

Improved Bounds and Schemes for the Declustering Problem^{*}

Benjamin Doerr, Nils Hebbinghaus, and 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 among the devices. Using deep results from discrepancy theory, we improve previous work of several authors concerning rectangular queries of higher-dimensional data. For this problem, 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 $d-1$ not larger than the smallest prime power in the canonical decomposition of M . Thus, in particular, our schemes work for arbitrary M in two and three dimensions, and arbitrary $M \geq d-1$ that is a power of two. These cases seem to be the most relevant in applications. For a lower bound, we show that a recent proof of a $\Omega_d(\log^{\frac{d-1}{2}} M)$ bound contains a critical error. Using an alternative approach, we establish this bound.

1 Introduction

The last decade saw dramatic improvements in computer processing speed and storage capacities. Nowadays, the bottleneck in data-intensive applications is disk I/O, the time needed to retrieve typically large amount of data from storage devices. One idea to overcome this obstacle is to spread the data on disks of multi-disk systems so that it can be retrieved in parallel. The data allocation is determined by so-called declustering schemes. Their aim is to allocate the data in such a manner that typical requests find their data evenly distributed on the disks.

A common example would be two dimensional geographic data. A typical request might ask for rectangular submap covering a particular region. The data blocks are associated with the tiles of a two dimensional grid and the queries are axis-parallel rectangles with borders along the grid, that request the data assigned to the tiles covered by the rectangle. The aim is to assign the tiles to the disks such that all disks have almost the same workload for all queries. A three dimensional application could regard the temperature distribution in a (human) body.

^{*} supported by the DFG-Graduiertenkolleg 357 “Effiziente Algorithmen und Mehrskalmethoden”.

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. We denote the number of requested blocks by $|Q|$. The response time of a query is the maximum number of blocks that are assigned to the same disk. In an ideal declustering scheme for a system with M disks, the response time of all disks for all queries Q would be exactly $|Q|/M$. The performance of a declustering scheme is measured by the worst case additive deviation from $|Q|/M$.

Declustering is an intensively studied problem and a lot of schemes with different approaches [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 introduction of discrepancy in declustering, no known declustering scheme had theoretical performance bounds in arbitrary dimension d . Such bounds were only known 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 via geometric discrepancy. 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 that their proof technique yields a bound of $\Omega(\log^{\frac{d-1}{2}} M)$ for arbitrary $d \geq 3$, but their proof contains a crucial error (cf. Section 3).

The first non-trivial upper bounds for declustering schemes in arbitrary dimension were proposed by Chen and Cheng [CC02]. They present two schemes for the d -dimensional declustering problem. The first one has an additive error of $O(\log^{d-1} M)$, 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.

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 $O(\log^{d-1} M)$ 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 view-point of application. We also show that the latin hypercube construction used by Chen and Cheng [CC02] is much better than proven there. Where they show that a latin hypercube coloring extended to the whole grid has an error of at most 2^d times the one of the latin hypercube, we show that both errors are the same.

For the lower bound, we present the first correct proof of the $\Omega(\log^{\frac{d-1}{2}} M)$ bound. Again, a more careful analysis shows that the positive discrepancy is at least $1/2d$ times the normal discrepancy instead of 3^{-d} as in [SBC03].

2 Discrepancy Theory

In this section, we sketch the connection between the declustering problem and discrepancy theory. We start by noting that declustering is in fact a combinatorial discrepancy problem.

2.1 Combinatorial Discrepancy

Recall that the declustering problem is to assign data blocks from a multi-dimensional grid system to one of M storage devices in a balanced manner. The aim is that queries to a rectangular sub-grid 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, \dots, n_d .¹ A query Q requests the data assigned to a sub-grid $[x_1..y_1] \times \cdots \times [x_d..y_d]$ for some integers $1 \leq x_i \leq y_i \leq n_i$. We assume that the time to process such a query is proportional to the maximum number of requested data blocks that are stored in a single device. If we represent the assignment of the data blocks to the devices through a mapping $\chi : V \rightarrow [M]$, then the query time of the query above is $\max_{i \in [M]} |\chi^{-1}(i) \cap Q|$, where we identify the query Q with its associated sub-grid. Clearly, no declustering scheme can do better than $|Q|/M$. Hence a natural performance measure is the additive deviation from this lower bound.

