

Matrix Rounding with Respect to Small Submatrices

Benjamin Doerr*

January 4, 2005

Abstract

We show that any real valued matrix A can be rounded to an integer one B such that the error in all 2×2 (geometric) submatrices is less than 1.5, that is, we have $|a_{ij} - b_{ij}| < 1$ and $|\sum_{k=i}^{i+1} \sum_{\ell=j}^{j+1} (a_{k\ell} - b_{k\ell})| < 1.5$ for all i, j .

1 Introduction and Results

Let m, n be non-negative integers. We write $[n] = \{i \in \mathbb{N} \mid i \leq n\}$. Let A and B be real-valued $m \times n$ matrices. We call B a *rounding* of A if $b_{ij} \in \{[a_{ij}], \lceil a_{ij} \rceil\}$ for all $i \in [m], j \in [n]$. For $R \subseteq [m] \times [n]$ let $d(A, B, R) = \sum_{(i,j) \in R} (a_{ij} - b_{ij})$. A set $\{i, i+1\} \times \{j, j+1\}$ for some $i \in [m-1], j \in [n-1]$ shall be called a 2×2 *box*. Denote by \mathcal{R} the set of all 2×2 boxes and put

$$d(A, B) = \max_{R \in \mathcal{R}} |d(A, B, R)|.$$

In the context of an image processing application (digital halftoning), Asano, Matsui and Tokuyama [AMT00] proved that for any A there is a rounding

*Mathematisches Seminar II, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany, bed@numerik.uni-kiel.de

B such that $d(A, B) \leq 1.75$. This was improved to a bound of $5/3$ by Asano and Tokuyama [AT01]. Both proofs are relatively complicated and involve difficult case distinctions.

The object of this note is to give a short proof of an upper bound of 1.5 , and a slightly longer one showing that this bound is never attained.

Theorem 1. *For any $A \in \mathbb{R}^{m \times n}$ there is a rounding $B \in \mathbb{Z}^{m \times n}$ such that $d(A, B) < 1.5$.*

We note that there is a lower bound of 1 stemming from an odd cycle argument (cf. the two papers cited above and the example at the end of this paper).

2 Proof of the Bound $d(A, B) \leq 1.5$

We start with two elementary lemmas.

Lemma 2. *Let $m, n \in \mathbb{N}$. Let \mathcal{E} be a set of subsets of $[m] \times [n]$ such that*

- (i) $|E| = 3$ for all $E \in \mathcal{E}$,
- (ii) for each $E \in \mathcal{E}$ there is an $i \in [m]$ (which shall be denoted by $i(E)$) such that $E \subseteq \{i\} \times [n]$,
- (iii) for each two $E_1, E_2 \in \mathcal{E}$ such that $i := i(E_1) = i(E_2)$, we have $e_1 < e_2$ for all $(i, e_1) \in E_1, (i, e_2) \in E_2$, or $e_1 > e_2$ for all $(i, e_1) \in E_1, (i, e_2) \in E_2$. We write $E_1 < E_2$ in the first case, and $E_1 > E_2$ in the second.

Then there is a $T \subseteq [m] \times [n]$ such that $|T \cap E| = 1$ for all $E \in \mathcal{E}$ and such that for all $(s_1, s_2), (t_1, t_2) \in T$ we have $s_2 \neq t_2$ whenever $|s_1 - t_1| = 1$.

Proof. We use induction on $|\mathcal{E}|$. For $|\mathcal{E}| = 0$, there is nothing to show. Hence let us assume that $|\mathcal{E}| \geq 1$ and that the assertion of the lemma is true for smaller set systems. Let $E^* \in \mathcal{E}$ have least extension to the right,

i.e., $\max\{j \mid (i(E^*), j) \in E^*\} \leq \max\{j \mid (i(E), j) \in E\}$ for all $E \in \mathcal{E}$. Let $i^* = i(E^*)$.

