

Typical Rounding Problems

Benjamin Doerr

Mathematisches Seminar II, Christian-Albrechts-Universität zu Kiel,
Ludewig-Meyn-Str. 4, D-24098 Kiel, Germany,
bed@numerik.uni-kiel.de,
WWW home page: <http://www.numerik.uni-kiel.de/~bed/>

Abstract. The linear discrepancy problem is to round a given $[0, 1]$ -vector x to a binary vector y such that the rounding error with respect to a linear form is small, i.e., such that $\|A(x - y)\|_\infty$ is small for some given matrix A . The discrepancy problem is the special case of $x = (\frac{1}{2}, \dots, \frac{1}{2})$. A famous result of Beck and Spencer (1984) as well as Lovász, Spencer and Vesztergombi (1986) shows that the linear discrepancy problem is not much harder than this special case: Any linear discrepancy problem can be solved with at most twice the maximum rounding error among the discrepancy problems of the submatrices of A .

In this paper we strengthen this result for the common situation that the discrepancy of submatrices having n_0 columns is bounded by Cn_0^α for some $C > 0, \alpha \in (0, 1]$. In this case, we improve the constant by which the general problem is harder than the discrepancy one, down to $2(\frac{2}{3})^\alpha$. We also find that a random vector x has expected linear discrepancy $2(\frac{1}{2})^\alpha Cn^\alpha$ only. Hence in the typical situation that the discrepancy is decreasing for smaller matrices, the linear discrepancy problem is even less difficult compared to the discrepancy one than assured by the results of Beck and Spencer and Lovász, Spencer and Vesztergombi.

Key words: rounding, discrepancy, games.

1 Introduction

In this paper we deal with rounding problems, and in particular with the question how much easier it is to round a vector with all entries $\frac{1}{2}$ compared to the general case of $[0, 1]$ -vectors. A famous result of Beck and Spencer [4] and Lovász, Spencer and Vesztergombi [8] shows that the general problem can be reduced to the $\frac{1}{2}$ -case. In this paper we refine their result for the typical case that the rounding problem for smaller matrices can be solved better than for larger ones.

Let us be more precise: For a given matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$ we are interested in finding a vector $y \in \mathbb{Z}^n$ such that (1) $\|x - y\|_\infty \leq 1$ and (2) the rounding error $\|A(x - y)\|_\infty$ is small. y is sometimes called *approximate integer solution* for the linear system $Ay = Ax$.

It is easy to see from the problem statement that only the fractional part of x is important. Therefore we usually assume $x \in [0, 1]^n$ and consequently have $y \in \{0, 1\}$. It is also clear that rescaling A does not change the problem substantially: If we replace A by λA for some $\lambda > 0$, the set of optimal solutions

is not changed and their rounding error just changes by a factor of λ . Thus we lose nothing by assuming $A \in [-1, 1]^{m \times n}$.

In discrepancy theory, this problem is known under the term *linear discrepancy problem*:

$$\begin{aligned} \text{lindisc}(A, x) &= \min_{y \in \{0,1\}^n} \|A(x - y)\|_\infty, \\ \text{lindisc}(A) &= \max_{x \in [0,1]^n} \text{lindisc}(A, x). \end{aligned}$$

The special case that $x = \frac{1}{2}\mathbf{1}_n$ is called *combinatorial discrepancy problem*. It can be seen as the problem to partition the columns of A into two groups such that the row sums within each group are similar. We write

$$\text{disc}(A) := \min_{y \in \{0,1\}^n} \|A(\frac{1}{2}\mathbf{1}_n - y)\|_\infty = \frac{1}{2} \min_{y \in \{-1,1\}^n} \|Ay\|_\infty.^1$$

Note that already the combinatorial discrepancy problem is far from being easy: It is *NP*-hard to decide whether a 0, 1 matrix has discrepancy zero or not. On the other hand, a number of results and algorithms are known:

- If all column vectors have l_1 -norm at most t , then $\text{disc}(A) \leq t$ (Beck, Fiala [2]).
- A $y \in \{-1, 1\}^n$ such that $\|Ay\|_\infty \leq \sqrt{2n \ln(2m)}$ can be computed in time polynomial in n and m . In particular, $\text{disc}(A) \leq \sqrt{\frac{1}{2}n \ln(2m)}$ (Alon, Spencer [1]).
- If $m \geq n$, then $\text{disc}(A) \leq 3\sqrt{n \ln(2m/n)}$ (Spencer [12]).
- If A is the incidence matrix of a hypergraph \mathcal{H} having primal shatter function $\pi_{\mathcal{H}} = O(n^d)$, then $\text{disc}(A) = O(n^{\frac{1}{2} - \frac{1}{2d}})$ (Matoušek [10]). Hence this bound in particular holds if \mathcal{H} has VC-dimension d .
- If the dual shatter function satisfies $\pi_{\mathcal{H}}^* = O(n^d)$, then the discrepancy is $\text{disc}(A) = O(n^{\frac{1}{2} - \frac{1}{2d}} \sqrt{\log(n)})$ (Matoušek, Welzl, Wernisch [11]).