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

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

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

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

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

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

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

These notions were introduced by Srivastav and the first author in [DS99,DS03] extending the well-known notion of combinatorial discrepancy to arbitrary numbers of colors. Similar notions concerning this problem 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:

$$\begin{aligned} \text{disc}^+(\mathcal{H}, \chi) &:= \max_{i \in [M], E \in \mathcal{E}} (|\chi^{-1}(i) \cap E| - \frac{1}{M}|E|) \\ \text{disc}^+(\mathcal{H}, M) &:= \min_{\chi: V \rightarrow [M]} \text{disc}^+(\mathcal{H}, \chi) \end{aligned}$$

For many problems a distinction of these two concepts is not necessary as $\frac{1}{M-1} \text{disc}(\mathcal{H}) \leq \text{disc}^+(\mathcal{H}) \leq \text{disc}(\mathcal{H})$ holds for all hypergraphs \mathcal{H} , and the influence of the number of colors is not known for many classes of hypergraphs. This is different for the declustering problem. Summarizing the above discussion, we have

Theorem 1 *The additive error of an optimal declustering scheme for the higher-dimensional interval query problem is $\text{disc}^+(\mathcal{H}, M)$.*

Since a central result of this paper are discrepancy bounds that are 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 the case $M = 2$, a multi-dimensional 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 positive integers and q_1 the smallest prime power in the canonical factorization of M into prime powers. We have*

- (i) $\text{disc}^+(\mathcal{H}_N^d, M) = O(\log^{d-1} M)$ for $d \leq q_1 + 1$, independently of $N \in \mathbb{N}$,
- (ii) $\text{disc}^+(\mathcal{H}_N^d, M) = \Omega(\log^{\frac{d-1}{2}} M)$ for $N \geq M$,
- (iii) $\text{disc}^+(\mathcal{H}_N^d, M) = \Theta(\log M)$ 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 this area.

The problem of geometric discrepancy in the unit cube $[0, 1]^d$ is to distribute $n \in \mathbb{N}$ points evenly with respect to axis-parallel boxes: In every box R should be approximately $n \text{vol}(R)$ points, where $\text{vol}(R)$ denotes the volume of R . Again, discrepancy quantifies the distance to a perfect distribution. The discrepancy of a point set \mathcal{P} with respect to a box $R \subseteq [0, 1]^d$ is defined by

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

the discrepancy of \mathcal{P} for the set of all axis-parallel boxes \mathcal{R}_d 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

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

3 The Lower Bound

The general idea in the proofs of the lower bound in Sinha et al. [SBC03] and Anstee et al. [ADKS00] is the same, here described in two dimensions:

Starting with an arbitrary M -coloring of $[M]^2$, there is a monochromatic set \hat{P} with M vertices. Based on this set, an M -point set \mathcal{P} in $[0, 1]^2$ is constructed. Schmidt's lower bound [Sch72] ensures the existence of a rectangle R such that $D(\mathcal{P}, R) = \Omega(\log M)$. Rounding R to the $[M]^2$ grid, they construct a hyperedge \hat{R} that has approximately the volume as R . Additionally \hat{R} contains as many vertices of \hat{P} as R points of \mathcal{P} . With the help of \hat{R} and a short calculation the lower bound of the additive error $\Omega(\log M)$ is shown.

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. Recall that the authors started with a color class of exactly M^{d-1} points (we lift their analysis to arbitrary dimension). They down-scaled it by a factor of M to a set in the unit cube (that, note this fact, is a subset of $\{0, \frac{1}{M}, \frac{2}{M}, \dots, \frac{M-1}{M}\}^d$). Then their geometric discrepancy argument yields a rectangle of polylogarithmic discrepancy, which is "rounded" to obtain a subgrid with polylogarithmic discrepancy in the combinatorial setting. However, the rectangle $[0, \frac{M-1}{M}]^d$ has a much larger discrepancy: It contains all M^{d-1} points, but has a volume of $(\frac{M-1}{M})^d$ only.

