Preface

The use of random sampling in the field of discrete and computational geometry started in the 1980s, motivated by the challenges in designing efficient algorithms for geometric problems. While these earlier uses were tightly coupled with specific geometric scenarios, soon the key problems were formulated abstractly in the framework of combinatorial and geometric set systems. We state one of the principal structures that will be studied in this framework. Let X be a set of n elements and $\mathcal{F}$ a collection of subsets of X; the pair $(X, \mathcal{F})$ forms a set system.

> **Epsilon-nets:** Given a parameter $\epsilon \in (0, 1]$, a set $N \subseteq X$ is an ϵ-net of $(X, \mathcal{F})$ if each $S \in \mathcal{F}$ of size at least $\epsilon |X|$ has non-empty intersection with N.

The goal is to find ϵ-nets of small size; this of course depends on the structure and complexity of $(X, \mathcal{F})$. A classical geometric instance of this question—first studied in 1987 and settled conclusively in 2017—is to determine, given any set P of n points in $\mathbb{R}^d$, the smallest $N \subseteq P$ such that any half-space containing at least ϵn points of P contains at least one point of N.

Selective aspects of ϵ-nets have been presented in earlier texts (*Combinatorial Geometry*, Pach and Agarwal, 1995; *The Discrepancy Method*, Chazelle, 2000; *Lectures on Discrete Geometry*, Matoušek, 2004; *Geometric Approximation Algorithms*, Har-Peled, 2011). However, the last ten years have seen significant progress with many open problems in the area having been resolved during this time. These include optimal lower bounds for ϵ-nets for most geometric set systems, the use of shallow-cell complexity to unify proofs, simpler algorithms to construct ϵ-nets, and the use of ϵ-approximations for construction of coresets via sensitivity analysis, to name a few. This book presents a didactic account of these recent developments. We will revisit classical results, but with new and more elegant proofs which unify earlier work.

Chapter 1 introduces the two key technical ingredients that lie at the heart of the analysis of random sampling methods in this book: the complexity of certain combinatorial structures arising in geometric configurations and the probability of a random variable deviating far from its expectation. While historically these two have been considered separate statements with entirely different proofs, we present a powerful probabilistic technique from which both of these bounds can be deduced in a uniform way.

Chapters 2 and 3 initiate the study of ϵ-nets for some basic geometric set systems in $\mathbb{R}^2$, delineating the precise geometric properties that are relevant to the construction of ϵ-nets; these are then combined with probabilistic techniques to derive asymptotically optimal bounds on the size of ϵ-nets.

While these will be superseded later by more general and powerful combinatorial machinery, they are important in understanding the intuition, ideas, and analysis at their most elementary level.

The move from geometric to combinatorial set systems requires formulating and proving analogs of geometric properties for combinatorial systems. Chapters 4 and 5 are devoted to building this technical foundation. First, the VC-dimension and the shallow-cell complexity of combinatorial set systems are introduced and studied as measures of complexity of a set system. These are then used to construct combinatorial equivalents of geometric properties relevant to ϵ-net constructions.

Chapters 6 to 8 present the current best bounds for ϵ-nets for combinatorial set systems $(X, \mathcal{F})$, where the bounds depend on the VC-dimension and the shallow-cell complexity of $\mathcal{F}$. Together these chapters contain the insight that one can derive optimal bounds for geometric set systems from combinatorial bounds based on the shallow-cell complexity of the corresponding set system. Chapter 9 studies a geometric case where small ϵ-nets do not exist, set systems induced by convex sets in $\mathbb{R}^d$, and where one has to turn to the notion of a *weak* ϵ-net.

Chapters 10 and 11 are concerned with lower bounds on sizes of ϵ-nets, based on the insight that a lower bound on the size of an ϵ-net for a given set system $(X, \mathcal{F})$ follows from a lower bound on the VC-dimension of a related set system, the k-fold union of $\mathcal{F}$. These lower bounds are then used to show optimality of the ϵ-net bounds presented earlier.

Chapters 12 to 15 study another notion of samples, *ϵ-approximations*, for both geometric and combinatorial set systems. It also includes an application of ϵ-approximations for constructing small coresets for some geometric optimization problems.

Chapter 16 concludes the book with a list of bounds on the VC-dimension and shallow-cell complexity for most commonly studied geometric set systems, as well as on sizes of their ϵ-nets and ϵ-approximations. This will serve as a reference for those looking for the state-of-the-art bounds on these topics.

We now briefly list some topics which are not in this book: algorithms tailored to construct ϵ-nets efficiently for specific geometric set systems, efficient deterministic versions of the probabilistic algorithms, range searching and other classic algorithmic applications, bounds for combinatorial discrepancy of geometric set systems. This choice was guided by two factors. First, the techniques involved in these are rather different, often relying on detailed geometric data-structures; doing justice to this essentially requires another book. Second, parts of it have been covered very nicely in earlier texts (e.g., in *The Discrepancy Method* by Chazelle).