Let T be as assured by the lemma with respect to the set system $\mathcal{E} \setminus \{E^*\}$. By construction, there is at most one set $E \in \mathcal{E}$ in each of the rows $i^* - 1$ and $i^* + 1$ such that $\{j \mid (i(E), j) \in E\}$ intersects $\{j \mid (i^*, j) \in E^*\}$ non-trivially. Since T has exactly one vertex in these (at most) two sets and $|E^*| = 3$, there is a $j \in [n]$ such that $(i^* - 1, j) \notin T$, $(i^* + 1, j) \notin T$ and $(i^*, j) \in E^*$. Thus $T \cup \{(i^*, j)\}$ satisfies the claim. \square

Lemma 3. *Let $n \in \mathbb{N}$. Let $a_1, a_n \in \{0, 1\}$ and $a_2, \dots, a_{n-1} \in \{\frac{1}{3}, \frac{2}{3}\}$. Then*

- (i) *there are $b_1, \dots, b_n \in \{0, 1\}$ such that $b_1 = a_1$, $b_n = a_n$ and $|b_i - a_i + b_{i+1} - a_{i+1}| \leq \frac{1}{3}$ for all $i \in [n - 1]$,*
- (ii) *or there are three distinct numbers $x^{(1)}, x^{(2)}, x^{(3)} \in [n - 1]$ and for each $k \in [3]$ there are $b_1^{(k)}, \dots, b_n^{(k)} \in \{0, 1\}$ such that $b_1^{(k)} = a_1$, $b_n^{(k)} = a_n$ and for all $i \in [n - 1] \setminus \{x^{(k)}\}$ we have $|b_i^{(k)} - a_i + b_{i+1}^{(k)} - a_{i+1}| \leq \frac{1}{3}$, whereas $|b_{x^{(k)}}^{(k)} - a_{x^{(k)}} + b_{x^{(k)}+1}^{(k)} - a_{x^{(k)}+1}| = \frac{2}{3}$.*

Proof. Assume that (i) does not hold. Then there is a $j \in \{2, \dots, n - 2\}$ such that $a_j = a_{j+1}$, as otherwise putting $b_i = \lfloor a_i + \frac{1}{2} \rfloor$ satisfies (i). Choose j minimal subject to these conditions. For all $k \in [3]$, let $b_1^{(k)} = a_1$ and $b_n^{(k)} = a_n$. For all $i \in \{2, \dots, n - 1\}$ let $b_i^{(1)} = 0$, if i is even, and $b_i^{(1)} = 1$, if i is odd. Put $b_i^{(2)} = 1 - b_i^{(1)}$. Then for exactly one $k \in [2]$, $|b_1^{(k)} - a_1 + b_2^{(k)} - a_2| = \frac{2}{3}$, and for exactly one $\ell \in [2]$, $|b_{n-1}^{(\ell)} - a_{n-1} + b_n^{(\ell)} - a_n| = \frac{2}{3}$. We have $k \neq \ell$, as otherwise one of the two roundings satisfies (i). By construction, we have $|b_i^{(k)} - a_i + b_{i+1}^{(k)} - a_{i+1}| \leq \frac{1}{3}$ for all other $(k, i) \in [2] \times [n - 1]$.

Now define $(b_i^{(3)})$ as follows: For $i \leq j + 1$, let $b_i^{(3)} = \lfloor a_i + \frac{1}{2} \rfloor$ be a rounding of a_i . For $i \in \{j + 2, \dots, n - 2\}$, let $b_i^{(3)} = 1 - b_{i-1}^{(3)}$ be defined inductively. Note that $|b_{n-1}^{(3)} - a_{n-1} + b_n^{(3)} - a_n| = \frac{1}{3}$, as otherwise replacing $b_i^{(3)}$ by $1 - b_i^{(3)}$ for all $i \geq j + 1$ would satisfy (i) again. Hence we have $|b_i^{(3)} - a_i + b_{i+1}^{(3)} - a_{i+1}| \leq \frac{1}{3}$ for all $i \in [n - 1] \setminus \{j\}$ and $|b_j^{(3)} - a_j + b_{j+1}^{(3)} - a_{j+1}| = \frac{2}{3}$. This proves the claim. \square

The heart of the proof is the following lemma concerning roundings of matrices with entries in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$.

Lemma 4. *For any $A \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}^{m \times n}$ there is a rounding $B \in \{0, 1\}^{m \times n}$ such that $d(A, B) \leq 1$.*

Proof. Let $A \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}^{m \times n}$. We briefly sketch the main idea. The rows of A are almost treated separately. Each row consists of strings of $\frac{1}{3}$ s and $\frac{2}{3}$ s, separated by zeroes and ones. All zeroes and ones shall be unchanged in B . For the intermediate strings of $\frac{1}{3}$ and $\frac{2}{3}$, Lemma 3 yields a rounding with error at most $\frac{1}{3}$ in each two consecutive entries except for possibly one location, where an error of $\frac{2}{3}$ cannot be avoided. Should this occur, however, we may choose this location out of at least three different possibilities. Invoking Lemma 2, we shall round the rows in such a way that this large error never occurs at the same location of two adjacent rows. Thus each 2×2 box contains at most one error of $\frac{2}{3}$ in one of its (two) rows, leading to a total error of at most 1 in the box.