This yields a discrepancy of $M^{d-1}(1 - (\frac{M-1}{M})^d) = \Omega(M^{d-2})$. If the rounding argument of Sinha et al. [SBC03] was correct, it would yield a subgrid with a discrepancy polynomial in M (for $d \geq 3$), which contradicts the known and new upper bounds. The problem is that rounding an arbitrary box to a box in the grid can cause a roundoff error which is of magnitude larger than the discrepancy. For this reason, a straight generalization of the proof of Anstee et al. [ADKS00] of the lower bound in two dimensions is not possible. In particular, we have to ensure the existence of a *small* box having large discrepancy. Beck and Chen [BC87] showed a lower bound for cubes with side at most $s := n^{-2/(2d+1)}$, where n is the number of points distributed in the unit cube $[0, 1]^d$. Still, this is too large to control the rounding error. Following the notation introduced in Beck and Chen [BC87], the cube $[-s, s]^d$ has side s , we show

Theorem 3 *For any n -point set \mathcal{P} in the unit cube $[0, 1]^d$, there is an axis-parallel cube Q with side at most $n^{-\frac{(2d-3)d}{(d-1)^2(2d+1)}}$ fully contained in $[0, 1]^d$ with*

$$D(\mathcal{P}, Q) = \Omega(\log^{\frac{d-1}{2}} n).$$

We first deduce Theorem 2 (ii) from Theorem 3.

Proof (Theorem 2 (ii)). We show the claim for $N = M$, which clearly implies the result for arbitrary $N \geq M$. Let $\chi : [M]^d \rightarrow [M]$ be a M -coloring of \mathcal{H}_M^d . Without loss of generality we may assume $|\chi^{-1}(1)| \geq M^{d-1}$. In the case $|\chi^{-1}(1)| \geq M^{d-1} + \frac{k}{2} \log^{\frac{d-1}{2}} M$, where k is the constant implicitly given in Theorem 3, we have $\text{disc}(\mathcal{H}_M^d, \chi) \geq ||\chi^{-1}(1)| - M^{d-1}| \geq \frac{k}{2} \log^{\frac{d-1}{2}} M$. Therefore, we may assume $|\chi^{-1}(1)| < M^{d-1} + \frac{k}{2} \log^{\frac{d-1}{2}} M$. For every vertex $z = (z_1, z_2, \dots, z_d) \in \chi^{-1}(1)$ we define $x_z := (\frac{2z_1-1}{2M}, \frac{2z_2-1}{2M}, \dots, \frac{2z_d-1}{2M})$. Let $\mathcal{P} := \{x_z \mid z \in \chi^{-1}(1)\}$ and $n := |\mathcal{P}|$. By Theorem 3, there is a cube $Q = \prod_{i=1}^d [x_i, x_i + 2s)$ such that the side s is at most $n^{-\frac{(2d-3)d}{(d-1)^2(2d+1)}}$ and

$$D(\mathcal{P}, Q) = ||\mathcal{P} \cap Q| - n \text{vol}(Q)| \geq k \log^{\frac{d-1}{2}} M.$$

Now we construct a box B by rounding the x_i and $x_i + 2s$ to the nearest multiple of $\frac{1}{M}$. We ensure $\mathcal{P} \cap B = \mathcal{P} \cap Q$ by rounding up $x_i + 2s$ if $x_i + 2s = \frac{h}{2M}$ and rounding x_i down if $x_i = \frac{h}{2M}$ for an odd h .

Since we have chosen a relatively small cube Q , our rounding changes the volume not to much. Using $n \geq M^{d-1}$, we get

$$|\text{vol}(Q) - \text{vol}(B)| \leq 2d \frac{1}{2M} (\frac{1}{M} + 2s)^{d-1} < d3^{d-1} M^{-(d-1)}.$$

The combinatorial counterpart of B is the box

$$\hat{B} := \{x \in [M]^d \mid (\frac{2x_1-1}{2M}, \dots, \frac{2x_d-1}{2M}) \in B\}.$$

One can easily check that $M^d \text{vol}(B) = |\hat{B}|$. By construction,

$$\begin{aligned} \text{disc}(\mathcal{H}_M^d, \chi) &\geq \left| |\chi^{-1}(1) \cap \hat{B}| - \frac{1}{M} |\hat{B}| \right| \\ &= \left| |\mathcal{P} \cap Q| - M^{d-1} \text{vol}(B) \right| \\ &= \left| (|\mathcal{P} \cap Q| - n \text{vol}(Q)) + (n \text{vol}(Q) - M^{d-1} \text{vol}(Q)) \right. \\ &\quad \left. + M^{d-1} (\text{vol}(Q) - \text{vol}(B)) \right| \\ &\geq \frac{k}{2} \log^{\frac{d-1}{2}} M - O(1) = \Omega \left(\log^{\frac{d-1}{2}} M \right). \end{aligned}$$