We refer to the chapter Beck and Sós [3] and the book Matoušek [9] for further discrepancy results. For our purposes a result of Beck and Spencer [4] and Lovász, Spencer and Vesztergombi [8] is crucial: It shows that the linear discrepancy problem is not much harder than the combinatorial one:

¹ Note that some papers define the linear and combinatorial discrepancy to be twice our values. This is motivated by the notion of hypergraph discrepancy: The discrepancy of a hypergraph is the least $k \in \mathbb{N}_0$ such that there is a 2-coloring of the vertex set such in each hyperedge the number of vertices in one color deviates from that in the other by at most k . If a hypergraph has discrepancy k , its incidence matrix has discrepancy $\frac{1}{2}k$ (in our notation), and vice versa. This motivates to define the discrepancy of a matrix A by $\min_{y \in \{-1,1\}^n} \|Ay\|_\infty$. On the other hand, from the viewpoint of rounding problems, our notation seems more appropriate.

Theorem 1. For any $A \in [-1, 1]^{m \times n}$ and $x \in [0, 1]^n$, there is a $y \in \{0, 1\}^n$ such that

$$\|A(x - y)\|_\infty \leq 2 \max_{A_0 \leq A} \text{disc}(A_0).^2$$

A $y \in \{0, 1\}^n$ such that $\|A(x - y)\|_\infty \leq 2D + O(2^{-k}n)$ can be computed by k times solving a combinatorial discrepancy problem for a submatrix of A with discrepancy at most D .

The constant of 2 in Theorem 1 cannot be improved in general: For arbitrary $n \in \mathbb{N}$, Lovász, Spencer and Vesztergombi [8] provide an example $A \in \{0, 1\}^{(n+1) \times n}$, $x \in [0, 1]^n$ such that any $y \in \{0, 1\}^n$ fulfills

$$\|A(x - y)\|_\infty = 2(1 - \frac{1}{n+1}) \max_{A_0 \leq A} \text{disc}(A_0).$$

On the other hand, Theorem 1 is known to be not sharp: The factor of 2 can be replaced by $2(1 - \frac{1}{2m})$ as shown in [5]. In between these two results little seems to be known. Spencer conjectures that $2(1 - \frac{1}{n+1})$ is the right constant in Theorem 1. This has been proven for totally unimodular matrices in [6].

Before explaining our results, we would like to point out that Theorem 1 requires understanding not only the discrepancy problem for A , but also for all submatrices of A . This is known under the term ‘hereditary discrepancy problem’, the corresponding notion is the *hereditary discrepancy* of A defined by

$$\text{herdisc}(A) := \max_{A_0 \leq A} \text{disc}(A_0).$$

It is not difficult to construct examples where the discrepancy of a submatrix (and thus the hereditary discrepancy) is much larger than the discrepancy of the matrix itself (which might even be zero). However, all these examples have the flavor of being artificially designed for this purpose. The situation usually encountered (both when looking at examples or results like the ones above) is that the discrepancy or the upper bound given by a result does not deviate significantly from the respective maximum taken over all submatrices.

In many cases the discrepancy behavior is even more regular: Smaller matrices tend to have lower discrepancies. To formalize this we introduce the (hereditary) discrepancy function of A : Define $h_A(n_0)$ to be the largest discrepancy among all submatrices of A having at most n_0 columns (to save some floors, we regard h_A as a function on the non-negative real numbers). Then most of the results above show $h_A(n_0) \leq Cn_0^\alpha$ for some $C > 0$ and $\alpha \leq 1$.

The main result of this paper is that this stronger discrepancy assumption can be exploited for the linear discrepancy problem. This reduces the constant of 2 in Theorem 1, the factor by which the general problem can be harder than the combinatorial one, down to $2(\frac{2}{3})^\alpha$ (e.g., 1.63 for $h_A = O(\sqrt{n_0})$):

² We write $A_0 \leq A$ to denote that A_0 is a submatrix of A .

Theorem 2. *If $h_A(n_0) \leq Cn_0^\alpha$ for all $n_0 \in \{1, \dots, n\}$, then*

$$\text{lindisc}(A) \leq 2 \left(\frac{2}{3}\right)^\alpha Cn^\alpha.$$

A more detailed analysis yields bounds for $\text{lindisc}(A, x)$ that take into account the vector x . We present a function $w : [0, 1] \rightarrow [0, \frac{2}{3}]$ such that

$$\text{lindisc}(A, x) \leq 2 \left(\sum_{i=1}^n w(x_i) \right)^\alpha Cn^\alpha$$

holds for all $x \in [0, 1]^n$. This allows an average case analysis showing that an x picked uniformly at random has expected $\text{lindisc}(A, x)$ at most $2(\frac{1}{2})^\alpha Cn^\alpha$. It also shows that ‘small’ x have lower linear discrepancies: We prove $\text{lindisc}(A, x) \leq 2(2\|x\|_1(-\log_2(\frac{1}{n}\|x\|_1) + 1))^\alpha Cn^\alpha$. This might seem natural at first, but recall that in the example (A, x) such that $\text{lindisc}(A, x) = 2(1 - \frac{1}{n+1}) \text{herdisc}(A)$ in [8] we have $x = (\frac{1}{n+1}, \dots, \frac{1}{n+1})$.

