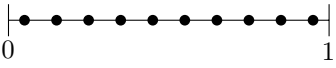
1. Introduction

In this chapter, we introduce the concept of discrepancy. We formulate a basic problem concerning discrepancy for rectangles, we show its connections to the discrepancy of infinite sequences in the unit interval, and we briefly comment on the historical roots of discrepancy in the theory of uniform distribution (Section 1.1). In Section 1.2, we introduce discrepancy in a general geometric setting, as well as some variations of the basic definition. Section 1.3 defines discrepancy in a seemingly different situation, namely for set systems on finite sets, and shows a close relationship to the previously discussed "Lebesgue-measure" discrepancy. Finally, Section 1.4 is a mosaic of notions, results, and comments illustrating the numerous and diverse connections and applications of discrepancy theory. Most of the space in that section is devoted to applications in numerical integration and similar problems, which by now constitute an extensive branch of applied mathematics, with conventions and methods quite different from "pure" discrepancy theory.

1.1 Discrepancy for Rectangles and Uniform Distribution

The word *discrepancy* means "disagreement" (from Latin *discrepare*—to sound discordantly). In our case it is a "disagreement between the ideal situation and the actual one," namely a "deviation from a perfectly uniform distribution."

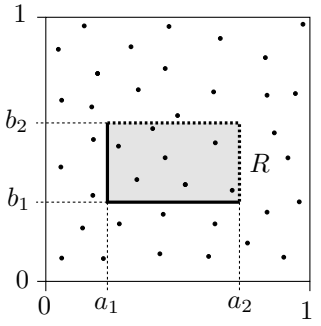
We will investigate how uniformly an n-point set can be distributed in the d-dimensional unit cube $[0,1]^d$. For $d = 1$, the set of n equidistant points as in the following picture



hardly finds serious competitors as a candidate for the most uniformly distributed n-point set in the unit interval. But already in dimension 2, one can come up with several reasonable criteria of uniform distribution, and sets that are very good for some may be quite bad for others.

J. Matoušek, *Geometric Discrepancy*, Algorithms and Combinatorics 18,
DOI 10.1007/978-3-642-03942-3_1,

Here is one such criterion: "uniformly" means, for the moment, "uniformly with respect to axis-parallel rectangles." Let P be an n-point set in the unit square $[0,1]^2$. Let us consider an axis-parallel rectangle[1] $R = [a_1, b_1) \times [a_2, b_2) \subseteq [0,1]^2$:



For a uniformly distributed set P, we expect that the number of points of P that fall in the rectangle R is approximately $n \cdot \mathrm{vol}(R)$, where $\mathrm{vol}(R)$ denotes the area of R. (Note that $n \cdot \mathrm{vol}(R)$ is the expected number of points hitting R if we pick n points in the unit square uniformly and independently at random.) Let us call P *justly distributed* if the deviation

$$\big|n \cdot \mathrm{vol}(R) - |P \cap R|\big|$$

is at most 100 for all axis-parallel rectangles R. Do arbitrarily large justly distributed set exist? (Or, should the constant 100 be too small, we can ask if the deviation can be bounded by some other constant, possibly large but independent of n, P, and R.) This is one of the fundamental questions that gave birth to discrepancy theory. Since we do not hope to keep the reader in suspense until the end of the book by postponing the answer, we can just as well state it right away: no, just distribution is impossible for sufficiently large sets. Any distribution of n points in the unit square has to display a significant irregularity for some rectangle R, and the magnitude of the irregularity must grow to infinity as $n \to \infty$. For this particular two-dimensional problem, it is even known fairly precisely how large this irregularity must be, and we will see the corresponding lower and upper bound proofs later in this book. The proofs may perhaps seem simple, but one should not forget that the presentation is based on the work of outstanding mathematicians and that originally the problem looked formidably difficult. To put these results into a better perspective, we remark that already the obvious generalization of the problem in dimension 3 has so far defied all attempts at obtaining a quantitatively precise answer.

[1] For technical reasons, we take semi-open rectangles—the left side and the bottom side are included, the right and top sides are not. For the discrepancy this doesn't matter much; we only accept this convention for simplifying some formulas in the sequel.

Here is some notation for expressing these questions and answers. First we introduce the symbol $D(P, R)$ for the deviation of P from uniform distribution on a particular rectangle R, namely

$$D(P, R) = n \cdot \mathrm{vol}(R) - |P \cap R|.$$

Let $\mathcal{R}_2$ denote the set of all axis-parallel rectangles in the unit square. The quantity

$$D(P, \mathcal{R}_2) = \sup_{R \in \mathcal{R}_2} |D(P, R)|$$

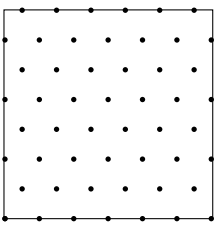
is called the *discrepancy of* P for axis-parallel rectangles, and the function

$$D(n, \mathcal{R}_2) = \inf_{\substack{P \subset [0,1]^2 \\ |P| = n}} D(P, \mathcal{R}_2)$$

quantifies the smallest possible discrepancy of an n-point set. The above question about a just distribution can thus be re-formulated as follows:

1.1 Problem. *Is $D(n, \mathcal{R}_2)$ bounded above by a constant for all n, or does $\limsup_{n\to\infty} D(n, \mathcal{R}_2) = \infty$ hold?*

In this book, we will judge the uniformity of distribution exclusively in terms of discrepancy, but we should remark that there are also other sensible criteria of uniform distribution. For example, one such criterion might be the minimum distance of two points in the considered set. This concept is also studied quite extensively (in the theory of ball packings, in coding theory, and so on), but it is quite distinct from the uniform distribution measured by discrepancy. For example, the set in the unit square maximizing the minimum interpoint distance is (essentially) a triangular lattice:



As it turns out, this set is quite bad from the discrepancy point of view: it has discrepancy about $\sqrt{n}$, while in Chapter 2 we will learn how to produce sets with only $O(\log n)$ discrepancy. On the other hand, a set with a very good discrepancy may contain two very close points.

Uniform Distribution of Infinite Sequences. The question about the "most uniform" distribution in the one-dimensional interval $[0, 1]$ is trivial for an n-point set, but it becomes quite interesting for an infinite sequence $u = (u_1, u_2, \ldots)$ of points in $[0, 1]$. Here we want that if the points of u are added one by one in the natural order, they "sweep out" all subintervals

of $[0,1]$ as evenly as possible. This is actually the setting where discrepancy theory began. So let us outline the definitions concerning uniform distribution of sequences.

The sequence $u = (u_1, u_2, \ldots)$ is called *uniformly distributed* in $[0,1]$ if we have, for each subinterval $[a,b) \subset [0,1]$,

$$\lim_{n\to\infty} \left(\frac{1}{n} \big| \{u_1, \ldots, u_n\} \cap [a,b) \big| \right) = b - a. \tag{1.1}$$

Uniformly distributed sequences have the following seemingly stronger property (which is actually not difficult to prove from the just given definition of uniform distribution). For any Riemann-integrable function $f\colon [0,1] \to \mathbf{R}$, we have

$$\lim_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} f(u_i) \right) = \int_0^1 f(x)\,\mathrm{d}x. \tag{1.2}$$

Note that (1.1) is a particular case of the last equation, with the characteristic function of the interval $[a,b)$ in the role of f. Thus, in order to test the validity of (1.2) for all Riemann-integrable functions f, it suffices to consider all characteristic functions of intervals in the role of f.

Another interesting class of functions which are sufficient for testing (1.1) are the trigonometric polynomials, i.e. functions of the form

$$f(x) = \sum_{k=0}^{n} \Big(a_k \sin(2\pi k x) + b_k \cos(2\pi k x) \Big)$$

with real or complex coefficients $a_0, a_1, \ldots, a_n$ and $b_0, b_1, \ldots, b_n$. More conveniently, a trigonometric polynomial can be written using the complex exponential: $f(x) = \sum_{k=-n}^{n} c_k e^{2\pi \mathrm{i} k x}$, with i standing for the imaginary unit. From a basic approximation theorem involving trigonometric polynomials (a suitable version of Weierstrass' approximation theorem), it can be shown that if (1.2) holds for all trigonometric polynomials f, then the sequence u is uniformly distributed. Since any trigonometric polynomial is a linear combination of the functions $x \mapsto e^{2\pi \mathrm{i} k x}$ for various integers k, and since for $k = 0$, the condition (1.2) with the function $f(x) = e^{2\pi \mathrm{i} 0 x} = 1$ is trivially satisfied by any sequence u, the following criterion is obtained: a sequence $u = (u_1, u_2, \ldots)$ is uniformly distributed in $[0,1]$ if and only if we have, for all integers $k \neq 0$,

$$\lim_{n\to\infty} \left(\frac{1}{n} \sum_{j=1}^{n} e^{2\pi \mathrm{i} k u_j} \right) = \int_0^1 e^{2\pi \mathrm{i} k x}\,\mathrm{d}x = 0.$$

This result is called *Weyl's criterion*. Here is a simple but lovely application:

1.2 Theorem. *For each irrational number α, the sequence $u = (u_1, u_2, \ldots)$ given by $u_n = \{\alpha n\}$ is uniformly distributed in $[0,1]$. (Here $\{x\}$ denotes the fractional part of x.)*

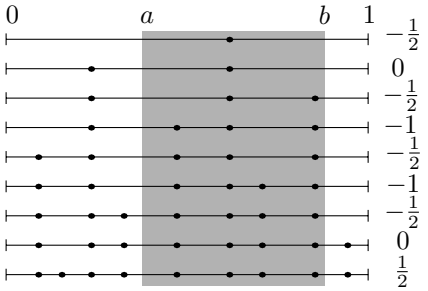


Fig. 1.1. Adding the terms of a sequence one by one; the numbers on the right are the deviations $n(b-a) - |\{u_1, \dots, u_n\} \cap [a,b)|$ for the marked interval $[a,b)$ of length $\frac{1}{2}$.

Proof. We use Weyl's criterion. This is particularly advantageous here since we have $e^{2\pi i k u_n} = e^{2\pi i k \alpha n}$, with the unpleasant "fractional part" operation disappearing. Putting $A_k = e^{2\pi i k \alpha}$, we calculate

$$\sum_{j=1}^{n} e^{2\pi i k u_j} = \sum_{j=1}^{n} A_k^j = \frac{A_k^{n+1} - A_k}{A_k - 1}.$$

We have $|A_k| = 1$, and since α is irrational, $k\alpha$ is never an integer for a nonzero k, and so $A_k \neq 1$. Therefore, $\left|\frac{A_k^{n+1}-A_k}{A_k-1}\right| \leq \frac{2}{|A_k-1|}$ is bounded by a number independent of n, and we have

$$\lim_{n\to\infty} \left(\frac{1}{n} \sum_{j=1}^{n} e^{2\pi i k u_j} \right) = 0$$

as required. □

Discrepancy of Sequences: a "Dynamic" Setting. We now know that all the sequences $(\{n\alpha\})$ with α irrational are uniformly distributed, but if one looks into the matter more closely, one finds that some are more uniformly distributed than the others. Discrepancy was first introduced as a quantitative measure of non-uniformity of distribution for infinite sequences. We define the *discrepancy of an infinite sequence* u in $[0,1]$ as the function

$$\Delta(u,n) = \sup_{0\leq a\leq b\leq 1} \Big| n(b-a) - \big|\{u_1, \dots, u_n\} \cap [a,b)\big| \Big|$$

(see Fig. 1.1). The original formulation of Problem 1.1 actually was: does there exist a sequence u with $\Delta(u,n)$ bounded by a constant for all n?