Thus, $\text{disc}(\mathcal{H}_M^d, M) = \Omega(\log^{\frac{d-1}{2}} M)$. 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{B} 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{B} in $[M]^d$ has at least the same discrepancy as \hat{B} , 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{B} in color 1. \square

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

To prove Theorem 3, we need some notions from Fourier analysis. Let $\mathcal{P} := \{p_1, p_2, \dots, p_n\} \subseteq \mathbb{R}^d$ and $\nu := \sum_{i=1}^n \delta_{p_i} - n\mu$, where δ_{p_i} denotes the Dirac measure concentrated on p_i and μ is the d -dimensional Lebesgue measure on $[0, 1]^d$ with $\mu([0, 1]^d) = 1$. For any $\lambda \in (0, 1]$ and $g \in L^2(\mathbb{R}^d)$ write $g_\lambda(x) := g(\lambda^{-1}x)$ for all $x \in \mathbb{R}^d$. Put $F_g := g * \nu$. Then we have

$$F_g(x) = \int_{\mathbb{R}^d} g(x-y) d\nu(y) = \sum_{i=1}^n g(x-p_i) - n \int_{\mathbb{R}^d} g(x-y) d\mu(y).$$

Let $\mathbb{1}_r$ be the characteristic function of the cube $[-r, r]^d$. Then $|F_{\mathbb{1}_r}(x)|$ is the discrepancy of $Q_r(x) := (x + [-r, r]^d) \cap [0, 1]^d$ with respect to the set \mathcal{P} :

$$|F_{\mathbb{1}_r}(x)| = \left| |\mathcal{P} \cap Q_r(x)| - n \text{vol}(Q_r(x)) \right| = \text{disc}(\mathcal{P}, Q_r(x)).$$

Let $\Delta_1(g) := \int_{\mathbb{R}^d} |F_g(x)|^2 dx$ and $\Delta_2(g) := \int_0^1 \int_{\mathbb{R}^d} |F_{g_\lambda}(x)|^2 dx d\lambda$. By Parseval's theorem for Fourier transforms we have $\Delta_1(g) := \int_{\mathbb{R}^d} |\hat{g}(t)|^2 |\hat{\nu}(t)|^2 dt$ and $\Delta_2(g) := \int_{\mathbb{R}^d} \left(\int_0^1 |\hat{g}_\lambda(t)|^2 d\lambda \right) |\hat{\nu}(t)|^2 dt$. Here \hat{f} denotes the Fourier transform

$$\hat{f} : \mathbb{R}^d \rightarrow \mathbb{C}, t \mapsto \hat{f}(t) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} f(x) e^{-ix \cdot t} dx$$

of $f : \mathbb{R}^d \rightarrow \mathbb{C}$. Let $m := n^{\frac{(2d-3)d}{(d-1)^2(2d+1)}}$. Note that $m > 1$. For the proof of Theorem 3 we need the following main lemma, which determines an average discrepancy for all cubes of side at most $\frac{1}{m}$ that intersect the unit cube $[0, 1]^d$.

Lemma 4. *We have $\Delta_2(\mathbb{1}_{\frac{1}{m}}) = \Omega(\log^{d-1} n)$.*

Let us first derive Theorem 3 from Lemma 4.

Proof (Theorem 3). We distinguish two cases. Either there exists some $r \in [0, \frac{1}{m}]$ and $x_0 \in \mathbb{R}^d$ with $|F_{\mathbb{1}_r}(x_0)| > 2n(\frac{2}{m})^d$ or there does not. In the former case, the cube Q_0 with center x_0 and side r has discrepancy at least $2n(\frac{2}{m})^d$, as we have mentioned above. This cube may cross the border of $[0, 1]^d$, but we can find a cube Q with side $\frac{1}{m}$ and $Q_0 \cap [0, 1]^d \subseteq Q$ fully contained in $[0, 1]^d$. With $n \text{vol}(Q_0) = n(2r)^d \leq n(\frac{2}{m})^d$, we see that the discrepancy of Q_0 must be caused by the excess of points in Q_0 . Therefore we have

$$D(\mathcal{P}, Q) \geq |\mathcal{P} \cap Q| - n \text{vol}(Q) \geq n(\frac{2}{m})^d = 2^d n^{\frac{1}{(d-1)^2(2d+1)}} = \Omega(\log^{\frac{d-1}{2}} n).$$