All our results are constructive in the following sense: Let A be given. Assume that we can solve discrepancy problems for submatrices of A having n_0 columns with rounding error at most Cn_0^α . Then for any $x \in [0, 1]^n$ we can compute a $y \in \{0, 1\}^n$ such that $\|A(x - y)\|_\infty \leq 2(\sum_{i=1}^n w(x_i))^\alpha Cn^\alpha + O(2^{-k}n)$ by solving k discrepancy problems for submatrices of A .

2 Reduction to Game Theory

Our proofs are based on the proof of Theorem 1, which we state here in a language suitable for our further work. Here and in the remainder we use the shorthand $[n]$ to denote the set $\{1, \dots, n\}$.

Proof (of Theorem 1). Let $x \in [0, 1]^n$. We construct a $y \in \{0, 1\}^n$ such that $\|A(x - y)\|_\infty$ is small. As

$$x \mapsto \min_{y \in \{0, 1\}^n} \|A(x - y)\|_\infty$$

is a continuous function and $\{\sum_{i=1}^n b_i 2^{-i} \mid n \in \mathbb{N}, b_1, \dots, b_n \in \{0, 1\}\}$ is dense in $[0, 1]$, we may assume that x has finite binary expansion of length k , i.e., there is a $k \in \mathbb{N}$ such that $2^k x$ is integral.

Set $a^{(0)} := x$. We define a series of intermediate ‘roundings’ $a^{(l)}, l = 1, \dots, k$ having binary length at most $k - l$. Suppose that for $l \in \{1, \dots, k\}$, $a^{(l-1)}$ is already defined and satisfies $a^{(l-1)} 2^{k-l+1} \in \mathbb{Z}^n$. Set

$$X^{(l)} := \{j \in [n] \mid a_j^{(l-1)} 2^{k-l+1} \text{ odd}\},$$

the set of all j such that the binary expansion of $a_j^{(l-1)} 2^{k-l+1}$ ends in 1. By the definition of combinatorial discrepancy, there is an $\varepsilon^{(l)} : X^{(l)} \rightarrow \{-1, +1\}$ such that

$$d_i^{(l)} := \frac{1}{2} \sum_{j \in X^{(l)}} a_{ij} \varepsilon^{(l)}(j)$$

satisfies $|d_i^{(l)}| \leq h_A(|X^{(l)}|)$ for all $i \in [m]$. Define

$$a_j^{(l)} := \begin{cases} a_j^{(l-1)} - 2^{-(k-l+1)}\varepsilon^{(l)}(j) & \text{if } j \in X^{(l)} \\ a_j^{(l-1)} & \text{otherwise.} \end{cases}$$

Then $a^{(l)} 2^{k-l} \in \mathbb{Z}^n$ and $A(a^{(l-1)} - a^{(l)}) = 2^{-(k-l)}d^{(l)}$. Having defined $a^{(l)}$ for all $l \in \{0, \dots, k\}$, we put $y = a^{(k)}$ and compute

$$\begin{aligned} \|A(x - y)\|_\infty &= \left\| A \left(\sum_{l=1}^k (a^{(l-1)} - a^{(l)}) \right) \right\|_\infty \\ &= \left\| \sum_{l=1}^k 2^{-(k-l)} d^{(l)} \right\|_\infty \\ &\leq \sum_{l=1}^k 2^{-(k-l)} h_A(|X^{(l)}|). \end{aligned}$$

From $h_A(|X^{(l)}|) \leq \text{herdisc}(A)$ for all $l \in [k]$ we get the original result $\|A(x - y)\|_\infty \leq 2 \text{herdisc}(A)$. \square

There is one option we did not use in the above algorithm: At any time during the above rounding process, we may replace $\varepsilon^{(l)}$ by $-\varepsilon^{(l)}$. This changes the resulting $a^{(l)}$, but does not violate our discrepancy guarantee as we just replace $d^{(l)}$ by $-d^{(l)}$. By choosing signs for the $\varepsilon^{(l)}$, $l \in [k]$ in a clever way, we try to keep the sets $X^{(l)}$, $l \in [k]$ small and thus improve the discrepancy bound.

Note that if we change the sign of one $\varepsilon^{(l)}$, this does not only change the last digit of the binary expansion of the $a^{(l)}$, but may change any digit. Furthermore, it is very difficult to get suitable information about the $\varepsilon^{(l)}$ in the general case. We therefore regard the sign-choosing problem as an on-line problem, i.e., we analyze what can be achieved by choosing the sign of $\varepsilon^{(l)}$ without knowing the possible colorings $\varepsilon^{(l+1)}$.

Worst-case analyses of on-line problems naturally lead to games. One player represents the on-line algorithm and the other one the data not known to the algorithm. Our problem is modeled by the following two-player game. For obvious reasons we call the players ‘Pusher’ and ‘Chooser’. Let f be any real function with domain containing $\{0, \dots, n\}$.

