Improved Bounds and Schemes for the Declustering Problem*

Benjamin Doerr Nils Hebbinghaus
Sören Werth
Mathematisches Seminar, Bereich II,
Christian-Albrechts-Universität zu Kiel
Christian-Albrechts-Platz 4, 24118 Kiel, Germany.
{bed,nhe,swe}@numerik.uni-kiel.de

Abstract

The declustering problem is to allocate given data on parallel working storage devices in such a manner that typical requests find their data evenly distributed on the devices. Using deep results from discrepancy theory, we improve previous work of several authors concerning rectangular queries to higher-dimensional data. We give a declustering scheme with an additive error of $O_d(\log^{d-1} M)$ independent of the data size, where d is the dimension, M the number of storage devices and $d-1$ does not exceed the smallest prime power in the canonical decomposition of M into prime powers. In particular, our schemes work for arbitrary M in two and three dimensions, and arbitrary M that is a power of two and at least $d-1$. For a lower bound, we show that a recent proof of a $\Omega_d(\log^{\frac{d-1}{2}} M)$ bound contains an error. We close the gap in the proof and thus establish the bound.

*supported by the DFG-Graduiertenkolleg 357 "Effiziente Algorithmen und Mehrskalenmethoden".

1 Introduction

The last decade saw dramatic improvements in computer processing speeds and storage capacities. Nowadays, the bottleneck in data-intensive applications is the time needed to retrieve typically large amounts of data from external storage devices. One idea to overcome this obstacle is to distribute the data on disks of multi-disk systems so that it can be retrieved in parallel. Hopefully, this declustering reduces the retrieval time by a factor equal to the number of disks. The data allocation is determined by so-called declustering schemes. The schemes should allocate the data in such a manner that typical requests find their data evenly distributed on the disks.

We consider the problem of declustering uniform multi-dimensional data that is arranged in a multi-dimensional grid. There are many data-intensive applications that deal with this kind of data, especially multi-dimensional databases as remote-sensing databases [CMA$^+$97]. A range query Q requests the data blocks that are associated with a hyper-rectangular subspace of the grid. Since we will not deal with syntactic issues of queries, we may identify a query with the set of requested block. In consequence, $|Q|$ denotes the number of requested blocks.

The response time of a query Q is (proportional to) the maximum number of blocks of Q that are assigned to the same disk (hence we assume identical disks). In an ideal declustering scheme for a system with M disks, this would be $|Q|/M$ for all queries Q. As we will see, this aim cannot be achieved. The quality of a declustering scheme is measured by the worst case (over all queries Q) additive deviation of the response time from the ideal value $|Q|/M$.

Declustering is an intensively studied problem and a number of schemes [CBS03, PAGAA98, AP00, DS82, FB93] have been developed in the last twenty years. It was an important turning point when discrepancy theory was connected to declustering.

Before the use of discrepancy theory, no provable performance bounds were known for arbitrary dimension d. Such bounds existed only for a few declustering schemes in two dimensions: The known results for these schemes considered only special cases, e. g., for the scheme proposed in [CBS03] a proof for the average performance is given if the number M of disks is a Fibonacci number, and for the construction of the scheme in [AP00] M has to be a power of 2.

A breakthrough was marked by noting that the declustering problem is a discrepancy problem. Sinha, Bhatia and Chen [SBC03] and Anstee, Demetrovics, Katona and Sali [ADKS00] developed declustering schemes for all M for two dimensional problems and proved their asymptotically optimal behavior. The schemes of Sinha et al. [SBC03] are based on two dimensional low discrepancy point sets. They also give generalizations to arbitrary dimension d, but without bounds on the error. Both papers show a lower bound of $\Omega(\log M)$ for the additive error of any declustering scheme in dimension two. The result of Anstee et al. [ADKS00] applies to latin square type colorings only, but their proof can easily be extended to the general case as well. Sinha et al. [SBC03] claim a bound of $\Omega_d(\log^{\frac{d-1}{2}} M)$ for arbitrary dimension d, but their proof contains an error for $d \geq 3$ (cf. Section 3).

The first non-trivial upper bounds for declustering schemes in arbitrary dimension were proposed by Chen and Cheng [CC02], who present two schemes for the d–dimensional declustering problem. The first one has an additive error of $O_d(\log^{d-1} M)$, but works only if $M = p^k$ for some $k \in \mathbb{N}$ and p a prime such that $d \leq p$. The second works for arbitrary M, but the error increases with the size of the data. (Note that all other bounds stated in this paper are independent of the data size.)