Let us assume the latter case. Lemma 4 gives us a lower bound for the average square discrepancy of all cubes of side at most $\frac{1}{m}$. Since the contribution of cubes

intersecting the border of $[0, 1]^d$ to this average square discrepancy is

$$O\left(\frac{1}{m} \left(n\left(\frac{1}{m}\right)^d\right)^2\right) = O\left(n^{-\frac{d-2}{(d-1)^2}}\right) = O(1),$$

there is a cube Q with side at most $\frac{1}{m}$ and discrepancy $\Omega(\log^{\frac{d-1}{2}} n)$ fully contained in $[0, 1]^d$. \square

It remains to prove Lemma 4. We set for all $l = (l_1, l_2, \dots, l_d) \in \mathbb{Z}^d$

$$h_l(x) := \prod_{i=1}^d \exp\left(-\frac{1}{2} l_i^2 x_i^2\right).$$

By the fact that $\hat{f}(t) = a^{-1} \exp\left(-\frac{t^2}{2a^2}\right)$ for $f(x) = \exp\left(-\frac{1}{2} a^2 x^2\right)$, the Fourier transform of h_l is $\hat{h}_l(t) = \prod_{i=1}^d \frac{1}{l_i} \exp\left(-\frac{t_i^2}{2l_i^2}\right)$. Now let L be the integer power of 2 satisfying $4(2\pi)^{\frac{d}{2}} n \leq L < 8(2\pi)^{\frac{d}{2}} n$ and

$$\mathbb{Z}^d(L, m) := \left\{ l \in \mathbb{Z}^d \mid l_i = 2^{s_i} \geq m, s_i \in \mathbb{Z}, \prod_{i=1}^d l_i = L \right\}.$$

The following three lemmas yield the Lemma 4.

Lemma 5. $|\mathbb{Z}^d(L, m)| > \Omega(\log^{d-1} n)$.

Proof. Set $L' := \log_2 L$ and $m' := \lceil \log_2 m \rceil$. Then $|\mathbb{Z}^d(L, m)|$ is the number of integral lattice points (s_1, s_2, \dots, s_d) with $\sum_{i=1}^d s_i = L'$ and $s_i \geq m'$ for all $1 \leq i \leq d$. Hence

$$|\mathbb{Z}^d(L, m)| = \binom{L' - (m' - 1)d - 1}{d - 1} \geq \frac{(L' - m'd + 1)^{d-1}}{(d-1)!}.$$

With $L' \geq \log_2 \left(4(2\pi)^{\frac{d}{2}} n\right) > \log_2 n + d + 1$ and $m' < \frac{(2d-3)d \log_2 n}{(d-1)^2(2d+1)} + 1$ we get

$$|\mathbb{Z}^d(L, m)| = \Omega(\log^{d-1} n).$$

\square

The following two lemmas are taken from Beck and Chen [BC87]:

Lemma 6 ([BC87], Lemma 6.3). $\Delta_2(\mathbb{1}_{\frac{1}{m}}) = \Omega\left(\sum_{l \in \mathbb{Z}^d(L, m)} \Delta_1(h_l)\right)$.

Lemma 7 ([BC87], Lemma 6.4). For every $l \in \mathbb{Z}^d(L, m)$ we have

$$\Delta_1(h_l) = \Omega(1).$$

Now Lemma 4 is a direct consequence of Lemma 5, 6 and 7. We get

$$\Delta_2(\mathbb{1}_{\frac{1}{m}}) = \Omega \left(\sum_{l \in \mathbb{Z}^d(L,m)} \Delta_1(h_l) \right) = \sum_{l \in \mathbb{Z}^d(L,m)} \Omega(1) = \Omega(\log^{d-1} n).$$

It remains to prove the lower bound of Theorem 2 (iii). Anstee et al. [ADKS00] only treated latin square type colorings of $[M]^2$. However, the proof is easily extended through the triangle inequality argument used in the proof of Theorem 2 (ii).

4 The Upper Bound

In this section, we present a declustering scheme showing our upper bound. As in previous work, we use geometric discrepancies 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 $E = \prod_{i=1}^d [a_i b^{-d_i}, (a_i + 1) b^{-d_i} [$, 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 $O(\log^{d-1} n)$ elementary intervals of volume b^{t-m} , the following is immediate:

Theorem 8 *A (t, m, d) -net \mathcal{P}_{net} in base b with $n = b^m$ points has discrepancy*

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

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 9 *Let $b \geq 2$ be an arbitrary base and $b = q_1 q_2 \dots q_u$ be the canonical factorization of b into prime powers such that $q_1 < \dots < q_u$. 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 10. 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 11 we use the M -coloring of \mathcal{H}_M^d to construct an M -coloring of \mathcal{H}_N^d with same discrepancy.