Let us sketch the connection of this formulation concerning infinite sequences to the formulation with axis-parallel rectangles. First, suppose that u is some given sequence in $[0,1]$. We claim that for every natural number n, there exists an n-point set $P \subset [0,1]^2$ with

$$D(P, \mathcal{R}_2) \le 2 \max\{\Delta(u,k)\colon k = 1, 2, \ldots, n\} + 2. \tag{1.3}$$

A suitable set P can be defined as the "graph" of the first n terms of u. Namely, we put

$$P = \left\{\left(\tfrac{1}{n}, u_1\right), \left(\tfrac{2}{n}, u_2\right), \left(\tfrac{3}{n}, u_3\right), \ldots, \left(\tfrac{n}{n}, u_n\right)\right\}.$$

We leave it as Exercise 1(a) to verify that (1.3) indeed holds for this P. Conversely, suppose that we have an n-point set P in $[0,1]^2$. Let us list the points of P in the order of increasing x-coordinates; that is, write $P = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_1 \le x_2 \le \cdots \le x_n$. Then it is not difficult to verify that if u is a sequence with $y_1, y_2, \ldots, y_n$ as the first n terms, we have

$$\Delta(u,k) \le 2D(P, \mathcal{R}_2) \quad \text{for all } k = 1, 2, \ldots, n \tag{1.4}$$

(Exercise 1(b)). Finally, with a little more work one can show that if we have $D(n, \mathcal{R}_2) \le f(n)$ for some nondecreasing function f and for all n, then there exists a sequence u with $\Delta(u,n) = O(f(n))$. Therefore, the question about an infinite sequence with bounded discrepancy and Problem 1.1 are equivalent in a strong sense—even the quantitative bounds are the same up to a small multiplicative constant.

The difference between the discrepancy $D(P, \mathcal{R}_2)$ of a finite point set and the discrepancy $\Delta(u,n)$ of an infinite sequence is not so much in the finite/infinite distinction (note that $\Delta(u,n)$ is well-defined even for a finite sequence with at least n terms), but rather, it distinguishes a "static" and a "dynamic" setting. In the definition of the discrepancy for rectangles, we deal with the behavior of the whole set P, whereas in the definition of $\Delta(u,n)$, we look at all the initial segments $\{u_1\}$, $\{u_1, u_2\}$,..., $\{u_1, u_2, \ldots, u_n\}$ simultaneously. If we start with the empty interval $[0,1]$ and add the points of the sequence one by one in the natural order, the current set should be uniformly distributed all the time. Note that the discrepancy of a sequence can change drastically by rearranging the terms into a different order (while the discrepancy of a set does not depend on any ordering of the points). As the above reductions show, the dynamic problem in dimension 1 is more or less equivalent to the static problem in dimension 2, and similar reductions are possible between dynamic settings in dimension d and static settings in dimension $d+1$. In this book, we will mostly treat the static case.

Bibliography and Remarks. Discrepancy theory grew out of the theory of uniform distribution. A nice and accessible book where this development can be followed is Hlawka [Hla84]. The fact that the discrepancy for axis-parallel rectangles grows to infinity, in the equivalent formulation dealing with one-dimensional infinite sequences, was conjectured by Van der Corput [Cor35a], [Cor35b] and first proved by Van Aardenne-Ehrenfest [AE45], [AE49]. Her lower bound for the discrepancy was improved by Roth [Rot54], who invented the two-dimensional

formulation of Problem 1.1 and used it to establish a much stronger lower bound for the discrepancy in question[2] (see Section 6.1).

A foundational paper in the theory of uniform distribution is due to Weyl [Wey16]. Earlier, uniform distribution of one-dimensional sequences $(\{n\alpha\})$ with irrational α was proved by several authors by elementary means, but the criterion involving exponential sums enabled Weyl to establish a multidimensional analogue—uniform distribution of Kronecker sequences; see Section 2.5.

Weyl's criterion uses trigonometric polynomials for testing uniform distribution of a sequence. There are also sophisticated results in this spirit bounding the discrepancy of a sequence in terms of certain trigonometric sums. The most famous of such theorems is perhaps the Erdős–Turán inequality: for any sequence $u = (u_1, u_2, \ldots)$ of points in $[0, 1]$ and any integer $H \geq 1$, we have

$$\Delta(u, n) \leq \frac{10n}{H+1} + \frac{4}{\pi} \sum_{h=1}^{H} \frac{1}{h} \left| \sum_{k=1}^{n} e^{2\pi i h u_k} \right|$$

(Hlawka [Hla84] has a masterly exposition). A multidimensional version of this inequality is due to Koksma, and various other estimates of this type are known (see e.g. [DT97]). Such inequalities are useful but in general they need not give tight bounds and sometimes the trigonometric sums may be too difficult to estimate.

There is an extensive literature and many beautiful results concerning the uniform distribution and various kinds of discrepancy of specific sequences, such as the sequences $(\{n\alpha\})$ for irrational α and their higher-dimensional analogues. A minor sample of theorems will be mentioned in Section 2.5; much more material and citations can be found in the books Drmota and Tichy [DT97] or Kuipers and Niederreiter [KN74], or also in the lively surveys Sós [Sós83a] and Beck and Sós [BS95].

Some of these results are closely connected to ergodic theory and similar branches of mathematics. Some well-known low-discrepancy sequences can be obtained from the initial point by iterating a suitable ergodic transform, and the ergodicity of the transform is directly related to the uniform distribution of the sequence. For example, for α irrational and for any $x_0 \in \{0, 1\}$, the sequence $(\{n\alpha + x_0\})_{n=0}^{\infty}$ is uniformly distributed in $[0, 1)$. Consequently we have, for any Riemann-integrable function f and all $x_0 \in [0, 1)$,

$$\lim_{n\to\infty} \left(\frac{1}{n} \sum_{i=1}^{n} f(T^i x_0) \right) = \int_0^1 f(x)\, \mathrm{d}x, \tag{1.5}$$

[2] The significance of this paper of Roth is also documented by the subsequent popularity of its title in discrepancy theory—look, for instance, at the list of references in [BC87].

where $T\colon [0,1) \to [0,1)$ is given by $Tx = \{x + \alpha\}$. This T is obviously a measure-preserving transform of $[0,1)$, and (1.5) is the conclusion of Birkhoff's ergodic theorem for this T (more precisely, the ergodic theorem would only imply (1.5) for *almost all* x_0). And indeed, T is and important example of an ergodic transform. The connection of other low-discrepancy sequences to ergodic transforms has been investigated by Lambert [Lam85]. On the other hand, some results first discovered for the $(\{n\alpha\})$ sequences were later generalized to flows (see [DT97] or [Sós83a] for some references).

The notion of uniform distribution of a sequence can be generalized considerably, for instance to sequences in a compact topological group X, by requiring that $\lim_{n\to\infty}\bigl(\frac{1}{n}\sum_{i=1}^{n} f(u_i)\bigr) = \int_X f(x)\,\mathrm{d}x$ for all continuous functions f. Or, instead of a discrete sequence u, one can look at the uniform distribution of a function $u\colon [0,\infty) \to \mathbf{R}^d$, where uniform distribution can be defined by the condition $\lim_{t\to\infty}\bigl(\frac{1}{t}\int_0^t f(u(t))\,\mathrm{d}t\bigr) = \int_{\mathbf{R}^d} f(x)\,\mathrm{d}x$, and so on.

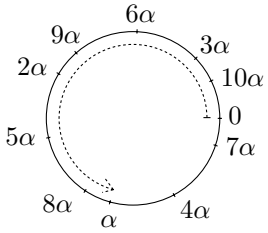
Books and Surveys. A basic source on older results in geometric discrepancy theory, more comprehensive in this respect than the present text, is a book by Beck and Chen [BC87]. A newer excellent overview, with many references but only a few proofs, is a handbook chapter by Beck and Sós [BS95]. Alexander et al. [ABC97] also give a brief but delightful survey. An extensive recent monograph with an incredible count of 2000 bibliography references is Drmota and Tichy [DT97]. It covers results from many directions, but its main focus is the classical uniform distribution theory (investigation of the discrepancy of particular sequences etc.). The books Spencer [Spe87], Alon and Spencer [AS00], Montgomery [Mon94], and Pach and Agarwal [PA95] have nice but more narrowly focused chapters on discrepancy. Chazelle [Cha00] is a monograph on discrepancy and its relations to theoretical computer science. Discrepancy theory is now an extensive subject with many facets, reaching to a number of mathematical disciplines. The amount of available material and literature makes any account of the size of a normal book necessarily incomplete. It is no wonder that more narrowly focused subfields tend to single out and the communication and flow of results and ideas between these areas are often nontrivial.

Exercises

1. Let $u = (u_1, u_2, \ldots)$ be an infinite sequence of real numbers in the interval $[0,1]$.
   (a) Verify that if an n-point set P is constructed from u as in the text above then (1.3) holds. (Consider the rectangles $[0, \frac{i}{n}) \times [a,b)$ first.)

(b)* Show that if P is a given n-point set in $[0,1]^2$ and u is a sequence with the first n terms defined as in the text above then (1.4) holds.
(c)* Show that if $D(n,\mathcal{R}_2) \le f(n)$ for some nondecreasing function f and for all n, then there exists a sequence u with $\Delta(u,n) = O(f(n))$, where the constant of proportionality depends on f.

2. (Three-distance theorem)
(a)** Let α be a real number, let n be a natural number, and let $0 \le z_1 \le z_2 \le \cdots \le z_n < 1$ be the first n terms of the sequence $(\{i\alpha\})_{i=1}^{\infty}$ listed in a nondecreasing order. Prove that the differences $|z_{j+1} - z_j|$, $j = 1, 2, \ldots, n-1$, attain at most three distinct values. Moreover, if there are three values $\delta_1 < \delta_2 < \delta_3$, then $\delta_3 = \delta_1 + \delta_2$. It may be instructive to imagine that the real axis with the numbers 0, α, 2α,..., $n\alpha$ on it is wound around a circle of unit length, which produces a picture similar to the following one (here $\alpha = 1/\sqrt{2}$):



(b)** Let α be irrational, and p be the permutation of the set $\{1, 2, \ldots, n\}$ such that $0 < \{p(1)\alpha\} < \{p(2)\alpha\} < \cdots < \{p(n)\alpha\} < 1$. Show that the whole of p can be determined by the knowledge of $p(1)$ and $p(n)$ (without knowing α). (This illustrates that the sequence $(\{i\alpha\})_{i=1}^{\infty}$ is highly non-random in many respects, although it might perhaps look random-like at first sight.)

These results are due to Sós [Sós58], and we refer to that paper for a solution of this exercise.

1.2 Geometric Discrepancy in a More General Setting

Discrepancy is also studied for classes of geometric figures other than the axis-parallel rectangles, such as the set of all balls, or the set of all boxes, and so on. For discrepancy, only the part of a set $A \in \mathcal{A}$ lying in the unit cube $[0,1]^d$ is important. We are interested in finding an n-point set $P \subset [0,1]^d$ such that the fraction of points of P lying in A is a good approximation of the volume of $A \cap [0,1]^d$, and the discrepancy measures the accuracy of such an approximation. For more convenient notation, let us write $\mathrm{vol}_\square(A)$ for $\mathrm{vol}(A \cap [0,1]^d)$.

For an n-point set $P \subset [0,1]^d$ and $A \in \mathcal{A}$, we put

$$\begin{aligned} D(P,A) &= n \cdot \mathrm{vol}_\square(A) - |P \cap A| \\ D(P,\mathcal{A}) &= \sup_{A \in \mathcal{A}} |D(P,A)|. \end{aligned}$$

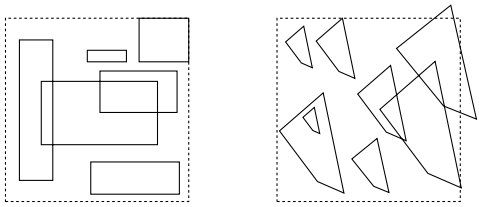
The quantity $D(P,\mathcal{A})$ is called the *discrepancy of P for $\mathcal{A}$*. Further we define the *discrepancy function* of $\mathcal{A}$:

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

Hence, in order to show that $D(n,\mathcal{A})$ is small (an upper bound), we must exhibit one n-point set and verify that it is good for all $A \in \mathcal{A}$. To prove that $D(n,\mathcal{A})$ is large (lower bound), we have to demonstrate that for any n-point set P, given by the enemy, there exists a bad $A \in \mathcal{A}$.