Our Results: We work both on upper and lower bounds. For the upper bound, we present an improved scheme that yields an additive error of $O_d(\log^{d-1} M)$ for all values of M (independent of the data size) and all d such that $d \leq q_1 + 1$, where q_1 is the smallest factor in the canonical decomposition of M into prime powers. Thus, in particular, our schemes work for M being a power of two such that $M \geq d - 1$ and for all M in dimension 2 and 3, which is very useful from the viewpoint of application. We also show that the latin hypercube construction used by Chen and Cheng [CC02] and in our work is much better than proven there. Where they show that the final scheme has an error of at most 2^d times the one of the latin hypercube coloring, we show that both errors are the same.

For the lower bound, we present the first correct proof of the $\Omega_d(\log^{\frac{d-1}{2}} M)$ bound.

2 Discrepancy Theory

In this section, we sketch the connection between the declustering problem and discrepancy theory.

2.1 Combinatorial Discrepancy

Recall that the declustering problem is to assign data blocks from a multi-dimensional grid to M storage devices (disks) in a balanced manner. The aim is that range queries use all storage devices in a similar amount. More precisely, our grid is $V = [n_1] \times \cdots \times [n_d]$ for some positive integers $n_1, \ldots, n_d$.[1] A query Q requests the data assigned to a *rectangle* (or *box*) $[x_1..y_1] \times \cdots \times [x_d..y_d]$ for some integers $1 \le x_i \le y_i \le n_i$. We identify a query with the set of blocks it requests, i.e., $Q = [x_1..y_1] \times \cdots \times [x_d..y_d]$.

We assume that the time to process a query is proportional to the maximum number of requested data blocks that are stored on a single device. We represent the assignment of data blocks to devices through a mapping $\chi : V \to [M]$. The processing time of the query Q then is $\max_{i \in [M]} |\chi^{-1}(i) \cap Q|$. Clearly, no declustering scheme can do better than $|Q|/M$. Hence a natural performance measure is the additive deviation from this lower bound. We are interested in the worst-case behavior. Thus we are looking for declustering schemes such that $\max_Q \max_{i \in [M]} |\chi^{-1}(i) \cap Q|$ is small.

This makes the problem a combinatorial discrepancy problem in M colors. Denote by $\mathcal{E}$ the set of all rectangles in V. Then $\mathcal{H} = (V, \mathcal{E})$ is a hypergraph. For a coloring $\chi : V \to [M]$, the discrepancy of a hyperedge $E \in \mathcal{E}$ with respect to χ is

$$\operatorname{disc}(E, \chi) := \max_{i \in [M]} \left| |\chi^{-1}(i) \cap E| - \tfrac{1}{M}|E| \right|,$$

the discrepancy of $\mathcal{H}$ with respect to χ is

$$\operatorname{disc}(\mathcal{H}, \chi) := \max_{i \in [M], E \in \mathcal{E}} \left| |\chi^{-1}(i) \cap E| - \tfrac{1}{M}|E| \right|,$$

[1] We use the notations $[n] := \{1, 2, \ldots, n\}$ and $[n..m] := \{n, \ldots, m\}$ for $n, m \in \mathbb{N}$, $n \le m$.

and the discrepancy of $\mathcal{H}$ in M colors is

$$\operatorname{disc}(\mathcal{H}, M) := \min_{\chi: V \to [M]} \operatorname{disc}(\mathcal{H}, \chi).$$

These definitions were introduced by Srivastav and the first author in [DS99, DS03] extending the well-known notion of combinatorial discrepancy to any number of colors. Similar notions were used by Biedl et al. [BČC+02] and Babai, Hayes and Kimmel [BHK01]. For our purposes, only a positive deviation has to be regarded. We adapt the multi-color discrepancy notion in the obvious way and define the positive discrepancy by

$$\begin{aligned} \operatorname{disc}^+(\mathcal{H}, \chi) &:= \max_{i \in [M], E \in \mathcal{E}} \left(|\chi^{-1}(i) \cap E| - \tfrac{1}{M}|E| \right), \\ \operatorname{disc}^+(\mathcal{H}, M) &:= \min_{\chi: V \to [M]} \operatorname{disc}^+(\mathcal{H}, \chi). \end{aligned}$$

Clearly, we have $\frac{1}{M-1} \operatorname{disc}(\mathcal{H}) \leq \operatorname{disc}^+(\mathcal{H}) \leq \operatorname{disc}(\mathcal{H})$ for all hypergraphs $\mathcal{H}$. Summarizing the discussion above, we have

Theorem 1. *The additive error of an optimal declustering scheme for range queries is* $\operatorname{disc}^+(\mathcal{H}, M)$*.*

Since a central result of this paper are discrepancy bounds independent of the size of the grid, we usually work with the hypergraph $\mathcal{H}_N^d = ([N]^d, \mathcal{E}_N^d)$, $\mathcal{E}_N^d = \{\prod_{i=1}^d [x_i..y_i] \mid 1 \leq x_i \leq y_i \leq N\}$ for some sufficiently large integer N. Furthermore, we regard only the case that $M \geq 3$. For $M = 2$, a checkerboard coloring yields a declustering scheme with an additive error of $1/2$. We prove the following result.