The Game $G(a^{(0)}, f)$: The starting position of the game is a vector $a^{(0)} \in [0, 1]^n$ having a finite binary expansion of length at most k , i.e., $2^k a^{(0)}$ is integral. The game then consists of k rounds of the following structure:

Round l :

- Set $X^{(l)} := \{j \in [n] \mid 2^{k-l+1} a_j^{(l-1)} \text{ odd}\}$.
- Pusher selects a partition $S^{(l)} \dot{\cup} T^{(l)} = X^{(l)}$.
- Chooser chooses one partition class $Y^{(l)} \in \{S^{(l)}, T^{(l)}\}$.

– The position is updated according to

$$a_j^{(l)} := \begin{cases} a_j^{(l-1)} - 2^{-(k-l+1)} & \text{if } j \in X^{(l)} \setminus Y^{(l)} \\ a_j^{(l-1)} + 2^{-(k-l+1)} & \text{if } j \in Y^{(l)} \\ a_j^{(l-1)} & \text{otherwise.} \end{cases}.$$

Objective of the game: We call the value $\sum_{l=1}^k 2^{-k+l} f(|X^{(l)}|)$ the pay-off (for Pusher). As the name suggests, it is Pusher's aim to maximize this value (and Chooser's, to keep it small). The maximum pay-off Pusher can enforce in a game started in position $a^{(0)}$ is the value $v(a^{(0)}, f)$ of this game.

From the discussion above the following connection between the game $G(x, f)$ and our rounding problem is obvious:

Lemma 1. $\text{lindisc}(A, x) \leq v(x, h_A)$.

We complete the proof of our main results by estimating the values of the corresponding games. Note that we may always replace h_A by a pointwise not smaller function f : Then $v(x, h_A) \leq v(x, f)$ implies $\text{lindisc}(A, x) \leq v(x, f)$. We call such an f an upper bound for h_A . The results cited in the introduction also indicate that we lose little by assuming f to be concave, non-decreasing and non-negative.

3 Worst-Case Analysis

In this section we prove an upper bound on the game values, which by Lemma 1 yields an upper bound on the linear discrepancy of A .

Lemma 2. *Let $f : [0, n] \rightarrow \mathbb{R}$ be concave and non-decreasing. Then $v(a^{(0)}, f) \leq 2f(\frac{2}{3}n)$ holds for all starting positions $a^{(0)}$.*

Proof. To give an upper bound on $v(a^{(0)}, f)$, we have to show that Chooser has a strategy such that no matter what partitions Pusher selects, the pay-off will never exceed this bound. We analyze the following strategy.

Chooser's strategy: Assume all notation given as in the definition of the game. We may assume that k is even. In an even numbered round l , Chooser chooses the partition class arbitrarily. If l is odd, this is in particular not the last round, Chooser proceeds like this: He chooses that one of the two alternatives, that minimizes the size of $X^{(l+1)}$.

Analysis: Let $l \in [k]$ be odd. Let $X^{(l)} = S^{(l)} \dot{\cup} T^{(l)}$ be the partition given by Pusher. Denote by $X^{(l+1)} \circ Y^{(l)}$ the value of $X^{(l+1)}$ resulting from Chooser's move $Y^{(l)}$. Now we easily see that $(X^{(l+1)} \circ S^{(l)}) \cap X^{(l)}$ and $(X^{(l+1)} \circ T^{(l)}) \cap X^{(l)}$ form a partition of $X^{(l)}$. On the other hand, $(X^{(l+1)} \circ S^{(l)}) \setminus X^{(l)} = (X^{(l+1)} \circ T^{(l)}) \setminus X^{(l)}$, that is, the complement of $X^{(l)}$ is not affected by Chooser's move. Hence Chooser

has to choose $Y^{(l)}$ in such a way that $(X^{(l+1)} \circ Y^{(l)}) \cap X^{(l)}$ is minimized. Then $|(X^{(l+1)} \circ Y^{(l)}) \cap X^{(l)}| \leq \frac{1}{2}|X^{(l)}|$, and thus

$$\begin{aligned} |X^{(l+1)} \circ Y^{(l)}| &= |(X^{(l+1)} \circ Y^{(l)}) \setminus X^{(l)}| + |(X^{(l+1)} \circ Y^{(l)}) \cap X^{(l)}| \\ &\leq |[n] \setminus X^{(l)}| + \frac{1}{2}|X^{(l)}| \\ &= n - \frac{1}{2}|X^{(l)}|. \end{aligned}$$

We conclude that if Chooser follows the strategy proposed above, we have $|X^{(l+1)}| \leq n - \frac{1}{2}|X^{(l)}|$ for all odd $l \in [k]$. Let

$$\bar{f} : [0, n] \rightarrow \mathbb{R}; x \mapsto 2f(n - \frac{1}{2}x) + f(x).$$