Lemma 10. *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 once² and*

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

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 elementary ‘‘rows’’, i.e., 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 $\hat{\mathcal{P}} := \left\{ x \in [M]^d \mid \mathcal{P}_{net} \cap \prod_{i=1}^d [\frac{x_i-1}{M}, \frac{x_i}{M}) \neq \emptyset \right\}$. Then each row of $[M]^d$ contains exactly one point of $\hat{\mathcal{P}}$. We define the coloring $\chi_M : [M]^d \rightarrow [M]$ by $\chi_M(y, x_2, \dots, x_d) = i$ for all $x = (x_1, x_2, \dots, x_d) \in \hat{\mathcal{P}}$, $i, y \in [M]$ such that $y \equiv x_1 + (i-1) \pmod{M}$. 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

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

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 $R = \prod_{i=1}^d [\frac{x_i-1}{M}, \frac{y_i}{M})$. Then $|\hat{\mathcal{P}} \cap \hat{R}| = |\mathcal{P}_{net} \cap R|$ and $|\hat{R}| = M^d \text{vol}(R)$. Thus the combinatorial discrepancy of \hat{R} equals the geometric one of R . We have

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

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

Lemma 11. *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 $\chi(x_1, \dots, x_d) = \chi_M(y_1, \dots, y_d)$ with $x_i \equiv y_i \pmod{M}$ for $i \in [d]$, $x_i \in [N]$, $y_i \in [M]$. Then*

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

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

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 $\tilde{x}_i, \tilde{y}_i \in [M]$ with $x_i \equiv \tilde{x}_i \pmod{M}$ respectively $y_i \equiv \tilde{y}_i \pmod{M}$. Set $\bar{x}_i := \min\{\tilde{x}_i, \tilde{y}_i\}$ and $\bar{y}_i := \max\{\tilde{x}_i, \tilde{y}_i\}$ for all $i \in [d]$. We have $\text{disc}(\hat{R}, \chi) = \text{disc}([x_1..y_1] \times [x_2..y_2] \times \dots \times [x_d..y_d], \chi)$, since whole rows have discrepancy zero. Applying this successively in every coordinate we get

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

□

Lemma 11 is a remarkable improvement of Theorem 4.2 in [CC02], where $\text{disc}(\mathcal{H}_N^d, \chi) \leq 2^d \text{disc}(\mathcal{H}_M^d, \chi_M)$ 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 10 and Lemma 11.

Proof (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 9 provides a $(0, d-1, d)$ -net \mathcal{P}_{net} in base M in $[0, 1)^d$. Using Lemma 10, we get an M -coloring χ_M of \mathcal{H}_M^d such that all rows contain each color exactly once and $\text{disc}(\mathcal{H}_M^d, \chi_M) \leq D(\mathcal{P}_{net}, \mathcal{R}_d)$. With Lemma 11 and Theorem 8, we have $\text{disc}(\mathcal{H}_N^d, M) \leq D(\mathcal{P}_{net}, \mathcal{R}_d) = O(\log^{d-1} M)$. □

5 Conclusion

We gave lower and upper bounds for the declustering problem. This paper contains the first complete and correct proof of the lower bound $\Omega(\log^{\frac{d-1}{2}} M)$ for arbitrary values of M and d . Moreover, the implicit constant was improved by a factor of $\frac{3^d}{2^d}$.

We propose a declustering scheme that has an additive error of $O(\log^{d-1} M)$ 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 11 improves the analysis of Chen and Cheng [CC02] of the discrepancy of latin square colorings by a factor of 2^{-d} .

The natural problem to close the gap between the lower and upper bound is probably a very hard one. The reason is that the corresponding problem of geometric discrepancies of rectangles seems to be extremely difficult. Closing the gap between the $\Omega(\log^{\frac{d-1}{2}} n)$ lower and the $O(\log^{d-1} n)$ upper bound 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 bug was found [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, Berlin–Heidelberg, 1999. Springer Verlag.
- [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.
- [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.