Theorem 2. *Let $M \geq 3$ and $d \geq 2$ be integers and q_1 the smallest prime power in the canonical factorization of M into prime powers. Then*

(i) $\operatorname{disc}^+(\mathcal{H}_N^d, M) = O_d(\log^{d-1} M)$ *for $d \leq q_1 + 1$, independently of $N \in \mathbb{N}$,*

(ii) $\operatorname{disc}^+(\mathcal{H}_N^d, M) = \Omega_d(\log^{\frac{d-1}{2}} M)$ *for $N \geq M$,*

(iii) $\operatorname{disc}^+(\mathcal{H}_N^d, M) = \Theta(\log M)$ *for $d = 2$.*

2.2 Geometric Discrepancy

As mentioned before, the use of geometric discrepancies in [SBC03, ADKS00] in the analysis of declustering problems was a major breakthrough in this area. We refer to the recent book of Matoušek [Mat99] for both a great introduction and a thorough treatment of geometric discrepancies.

The geometric discrepancy problem is to distribute n points evenly in a geometric setting. For our purposes, we regard discrepancies of point sets in $[0,1]^d$ with respect to axis-parallel boxes. Such a box R is the product $R = \prod_{i=1}^{d} [x_i, y_i)$ with $0 \le x_i \le y_i \le 1$ for all $i \in [d]$. Our aim is that each box R shall contain approximately $n \operatorname{vol}(R)$ points, where $\operatorname{vol}(R) = \prod_{i=1}^{d} (y_i - x_i)$ denotes the volume of R. Again, discrepancy quantifies the distance to a perfect distribution. The discrepancy of an n–point set $\mathcal{P}$ with respect to a box R is defined by

$$D(\mathcal{P}, R) = \big| |\mathcal{P} \cap R| - n \operatorname{vol}(R) \big|,$$

the discrepancy of $\mathcal{P}$ with respect to the set $\mathcal{R}_d$ of all axis-parallel boxes is

$$D(\mathcal{P}, \mathcal{R}_d) = \sup_{R \in \mathcal{R}_d} |D(\mathcal{P}, R)|$$

and the discrepancy of $\mathcal{R}_d$ for n-point sets is

$$D(n, \mathcal{R}_d) = \inf_{\substack{\mathcal{P} \subset [0,1)^d \\ |\mathcal{P}| = n}} D(\mathcal{P}, \mathcal{R}_d).$$

3 The Lower Bound

To prove our lower bounds, we use classical lower bounds for geometric discrepancies. Roth's [Rot54] famous lower bound for the L_2 discrepancy of the axis parallel boxes immediately implies the following.

Theorem 3 (Roth's lower bound). *Let $d \ge 2$. There exists a constant $k > 0$ (depending on d) such that for any n–point set $\mathcal{P}$ in the unit cube $[0,1)^d$, there is an axis-parallel box R in $[0,1)^d$ with*

$$D(\mathcal{P}, R) \ge k \log^{\frac{d-1}{2}} n.$$

It was Schmidt [Sch72] who came up with the sharp lower bound in two dimensions.

Theorem 4 (Schmidt's lower bound). *There is a constant $k > 0$ such that for any n–point set $\mathcal{P}$ in the unit square $[0,1)^2$, there is an axis-parallel rectangle R in $[0,1)^2$ with*

$$D(\mathcal{P}, R) \geq k \log n.$$

The general idea in the proofs of the lower bound for declustering schemes in Sinha et al. [SBC03] and Anstee et al. [ADKS00] (for $d = 2$ only) is the following.

Any M–coloring of $[M]^d$ produces a monochromatic set of M^{d-1} vertices. By scaling, this yields an M^{d-1}–point set $\mathcal{P}$ in $[0,1)^d$. The lower bounds above give a box R with polylogarithmic discrepancy. Round R to a box R_r with corners in $\{0, \frac{1}{M}, \cdots, \frac{M-1}{M}, 1\}^d$ in such a way that $R \cap \mathcal{P} = R_r \cap \mathcal{P}$. Then R and R_r have similar volume and hence similar discrepancy. Rescaling R_r yields a hyperedge $\hat{R}$ with combinatorial discrepancy equal to the geometric one of R_r.