Since f is concave, we have $\bar{f}(x) = 3(\frac{2}{3}f(n - \frac{1}{2}x) + \frac{1}{3}f(x)) \leq 3f(\frac{2}{3}n)$ for all $x \in [0, n]$. By definition, $\bar{f}(\frac{2}{3}n) = 3f(\frac{2}{3}n)$. Hence $\frac{2}{3}n$ is a global maximum of \bar{f} . Using this we bound the pay-off:

$$\begin{aligned} \sum_{l=1}^k 2^{-k+l} f(|X^{(l)}|) &= \sum_{\substack{l \in [k] \\ l \text{ odd}}} 2^{-k+l} \left(2f(|X^{(l+1)}|) + f(|X^{(l)}|) \right) \\ &\leq \sum_{\substack{l \in [k] \\ l \text{ odd}}} 2^{-k+l} \left(2f(n - \frac{1}{2}|X^{(l)}|) + f(|X^{(l)}|) \right) \\ &\leq \sum_{\substack{l \in [k] \\ l \text{ odd}}} 2^{-k+l} \left(2f(n - \frac{1}{3}n) + f(\frac{2}{3}n) \right) \\ &\leq \sum_{l=1}^k 2^{-k+l} f(\frac{2}{3}n) \leq 2f(\frac{2}{3}n). \end{aligned}$$

□

Lemma 1 and 2 give the following theorem, a slight generalization of Theorem 2 in the introduction.

Theorem 3. *If f is a concave and non-decreasing upper bound for h_A , then*

$$\text{lindisc}(A) \leq 2f(\frac{2}{3}n).$$

It may seem that our on-line strategy is very simple: Every second decision is chosen arbitrarily, the remaining ones only take into account the next move. Nevertheless, the game-theoretic analysis is tight in the worst-case:

Lemma 3. *For any f and any $k \in \mathbb{N}$ there is a starting position $a^{(0)}$ such that Pusher can enforce a pay-off of $2(1 - 2^{-k})f(\frac{2}{3}n)$ in a k -round game of $G(a^{(0)}, f)$.*

Proof. Let n be a multiple of 3. Put $x_k := \sum_{i=1}^{\lfloor k/2 \rfloor} 2^{-2i}$. If k is odd, let $a^{(0)}$ be such that one third of its components equal x_k and two thirds are $x_k + 2^{-k}$. If k is even, two thirds of the $a_j^{(0)}$, $j \in [n]$ shall equal x_k , and one third $x_{k-2} + 2^{-(k+2)}$.

If k is odd, $X^{(1)} := \{j \in [n] \mid 2^k a_j^{(0)} \text{ odd}\}$ has cardinality $\frac{2}{3}n$ by definition of $a^{(0)}$. Let $S^{(1)} \dot{\cup} T^{(1)}$ be any partition of $X^{(1)}$ such that $|S^{(1)}| = |T^{(1)}| = \frac{1}{3}n$. Regardless of Chooser's choice, half of the $x_k + 2^{-k}$ -values (and thus a total of $\frac{1}{3}n$) are rounded up to $x_{k-2} + 2^{-(k+2)}$, the remaining ones are rounded down to x_k . Hence we end up with the starting position for the game lasting $k-1$ rounds.

Similarly, if k is even, we have $|X^{(1)}| = \frac{2}{3}n$, and partitioning $X^{(1)}$ into equally sized classes proceeds the game to the starting position for the game lasting $k-1$ rounds.

Hence by induction we conclude that this strategy ensures $|X^{(l)}| = \frac{2}{3}n$ for all $l \in [k]$, and thus a pay-off of $2(1 - 2^{-k})h_A(\frac{2}{3}n)$. \square

The strategy described in the proof of Lemma 2 can also be applied to non-concave functions f . The following corollary shows that an example $A \in \{0, 1\}^{m \times n}$, $x \in [0, 1]^n$ having $\text{lindisc}(A, x) > 2(1 - \frac{1}{n+1}) \text{herdisc}(A)$ (thus being stronger than the currently best known ones of Lovász, Spencer and Vesztergombi) must have constant discrepancy function on $[\frac{2}{3}n, n]$. Hence such examples — should they exist — do not display the regular discrepancy behavior we investigate in this paper, and thus must have a rather particular structure.

Corollary 1. *If $A \in \{0, 1\}^{m \times n}$ and $h_A(\frac{2}{3}n) < \text{herdisc}(A)$, then*

$$\text{lindisc}(A, x) \leq 2(1 - \frac{1}{n+1}) \text{herdisc}(A).$$

Proof. If Chooser follows the strategy proposed in the proof of Lemma 2, he ensures that $|X^{(l)}| \leq \frac{2}{3}n$ or $|X^{(l-1)}| \leq \frac{2}{3}n$ holds for all $l \in [k]$. Hence

$$\text{lindisc}(A) \leq \sum_{l=0}^{\infty} 2^{-2l} (\text{herdisc}(A) + \frac{1}{2}h_A(\frac{2}{3}n))$$

follows along the lines of the proof of Lemma 2. If $h_A(\frac{2}{3}n) < \text{herdisc}(A)$, then $h_A(\frac{2}{3}n) \leq \text{herdisc}(A) - \frac{1}{2}$. This yields $\text{lindisc}(A) \leq 2 \text{herdisc}(A) - \frac{1}{3}$, proving the claim for the case that $\text{herdisc}(A) < \frac{1}{6}(n+1)$. The case $\text{herdisc}(A) \geq \frac{1}{6}(n+1)$ is trivial, since $\text{lindisc}(A) \leq \frac{1}{4}(n+1)$ holds for any $A \in \{0, 1\}^{m \times n}$. \square