Let us be precise. Without loss of generality, we may assume that $a_{i1} = a_{in} = 0$ for all $i \in [m]$. A *non-integral string* in A is a set $S = \{i\} \times \{j_1, \dots, j_2\} \subseteq [m] \times [n]$ such that $j_2 \geq j_1 + 2$ and for all $(i, j) \in S$, a_{ij} is integral if and only if $j \in \{j_1, j_2\}$. Write $i(S) := i$ for this unique row i . For each such non-integral string S , apply Lemma 3 to the sequence $a_{ij_1}, \dots, a_{ij_2}$. If there is a rounding of this sequence with all errors in two consecutive elements being at most $\frac{1}{3}$ (option (i) in the lemma), then choose $b_{ij_1}, \dots, b_{ij_2}$ accordingly. Otherwise there are three sequences $b_{ij_1}^{(k,S)}, \dots, b_{ij_2}^{(k,S)}$, $k \in [3]$, together with distinct $x^{(1,S)}, x^{(2,S)}, x^{(3,S)} \in \{j_1, \dots, j_2 - 1\}$ such that $|b_{ij}^{(k,S)} - a_{ij} + b_{i,j+1}^{(k,S)} - a_{i,j+1}| \leq \frac{1}{3}$ for all $j \in \{j_1, \dots, j_2 - 1\} \setminus \{x^{(k,S)}\}$ and $|b_{ix^{(k,S)}}^{(k,S)} - a_{ix^{(k,S)}} + b_{i,x^{(k,S)}+1}^{(k,S)} - a_{i,x^{(k,S)}+1}| = \frac{2}{3}$. Denote by \mathcal{S} the set of non-integral strings where option (i) in Lemma 3 is not satisfied.

Let $\mathcal{E} = \{\{i(S)\} \times \{x^{(1,S)}, x^{(2,S)}, x^{(3,S)}\} \mid S \in \mathcal{S}\}$. Let T be as in Lemma 2. For each $S = \{i\} \times \{j_1, \dots, j_2\} \in \mathcal{S}$ put $(b_{ij_1}, \dots, b_{ij_2}) = (b_{ij_1}^{(k,S)}, \dots, b_{ij_2}^{(k,S)})$, where k is such that $(i(S), x^{(k,S)}) \in T$.

Finally, put $b_{ij} = a_{ij}$ for all remaining i, j . Let $R = \{i, i+1\} \times \{j, j+1\}$ be a box. If $|b_{ij} - a_{ij} + b_{i,j+1} - a_{i,j+1}| \geq \frac{2}{3}$, then $(i, j) \in T$. Since in this case

$(i+1, j) \notin T$, we have $|b_{i+1,j} - a_{i+1,j} + b_{i+1,j+1} - a_{i+1,j+1}| \leq \frac{1}{3}$ by construction. Hence $|d(A, B, R)| \leq 1$. \square

The proof of the weaker bound now follows easily from regarding the ternary expansion of A and rounding “digit by digit” in a similar manner as in Beck and Spencer [BS84] (there done with binary expansions). We say that a matrix $A \in [0, 1]^{m \times n}$ has ternary length ℓ for some $\ell \in \mathbb{N}$, if there are $A_1, \dots, A_\ell \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}^{m \times n}$ such that $A = \sum_{i=1}^{\ell} 3^{-i+1} A_i$ and $A_\ell \neq 0$. It has ternary length zero, if it is binary.

Proof of the bound $d(A, B) \leq 1.5$. Ignoring the integral parts of A , we may assume that $A \in [0, 1]^{m \times n}$ and thus have to show the existence of a binary B such that $d(A, B) \leq 1.5$ and $b_{ij} = a_{ij}$ whenever a_{ij} is integral. Assume that A has a finite ternary expansion of length ℓ . Let $A_\ell \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}^{m \times n}$ such that $A - 3^{-\ell+1} A_\ell$ has ternary length less than ℓ . Let B_ℓ be a rounding of A_ℓ as in Lemma 4. Then $\tilde{A} = A - 3^{-\ell+1}(A_\ell - B_\ell)$ has ternary length less than ℓ and $d(A, \tilde{A}) = 3^{-\ell+1} d(A_\ell, B_\ell) \leq 3^{-\ell+1}$. Hence an easy induction yields a binary B such that $d(A, B) \leq \sum_{i=0}^{\ell-1} 3^{-i} \leq 1.5$. Note that by Lemma 4 we also have $b_{ij} = a_{ij}$ whenever a_{ij} is integral. Density of the set of all matrices having finite ternary expansion and the fact that $A' \mapsto \min\{d(A', B) \mid B \text{ is a rounding of } A\}$ is continuous in A , yield the claim for arbitrary A . \square