The small, but crucial mistake in the proof of Sinha et al. [SBC03] is in the transfer from the geometric discrepancy setting back to the combinatorial one. Unlike in dimension $d = 2$, rounding R to R_r does not yield a constant change in the discrepancy in higher dimensions. The volume difference $|\operatorname{vol}(R) - \operatorname{vol}(R_r)|$ is still $O_d(\frac{1}{M})$. However, since the number of points is M^{d-1}, the discrepancy may change by something of order $\Theta_d(M^{d-2})$. This is way too large for $d > 2$.

For this reason, a straight generalization of the proof of Anstee et al. [ADKS00] of the lower bound in two dimensions (as attempted in [SBC03]) is not possible. We solve this problem by finding a *small* box R having large discrepancy. This keeps the change of volume due to rounding small enough so that the discrepancies of R and R_r differ by at most a constant.

Proof of Theorem 2 (ii). We show the claim for $N = M$, which clearly implies the result for arbitrary $N \geq M$. Let $\chi : [M]^d \to [M]$ be an M–coloring of $\mathcal{H}^d_M$. Choose an $s \in [M^{-\frac{d-2}{d-1}}, 2M^{-\frac{d-2}{d-1}}) \cap [0,1]$ such that s is a multiple of

$\frac{1}{M}$. Such an s exists since $M^{-\frac{d-2}{d-1}} > \frac{1}{M}$. Without loss of generality, we may assume that $|\chi^{-1}(1) \cap [1..sM]^d| \geq s^d M^{d-1}$. By Theorem 3 for every n–point set $\mathcal{P}$ in the unit cube $[0,1)^d$, there is an axis-parallel box R in $[0,1)^d$ with $D(\mathcal{P}, R) \geq k \log^{\frac{d-1}{2}} n$.

If $|\chi^{-1}(1) \cap [1..sM]^d| \geq s^d M^{d-1} + \frac{k}{2} \left(\frac{1}{d-1}\right)^{\frac{d-1}{2}} \log^{\frac{d-1}{2}} M$, we clearly have $\mathrm{disc}(\mathcal{H}_M^d, \chi) \geq \left||\chi^{-1}(1) \cap [1..sM]^d| - s^d M^{d-1}\right| \geq \frac{k}{2} \left(\frac{1}{d-1}\right)^{\frac{d-1}{2}} \log^{\frac{d-1}{2}} M$. Therefore, we may assume

$$|\chi^{-1}(1) \cap [1..sM]^d| < s^d M^{d-1} + \frac{k}{2} \left(\tfrac{1}{d-1}\right)^{\frac{d-1}{2}} \log^{\frac{d-1}{2}} M.$$

For every vertex $z = (z_1, z_2, \ldots, z_d) \in \chi^{-1}(1) \cap [1..sM]^d$ we define $x_z := \left(\frac{2z_1-1}{2M}, \frac{2z_2-1}{2M}, \ldots, \frac{2z_d-1}{2M}\right)$. Let $\mathcal{P} := \{x_z \mid z \in \chi^{-1}(1) \cap [1..sM]^d\}$ and $n := |\mathcal{P}|$. Then $\widetilde{\mathcal{P}} := \frac{1}{s}(\mathcal{P})$ is an n-point set in the unit cube $[0,1)^d$. Estimating the cardinality of $\widetilde{\mathcal{P}}$, we get $n \geq s^d M^{d-1} \geq M^{\frac{1}{d-1}}$. By Theorem 3, there exists a box $\widetilde{R}$ in $(0,1)^d$ with

$$\left||\widetilde{R} \cap \widetilde{\mathcal{P}}| - n \,\mathrm{vol}(\widetilde{R})\right| \geq k \log^{\frac{d-1}{2}} n \geq k \left(\frac{1}{d-1}\right)^{\frac{d-1}{2}} \log^{\frac{d-1}{2}} M. \tag{1}$$

Let $R := s\widetilde{R} = \prod_{i=1}^{d} [x_i, y_i)$ be the corresponding box in $[0,s)^d$. Now we construct a box R_r by rounding the x_i and y_i to the nearest multiple of $\frac{1}{M}$. In case of ties, we round down. This ensures $\mathcal{P} \cap R_r = \mathcal{P} \cap R$.