4 Improved Strategies and Average Case

As Lemma 3 shows, the on-line sign-choosing algorithm presented in the proof of Lemma 2 is optimal in the worst case (given by a particular starting position). In the following we present a more complicated on-line strategy, that is optimal in the worst-case as well, but yields tighter bounds for other starting positions. It is a potential function strategy, that is, we define a potential function for all positions and Chooser's strategy is to minimize this potential.

For a finite binary sequence $b = (b_1, \dots, b_k) \in \{0, 1\}^k$ recursively define

$$w(b, k) := b_k, \\ w(b, i) := \begin{cases} \frac{1}{2}w(b, i+1) & \text{if } b_i = b_{i+1} \\ 1 - \frac{1}{2}w(b, i+1) & \text{otherwise.} \end{cases}$$

Put $w(b) = \sum_{i=1}^k 2^{-i} w(b, i)$. For a number $a \in [0, 1[$ having finite binary expansion $a = \sum_{i=1}^k 2^{-i} b_i$, we write $w(a) := w((b_1, \dots, b_k))$. Put $w(1) = 0$. We have

Lemma 4. *Let $a \in [0, 1]$ having finite binary expansion $a = \sum_{i=1}^k 2^{-i} b_i$ such that $b_k = 1$.*

- (i) $\sum_{b \in \{0,1\}^k} w(b) = 2^{k-1} - \frac{1}{2}$.
- (ii) $w(a) = w(1 - a)$.
- (iii) $w(a) \leq \frac{2}{3}$.
- (iv) $w(a) \leq 2a(-\log_2(a) + 1)$.
- (v) $w(a) = 2^{-k} + \frac{1}{2}(w(a + 2^{-k}) + w(a - 2^{-k}))$.

Proof. Ad (i): We use induction on k . For $k = 1$ we compute

$$\sum_{b \in \{0,1\}^k} w(b) = w((0)) + w((1)) = 0 + \frac{1}{2}.$$

Let $k \geq 2$. For $b = (b_1, \dots, b_k) \in \{0, 1\}^k$ put $\bar{b} = (1 - b_1, \dots, b_k)$. Then $w(b, i) = w(\bar{b}, i)$ for $i \geq 2$ and $w(b, 1) + w(\bar{b}, 1) = 1$. Further, we have $w(b) = \frac{1}{2}w(b, 1) + \frac{1}{2}w((b_2, \dots, b_k))$. Thus

$$w(b) + w(\bar{b}) = \frac{1}{2}(w(b, 1) + w(\bar{b}, 1)) + w((b_2, \dots, b_k)) = \frac{1}{2} + w((b_2, \dots, b_k)).$$

Hence

$$\begin{aligned} \sum_{b \in \{0,1\}^k} w(b) &= \sum_{\substack{b \in \{0,1\}^k \\ b_1=0}} (w(b) + w(\bar{b})) \\ &= \sum_{\substack{b \in \{0,1\}^k \\ b_1=0}} \left(\frac{1}{2} + w((b_2, \dots, b_k)) \right) \\ &= 2^{k-2} + \sum_{b \in \{0,1\}^{k-1}} w(b) = 2^{k-1} - \frac{1}{2}. \end{aligned}$$

Ad (ii): If $a = 0$, then (ii) is satisfied by definition. Hence assume $a \neq 0$ and $k \geq 1$. Let $b = (b_1, \dots, b_k)$. Define $\tilde{b} \in \{0, 1\}^k$ by $\tilde{b}_k = 1$ and $\tilde{b}_i = 1 - b_i$ for $i = 1, \dots, k-1$. Then $1 - a = \sum_{i=1}^k 2^{-i} \tilde{b}_i$. Now (ii) follows from $w(b, i) = w(\tilde{b}, i)$ for $i = 1, \dots, k$.

Ad (iii): By (ii), we may assume $b_1 = 0$. If $b_2 = 0$, then

$$w(b) = \frac{1}{4}w(b, 2) + \frac{1}{4}w(b, 2) + \frac{1}{4}w((b_3, \dots, b_k)) \leq \frac{1}{2} + \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{3}$$

by induction on k . If $b_2 = 1$, then

$$w(b) = \frac{1}{2}(1 - \frac{1}{2}w(b, 2)) + \frac{1}{4}w(b, 2) + \frac{1}{4}w((b_3, \dots, b_k)) \leq \frac{1}{2} + \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{3}$$

again by induction.