A Warning Concerning Notational Conventions. Let us stress that in this book, the discrepancy is measured in units of *points of P*. Often it is more natural to work with the *relative* error, i.e. with the quantity $\frac{1}{n}D(P,A)$, and this is also what one finds in a significant part of the literature. Indeed, $\frac{1}{n}D(P,A)$ is the relative error made by approximating $\mathrm{vol}_\square(A)$ by the fraction of points of P falling into A, and in many applications, the relative error is prescribed and we are looking for the smallest point set providing the desired accuracy. This interpretation becomes somewhat obscured by the definition of discrepancy we use, with the unit of one point. Nevertheless, we stick to the more traditional way, which usually leads to nicer formulas.

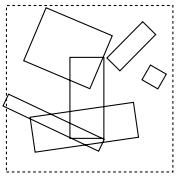
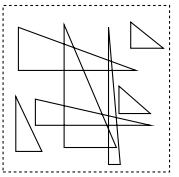
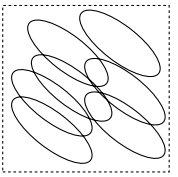
Two Basic Types of Behavior of the Discrepancy Function. One can try to classify the various classes $\mathcal{A}$ of geometric shapes according to the behavior of their discrepancy function. Here is perhaps the most significant (and a bit vague) division. On the one hand, we have classes consisting of scaled and translated copies of a fixed polygon or polytope, such as the class $\mathcal{R}_d$ of all axis-parallel boxes (no rotation is allowed). Two such families in the plane are indicated below:



discrepancy
about $\log n$

For such classes, as a rule, the discrepancy function is bounded from above and from below by some constant powers of $\log n$. On the other hand, for rotationally invariant classes, such as halfspaces or rectangular boxes in arbitrarily rotated positions, the discrepancy function behaves like a fractional

power of n, and in higher dimensions it is quite close to $\sqrt{n}$. Similar behavior occurs for translated, or translated and scaled, copies of a set with a smooth curved boundary, such as a disc or an ellipsoid. Three examples are schematically depicted below:

discrepancy about $n^{1/4}$

The middle example, the family of all triangles with two sides parallel to the axes, is particularly striking when compared with the case of axis-parallel rectangles. There are point sets giving small discrepancy, of the order $\log n$, for all axis-parallel rectangles, but if we slice each rectangle by its diagonal, some of the resulting triangles have *much* larger discrepancy.

Surprisingly, no natural classes of geometric objects are known with an intermediate behavior of discrepancy (larger than a power of $\log n$ but smaller than any fixed power of n).

The just indicated basic classification of shapes also strongly influences the subdivision of this book into chapters and sections: the two cases, classes with polylogarithmic discrepancy and classes with much larger discrepancy, usually involve distinct techniques and are mostly treated separately. Another general wisdom to remember for the study of discrepancy is this: *look at the boundary*. The irregularity of distribution always "happens" close to the boundary of the considered set, and the boundary length and shape influence the magnitude of the irregularity. The area of the considered sets, for example, is much less significant. Again, I know of no suitable exact formulation of this principle, but we will see some examples throughout the book.

More Generalizations and Variations. Clearly, discrepancy can be defined in yet more general situations. One obvious generalization is to replace the unit cube $[0,1]^d$ by other domains (a frequently investigated case is the d-dimensional unit sphere S^d), or even by complicated sets like fractals. In this book, we mostly keep working with the unit cube, since this setting seems appropriate for the first encounter with most of the ideas, and also most of the known results are formulated for the unit cube situation.

Later we will meet interesting generalizations of discrepancy in other directions, such as average discrepancy, combinatorial discrepancy, discrepancy of weighted point sets, discrepancies with respect to classes of functions, toroidal discrepancy, etc.

Decomposing Geometric Shapes for Bounding Discrepancy. We now mention a simple observation, which often allows us to simplify the class of sets for which the discrepancy is studied.

1.3 Observation. *If A, B are disjoint (and measurable) sets, then $|D(P, A \cup B)| = |D(P, A) + D(P, B)| \le |D(P, A)| + |D(P, B)|$, for an arbitrary finite set P. Similarly for $A \subseteq B$, we have $|D(P, B \setminus A)| \le |D(P, A)| + |D(P, B)|$.* □

As an example, we express axis-parallel rectangles (and boxes in higher dimensions) using simpler sets. For a point $x = (x_1, \dots, x_d) \in [0,1]^d$, we define the *corner*[3] *with vertex at* x as the set

$$C_x = [0, x_1) \times [0, x_2) \times \cdots \times [0, x_d).$$

This corner is also written $C_{(x_1, x_2, \dots, x_d)}$. Let $\mathcal{C}_d = \{C_x \colon x \in [0,1]^d\}$ be the set of all d-dimensional corners.

1.4 Observation. *For any finite set $P \subseteq [0,1]^d$ we have*

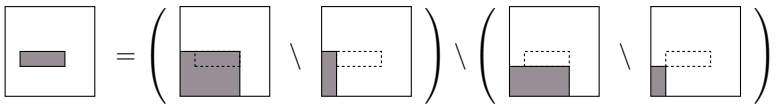
$$D(P, \mathcal{C}_d) \le D(P, \mathcal{R}_d) \le 2^d D(P, \mathcal{C}_d)$$

($\mathcal{R}_d$ stands for the set of all (semi-open) axis-parallel boxes in dimension d).

Sketch of Proof. The first inequality is obvious (each corner is an axis-parallel box). To see the second inequality, we express any axis-parallel box R using 2^d corners. For instance, in the plane we have

$$[a_1, b_1) \times [a_2, b_2) = \big(C_{(b_1,b_2)} \setminus C_{(a_1,b_2)}\big) \setminus \big(C_{(b_1,a_2)} \setminus C_{(a_1,a_2)}\big),$$

pictorially



Finding the expression for a d-dimensional box using 2^d corners is left as an exercise. □

Thus, if we are not interested in the exact constant of proportionality, we can estimate the discrepancy for corners instead of that for axis-parallel rectangles.

Let us remark that the discrepancy for corners is frequently treated in the literature, and it is often denoted by D^* and called, for historical reasons, the *star-discrepancy*.

Average Discrepancy. In our definition above, the discrepancy $D(P, \mathcal{A})$ is taken as a supremum over all sets $A \in \mathcal{A}$, so it is a discrepancy *in the worst case*. In order to show a lower bound for discrepancy of some point set, it suffices to exhibit a single bad set A from the class $\mathcal{A}$ of allowed shapes. In most of the known proofs, one actually shows that a "random" or "average"

[3] In the literature, corners are sometimes called *anchored boxes*.

set from $\mathcal{A}$ must be bad. For this, and also for various applications, we need to define an average discrepancy. Since the set $\mathcal{A}$ is, in general, infinite, in order to speak of an average over $\mathcal{A}$, we have to fix a measure ν on $\mathcal{A}$. For convenience we assume that it is a probability measure, i.e. that $\nu(\mathcal{A}) = 1$. For the time being, we give only one example of such a measure ν, namely a measure on the set $\mathcal{C}_d$ of all d-dimensional corners. Since each corner C_x is determined by its vertex $x \in [0,1]^d$, we can define the measure of a set of corners $\mathcal{K} \subseteq \mathcal{C}_d$ as $\mathrm{vol}(\{x \in [0,1]^d \colon C_x \in \mathcal{K}\})$.

For a given number p, $1 \le p < \infty$, and for a probability measure ν on $\mathcal{A}$ we define the *p-th degree average discrepancy* (also called the *L_p-discrepancy*) as follows:

$$D_{p,\nu}(P,\mathcal{A}) = \Big(\int_{\mathcal{A}} |D(P,A)|^p \mathrm{d}\nu(A)\Big)^{1/p}$$
$$D_{p,\nu}(n,\mathcal{A}) = \inf_{\substack{P\subset[0,1]^d\\ |P|=n}} D_{p,\nu}(P,\mathcal{A}).$$

If the measure ν is clear from the context, we only write D_p instead of $D_{p,\nu}$. For example, the concrete formula for the L_p-discrepancy for corners is

$$D_p(P,\mathcal{C}_d) = \Big(\int_{[0,1]^d} |D(P,C_x)|^p \,\mathrm{d}x\Big)^{1/p}.$$

It is easy to see that for any p and any ν, we have $D_{p,\nu}(P,\mathcal{A}) \le D(P,\mathcal{A})$ (the integral over a region of unit measure is upper-bounded by the maximum of the integrated function). By a well-known inequality for L_p-norms, we also have $D_{p,\nu}(P,\mathcal{A}) \le D_{p',\nu}(P,\mathcal{A})$ whenever $p \le p'$.

Some people may find it convenient to think about the L_p-discrepancy using a probabilistic interpretation. If the set P is fixed and $A \in \mathcal{A}$ is chosen at random according to the probability measure ν, then the discrepancy $D(P,A)$ is a random variable, and $D_{p,\nu}(P,\mathcal{A})^p$ is its pth (absolute) moment. (Note that the expectation of $D(P,A)$ need not be 0 in general, and so the L_2-discrepancy is not the same as the variance.) The L_2-discrepancy is the most important one among the various average discrepancies. It is usually the easiest to handle analytically, mainly because we need not take any absolute values in the definition.

In many papers, mainly in more practically oriented ones, the L_2-discrepancy for corners is used as the main measure of non-uniformity of distribution of a point set. (Part of its popularity can be attributed to its efficient computability; see Section 2.4 and, in particular, Exercise 2.4.11.) However, it can be argued that the L_2-discrepancy for corners does not capture the intuitive notion of uniform distribution too well, especially in higher dimensions. Roughly speaking, it exaggerates the importance of points lying close to the vertex $(0,0,\ldots,0)$ of the unit cube, and, in high dimension, a "typical" corner has a very small volume. Moreover, the directions of the coordinate axes

play a significant role in the definition, and the L_2-discrepancy for corners says very little concerning the uniform distribution with respect to halfplanes, for instance. Modifications have been proposed that address some of these shortcomings; more details are given in the remarks below.

Bibliography and Remarks. The question of discrepancy for classes of shapes other than axis-parallel boxes was first raised by Erdős [Erd64]. For the special case where the point set is (a part of) the lattice $\mathbf{Z}^2$ or other lattices, discrepancy for right-angled triangles was considered much earlier (Hardy and Littlewood [HL22a], [HL22b]), and an extensive theory concerning the number of lattice points in various convex bodies has been developed ([Skr98] provides a list of references).

The rough classification of shapes according to the behavior of their discrepancy function emerged from fine works of several researchers, most notably of Roth, Schmidt, and Beck; references will be given in the subsequent chapters.

The L_2-discrepancy for corners was introduced by Roth [Rot54], first as a technical device for a lower-bound proof. Since then, it has been used widely in numerous theoretical and empirical studies. As was remarked above, it has some disadvantages. If the dimension is not very small in terms of the number of points, say if $n \le 2^d$ (which is often the case in applications), then the L_2-discrepancy for corners gives very little information about uniform distribution, essentially because the average volume of a corner is very small; see Exercise 2.4.5 or [Mat98c]. A notion of L_2-discrepancy favoring larger corners can be found in Hickernell [Hic98], [Hic96]. We will consider a particular instance of Hickernell's notion in the discussion of Zaremba's inequality (1.8) in Section 1.4 and in Exercise 2.4.6.