Since we have chosen a small box R, this rounding changes the volume not too much. Using $s \leq 2M^{-\frac{d-2}{d-1}}$, we get

$$|\,\mathrm{vol}(R) - \mathrm{vol}(R_r)| \leq 2d \tfrac{1}{2M} s^{d-1} < d 2^{d-1} M^{-(d-1)}.$$

The combinatorial counterpart of R_r is the box

$$\hat{R} := \left\{ x \in [M]^d \,\middle|\, \left(\tfrac{2x_1-1}{2M}, \ldots, \tfrac{2x_d-1}{2M}\right) \in R_r \right\}.$$

One easily checks that $M^d \operatorname{vol}(R_r) = |\hat{R}|$. By construction and (1),

$$\begin{aligned}
\operatorname{disc}(\mathcal{H}_M^d, \chi) &\geq \left| |\chi^{-1}(1) \cap \hat{R}| - \tfrac{1}{M}|\hat{R}| \right| \\
&= \left| |\mathcal{P} \cap R| - M^{d-1} \operatorname{vol}(R_r) \right| \\
&= \Big| |\widetilde{\mathcal{P}} \cap \widetilde{R}| - n \operatorname{vol}(\widetilde{R}) + \big(n - s^d M^{d-1}\big) \operatorname{vol}(\widetilde{R}) \\
&\quad + M^{d-1} \left(\operatorname{vol}(R) - \operatorname{vol}(R_r) \right) \Big| \\
&\geq \tfrac{k}{2} \left(\frac{1}{d-1} \right)^{\frac{d-1}{2}} \log^{\frac{d-1}{2}} M - O(1) = \Omega_d \left(\log^{\frac{d-1}{2}} M \right).
\end{aligned}$$

Thus, $\operatorname{disc}(\mathcal{H}_M^d, M) = \Omega_d(\log^{\frac{d-1}{2}} M)$. It remains to show that this bound also holds for the positive discrepancy. To this end, let us assume that the discrepancy of the box $\hat{R}$ in color 1 is caused by a lack of vertices in color 1. Since $|\chi^{-1}(1)| \geq M^{d-1}$, the complement of $\hat{R}$ in $[M]^d$ has at least the same discrepancy as $\hat{R}$, but caused by an excess of vertices in color 1. Though this complement is not a box, it is the union of at most $2d$ boxes. Therefore, one of these boxes has a positive discrepancy that is at least $\frac{1}{2d}$ times the discrepancy of $\hat{R}$ in color 1. $\square$

This last argument increases the implicit constant of the lower bound by a factor of $\frac{3^d}{2d}$ compared to the approach of Sinha et al. [SBC03].

We briefly show how to use the above to prove the $\Omega(\log M)$ bound for dimension $d = 2$. For this bound, two not completely satisfying bounds exist. Anstee et al. [ADKS00] only treated latin square type colorings of $[M]^2$ and posed it an open problem to extend their result to arbitrary colorings. The proof in [SBC03] does not have this restriction, but is not very precise, which in particular helped to hide the error for $d > 2$.

As a simple and clean proof we therefore propose the following: Use the same reasoning as in the case of arbitrary dimension $d \geq 2$, but apply Schmidt's lower bound instead of Roth's. The parameter s can be choosen as 1. In dimension $d = 2$ we do not need small boxes, because the roundoff error has an effect on the discrepancy which is of order $O(1)$.

4 The Upper Bound

In this section, we present a declustering scheme showing our upper bound. As in previous work, we use low discrepancy point sets to construct the declustering scheme. In the following we use the notation of Niederreiter [Nie87]. For an integer $b \geq 2$, an *elementary interval* in base b is an interval of the form $E = \prod_{i=1}^{d} \left[a_i b^{-d_i}, (a_i+1)b^{-d_i}\right)$, with integers $d_i \geq 0$ and $0 \leq a_i < b^{d_i}$ for $1 \leq i \leq d$. For integers t, m such that $0 \leq t \leq m$, a (t,m,d)*–net* in base b is a point set of b^m points in $[0,1)^d$ such that all elementary intervals with volume b^{t-m} contain exactly b^t points.

Note that any elementary interval with volume b^{t-m} has discrepancy zero in a (t,m,d)–net. Since any subset of an elementary interval of volume b^{t-m} has discrepancy at most b^t and any box can be packed with elementary intervals in a way that the uncovered part can be covered by $O_d(\log^{d-1} b^m)$ elementary intervals of volume b^{t-m}, the following is immediate:

Theorem 5. *A (t,m,d)–net $\mathcal{P}_{net}$ in base b with $n = b^m$ points has discrepancy*

$$D(\mathcal{P}_{net}, \mathcal{R}_d) = O_d(\log^{d-1} n).$$

The central argument in our proof of the upper bound is the following result of Niederreiter [Nie87] on the existence of $(0,m,d)$–nets. From the view-point of application it is important that his proof is constructive.

