} \end{aligned}$$

can be computed in time $O(mn\ell)$.

For a matrix with arbitrary entries $x_{ij} := \sum_{d=1}^{\ell} x_{ij}^{(d)} 2^{-d} + x'_{ij}$ where $x'_{ij} < 2^{-\ell}$ and $x_{ij}^{(d)} \in \{0, 1\}$ for $i \in [1..m], j \in [1..n], d \in [1..\ell]$, we may use the ℓ highest bits to get an approximate randomized rounding. If (before doing so) we round the remaining part x'_{ij} of each entry to $2^{-\ell}$ with probability $2^{\ell} x'_{ij}$ and to 0 otherwise, we still have that $Y \approx X$, but we introduce an additional error of at most $2^{-\ell} mn$ in the constraints of Theorem 3.2.

4 Other Applications

One of the most basic rounding results states that any sequence x_1, \dots, x_n of numbers can be rounded to an integer one y_1, \dots, y_n such that the rounding errors $|\sum_{j=a}^b (x_j - y_j)|$ are less than one for all $a, b \in [1..n]$. Such roundings can be computed efficiently in linear time by a one-pass algorithm resembling Kadane's scanning algorithm (described in Bentley's Programming Pearls [4]). Extensions in different directions have been obtained in [9,10,13,17,19]. This

rounding problem has found a number of applications, among others in image processing [1,18].

Theorem 1.1 extends this result to two-dimensional sequences. Here the rounding error in arbitrary intervals of a row or column is less than two. In [11] a lower bound of 1.5 is shown for this problem. Thus an error of less than one as in the one-dimensional case cannot be achieved.

Rounding a matrix while considering the errors in column sums and partial row sums also arises in scheduling [5,14,15,20]. For this, however, one does not need our result in full generality. It suffices to use the linear-time one-pass algorithm given in [11]. This algorithm rounds a matrix having unit column sums and can be extended to compute a quasi rounding for arbitrary matrices. While this algorithm keeps the error in all initial row intervals small, for columns only the error over the whole column is considered.

References

- [1] T. Asano. Digital halftoning: Algorithm engineering challenges. *IEICE Trans. on Inf. and Syst.*, E86-D:159–178, 2003.
- [2] M. Bacharach. Matrix rounding problems. *Management Science (Series A)*, 12:732–742, 1966.
- [3] Zs. Baranyai. On the factorization of the complete uniform hypergraph. In *Infinite and finite sets (Colloq., Keszthely, 1973; dedicated to P. Erdős on his 60th birthday)*, Vol. I, pages 91–108. Colloq. Math. Soc. János Bolyai, Vol. 10. North-Holland, Amsterdam, 1975.
- [4] J. L. Bentley. Algorithm design techniques. *Commun. ACM*, 27:865–871, 1984.
- [5] N. Brauner and Y. Crama. The maximum deviation just-in-time scheduling problem. *Discrete Appl. Math.*, 134:25–50, 2004.
- [6] B. D. Causey, L. H. Cox, and L. R. Ernst. Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, 80:903–909, 1985.
- [7] L. H. Cox. A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82:520–524, 1987.
- [8] L. H. Cox and L. R. Ernst. Controlled rounding. *Informes*, 20:423–432, 1982.
- [9] B. Doerr. Global roundings of sequences. *Information Processing Letters*, 92:113–116, 2004.

- [10] B. Doerr. Linear discrepancy of totally unimodular matrices. *Combinatorica*, 24:117–125, 2004.
- [11] B. Doerr, T. Friedrich, C. Klein, and R. Osbild. Rounding of sequences and matrices, with applications. In *Third Workshop on Approximation and Online Algorithms*, volume 3879 of *Lecture Notes in Computer Science*, pages 96–109. Springer, 2006.
- [12] I. P. Fellegi. Controlled random rounding. *Survey Methodology*, 1:123–133, 1975.
- [13] D. E. Knuth. Two-way rounding. *SIAM J. Discrete Math.*, 8:281–290, 1995.
- [14] Y. Monden. What makes the Toyota production system really tick? *Industrial Eng.*, 13:36–46, 1981.
- [15] Y. Monden. *Toyota Production System*. Industrial Engineering and Management Press, Norcross, GA, 1983.
- [16] P. Raghavan. Probabilistic construction of deterministic algorithms: Approximating packing integer programs. *J. Comput. Syst. Sci.*, 37:130–143, 1988.
- [17] K. Sadakane, N. Takki-Chebihi, and T. Tokuyama. Combinatorics and algorithms on low-discrepancy roundings of a real sequence. In *ICALP 2001*, volume 2076 of *Lecture Notes in Computer Science*, pages 166–177, Berlin Heidelberg, 2001. Springer-Verlag.
- [18] K. Sadakane, N. Takki-Chebihi, and T. Tokuyama. Discrepancy-based digital halftoning: Automatic evaluation and optimization. In *Geometry, Morphology, and Computational Imaging*, volume 2616 of *Lecture Notes in Computer Science*, pages 301–319, Berlin Heidelberg, 2003. Springer-Verlag.
- [19] J. Spencer. *Ten lectures on the probabilistic method*, volume 64 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.
- [20] G. Steiner and S. Yeomans. Level schedules for mixed-model, just-in-time processes. *Management Science*, 39:728–735, 1993.
- [21] L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*, volume 155 of *Lecture Notes in Statistics*. Springer, 2001.