Our result above can be extended to infinite matrices as well. The easiest way to do so involves the “compactness principle”. Alternatively, one can extend Lemma 2 and 4 in a suitable way to obtain a sequence of matrices of increasing size that contain earlier members of the sequence and that all are roundings with error at most 1.5 of the corresponding parts of A .

We may add that the analysis of Lemma 4 is sharp, even if we omit the rounding requirement. Let

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 & 0 & \frac{2}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then $d(A, B) \geq 1$ holds for all integral B : If not, then $b_{ij} = 0$ whenever $a_{ij} = 0$, as all zeros of A are contained in a 2×2 box containing zeroes only. Thus, in the circle of $\frac{1}{3}$ s and $\frac{2}{3}$ s in column 3 to 6, all $\frac{1}{3}$ s have to become zeroes in B and all $\frac{2}{3}$ s ones. The same holds in the circle in the right half of the matrix. Now any choice of values for the remaining two entries generates a 2×2 box R with $|d(A, B, R)| \geq 1$.

This example shows again the lower bound of 1 for the matrix rounding problem, and also that an improvement of Lemma 4 (which would immediately show an upper bound of 1 for the general case) is not possible.

3 Improvement to Strictly Less Than 1.5

In this section, we show how to extend the proof of the previous section to obtain a bound of strictly less than 1.5. Since we are only able to show $1.5 - \exp(-\Theta(mn))$, we decided not to spoil the proof above, but to do this in a separate section.

The key idea is to make sure that for each box R not all $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ intermediate roundings have the same error $d(A, B, R)$ of -1 or 1 . To this end, we modify Lemma 4 to ensure that a particular box R does not get a particular error $\varepsilon \in \{-1, 1\}$, i.e., that $d(A, B, R) \neq \varepsilon$. We start with a slight enhancement of Lemma 2 stating that we may exclude an arbitrary point from possibly being in the transversal.

Lemma 5. *In the situation of Lemma 2, for all $x \in [m] \times [n]$, there is a $T \subseteq ([m] \times [n]) \setminus \{x\}$ such that $|T \cap E| = 1$ for all $E \in \mathcal{E}$ and such that for all $(s_1, s_2), (t_1, t_2) \in T$ we have $s_2 \neq t_2$ whenever $|s_1 - t_1| = 1$.*

Proof. The proof is along the lines of the proof of Lemma 2. In the inductive step, if $x \notin E^*$, we face no problems. Hence the only difficulty occurs if we have $x \in E$ for all $E \in \mathcal{E}$ that have least extension to the left and (by symmetry) all $E \in \mathcal{E}$ that have least extension to the right. In this case, x is contained in some $E \in \mathcal{E}$ that is both the unique set in \mathcal{E} having least extension to the left and the unique set in \mathcal{E} having least extension to the right. However, if such a set exists, each row $\{i\} \times [n]$ contains at most one set of \mathcal{E} , in which case it is trivial to find a T as desired. \square

Lemma 6. *Let $A \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}^{m \times n}$, $R \in \mathcal{R}$ and $\varepsilon \in \{-1, 1\}$. Then there is a rounding B of A such that $d(A, B) \leq 1$ and $d(A, B, R) \neq \varepsilon$.*

Proof. Let T and B be as in the proof of Lemma 4. Assume that $d(A, B, R) = \varepsilon$. Let $R = \{i, i+1\} \times \{j, j+1\}$. By symmetry, assume that $b_{ij} - a_{ij} + b_{i,j+1} - a_{i,j+1} = \varepsilon \frac{2}{3}$ and $b_{i+1,j} - a_{i+1,j} + b_{i+1,j+1} - a_{i+1,j+1} = \varepsilon \frac{1}{3}$. Then $(i, j) \in T$. Apply Lemma 5 with $x = (i, j)$ to obtain T' . Use T' to construct B' as in the proof of Lemma 4. It remains to show that $d(A, B', R) \neq \varepsilon$. Since Lemma 3 yields only two possible values for $b'_{ij} - a_{ij} + b'_{i,j+1} - a_{i,j+1}$ and $|b'_{ij} - a_{ij} + b'_{i,j+1} - a_{i,j+1}| = \frac{2}{3}$ is not possible due to $(i, j) \notin T'$, we conclude $b'_{ij} - a_{ij} + b'_{i,j+1} - a_{i,j+1} = -\varepsilon \frac{1}{3}$. Also, we have $b'_{i+1,j} - a_{i+1,j} + b'_{i+1,j+1} - a_{i+1,j+1} \in \{\varepsilon \frac{1}{3}, -\varepsilon \frac{2}{3}\}$. Hence $d(A, B', R) \in \{0, -\varepsilon\}$. \square

Now it is easy to prove the stronger version of the theorem.