Another counterintuitive feature of the L_2-discrepancy for corners is the lack of translation invariance: $D_2(P, \mathcal{C}_d)$ may be very different from $D_2(\{P + x\}, \mathcal{C}_d)$, where $\{P + x\}$ arises from P by translation by the vector x and then reducing all coordinates of each point modulo 1 (Lev [Lev95] makes this observation and notes some other undesirable properties). In fact, a surprising result of [Lev96] shows that for any n-point set $P \subset [0,1]^d$, there exists a translation vector x such that

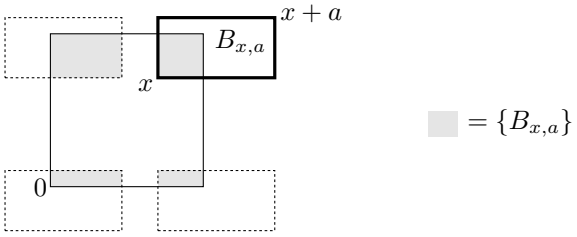
$$D_2(\{P + x\}, \mathcal{C}_d) = \Omega(D(P, \mathcal{C}_d)),$$

with the constant of proportionality depending on d. That is, for any point set there is a translated copy whose L_2-discrepancy is nearly as bad as the worst-case discrepancy!

An alternative notion, advocated in [Lev95], is the L_2-discrepancy with respect to the class

$$\tilde{\mathcal{R}}_d = \{\{B_{x,a}\} : x, a \in [0,1]^d\},$$

where $B_{x,a}$ stands for the box $[x_1, x_1 + a_1) \times [x_2, x_2 + a_2) \times \cdots \times [x_d, x_d + a_d)$ and $\{B_{x,a}\}$ is $B_{x,a}$ reduced modulo 1 in each coordinate:



The measure of a set $R \subseteq \tilde{\mathcal{R}}_d$ is the $2d$-dimensional Lebesgue measure of the corresponding set of pairs $(x, a) \in [0,1]^d \times [0,1]^d$. The discrepancy $D_2(P, \tilde{\mathcal{R}}_d)$ is translation-invariant. Moreover, it always lies between suitable constant multiples of another discrepancy-like quantity, the *diaphony* of P (where the constants depend on the dimension; see Exercise 7.1.6). This rather technical-looking notion, introduced by Zinterhof [Zin76], is motivated by many proofs where estimates on the (usual) discrepancy are obtained via Fourier analysis. The diaphony of P is

$$\left(\sum_{m \in \mathbf{Z}^d \setminus \{0\}} \frac{|\hat{P}(m)|^2}{\prod_{k=1}^d \max(|m_k|^2, 1)} \right)^{1/2},$$

where $\hat{P}(m)$ is the exponential sum $\sum_{p \in P} e^{-2\pi \mathrm{i} \langle m, p \rangle}$, with i denoting the imaginary unit and $\langle ., . \rangle$ the usual scalar product in $\mathbf{R}^d$. Thus, the L_2-discrepancy $D_2(P, \tilde{\mathcal{R}}_d)$ provides a convenient geometric interpretation of diaphony (up to a constant factor, that is). Lev [Lev95] suggests to call the discrepancy for $\tilde{\mathcal{R}}_d$ the *Weyl discrepancy*, because Weyl's foundational paper [Wey16] also considers the (worst-case) discrepancy for intervals taken modulo 1, i.e. on the unit circle. A formula for an efficient computation of this kind of discrepancy can be found in Exercise 7.1.8.

To conclude this discussion of alternatives to the L_2-discrepancy for corners, let us remark that the latter has its advantages too: it is well-established in the literature, and it is perhaps more intuitive and sometimes technically simpler than the alternative notions mentioned above (Hickernell's generalized discrepancy or the discrepancy for $\tilde{\mathcal{R}}_d$). For most of the questions studied in this book, the differences between these notions are not very important. In any case, for measuring the irregularity of distribution, the choice of the "right" discrepancy should be guided by the particular application, and there is probably no single optimal definition.

The discrepancy for the class $\tilde{\mathcal{R}}_d$ of boxes reduced modulo 1 is a special case of the so-called *toroidal discrepancy*. For an arbitrary class

$\mathcal{A}$ of shapes, we can define the corresponding class $\tilde{\mathcal{A}} = \{\{A\} : A \in \mathcal{A}\}$. That is, instead of cutting off the parts of a set A protruding out from the unit cube, we wrap them around. In other words, the unit cube is replaced by (or interpreted as) the torus $\mathbf{R}^d/\mathbf{Z}^d$, which has some technical advantages for methods involving Fourier analysis. Toroidal discrepancy has been used for a long time, especially in proofs of lower bounds (e.g., in Schmidt [Sch69c]). We will return to this in the remarks to Section 7.1.

The second most important domain in which discrepancy is studied, besides the unit cube, is probably the unit sphere S^d. Various notions of discrepancy for this situation, and their applications to numerical integration, are surveyed in Grabner et al. [GKT97]. A very natural and much investigated notion is the discrepancy for *spherical caps* (i.e. intersections of S^d with halfspaces). A little more about this will be said in the remarks to Section 3.1.

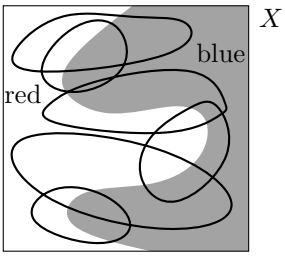
Exercises

1. Prove that any axis-parallel box $R = [a_1, b_1) \times \ldots \times [a_d, b_d)$ can be expressed by 2^d corners, using the operations of disjoint union and "encapsulated difference" (meaning the difference of two sets A, B with $B \subseteq A$).
2. Let $\mathcal{A}$ be some class of measurable sets in $\mathbf{R}^d$.
   (a) Prove that for each n, $D(n, \mathcal{A}) - 1 \le D(n+1, \mathcal{A}) \le D(n, \mathcal{A}) + 1$.
   (b) Is the function $D(n, \mathcal{A})$ necessarily nondecreasing in n?
3. Check that $D_p(P, \tilde{\mathcal{R}}_2) = D_p(\{P + x\}, \tilde{\mathcal{R}}_d)$ for any finite $P \subset [0,1]^d$, any $x \in \mathbf{R}^d$, and any $p \in [1, \infty)$, where $\tilde{\mathcal{R}}_d$ is as in the remarks above.

1.3 Combinatorial Discrepancy

In this section, we start considering a seemingly different problem. Let X be an n-element set, and let $\mathcal{S}$ be a system of subsets[4] of X. We want to color each point of X either red or blue, in such a way that any of the sets of $\mathcal{S}$ has roughly the same number of red points and blue points, as in the following (schematic and possibly misleading) picture:

[4] Sometimes we will write "a set system $(X, \mathcal{S})$," meaning that X is a set and $\mathcal{S}$ is a system of subsets of X. This notation is analogous to the standard notation (V, E) for graphs, where V is the vertex set and E is the edge set. In fact, our notion of "set system" is fully synonymous to the notion of "hypergraph," and for hypergraphs, the notation $(X, \mathcal{S})$ is quite standard. On the other hand, when the underlying set is understood (usually it is the union of all sets in $\mathcal{S}$), we will say "the set system $\mathcal{S}$" only.



Easy considerations reveal that it is not always possible to achieve an exact splitting of each set. The error for some sets may even be arbitrarily large—in fact, if we take all subsets of X for $\mathcal{S}$, then there will always be a completely monochromatic set of size at least $\frac{n}{2}$. The maximum deviation from an even splitting, over all sets of $\mathcal{S}$, is the *discrepancy* of the set system $\mathcal{S}$. We now express this formally. A *coloring* of X is any mapping $\chi\colon X \to \{-1, +1\}$. The *discrepancy* of $\mathcal{S}$, denoted by $\operatorname{disc}(\mathcal{S})$, is the minimum, over all colorings χ, of

$$\operatorname{disc}(\chi, \mathcal{S}) = \max_{S \in \mathcal{S}} |\chi(S)|,$$

where we use the shorthand $\chi(S)$ for $\sum_{x \in S} \chi(x)$. (If $+1$'s are red and -1's are blue then $\chi(S)$ is the number of red points in S minus the number of blue points in S.)

To distinguish this notion of discrepancy from the one introduced previously, we sometimes speak of *combinatorial discrepancy*.[5] Our earlier notion of discrepancy, where we approximate the continuous Lebesgue measure by a discrete point set, may be referred to as *Lebesgue-measure discrepancy* (also "measure-theoretic discrepancy" or "continuous discrepancy" in the literature). Here we mention just a few facts and definitions concerning combinatorial discrepancy; Chapter 4 is devoted to a more systematic treatment.

Combinatorial discrepancy can be transferred to a geometric setting as well. The following is a typical example of a geometrically defined problem in combinatorial discrepancy: given an n-point set P in the plane, we want to color each point of P red or blue in such a way that the maximum difference, over all halfplanes h, in the number of red points and blue points in h is as small as possible. Such a problem can be re-phrased using the combinatorial discrepancy of the set system induced by halfplanes.

If $(X, \mathcal{A})$ is a set system, with X possibly infinite, and $Y \subseteq X$ is a set, we define the *set system induced by $\mathcal{A}$ on Y* as the set system

$$\mathcal{A}|_Y = \{A \cap Y\colon A \in \mathcal{A}\}.$$

(We remark that $\mathcal{A}|_Y$ is sometimes called the *trace* of $\mathcal{A}$ on Y.) In a geometric setting, $\mathcal{A}$ is a system of subsets of $\mathbf{R}^d$, such as the system of all halfspaces, or the system of all balls, and so on, and we will investigate the combinatorial

[5] Also the name "red-blue discrepancy" is used in the literature.

discrepancy of set systems $\mathcal{A}|_P$, where $P \subseteq \mathbf{R}^d$ is a finite set. For a more convenient notation, we will also write $\operatorname{disc}(P, \mathcal{A})$ for $\operatorname{disc}(\mathcal{A}|_P)$. Explicitly, $\operatorname{disc}(P, \mathcal{A})$ is the minimum, over all colorings $\chi\colon P \to \{-1, 1\}$, of

$$\operatorname{disc}(\chi, P, \mathcal{A}) = \max_{A \in \mathcal{A}} \bigl|\chi(P \cap A)\bigr|.$$

Further we define the *discrepancy function* of $\mathcal{A}$ by

$$\operatorname{disc}(n, \mathcal{A}) = \max_{|P|=n} \operatorname{disc}(P, \mathcal{A}).$$

Combinatorial discrepancy in a geometric setting is worth investigating for its own sake, but moreover, there is a close connection with the Lebesgue-measure discrepancy. Roughly speaking, upper bounds for the combinatorial discrepancy for some class $\mathcal{A}$ imply upper bounds for the Lebesgue-measure discrepancy for $\mathcal{A}$. (The reverse direction does not work in general.) This relation is used in many proofs; currently it appears convenient to prove many upper bounds in the combinatorial setting, and some lower bounds, even for combinatorial discrepancy, are proved via the Lebesgue-measure setting. Before giving a precise formulation of this relationship, we introduce another useful notion.

A Common Generalization. The ε-approximation, a notion with origins in probability theory, can be regarded as a generalization of both the Lebesgue-measure discrepancy and the combinatorial discrepancy. It is defined in the following setting: X is some finite or infinite ground set, μ is a measure on X with $\mu(X) < \infty$, and $\mathcal{S}$ is a system of μ-measurable subsets of X. Let $\varepsilon \in [0, 1]$ be a real number. We say that a finite subset $Y \subseteq X$ is an *ε-approximation* for the set system $(X, \mathcal{S})$ with respect to the measure μ if we have, for all $S \in \mathcal{S}$,

$$\left| \frac{|Y \cap S|}{|Y|} - \frac{\mu(S)}{\mu(X)} \right| \le \varepsilon.$$

This means that the fraction of the points of Y lying in S should approximate the relative measure of S with accuracy no worse than ε. If the phrase "with respect to μ" is omitted, we always mean the counting measure on X given by $\mu(S) = |S|$ (we thus also assume that X is a finite set). For example, if X are the inhabitants of some country, the sets in $\mathcal{S}$ are various interest groups, and Y are the members of the parliament, then Y being a $\frac{1}{100}$-approximation means that all interest groups are represented proportionally, with deviation at most 1% of the total population.