Theorem 6. *Let $b \geq 2$ be an arbitrary base and $b = q_1 q_2 \ldots q_u$ be the canonical factorization of b into prime powers such that $q_1 < \cdots < q_u$. Then for any $m \geq 0$ and $d \leq q_1 + 1$ there exists a $(0,m,d)$–net in base b.*

We use $(0,m,d)$–nets to construct an M–coloring of $\mathcal{H}_M^d$ in Lemma 7. For the definition of these colorings, we need the following special elements of $\mathcal{E}_M^d$: A set $\prod_{j=1}^{d} I_j \in \mathcal{E}_M^d$ is called a *row* of $[M]^d$ if there is an $i \in [d]$ with $I_i = [1..M]$ and $|I_j| = 1$ for all $j \neq i$. In Lemma 8 we use the M–coloring of $\mathcal{H}_M^d$ to construct an M–coloring of $\mathcal{H}_N^d$ with same discrepancy.

Lemma 7. *Let $\mathcal{P}_{net}$ be a $(0,d-1,d)$–net in base M in $[0,1)^d$. Then there is an M–coloring χ_M of $\mathcal{H}_M^d = ([M]^d, \mathcal{E}_M^d)$ such that all rows of $[M]^d$ contain* *every color exactly once*[2] *and*

$$\mathrm{disc}(\mathcal{H}_M^d, \chi_M) \le D(\mathcal{P}_{net}, \mathcal{R}_d).$$

Proof. The net $\mathcal{P}_{net}$ consists of M^{d-1} points and all elementary intervals with volume M^{-d+1} contain exactly one point. In particular, all subsets $\prod_{j=1}^d I_j$ of $[0,1)^d$ such that there is an $i \in [d]$ with $I_i = [0,1)$ and for all $j \neq i$ there exist $a_j \in [0..M-1]$ with $I_j = [\frac{a_j}{M}, \frac{a_j+1}{M})$, contain exactly one point.

We construct a coloring χ_M of $\mathcal{H}_M^d = ([M]^d, \mathcal{E}_M^d)$ corresponding to the set $\mathcal{P}_{net}$. Let $\hat{\mathcal{P}} := \left\{ x \in [M]^d \,\middle|\, \mathcal{P}_{net} \cap \prod_{i=1}^d [\frac{x_i - 1}{M}, \frac{x_i}{M}) \neq \emptyset \right\}$. Then each row of $[M]^d$ contains exactly one point of $\hat{\mathcal{P}}$. We define the coloring $\chi_M : [M]^d \to [M]$ by $\chi_M(y, x_2, \ldots, x_d) = i$ for all $x = (x_1, x_2, \ldots, x_d) \in \hat{\mathcal{P}}$, $i, y \in [M]$ such that $y \equiv x_1 + (i-1) \mod M$. Hence $\hat{\mathcal{P}}$ receives color 1, color class 2 is obtained from shifting $\hat{\mathcal{P}}$ along the first coordinate and so on. This defines an M–coloring χ_M of $\mathcal{H}_M^d = ([M]^d, \mathcal{E}_M^d)$ such that each row of $\mathcal{H}_M^d$ contains every color exactly once.

For this coloring it is sufficient to calculate $\max_{\hat{R} \in \mathcal{E}_M^d} \left| |\chi_M^{-1}(1) \cap \hat{R}| - \frac{1}{M}|\hat{R}| \right|$, because for each color $i \in [M]$ and each box $\hat{R} \in \mathcal{E}_M^d$ we get the same discrepancy for the box $\hat{R}'$, which is a copy of $\hat{R}$ shifted along the first dimension by $i-1$ and wrapped around perhaps, with respect to the color 1. If $\hat{R}'$ is wrapped around, it is the union of two boxes. Since whole rows have discrepancy zero, the discrepancy of those boxes is the same as the discrepancy of the the box between them, and we have

$$\mathrm{disc}(\mathcal{H}_M^d, \chi_M) = \max_{\hat{R} \in \mathcal{E}_M^d} \left| |\hat{\mathcal{P}} \cap \hat{R}| - \tfrac{1}{M}|\hat{R}| \right|.$$