Proof. Let $M = 2(m-1)(n-1)$. Let $(R_1, \varepsilon_1), \dots, (R_M, \varepsilon_M)$ be an enumeration of $\mathcal{R} \times \{-1, 1\}$. For sake of notational convenience, let $(R_i, \varepsilon_i) \in \mathcal{R} \times \{-1, 1\}$ be arbitrary for $i > M$. Let $A^{(\ell)} := A \in [0, 1]^{m \times n}$ have finite ternary expansion of length ℓ . Inductively, we define a sequence $A^{(i)} \in [0, 1]^{m \times n}$, $i = \ell - 1, \dots, 0$ such that $A^{(i)}$ has ternary length at most i , $d(A^{(i+1)}, A^{(i)}) \leq 3^{-i}$ and $d(A^{(i+1)}, A^{(i)}, R_{i+1}) \neq \varepsilon_{i+1} 3^{-i}$. Assume that $A^{(i+1)}$ is already defined. Let A_{i+1} be a $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ matrix such that $A^{(i+1)} - 3^{-i} A_{i+1}$ has ternary length at most i . Apply Lemma 6 on A_{i+1} , R_{i+1} , ε_{i+1} to obtain a rounding B_{i+1} . Put $A^{(i)} = A^{(i+1)} - 3^{-i}(A_{i+1} - B_{i+1})$. Now $A^{(i)}$ has the

desired properties. In particular, $A^{(0)} \in \{0, 1\}^{m \times n}$. We compute $d(A^{(\ell)}, A^{(0)})$. Let $R \in \mathcal{R}$. Let $j \in [M]$ such that $R = R_j$ and $\varepsilon_j = 1$. Then

$$1.5 - d(A^{(\ell)}, A^{(0)}, R) \geq 1.5 - \sum_{i=0}^{\ell-1} 3^{-i} + \frac{1}{3}3^{-j+1} \geq 3^{-j}.$$

Similarly, $d(A^{(\ell)}, A^{(0)}, R) - 1.5 \leq -3^{-k}$, where $k \in [M]$ is such that $R = R_k$ and $\varepsilon_k = -1$. Hence $d(A^{(\ell)}, A^{(0)}) \leq 1.5 - 3^{-M}$. \square

We did not try to optimize the 3^{-M} term. A closer inspection of Lemma 4 shows that a positive fraction (say $\alpha|\mathcal{R}|$) of all boxes do not have an error $d(A, B, R) \in \{-1, 1\}$. Hence we can use Lemma 4 up to the definition of $A^{((1-\alpha)|\mathcal{R}|)}$ and then use Lemma 6 only in the last $(1-\alpha)|\mathcal{R}|$ rounds to ensure that the at most $(1-\alpha)|\mathcal{R}|$ boxes R such that $d(A^{((1-\alpha)|\mathcal{R}|+1)}, A^{((1-\alpha)|\mathcal{R}|)}, R) \in \{-1, 1\}$ do not get this same error in all remaining rounds.

Acknowledgments

This work was done while the author was visiting Joel Spencer at the Courant Institute of Mathematical Sciences, New York City. I would like to thank both him and the institute for providing me with this great opportunity. Thank you to Josh Cooper, Roberto Oliveira and Joel Spencer also for their stimulating interest in this problem.

References

- [AMT00] T. Asano, T. Matsui, and T. Tokuyama. Optimal roundings of sequences and matrices. *Nordic J. Comput.*, 7:241–256, 2000.
- [AT01] T. Asano and T. Tokuyama. How to color a checkerboard with a given distribution—matrix rounding achieving low 2×2 -discrepancy. In *Algorithms and Computation (Christchurch, 2001)*, volume 2223 of *Lecture Notes in Computer Science*, pages 636–648, Berlin, 2001. Springer.

- [BS84] J. Beck and J. Spencer. Integral approximation sequences. *Math. Programming*, 30:88–98, 1984.