The connection to the Lebesgue-measure discrepancy is fairly obvious:

1.5 Observation. *If $\mathcal{A}$ is a class of Lebesgue-measurable sets in $\mathbf{R}^d$ and $P \subset [0, 1]^d$ is an n-point set, then $D(P, \mathcal{A}) \le \varepsilon n$ if and only if P is an ε-approximation*[6] *for $(\mathbf{R}^d, \mathcal{A})$ with respect to the measure $\operatorname{vol}_\square$.* □

[6] If we measured discrepancy as a relative error, rather than in the units of points, and if the term ε-approximation were not well-established, we could naturally

The relationship of ε-approximations to combinatorial discrepancy is a bit more complicated.

1.6 Lemma (Combinatorial discrepancy and ε-approximations). *Let $\mathcal{S}$ be a system of subsets of a $2n$-point set X.*

(i) If $Y \subset X$ is an n-point set that is an ε-approximation for $(X, \mathcal{S})$ then $\operatorname{disc}(\mathcal{S}) \le 2\varepsilon n$. (By the above agreement, we mean ε-approximation with respect to the counting measure on X.)

(ii) If $\mathcal{S}$ is such that $X \in \mathcal{S}$ and $\operatorname{disc}(\mathcal{S}) \le \varepsilon n$ then there exists an n-point set $Y \subset X$ that is an ε-approximation for $(X, \mathcal{S})$.

Proof. In (i), the mapping χ with $\operatorname{disc}(\chi, \mathcal{S}) \le \varepsilon n$ is given by $\chi(x) = 1$ for $x \in Y$ and $\chi(x) = -1$ for $x \notin Y$. Indeed, for any $S \in \mathcal{S}$ we have

$$\chi(S) = |S \cap Y| - (|S| - |S \cap Y|) = 2|S \cap Y| - |S|, \tag{1.6}$$

and since we assume

$$\left| \frac{|Y \cap S|}{|Y|} - \frac{|S|}{|X|} \right| = \frac{1}{2n} \Big| 2|Y \cap S| - |S| \Big| \le \varepsilon,$$

the required bound $|\chi(S)| \le 2\varepsilon n$ follows.

As for (ii), consider a coloring χ with $\operatorname{disc}(\chi, \mathcal{S}) \le \varepsilon n$, and let Y_0 be the larger of the two color classes $\chi^{-1}(1)$ and $\chi^{-1}(-1)$. Since we assume $X \in \mathcal{S}$, we have, using (1.6) with $S = X$, $|\chi(X)| = \big|2|Y_0| - 2n\big| \le \varepsilon n$, and consequently $n \le |Y_0| \le n + \frac{\varepsilon}{2} n$. Let Y be a set of exactly n points arising from Y_0 by removing some arbitrary $|Y_0| - n \le \frac{\varepsilon}{2} n$ points. For $S \in \mathcal{S}$, we calculate

$$\begin{aligned} \left| \frac{|Y \cap S|}{|Y|} - \frac{|S|}{|X|} \right| &= \frac{1}{2n} \Big| 2|Y \cap S| - |S| \Big| \\ &\le \frac{1}{n} \Big| |Y \cap S| - |Y_0 \cap S| \Big| + \frac{1}{2n} \Big| 2|Y_0 \cap S| - |S| \Big| \\ &\le \frac{\varepsilon}{2} + \frac{1}{2n} |\chi(S)| \le \varepsilon. \end{aligned}$$

□

Somewhat imprecisely, this proof can be summarized by saying "if χ is a coloring with small discrepancy, then each of the color classes $\chi^{-1}(1)$ and $\chi^{-1}(-1)$ makes a good ε-approximation." But the two color classes of a coloring need not be exactly of the same size, and this is a technical nuisance in the proof.

Another fairly trivial but useful observation about ε-approximations is

call an ε-approximation for a set system $(X, \mathcal{S})$ with respect to a measure μ a *set with discrepancy at most ε for $(X, \mathcal{S})$ with respect to μ*. This gives a fairly general definition of discrepancy, although certainly not the most general reasonable definition.

1.7 Observation (Iterated approximation). *Let Y_0 be an ε-approximation for $(X, \mathcal{S})$ with respect to some measure μ, and let Y_1 be a δ-approximation for the set system $(Y_0, \mathcal{S}|_{Y_0})$. Then Y_1 is an $(\varepsilon+\delta)$-approximation for $(X, \mathcal{S})$ with respect to μ.* □

After this digression concerning ε-approximations, we return to discrepancy.

Upper Bounds for Combinatorial Discrepancy Imply Upper Bounds for Lebesgue-Measure Discrepancy. Here is one possible precise formulation of the relationship of the combinatorial and Lebesgue-measure discrepancies.

1.8 Proposition (Transference lemma). *Let $\mathcal{A}$ be a class of Lebesgue-measurable sets in $\mathbf{R}^d$ containing a set A_0 with $[0,1]^d \subseteq A_0$. Suppose that $D(n, \mathcal{A}) = o(n)$ as $n \to \infty$, and that $\operatorname{disc}(n, \mathcal{A}) \le f(n)$ for all n, where $f(n)$ is a function satisfying $f(2n) \le (2-\delta)f(n)$ for all n and some fixed $\delta > 0$. Then we have*

$$D(n, \mathcal{A}) = O(f(n)).$$

On the other hand, if we know that $D(n, \mathcal{A}) = o(n)$ and $D(n, \mathcal{A}) \ge f(n)$ for all n, with a class $\mathcal{A}$ and a function $f(n)$ as above, then $\operatorname{disc}(n, \mathcal{A}) \ge cf(n)$ holds for infinitely many n with a suitable constant $c = c(\delta) > 0$.

All sublinear bounds $f(n)$ for discrepancy we are likely to encounter, such as $n^{1/2}$, $\log n$, etc., satisfy the condition in the proposition. Also the requirement that $D(n, \mathcal{A}) = o(n)$ is usually quite weak: it only requires that the Lebesgue measure on the sets of $\mathcal{A}$ can be approximated with an arbitrarily good relative accuracy by the uniform measure concentrated on a finite point set, but there is no condition on the size of the finite set. Except for quite wild sets, a fine enough regular grid of points suffices for such an approximation. So essentially the proposition says that $D(n, \mathcal{A}) = O(\operatorname{disc}(n, \mathcal{A}))$, except possibly for some pathological situations.

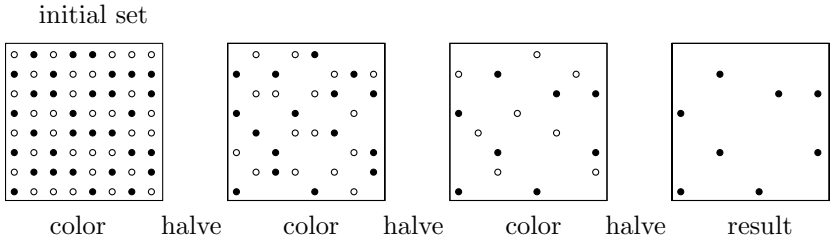
Proof of Proposition 1.8. Let $f(n)$ be a function as in the proposition and let n be a given number. We set $\varepsilon = \frac{f(n)}{n}$ and we choose a sufficiently large natural number k so that

$$\frac{D(2^k n, \mathcal{A})}{2^k n} \le \varepsilon.$$

In other words, there exists a set P_0 of $2^k n$ points that is an ε-approximation for $(\mathbf{R}^d, \mathcal{A})$ with respect to the measure $\mathrm{vol}_\square$. We have thus approximated the continuous measure $\mathrm{vol}_\square$ by the possibly very large but finite set P_0.

Next, we are going to reduce the size of this approximating set to n by a repeated halving, using Lemma 1.6(ii). Namely, we consider the set system $(P_0, \mathcal{A}|_{P_0})$ and we take a coloring χ_0 for it with discrepancy at most $f(|P_0|) = f(2^k n)$. By Lemma 1.6(ii), such a coloring yields a subset $P_1 \subset$

P_0 of $2^{k-1}n$ points that is an ε_0-approximation for $(P_0, \mathcal{A}|_{P_0})$, where $\varepsilon_0 = f(2^k n)/2^k n$. We repeat this step with the set system $(P_1, \mathcal{A}|_{P_1})$, obtaining a set $P_2 \subset P_1$ of $2^{k-2}n$ points that is an ε_1-approximation for $(P_1, \mathcal{A}|_{P_1})$ with $\varepsilon_1 = f(2^{k-1}n)/2^{k-1}n$, and so on. Schematically, this procedure is indicated in the following picture:



We make k such halving steps. The resulting set P_k has n points, and by Observation 1.7, it is an η-approximation for the original set system $(X, \mathcal{S})$ with respect to the measure $\mathrm{vol}_\square$, where

$$\begin{aligned}\eta &= \varepsilon + \sum_{i=0}^{k-1} \varepsilon_i = \frac{f(n)}{n} + \sum_{i=0}^{k-1} \frac{f(2^{k-i}n)}{2^{k-i}n} \\ &\le \frac{f(n)}{n}\left(1 + \sum_{j=1}^{\infty}\left(\frac{2-\delta}{2}\right)^j\right) = O\left(\frac{f(n)}{n}\right).\end{aligned}$$

In view of Observation 1.5, this implies the first part of the proposition. The second part is a contraposition of the first part and we leave it to the reader. □

Of course, the same proof could be phrased without introducing ε-approximations, but without such a notion, it would become somewhat obscure.

Combinatorial L_p-Discrepancy. This is similar to the L_p-discrepancy in the Lebesgue-measure setting. For a set system $\mathcal{S}$ on a finite set X and a coloring χ of X, we put

$$\mathrm{disc}_p(\chi, \mathcal{S}) = \left(\frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} |\chi(S)|^p\right)^{1/p}.$$

More generally, if $\mathcal{A}$ is a family of subsets of a set X, ν is a probability measure on $\mathcal{A}$, P is a finite subset of X, and χ is a coloring of P, we set

$$\mathrm{disc}_{p,\nu}(\chi, P, \mathcal{A}) = \left(\int_{\mathcal{A}} |\chi(P\cap A)|^p \mathrm{d}\nu(A)\right)^{1/p}.$$

Thus, each subset S of P induced by $\mathcal{A}$ is counted with weight equal to $\nu(\{A \in \mathcal{A}\colon A \cap P = S\})$.

Bibliography and Remarks. A close relation of the combinatorial discrepancy to the Lebesgue-measure one has been folklore for some time; M. Simmonovits (private communication) attributes this observation to V. T. Sós. A written version of this idea appears in Beck [Bec81a], where it is used to lower-bound the combinatorial discrepancy for axis-parallel rectangles in the plane (*Tusnády's problem*) using classical lower bounds on the Lebesgue-measure discrepancy. A quite general version of this "transference principle," dealing with classes of convex sets in the plane, was formulated by Lovász et al. [LSV86] (with the proof phrased slightly differently from our proof of Proposition 1.8).

The ε-approximations were defined and used by Vapnik and Chervonenkis [VC71] (the name itself was given by Haussler and Welzl [HW87]). We will hear more about them in Chapter 5.

Exercises

1. Prove the second part of Proposition 1.8 (beginning with "On the other hand,...") from the first part.
2. Let $\mathcal{K}_2$ denote the collection of all closed convex sets in the plane. Show that $D(n, \mathcal{K}_2) = o(n)$ and $\operatorname{disc}(n, \mathcal{K}_2) \geq \frac{n}{2}$.
3. Find a class $\mathcal{A}$ of measurable sets in the plane such that $D(n, \mathcal{A}) = \Omega(n)$.