Let $\hat{R} = \prod_{i=1}^d [x_i..y_i]$ an arbitrary hyperedge of $\mathcal{H}_M^d$. The associated box in $[0,1)^d$ is $R = \prod_{i=1}^d \left[\frac{x_i - 1}{M}, \frac{y_i}{M}\right)$. Then $|\hat{\mathcal{P}} \cap \hat{R}| = |\mathcal{P}_{net} \cap R|$ and $|\hat{R}| = M^d \,\mathrm{vol}(R)$. Thus the combinatorial discrepancy of $\hat{R}$ equals the geometric one of R. We have

$$\left| |\chi_M^{-1}(1) \cap \hat{R}| - \tfrac{1}{M}|\hat{R}| \right| = \left| |\mathcal{P}_{net} \cap R| - M^{d-1} \,\mathrm{vol}(R) \right| \le D(\mathcal{P}_{net}, \mathcal{R}_d).$$

Hence we get $\mathrm{disc}(\mathcal{H}_M^d, \chi_M) \le D(\mathcal{P}_{net}, \mathcal{R}_d)$. □

[2] Some authors call this a permutation scheme for $[M]^d$

Lemma 8. *Let χ_M be an M–coloring of $\mathcal{H}_M^d$ such that all rows of $[M]^d$ contain every color exactly once and χ a coloring of $\mathcal{H}_N^d$ defined by $\chi(x_1, \ldots, x_d) = \chi_M(y_1, \ldots, y_d)$ with $x_i \equiv y_i \mod M$ for $i \in [d]$, $x_i \in [N]$, $y_i \in [M]$. Then*

$$\mathrm{disc}(\mathcal{H}_N^d, \chi) = \mathrm{disc}(\mathcal{H}_M^d, \chi_M).$$

Proof. Let $\hat{R} = \prod_{i=1}^d [x_i..y_i]$ be an arbitrary hyperedge of $\mathcal{H}_N^d$. For all $i \in [d]$ there exist unique $\widetilde{x}_i, \widetilde{y}_i \in [M]$ with $x_i \equiv \widetilde{x}_i \mod M$ respectively $y_i \equiv \widetilde{y}_i \mod M$. Set $\bar{x}_i := \min\{\widetilde{x}_i, \widetilde{y}_i\}$ and $\bar{y}_i := \max\{\widetilde{x}_i, \widetilde{y}_i\}$ for all $i \in [d]$. We have $\mathrm{disc}(\hat{R}, \chi) = \mathrm{disc}([\bar{x}_1..\bar{y}_1] \times [x_2..y_2] \times \ldots \times [x_d..y_d], \chi)$, since whole rows have discrepancy zero. Applying this successively in every coordinate we get

$$\mathrm{disc}(\hat{R}, \chi) = \mathrm{disc}(\prod_{i=1}^{d} [\bar{x}_i..\bar{y}_i], \chi) = \mathrm{disc}(\prod_{i=1}^{d} [\bar{x}_i..\bar{y}_i], \chi_M).$$

□

Lemma 8 is a remarkable improvement of Theorem 4.2 in [CC02], where $\mathrm{disc}(\mathcal{H}_N^d, \chi) \leq 2^d \,\mathrm{disc}(\mathcal{H}_M^d, \chi_M)$ is shown. Note that this reduces the implicit constant in the upper bound by factor of 2^d.

It remains to show that the upper bound in Theorem 2 follows from Lemma 7 and Lemma 8.

Theorem 2(i). Let $M \geq 3$ and $d \geq 2$ be positive integers and $d \leq q_1 + 1$, where q_1 is the smallest prime power in the canonical factorization of M into prime powers. Theorem 6 provides a $(0, d-1, d)$–net $\mathcal{P}_{net}$ in base M in $[0,1)^d$. Using Lemma 7 , we get an M–coloring χ_M of $\mathcal{H}_M^d$ such that all rows contain each color exactly once and $\mathrm{disc}(\mathcal{H}_M^d, \chi_M) \leq D(\mathcal{P}_{net}, \mathcal{R}_d)$. With Lemma 8 and Theorem 5, we have $\mathrm{disc}(\mathcal{H}_N^d, M) \leq D(\mathcal{P}_{net}, \mathcal{R}_d) = O_d(\log^{d-1} M)$. □

5 Conclusion

We gave lower and upper bounds for the declustering problem. This paper contains the first complete proof of the lower bound $\Omega_d(\log^{\frac{d-1}{2}} M)$ for arbitrary values of M and d.

We propose a declustering scheme that has an additive error of $O_d(\log^{d-1} M)$ with the sole condition that $d \leq q_1 + 1$, where q_1 is the smallest prime power in the canonical factorization of M into prime powers. This improves the former best declustering schemes of Chen and Cheng [CC02], where either bounds depend on the data size N^d or $M = p^t$ and $p \geq d$ was required for a prime p and $t \in \mathbb{N}$. Furthermore, Lemma 8 improves the analysis of Chen and Cheng [CC02] of the discrepancy of latin square colorings by a factor of 2^{-d}.