Ad (iv): If $2^{-(\ell+1)} \leq a < 2^{-\ell}$, then $b_1 = \dots = b_\ell = 0$ and $b_{\ell+1} = 1$. Thus

$$w(a) = \ell 2^{-\ell} w(b, \ell) + 2^{-\ell} w((b_{\ell+1}, \dots, b_k)) \leq (\ell + 1) 2^{-\ell} < 2a(-\log_2 a + 1).$$

Ad (v): Assume for simplicity that $a + 2^{-k} \neq 1$ (this case is easily solved separately). Let $b^+ \in \{0, 1\}^{k-1}$ such that

$$a + 2^{-k} = \sum_{i=1}^{k-1} 2^{-i} b_i^+.$$

Note that $a - 2^{-k} = \sum_{i=1}^{k-1} 2^{-i} b_i$. Let $\ell \in [k-1]$ be minimal subject to $b_{\ell+1} = \dots = b_k = 1$. Then $b_i^+ = 0$ for $\ell+1 \leq i < k$, $b_\ell^+ = 1$ and $b_i^+ = b_i$ for all $i < \ell$. Thus we have $w((b_1, \dots, b_{k-1}), i) + w(b^+, i) = 2^{-k+i} + 0 = 2w(b, i)$ for $\ell+1 \leq i < k$. We also compute $w((b_1, \dots, b_{k-1}), \ell) + w(b^+, \ell) = 1 - 2^{-k+\ell} + 1 = 2w(b, \ell)$. Since $b_i^+ = b_i$ for all $i < \ell$, an easy induction yields $w((b_1, \dots, b_{k-1}), i) + w(b^+, i) = 2w(b, i)$ also for the remaining $i \in [k-1]$. Now (v) follows from the definition of w . \square

A reader familiar with probabilistic game analysis (cf. Spencer [13]) might prefer this randomized view: Given a , we repeat rounding the last non-zero digit of its binary expansion up or down with equal probabilities $\frac{1}{2}$. Then $w(b, i)$ is the probability that $b_i = 1$ when all higher bits are already rounded. Thus $w(x_i)$ is the expected contribution of a single entry of x to the pay-off $\sum_{l=1}^k 2^{-k+l} |X(l)|$ of the game $G(x, \text{id})$, if Chooser plays randomly.

By (v) of the lemma above, w is continuous on the set of numbers having finite binary expansion. Hence there is a unique continuation on $[0, 1]$, which we denote by w as well. Note that (ii) to (iv) of Lemma 4 now hold for arbitrary $a \in [0, 1]$. The inequality (iii) is sharp as shown by $a = \frac{1}{3}$ and $a = \frac{2}{3}$.

For a game position $x \in [0, 1]^n$ we put $w(x) = \sum_{i=1}^n w(x_i)$. Then

Lemma 5. *Let f be a concave, non-decreasing and non-negative function. Let $x \in [0, 1]^n$ be a starting position of a k -round game. If Chooser plays the strategy to minimize w , then the pay-off in the game $G(x, f)$ is at most $2f(w(x))$. Consequently, if f is an upper bound on h_A , then*

$$\text{lindisc}(A, x) \leq 2f(w(x))$$

holds for all $x \in [0, 1]^n$.

Proof. Assume first that $f = \text{id}_{\mathbb{R}}$. We proceed by induction on the length k of the binary expansion of a . If $k = 0$, the game ends before it started, and the pay-off is $0 = w(a)$. Hence let $k \geq 1$. Let $X = \{j \in [n] \mid 2^k a_j \text{ odd}\}$, and let $S \dot{\cup} T = X$ denote Pusher's move. Let s, t denote the positions that arise if Chooser chooses S, T . If Chooser takes S , by induction the pay-off is bounded by $2^{-k+1}|X| + 2w(s)$ (and an analogous statement holds for T). Let $y \in \{s, t\}$ be such that $w(y) = \min\{w(s), w(t)\}$. Then the pay-off resulting from choosing y is bounded by $2^{-k+1}|X| + 2w(y) \leq 2^{-k+1}|X| + w(s) + w(t) \leq w(a)$, where the latter inequality follows from Lemma 4.

In the notation of the definition of the game, we just showed that Chooser's strategy yields $\sum_{l=1}^k 2^{-k+l}|X^{(l)}| < 2w(a)$. Now let f be an arbitrary concave, non-decreasing and non-negative function. Then the pay-off of $G(a, f)$ is bounded by

$$\sum_{l=1}^k 2^{-k+l} f(|X^{(l)}|) \leq 2f\left(\sum_{l=1}^k 2^{-k+l-1}|X^{(l)}|\right) \leq 2f(w(a)).$$

□

Lemma 5 allows an average case analysis of the linear discrepancy problem.