1.4 On Applications and Connections

Sets with small discrepancy, that is, "very uniformly distributed," have considerable theoretical and practical significance. Moreover, discrepancy theory uses various nice and important mathematical ideas and techniques (some of which we intend to demonstrate in the subsequent chapters), and these ideas have numerous applications in other branches of mathematics. Also, in theoretical computer science, discrepancy theory methods became crucial in many results in recent years. In this section, we mainly discuss relations of discrepancy to numerical integration and to Ramsey theory. A few more applications and connections will be addressed in the remarks.

Numerical Integration. One of the most important applications of low-discrepancy sets is to numerical integration in higher dimensions. In numerical integration, the definite integral of a given function over some region, such as the unit cube, is approximated by the arithmetic mean of the function's values at suitably chosen points. A basic problem is which points are to be chosen for calculating the function's values so that the error of the approximation is as small as possible. The points of a regular grid, or other

straightforward generalizations of classical one-dimensional quadrature rules, do not work well in higher dimensions. On the other hand, point sets with small discrepancy are suitable candidates from both theoretical and practical points of view.

A well-known estimate for the integration error via discrepancy is the so-called *Koksma–Hlawka inequality*. Let $f\colon [0,1]^d \to \mathbf{R}$ be the integrated function and let $P \subset [0,1]^d$ be an n-point set used for the approximation. Then the inequality says

$$\left| \int_{[0,1]^d} f(x)\,\mathrm{d}x - \frac{1}{n}\sum_{p\in P} f(p) \right| \le \frac{1}{n} D(P,\mathcal{C}_d) V(f). \tag{1.7}$$

On the right-hand side, the first term (the discrepancy for corners) only depends on P, while the second term $V(f)$ is determined solely by f. For $d = 1$, $V(f)$ is the *variation* of f; for a continuously differentiable function f, we have $V(f) = \int_0^1 |f'(x)|\,\mathrm{d}x$. For higher dimensions, $V(f)$ denotes an appropriate generalization of variation, the so-called *variation in the sense of Hardy and Krause*, which we will not define here. Although the Koksma–Hlawka inequality is tight in the worst case, it is often very far from being tight for functions encountered in practice.

By now, there is a large body of theory concerning error estimates in the Koksma–Hlawka spirit. These inequalities bound the maximum (or average) integration error for functions from some class in terms of various kinds of discrepancy of the point set used to approximate the integral. Some of them even exactly characterize discrepancy as the worst-case integration error, or the average-case integration error, for very natural classes of functions.

Such results can be considered as a part of a general theory of optimal numerical integration. Here, roughly speaking, a function f from some suitable class is given by a black box, which is a hypothetical device computing $f(x)$ for any given input point x. The basic question is, what is the minimum necessary number of calls to the black box that allows one to calculate the integral of f with error at most ε. Here one need not restrict oneself to the particular algorithm approximating the integral of f by the average $\frac{1}{n}\sum_P f(p)$. It is allowed to combine the values of f obtained from the black box in any other, perhaps nonlinear, way. Moreover, the points are input to the black box one by one, with each point possibly depending on the values of f at the previous points (an *adaptive* algorithm). However, it turned out that for "reasonable" classes[7] of functions, neither nonlinearity nor adaptivity helps. However, it may be truly helpful to combine the computed function's values with weights other than $\frac{1}{n}$.

Discrepancy of Weighted Point Sets. As the reader may know from numerical analysis, more sophisticated one-dimensional quadrature rules (Simp-

[7] Here "reasonable" means closed on convex combinations and on the operation $f \mapsto -f$.

son's rule, Gauss quadrature, etc.) use non-uniform weights. They approximate the integral by

$$\sum_{p\in P} w(p)f(p),$$

where P is a suitable n-point set and $w(p) \in \mathbf{R}$ are real weights, generally distinct from $\frac{1}{n}$. Such formulas achieve error bounds that are not attainable with the uniform weights $\frac{1}{n}$. Not surprisingly, in the literature related to numerical integration, discrepancy is often investigated for *weighted* point sets. For a point set P with a weight function $w\colon P \to \mathbf{R}$, the quantity $|P\cap A|$ is replaced by $w(P\cap A) = \sum_{p\in P\cap A} w(p)$ in the definition of discrepancy in Section 1.2. Thus, we approximate the continuous measure $\mathrm{vol}_\square$ by a (signed) measure concentrated on an n-point set. Actually, there are at least four different notions of discrepancy involving weighted point sets: we may require the weights to be nonnegative and to sum up to $n = |P|$, or we may drop one of these two conditions or both (negative weights are not as absurd as it might seem, since some of the classical quadrature formulas, such as the Newton–Cotes rule, involve negative coefficients). For discrepancy theory, the generalization to weighted point sets is usually not too significant—most of the lower bounds in this book, say, go through for weighted point sets without much difficulty.

Discrepancy for Classes of Functions. The discrepancy of a point set, say the discrepancy for axis-parallel boxes, can obviously be viewed as the maximum integration error for a class of functions, namely for the class of characteristic functions of axis-parallel boxes. But in practice, one often integrates functions with much better smoothness properties, for example continuous functions, functions with continuous derivatives of rth order, or functions with "nice" Fourier series. In such cases, the integration method should ideally take some advantage of the nice behavior of the function. Therefore, it is natural to consider various "smoother analogues" of discrepancy as the maximum integration error for suitable classes of functions, hopefully resembling the functions we are likely to encounter in applications. Specifically, let $\mathcal{F}$ be a class of real Lebesgue-integrable functions on $[0,1]^d$. For a function $f \in \mathcal{F}$, we can set

$$D(P,f) = n\int_{[0,1]^d} f(x)\,\mathrm{d}x - \sum_{p\in P} f(p),$$

and proceed to define $D(P,\mathcal{F}) = \sup_{f\in\mathcal{F}} |D(P,f)|$ and so on. Note that this definition includes the discrepancy for a class $\mathcal{A}$ of sets as a special case: use the characteristic functions of the sets in $\mathcal{A}$ as $\mathcal{F}$. Interestingly, for some natural classes $\mathcal{F}$ of *smooth* functions, the standard notion of discrepancy for axis-parallel boxes is recovered (such an alternative characterization of the L_2-discrepancy for boxes is presented in the remarks below).

Another example of a class $\mathcal{F}$ considered in the literature are the characteristic functions of axis-parallel boxes smoothed out by r-fold integration:

for a parameter $r \geq 0$ and for a point $y \in [0,1]^d$, define a function h_y by setting $h_y(x) = \prod_{k=1}^{d} \max(0, x_k - y_k)^r$, and let $\mathcal{F} = \{h_y\colon y \in [0,1]^d\}$. The resulting notion of discrepancy is called the *r-smooth discrepancy*. In general, the goal is to choose the discrepancy-defining function class small and simple, so that the corresponding discrepancy notion can be handled reasonably, but in such a way that it provides strong "Koksma–Hlawka" type inequalities, i.e. supplies good error bounds for numerical integration of a possibly much wider class of functions. A modern approach to this issue uses the so-called *reproducing kernels* in Hilbert spaces of functions; a little more on this can be found in the remarks below.

Irregularities of Partitions and Ramsey Theory. The preceding part of this section discussed things quite close to practical applications. Now, for a change, let us mention relationship of discrepancy theory to some fast-growing areas of combinatorics. The fact that some set system $(X, \mathcal{S})$ has large combinatorial discrepancy can be rephrased as follows: for any coloring of X by two colors, there is a set where one color prevails significantly. This relates the discrepancy problem to the question of *2-colorability* of a set system. Namely, $(X, \mathcal{S})$ is 2-colorable if there is a coloring of X by two colors with no set of $\mathcal{S}$ completely monochromatic. So, in a sense, the lack of 2-colorability can be regarded as an ultimate case of large discrepancy: for any coloring by two colors, there is a set with one color prevailing completely.

As an example, let us consider the set $X = \{1, 2, \ldots, n\}$ and the set system $\mathcal{A}$ of all arithmetic progressions on X; that is, of all the sets $\{a, a+d, a+2d, \ldots\} \cap X$, $a, d \in \mathbf{N}$. A theorem of Roth states that the discrepancy of $\mathcal{A}$ is at least of the order $n^{1/4}$. On the other hand, if $\mathcal{A}_k$ denotes the subsystem of all $A \in \mathcal{A}$ of size at most k, a famous theorem of Van der Waerden asserts that for any k, there exists an $n = n(k)$ such that $\mathcal{A}_k$ is not 2-colorable. That is, if a sufficiently large initial segment of the natural numbers is colored by two colors, then there is always a long monochromatic arithmetic progression. Van der Waerden's theorem is one of the significant results in the so-called Ramsey theory. In a typical Ramsey-theory question, we consider some sufficiently large combinatorial or algebraic structure X (such as a graph, a finite vector space, etc.) and we color some small substructures of X by two colors (so we may color graph edges, lines in a vector space, etc.). We ask if there always exists a substructure of a given size with all the small substructures having the same color (so we may look for a large subgraph in the given graph with all edges monochromatic, or for a k-dimensional vector subspace with all lines of the same color, and so on). These problems can be formulated as questions about 2-colorability of certain set systems. (Of course, questions involving colorings with more than two colors are studied as well.)

Both discrepancy theory and Ramsey theory can thus be regarded as parts of theory of "irregularities of partition." For each Ramsey-theory question, we automatically get a corresponding discrepancy-theoretic question for the same set system, and vice versa. The case of arithmetic progressions is a

model example. Clearly, some questions that are interesting for Ramsey theory have a trivial or not so interesting discrepancy theory counterpart, and similarly for the other way round. Even if both versions are interesting, the methods of solution maybe vastly different (this is what happens for arithmetic progressions). Nevertheless, this connection can be inspiring and useful to keep in mind.

Another area related to combinatorial discrepancy theory but with mathematical life of its own is the theory of totally unimodular set systems and matrices. Here one is interested in set systems whose each subsystem has discrepancy at most 1. This subject will be briefly touched on in Section 4.4.

More Applications. Without going into details, let us mention that discrepancy theory has also been applied in such diverse areas as computer graphics, image processing, statistics, complexity of algorithms (in particular, replacing probabilistic algorithms by deterministic ones), graph theory, number theory, spectral theory of operators, and Tarski's problem of "squaring the circle" (partition the circle of area 1 into finitely many parts and move each part rigidly so that they together fill the unit square without overlap).

Bibliography and Remarks.

Quasi-Monte Carlo. Sets with low discrepancy can be used for numerical integration in higher dimensions, thus competing with (and often beating) random point sets employed in the popular *Monte Carlo method*. The replacement of random point sets by deterministic (or semi-random) constructions, with a presumably greater "uniformity," is usually called *quasi-Monte Carlo methods*. Such methods are not limited to integration; they can help in numerical solution to differential and integral equations, in optimization, and in other problems.

A concise survey of quasi-Monte Carlo methods is Spanier and Maize [SM94]; newer ones are James et al. [JHK97] and Morokoff and Caflisch [MC95], both written from a practical (computational physicist's) point of view. Tezuka [Tez95] has another brief introduction, also more on the practitioner's side. A more theoretically oriented and considerably more comprehensive (and also technically more demanding) is a monograph by Niederreiter [Nie92].

The study of efficient algorithms for approximating the integral of a function given by a black box is a part of the theory of *information-based complexity* . This theory considers the complexity of algorithms for "continuous" problems, such as computing derivatives, integrals, evaluating various linear operators on function spaces, etc. Two books covering this area are Traub et al. [TWW88] and the newer Traub and Werschulz [TW98].