The natural problem arising from this work is to close the gap between the lower and upper bound. However, this is probably a very hard one. The reason is that the corresponding problem of geometric discrepancies of boxes is extremely difficult. Closing the gap between the $\Omega_d(\log^{\frac{d-1}{2}} n)$ lower and the $O_d(\log^{d-1} n)$ upper bound for $D(n, \mathcal{R}_d)$ was baptized 'the great open problem' already in Beck and Chen [BC87]. Since then no further progress has been made for the general problem. Note that in the proof of a slight improvement due to Baker [Bak99] recently a serious error was found, so that the result was withdrawn by the author [talk of József Beck, Oberwolfach Seminar on Discrepancy Theory and Applications, March 2004].

References

[ADKS00] R. Anstee, J. Demetrovics, G. O. H. Katona, and A. Sali. Low discrepancy allocation of two-dimensional data. In *Foundations of Information and Knowledge Systems, First International Symposium*, volume 1762 of *Lecture Notes in Computer Science*, pages 1–12, 2000.

[AP00] M. J. Atallah and S. Prabhakar. (Almost) optimal parallel block access for range queries. In *Symposium on Principles of Database Systems*, pages 205–215, Dallas, 2000.

[Bak99] R. C. Baker. On irregularities of distribution II. *J. London Math. Soc.(2)*, 59:50–64, 1999.

[BC87] J. Beck and W. L. Chen. *Irregularities of distribution*, volume 89 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1987.

[BČC+02] T. Biedl, E. Čenek, T. Chan, E. Demaine, M. Demaine, R. Fleischer, and M. Wang. Balanced k-colorings. *Discrete Math.*, 254:19–32, 2002.

[BHK01] L. Babai, T. P. Hayes, and P. G. Kimmel. The cost of the missing bit: communication complexity with help. *Combinatorica*, 21:455–488, 2001.

[CBS03] C.-M. Chen, R. Bhatia, and R. K. Sinha. Multidimensional declustering schemes using golden ratio and kronecker sequences. In *IEEE Trans. on Knowledge and Data Engineering*, volume 15, 2003.

[CC02] C.-M. Chen and C. Cheng. From discrepancy to declustering: near optimal multidimensional declustering strategies for range queries. In *ACM Symp. on Database Principles*, pages 29–38, Madison, WI, 2002.

[CMA+97] C. Chang, B. Moob, A. Archarya, C. Shock, A. Sussman, and J. Saltz. Titan: a high performance remote-sensing database. In *Proc. of International Conference on Data Engineering*, pages 375–384, 1997.

[DS82] H. C. Du and J. S. Sobolewski. Disk allocation for cartesian product files on multiple disk systems. *ACM Trans. Database Systems*, 7:82–101, 1982.

[DS99] B. Doerr and A. Srivastav. Approximation of multi-color discrepancy. In D. Hochbaum, K. Jansen, J. D. P. Rolim, and A. Sinclair, editors, *Randomization, Approximation and Combinatorial Optimization (Proceedings of APPROX-RANDOM 1999)*, volume 1671 of *Lecture Notes in Computer Science*, pages 39–50, 1999.

[DS03] B. Doerr and A. Srivastav. Multicolour discrepancies. *Combinatorics, Probability and Computing*, 12:365–399, 2003.

[FB93] C. Faloutsos and P. Bhagwat. Declustering using fractals. In *Proceedings of the 2nd International Conference on Parallel and Distributed Information Systems*, pages 18 – 25, San Diego, CA, 1993.

[Mat99] J. Matoušek. *Geometric Discrepancy.* Springer-Verlag, Berlin, 1999.

[Nie87] H. Niederreiter. Point sets and sequences with small discrepancy. *Monatsh. Math.*, 104:273–337, 1987.

[PAGAA98] S. Prabhakar, K. Abdel-Ghaffar, D. Agrawal, and A. El Abbadi. Cyclic allocation of twodimensional data. In *14th International Conference on Data Engineering*, pages 94–101, Orlando, Florida, 1998.

[Rot54] K. F. Roth. On irregularities of distribution. *Mathematika*, 1:73–79, 1954.

[SBC03] R. K. Sinha, R. Bhatia, and C.-M. Chen. Asymptotically optimal declustering schemes for 2-dim range queries. *Theoret. Comput. Sci.*, 296:511–534, 2003.

[Sch72] W. M. Schmidt. On irregularities of distribution VII. *Acta Arith.*, 21:45–50, 1972.