Theorem 4. *Let f be a concave and non-decreasing upper bound for h_A . For an x chosen uniformly at random from $[0, 1]^n$, the expected linear discrepancy satisfies*

$$E(\text{lindisc}(A, x)) \leq 2f\left(\frac{1}{2}n\right).$$

Proof. Let $k \in \mathbb{N}$ and $B = \{\sum_{l=1}^k 2^{-l}b_l \mid b_1, \dots, b_k \in \{0, 1\}\}$. For a number $r \in [0, 1]$ denote by \tilde{r} the largest element of B not exceeding r . Put $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)$. Then

$$\begin{aligned} E(\text{lindisc}(A, x)) &\leq E(\text{lindisc}(A, \tilde{x})) + n2^{-k} \\ &= \sum_{\tilde{x} \in B^n} 2^{-nk} \text{lindisc}(A, \tilde{x}) + n2^{-k} \\ &\leq 2 \sum_{\tilde{x} \in B^n} 2^{-nk} f(w(\tilde{x})) + n2^{-k} \\ &\leq 2f\left(2^{-nk} \sum_{\tilde{x} \in B^n} w(\tilde{x})\right) + n2^{-k} \\ &\leq 2f\left(\frac{1}{2}n\right) + n2^{-k}, \end{aligned}$$

where the latter inequality follows from

$$\begin{aligned} \sum_{\tilde{x} \in B^n} w(\tilde{x}) &= \sum_{\tilde{x} \in B^n} \sum_{i=1}^n w(\tilde{x}_i) \\ &= \sum_{i=1}^n \sum_{\tilde{x} \in B^n} w(\tilde{x}_i) \\ &= \sum_{i=1}^n 2^{(n-1)k} \sum_{b \in B} w(b) \\ &= n2^{(n-1)k} \sum_{b \in B} w(b) \end{aligned}$$

and Lemma 4. □

Another consequence of Lemma 5 is that ‘small’ x have lower linear discrepancy:

Lemma 6. *Let f be a concave and non-decreasing upper bound for h_A . Let $x \in [0, 1]^n$ and $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$. Then*

$$\text{lindisc}(A, x) \leq 2f(2\|x\|_1(-\log_2(\bar{x}) + 1)).$$

Proof. From Lemma 4 and the concavity of $a \mapsto 2a(-\log_2(a) + 1)$, we conclude $w(x) \leq 2\|x\|_1(-\log_2(\bar{x}) + 1)$. Thus Lemma 5 proves the claim. \square

5 Conclusion

In this paper we investigated the relation between the linear discrepancy problem (rounding arbitrary vectors) and the combinatorial discrepancy problem (rounding vectors with entries $\frac{1}{2}$ only). We assumed that the discrepancy problem can be solved better for submatrices having fewer columns. This assumption is justified by the fact that most results are of this type. We showed that the classical results of Beck and Spencer and Lovász, Spencer and Vesztergombi on the relation of both rounding problems can be strengthened in this situation. We analyzed both the worst- and average case. Like in [7], our results indicate that the assumption of decreasing discrepancies is both natural and powerful.

We have to leave it as an open problem how tight our bounds are. Another open problem is for which vectors x the rounding problem is hardest. ‘Small’ vectors cause lower errors, and as $w(a) = \frac{2}{3}$ for some $a \in [0, 1]$ implies $a \in \{\frac{1}{3}, \frac{2}{3}\}$, our bound $\text{lindisc}(A, x) \leq 2f(\frac{2}{3}n)$ can only be tight if $x_i \in \{\frac{1}{3}, \frac{2}{3}\}$ for all $i \in [n]$. On the other hand, most examples seem to indicate that $x \approx \frac{1}{2}\mathbf{1}_n$ is the most difficult instance.

References

1. N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, Inc., 2nd edition, 2000.
2. J. Beck and T. Fiala. “Integer making” theorems. *Discrete Applied Mathematics*, 3:1–8, 1981.
3. J. Beck and V. T. Sós. Discrepancy theory. In R. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*. 1995.
4. J. Beck and J. Spencer. Integral approximation sequences. *Math. Programming*, 30:88–98, 1984.
5. B. Doerr. Linear and hereditary discrepancy. *Combinatorics, Probability and Computing*, 9:349–354, 2000.
6. B. Doerr. Lattice approximation and linear discrepancy of totally unimodular matrices. In *Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 119–125, 2001.
7. B. Doerr and A. Srivastav. Recursive randomized coloring beats fair dice random colorings. In A. Ferreira and H. Reichel, editors, *Proceedings of the 18th Annual Symposium on Theoretical Aspects of Computer Science (STACS) 2001*, volume 2010 of *Lecture Notes in Computer Science*, pages 183–194, Berlin–Heidelberg, 2001. Springer Verlag.

8. L. Lovász, J. Spencer, and K. Vesztergombi. Discrepancies of set-systems and matrices. *Europ. J. Combin.*, 7:151–160, 1986.
9. J. Matoušek. *Geometric Discrepancy*. Springer-Verlag, Berlin, 1999.
10. J. Matoušek. Tight upper bound for the discrepancy of half-spaces. *Discr. & Comput. Geom.*, 13:593–601, 1995.
11. J. Matoušek, E. Welzl, and L. Wernisch. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 13:455–466, 1984.
12. J. Spencer. Six standard deviations suffice. *Trans. Amer. Math. Soc.*, 289:679–706, 1985.
13. J. Spencer. Randomization, derandomization and antirandomization: Three games. *Theor. Comput. Sci.*, 131:415–429, 1994.