The area of quasi-Monte Carlo methods is certainly related to discrepancy, but it has somewhat distinct flavor and distinct goals. In "pure" discrepancy theory, as it has been developing so far, one is

mainly interested in asymptotic results, such as that for any fixed dimension d, one can construct an n-point set in $[0,1]^d$ with discrepancy $O(\log^{d-1} n)$ for axis-parallel boxes, where the constant of proportionality depends on d. Since a random point set would only give about $O(\sqrt{n})$ discrepancy, the just mentioned construction is better—period. But if one wants to use such a construction for numerical integration, asymptotic results do not suffice. One has to ask—for how large n is the discrepancy of the constructed set significantly better than the discrepancy of a random point set? Even for not too large dimension, such as 10, an astronomically large n may be required to show the superiority over the random points for some asymptotically good constructions. Moreover, one cannot simply say "the smaller discrepancy, the better set," since the Koksma–Hlawka inequality and its relatives often grossly overestimate the error. Nevertheless, point sets constructed as examples of low-discrepancy sets for axis-parallel boxes proved quite successful in many practical applications.

In this book, we restrict ourselves to a few occasional remarks concerning relations of discrepancy to quasi-Monte Carlo methods. For studying the quasi-Monte Carlo papers, a warning concerning different conventions might perhaps be helpful. In the discrepancy-theory literature, one usually looks at the discrepancy of *sets* (as defined above), while for quasi-Monte Carlo methods, the authors more often work with low-discrepancy *sequences* (where every initial segment is required to be a low-discrepancy set). There is a simple theoretical relation between these two settings. Essentially, good sets in dimension d correspond to good sequences in dimension $d-1$ (see Section 1.1 for the case $d = 2$). But from the practical point of view, the sequences are often preferable.

What Dimension? For various applications in physics, several authors have argued that the advantage of quasi-Monte Carlo methods over the Monte Carlo method (random points) becomes negligible from the practical point of view for dimensions over 20, say (e.g., [JT93]). Also, Sloan and Woźniakowski [SW97] show that numerical integration using fewer than 2^d sample points in dimension d is hopeless in the worst case for certain quite nice classes of functions: no algorithm can do better in the worst case than the trivial algorithm that always outputs 0 as the answer! The threshold 2^d is very sharp, since there exist algorithms with much smaller error using exactly 2^d points.

On the other hand, quasi-Monte Carlo methods have recently been applied successfully for problems of very high dimensions in financial computations (where even small errors may cost big money!); see, for instance, [PT95], [NT96]. A typical dimension appearing in these applications is 360, which is the number of months in 30 years—a typical period for which U.S. banks provide loans. These very

high-dimensional integrals can be seen as approximations to infinite-dimensional path-integrals (also some path-integrals in physics have been handled successfully; see [MC95] for references). Here the success of the quasi-Monte Carlo approach should probably be attributed to a special low-dimensional structure of the integrated functions. A partial theoretical explanation of this phenomenon was found by Sloan and Woźniakowski [SW98].

Error of Integration and Discrepancies. The notion of discrepancy with respect to a given class of functions is very natural in the context of quasi-Monte Carlo methods. It appears in numerous papers, often without references to previous literature with similar concepts. The earliest reference I found is Hlawka [Hla75], who considered the one-dimensional case with $\mathcal{F} = \{x \mapsto x^k \colon k = 0, 1, 2, \ldots\}$. This *polynomial discrepancy* and its higher-dimensional analogues have been studied further by Schmidt, Klinger, Tichy, and others; recent results and references can be found in [KT97]. The r-smooth discrepancy mentioned in the text was considered by Paskov [Pas93].

The one-dimensional Koksma–Hlawka inequality is due to Koksma [Kok43], and the multidimensional version was derived by Hlawka [Hla61].

We have not defined the variation in the sense of Hardy and Krause occurring in the Koksma–Hlawka inequality. Now we pick another among the numerous generalizations and modifications of the Koksma–Hlawka inequality and state it precisely. We begin with some notation, which will allow us to formulate the results more compactly and make them look less frightening. Through this and a few subsequent paragraphs, f is a real function on $[0,1]^d$. We recall the notation, for a finite $P \subset [0,1]^d$, $D(P,f) = n \int_{[0,1]^d} f(x)\,\mathrm{d}x - \sum_{p\in P} f(p)$, which is n-times the integration error. We let $[d] = \{1, 2, \ldots, d\}$, and for an index set $I = \{i_1, i_2, \ldots, i_k\} \subseteq [d]$, we put

$$\frac{\partial^{|I|} f(x)}{\partial x_I} = \frac{\partial^k f(x_1, x_2, \ldots, x_d)}{\partial x_{i_1} \partial x_{i_2} \cdots \partial x_{i_k}}.$$

The notation Q_I stands for the $|I|$-dimensional cube

$$Q_I = \{x \in [0,1]^d \colon x_i = 1 \text{ for all } i \notin I\}.$$

And here is the promised inequality, derived by Zaremba [Zar68], involving the L_2-discrepancy for corners:

$$|D(P,f)| \le D_{2,proj}(P) V_2(f). \tag{1.8}$$

The quantity $V_2(f)$ only depends on f:

$$V_2(f) = \left(\sum_{\emptyset \ne I \subseteq [d]} \int_{Q_I} \left(\frac{\partial^{|I|} f(x)}{\partial x_I} \right)^2 \mathrm{d}x \right)^{1/2}.$$

And $D_{2,proj}$ is a certain L_2-discrepancy of P for corners, taking into account all the coordinate projections of P:

$$D_{2,proj}(P)^2 = \sum_{\emptyset \neq I \subseteq [d]} D_2(\pi_I(P), \mathcal{C}_{|I|})^2,$$

with π_I denoting the projection on the coordinates $(x_i\colon i \in I)$. Some assumptions on f are needed for Zaremba's inequality, of course; for instance, it is enough that the mixed partial derivative $\frac{\partial^d}{\partial x_{[d]}}$ exist and be continuous on $[0,1]^d$, but this requirement can be further relaxed. A proof is indicated in Exercise 1.

Reproducing Kernels. Next, we indicate a fairly general approach to deriving Koksma–Hlawka type inequalities, which subsumes many earlier results and notions of discrepancy. We essentially follow Hickernell [Hic98] and Sloan and Woźniakowski [SW98] (the exposition in [SW98] is somewhat simpler). Hoogland and Kleiss [HK96] and James et al. [JHK97] present interesting and somewhat related ideas (using generating functions and Feynmann diagrams).

Let $(X, \langle .,. \rangle)$ be a Hilbert space of (some) real-valued functions on $[0,1]^d$. A *reproducing kernel* on X is a bivariate function $\eta\colon X \times X \to \mathbf{R}$ such that the function $\eta_x\colon y \mapsto \eta(y,x)$ is in X for all $x \in [0,1]^d$, and the scalar product with η_x represents the evaluation at x: for all $f \in X$ and $x \in [0,1]^d$, we have $f(x) = \langle f, \eta_x \rangle$. (To see what is going on here, one can work out simple examples in Exercise 2 below, and, for instance, [Wah90] provides a more comprehensive introduction to reproducing kernels.) For a reproducing kernel to exist, it is necessary and sufficient that the evaluation operators $T_x\colon f \mapsto f(x)$ be all bounded (by the Riesz representation theorem). For example, the perhaps most usual function space $L_2([0,1])$ with scalar product $\langle f, g \rangle = \int_0^1 f(x)g(x)\,\mathrm{d}x$ does not have any reproducing kernel (why?). Spaces with reproducing kernels mostly involve functions with some smoothness requirements (such as various Sobolev spaces), and the formulas for the scalar product usually contain derivatives.

For a fixed point set $P \subset [0,1]^d$, the integration error $D(P,f)$ is a linear functional on X, and it can be represented as $D(P,f) = \langle \xi_P, f \rangle$, where $\xi_P(x) = D(P, \eta_x)$. The Cauchy–Schwarz inequality then gives

$$|D(P,f)| \leq \|\xi_P\|_X \cdot \|f\|_X.$$

Here $\|.\|_X$ is the norm derived from the scalar product in X. The quantity $\|f\|_X$ is an abstract version of $V(f)$ from the Koksma–Hlawka inequality, and $\|\xi_P\|_X$ can be interpreted as a discrepancy of P. Moreover, ξ_P is a worst-case integrand, where the inequality holds with equality, and so we get a characterization of the discrepancy as a worst-case integration error. These ideas mechanize the process of deriving

Koksma–Hlawka type bounds greatly, but one has to find interesting spaces and reproducing kernels and calculate concrete formulas.

Using the Cauchy–Schwarz inequality in the above considerations leads to various notions of L_2-discrepancy; to obtain notions of L_p-discrepancy, one uses Hölder's inequality (see [Hic98], [SW98]).

Characterizations of discrepancy as an integration error found nice applications in discrepancy theory. Examples are Frolov [Fro80] and, in particular, Wasilkowski and Woźniakowski [WW95], [WW97], who upper-bound the L_2-discrepancy for corners using algorithms for approximate numerical integration.

As an example, we state a characterization of the usual L_2-discrepancy for corners as integration error for a natural class of smooth functions:

$$D_2(P, \mathcal{C}_d) = \sup_f |D(P, f)|. \tag{1.9}$$

The supremum is taken over all functions f such that $f(x) = 0$ for any x with at least one component equal to 1, the mixed partial derivative $\frac{\partial^d f}{\partial x_{[d]}}$ exists and is continuous, and

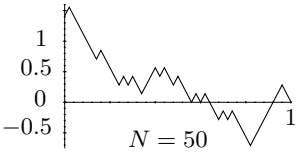
$$\int_{[0,1]^d} \left(\frac{\partial^d f(x)}{\partial x_{[d]}} \right)^2 \mathrm{d}x \le 1.$$

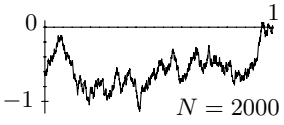
The scalar product is $\langle f, g \rangle = \int_{[0,1]^d} \frac{\partial^d f(x)}{\partial x_{[d]}} \cdot \frac{\partial^d f(x)}{\partial x_{[d]}} \,\mathrm{d}x$, and the reproducing kernel is very simple: $\eta(x, y) = \prod_{k=1}^d \min(1 - x_k, 1 - y_k)$ (Exercise 2). Zaremba's inequality (1.8), for instance, can be obtained by this approach as well, together with an example showing it to be tight. The appropriate the scalar product has to consider other mixed derivatives as well: $\langle f, g \rangle = \sum_{I \subseteq [d]} \int_{Q_I} \frac{\partial^{|I|} f(x)}{\partial x_I} \cdot \frac{\partial^{|I|} g(x)}{\partial x_I} \,\mathrm{d}x$. The reproducing kernel is then $\prod_{k=1}^d \min(2 - x_k, 2 - y_k)$; see [SW98].

Random Functions and Average-Case Error. There are also characterizations of discrepancy as the expected (average-case) integration error for a *random* function. Let us begin with some motivation of this approach. For the Monte Carlo method of integration, one can estimate the error (with a reasonable confidence) by choosing several random sets and comparing the results. This cannot easily be done for a quasi-Monte Carlo method that produces just one set of a given size, say. For this reason, error estimates have been theoretically investigated in another setting, namely when the point set is fixed and the integrated function is chosen "at random." Since natural classes of functions usually form infinite-dimensional spaces with no "canonical" measure on them, it is not clear what should a random function mean.

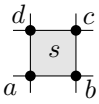
Woźniakowski [Woź91] obtained a very nice result for one possible definition of a "random function," the so-called *Wiener sheet measure*

or multidimensional *Brownian motion*. Instead of a precise definition of the appropriate class of functions and of the measure on it, we present an informal description of a random function from this class. To approximately plot the graph of a one-dimensional random function, start at the point $(1, 0)$ and proceed in N steps. In each step, choose one of the possibilities "up" or "down" at random with equal probability, and go left by $\frac{1}{N}$ and either up or down, according to the random choice, by $\frac{1}{\sqrt{N}}$:



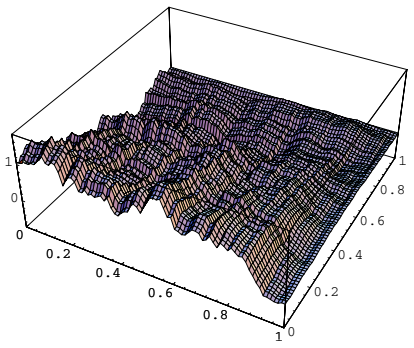


For large N, the resulting plot is approximately the graph of a random f. (The boundary condition $f(1) = 0$ is a consequence of the choice of boxes anchored at 0 in the definition of discrepancy.) For a 2-dimensional random function, subdivide the unit square $[0, 1]^2$ into an $N \times N$ square grid, and for each square s of this grid, independently choose a number $\delta_s \in \{-\frac{1}{N}, +\frac{1}{N}\}$ at random, both possibilities having probability $\frac{1}{2}$. Now define the values of f at all the vertices of the little squares: for each grid square, require the condition indicated in the following picture



$$f(a) - f(b) + f(c) - f(d) = \delta_s$$

and also use the boundary condition $f(x, 1) = f(1, y) = 0$ for all x, y. (It is easy to see that given the δ_s's, these conditions determine the values of f at all the vertices uniquely.) A result for $N = 80$ is shown below:



This can be generalized to dimension d in a straightforward manner.

Woźniakowski proved that for any fixed n-point set $P \subset [0,1]^d$ and f random in this sense, the expected integration error satisfies

$$\mathbf{E}\,[D(P,f)] = D_2(P,\mathcal{C}_d).$$

Earlier, similar results for various notions of a random function in the one-dimensional case were established by Sacks and Ylvisaker [SY70]. Woźniakowski [Woź91] has the d-dimensional statement and, moreover, directly relates the L_2-discrepancy to the average-case algorithmic complexity of numerical integration. Alternative derivations of this average-case characterization of discrepancy, as well as some generalizations, can be found in [MC94] and [JHK97].

Functions occurring in practical problems seldom resemble random functions in the sense discussed above; for instance, the latter ones are continuous but typically nowhere differentiable. Paskov [Pas93] derived an analogue of Woźniakowski's result for random functions with a given degree r of smoothness.

A general relation of the worst-case and average-case error estimates is considered in Wahba [Wah90] or in Traub et al. [TWW88].

Ramsey Theory. Nice overviews of Ramsey theory are Graham et al. [GRS90] or Nešetřil [Neš95]. An inspiring account of the connections of discrepancy theory to Ramsey theory is Sós [Sós83a] (a shorter version is in [BS95]). The $\Omega(n^{1/4})$ lower bound for discrepancy of arithmetic progressions is from Roth [Rot64]. This bound is asymptotically tight; we will say more about this problem in Sections 4.2, 4.5, and 4.6.

Number Theory. As an example of a result related to number theory, we can quote Beck's solution of a problem of Erdős concerning "flat" polynomials on the unit circle. Using discrepancy theory methods, Beck [Bec91a] proved that there are constants $c, \alpha > 0$ such that whenever $\xi_1, \xi_2, \ldots, \xi_n$ are complex numbers with $|\xi_i| = 1$ for all i and we define polynomials $p_1(z)$, $p_2(z)$,..., $p_n(z)$ by setting $p_i(z) = \prod_{j=1}^{i}(z - \xi_j)$, then

$$\max_{1\le i\le n} \max_{|z|=1} |p_i(z)| \ge cn^{\alpha}.$$

Geometry. The beautiful "squaring the circle" result mentioned in the text is due to Laczkovich [Lac90].

Here are two combinatorial geometry problems related to discrepancy. One of them asks for an n-point set on the unit d-dimensional sphere such that the sum of all the $\binom{n}{2}$ Euclidean distances determined by these points is maximal. An exact solution appears very difficult in most cases. Stolarsky [Sto73] discovered a relation of this problem to a

certain kind of discrepancy, and Beck [Bec84] used results in discrepancy theory to give good asymptotic bounds for the maximum sum of distances.

Another problem concerns the approximation of the unit ball in $\mathbf{R}^d$ by a *zonotope*. A zonotope is a special type of a convex polytope that can be defined as a d-dimensional projection of an n-dimensional cube, and for the approximation we want to have n as small as possible. This problem has several equivalent formulations, one of them being a "tomography" question (Betke and McMullen [BM83]): find the minimum number n of directions $y_1, \ldots, y_n \in S^{d-1}$ (where S^{d-1} is the unit sphere in $\mathbf{R}^d$), such that the surface area of any convex body K in $\mathbf{R}^d$ can be determined, up to a relative error of ε, by the knowledge of the volumes of the $(d-1)$-dimensional projections of K on the hyperplanes $\{x \in \mathbf{R}^d \colon \langle y_i, x\rangle = 0\}$.[8] Using harmonic analysis techniques similar to those employed for discrepancy lower bounds, Bourgain et al. [BLM89] established lower bounds for this problem, and these were shown to be asymptotically tight or almost tight in a sequence of papers (also applying various discrepancy theory methods): Bourgain and Lindenstrauss [BL88], Wagner [Wag93], Bourgain and Lindenstrauss [BL93], and Matoušek [Mat96b].

Graph Theory. Discrepancy of a certain kind also appears in graph theory. For instance, Chung [Chu97] defines the discrepancy of a graph G as

$$\max_{S \subseteq V} \bigl| e(S,S) - \rho |S|^2 \bigr|,$$

where V is the vertex set of G, $e(S,S)$ is the number of ordered pairs $(u,v) \in S \times S$ such that $\{u,v\}$ is an edge of G, and $\rho = e(V,V)/|V|^2$ is the *density* of G. Thus, in a graph with small discrepancy, the number of edges on each subset S is close to the expected number of edges on a random subset of size $|S|$. The discrepancy of a graph can be bounded in terms of the second largest eigenvalue of its adjacency matrix. For graphs of density about $\frac{1}{2}$, the best possible discrepancy is of the order $n^{3/2}$. If a graph is a good *expander* then it has small discrepancy. Expanders are a very important type of "random-like" graphs, with numerous applications (in communication networks, parallel computing, sorting networks, pseudorandom generators, error-correcting codes, and so on), and the reader can learn about them in [AS00] or in [Chu97], for instance.

Computer Science. As we mentioned at the beginning of this section, discrepancy theory methods gained in importance in computer sci-

[8] To appreciate this formulation, one should know that if we are given the volumes of all the $(d-1)$-dimensional projections of a convex body K then the surface area is determined exactly by *Cauchy's surface area formula*. For instance, in $\mathbf{R}^3$, the surface area equals 4 times the expected area of the projection in a random direction; see [San76].

ence in recent years, as is successfully illustrated by the book [Cha00]. Whenever a small sample is needed that represents well a large collection of objects, which is a very frequent situation in the search for efficient algorithms, connections to discrepancy theory may appear, and they often do.

For instance, geometric discrepancy turned out to be relevant in several results in computational geometry. This field of computer science considers the design of efficient algorithms for computing with geometric objects in the Euclidean space, usually of a low dimension (see also the remarks to Section 5.2). Lower bounds for geometric discrepancy have been used by Chazelle [Cha98] to show lower bounds for the computational complexity of a geometric database problem (the so-called range searching), and for another version of this problem, a set with low discrepancy for axis-parallel rectangles has been employed in a lower-bound proof (see [Cha00]). The ε-approximations in geometrically defined set systems play a key role in the so-called *derandomization* of computational geometry algorithms; that is, replacing probabilistic algorithms by deterministic ones. For more information see the survey [Mat96a] or the book [Cha00].

In another subfield of computer science, derandomizing combinatorial algorithms, the questions have discrepancy-theory flavor too but often they concern spaces of very high dimensions. For example, Linial et al. [LLSZ97] construct n-point subsets of the d-dimensional grid $\{1, 2, \ldots, q\}^d$ uniformly distributed with respect to the *combinatorial rectangles*, where a combinatorial rectangle is a set of the form $S_1 \times S_2 \times \cdots \times S_d$, with $S_1, \ldots, S_d \subseteq \{1, 2, \ldots, q\}$ being arbitrary subsets. (In out terminology, they construct an ε-approximation for combinatorial rectangles.) In contrast to the "classical" discrepancy theory setting, where the dimension is considered fixed, they need to investigate the situation where d is large (comparable to n and q, say). The main challenge here is to approach the quality of a random set by a deterministic construction, while in classical discrepancy, one can usually beat random sets. Also various constructions of *approximately k-wise independent* random variables on small probability spaces can be viewed as (explicit) constructions of small ε-approximations for certain set systems. An introduction to k-wise independence in derandomization can be found in Alon and Spencer [AS00] or in Motwani and Raghavan [MR95], and a sample of papers devoted to such constructions are [AGHP92], [EGL$^+$92].

Another interesting example of an explicit construction of an ε-approximation is provided by Razborov et al. [RSW93], who describe a set $A \subset \{1, 2, \ldots, n-1\}$ of size bounded by a polynomial in $\log n$ and in $\frac{1}{\varepsilon}$ that is an ε-approximation for the system of all arithmetic progressions (modulo n). Here it is easy to show that a random A will

work with high probability; the point is to avoid randomness. This result was applied by Alon and Mansour [AM95] in a fast deterministic algorithm for interpolating multivariate polynomials.

Exercises

1. (Zaremba's inequality) Let $f\colon [0,1]^d \to \mathbf{R}$ have a continuous mixed partial derivative $\frac{\partial^d f}{\partial x_{[d]}}$ (notation as in the remarks above).
   (a)* Derive the following identity for the integration error, by repeated integration by parts:

$$D(P,f) = \sum_{\emptyset \neq I \subseteq [d]} (-1)^{|I|} \int_{Q_I} D(P, C_x) \cdot \frac{\partial^{|I|} f(x)}{\partial x_I} \, \mathrm{d}x.$$

   Try to get at least the cases $d = 1$ (where nothing too interesting happens) and $d = 2$.
   (b) Using Cauchy–Schwarz, derive Zaremba's inequality (1.8) from (a).
2. (Reproducing kernels) Consider the Hilbert space X of absolutely continuous functions $f\colon [0,1] \to \mathbf{R}$ such that $f(1) = 0$ and $f' \in L_2(0,1)$ (i.e. $\int_0^1 f'(x)^2 \, \mathrm{d}x < \infty$). The scalar product is $\langle f, g \rangle = \int_0^1 f'(x) g'(x) \, \mathrm{d}x$. Recall that an absolutely continuous function f is differentiable almost everywhere, and we have $\int_a^b f'(x) \, \mathrm{d}x = f(b) - f(a)$. Functions with a continuous first derivative form a dense subspace in X.
   (a) Check that $\eta(x,y) = \min(1-x, 1-y)$ is a reproducing kernel in X.
   (b) Calculate ξ_P, and check that the corresponding discrepancy is just the L_2-discrepancy of P for corners.
   (c)* Generalize (a) and (b) to an arbitrary dimension d (try at least $d = 2$), with the reproducing kernel $\eta(x,y) = \prod_{k=1}^d \min(1-x_k, 1-y_k)$, scalar product $\langle f, g \rangle = \int_{[0,1]^d} \frac{\partial^d f(x)}{\partial x_{[d]}} \cdot \frac{\partial^d g(x)}{\partial x_{[d]}} \, \mathrm{d}x$, and functions f satisfying $f(x) = 0$ for all x with at least one component equal to 1. Derive (1.9).

   Remark. The functions with $\frac{\partial^d f}{\partial x_{[d]}}$ continuous form a dense set in the appropriate Hilbert space in (c). To describe the functions in the resulting (Sobolev) space, one needs the notion of distributional derivatives, and the definitions are not entirely simple (see a book dealing with Sobolev spaces, such as [Ada75], [Wah90]). But for this exercise, such a description is not really needed, and all the functions actually encountered in the proof are piecewise polynomial.