While our key objective is to give a clear account of the ideas (as much as is possible by us), we also hope that reading this book is a pleasant experience (the wonderful texts *Combinatorial Geometry* by Pach and Agarwal and *Lectures in Discrete Geometry* by Matoušek being exemplary in this regard). We have also taken care to make the present text useful for teaching: all calculations are written in sufficient detail; each section begins with an "overview of ideas" which gives intuition into the proof; wherever possible we first present the simplest non-trivial instance of the

idea before dealing with the more general case; each chapter can be read mostly independently (though it might use earlier results); additional insights, ideas and calculations that are not crucial to the main text are interspersed throughout in small font size. Each section typically contains one or two results, and is carved up into smaller themes that are delineated by the symbol ∞.

The text should be suitable as a first introduction to sampling aspects for senior undergraduate and graduate students in computer science, mathematics and statistics. While mathematical maturity will certainly help in appreciating the ideas presented here, only a basic familiarity with discrete mathematics, probability and combinatorics is required to understand the material. For background on these topics, the following books are recommended:

H. Tijms. *Understanding Probability: Chance Rules in Everyday Life*. Cambridge University Press, 2007.

J. Matoušek and J. Nešetřil. *Invitation to Discrete Mathematics*. Oxford University Press, 2008.

M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2017.

For teaching a course on these topics, we recommend that around 2 hours of class time be devoted to each chapter; this text is suitable for a 30 to 40 hour course on the subject (there is considerable freedom in the choice of topics to cover). We also hope that this book will be useful for researchers in the field as a reference text for looking up specific bounds as well as learning quickly the ideas and techniques behind specific results.

We would be grateful if any errors are reported to `nabilhmustafa@gmail.com`.

Acknowledgments. This text benefited greatly from feedback and discussions with several mentors, colleagues and students. In particular, I am grateful to the following people: Imre Bárány, Victor-Emmanuel Brunel, Jean Cardinal, Timothy Chan, Bernard Chazelle, Mónika Csikós, Kunal Dutta, Fritz Eisenbrand, David Eppstein, Jeff Erickson, Martina Gallato, Arijit Ghosh, Andrey Kupavskii, Jesús De Loera, Frédéric Meunier, Wolfgang Mulzer, Márton Naszódi, János Pach, Dömötör Pálvölgyi, Dominique Perrin, Jeff Phillips, Saurabh Ray, Güntor Rote, Gabor Tardos, Csaba Tóth, Kasturi Varadarajan, and Emo Welzl.

I am also grateful to the Agence Nationale de la Recherche (ANR) for funding my research for the past ten years, and to my colleagues at LIPN, Villetaneuse and LIGM, Marne-la-Vallée.

It was a pleasure to work the people from the AMS Publishing. I would like to especially thank Ina Mette for her great help and patience.

Nabil H. Mustafa
Nogent-sur-Marne, September 2021

Background

Basic notation. The cardinality of a finite set X is denoted by $|X|$. For a real number a, $\lfloor a \rfloor$ denotes the largest integer less than or equal to a; similarly $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. We use the notation $[n]$ for the set $\{1, \ldots, n\}$, where n is a positive integer. We will use the notation $A = B \pm C$, where $A, B \in \mathbb{R}$ and $C \in \mathbb{R}^+$, as a shorthand for $A \in [B - C, B + C]$. We will use $\log n$ for logarithms with base 2, and $\ln n$ for logarithms with base e. The letters $\mathbb{N}$, $\mathbb{R}$, $\mathbb{Z}$ are reserved for the set of all natural numbers, reals and integers, respectively.

Asymptotic notation. For two real valued functions f and g, we say that $f = O(g)$ if there exist large-enough constants $n, C > 0$ such that $f(x) \leq Cg(x)$ for all $x \geq n$. Here the values of n, C might depend on other quantities considered as constants; we will explicitly point out such dependencies when they occur. The notation $f = \Omega(g)$ is equivalent to $g = O(f)$, $f = \Theta(g)$ if and only if $f = O(g)$ and $f = \Omega(g)$, and $f = o(g)$ if $\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$.

Set systems. Given a finite set X, 2^X will denote the collection of all subsets of X. Similarly, for $0 \leq k|X|$, $\binom{X}{k}$ will denote the collection of all subsets of X of size k, and so $\left|\binom{X}{k}\right| = \binom{|X|}{k}$. A set system is a pair $(X, \mathcal{F})$, where X is a set and $\mathcal{F}$ is a collection of subsets of X. When X is clear from the context, we will simply use $\mathcal{F}$ to denote the set system.

Geometric notions. $\mathbb{R}^d$ will denote the d-dimensional Euclidean space. For a measurable set $X \subseteq \mathbb{R}^d$, $\mathrm{vol}(X)$ denotes the d-dimensional Lebesgue measure of X. The symbol ∂X denotes the boundary of $X \subseteq \mathbb{R}^d$, and $\mathrm{int}(X)$ the interior of X. For $p \in \mathbb{R}^d$ and $r > 0$, $\mathrm{Ball}(p, r)$ denotes the closed ball of radius r centered at p. A set $X \subset \mathbb{R}^d$ is convex if for every $p, q \in X$, the segment pq is contained in X. The set $\mathrm{conv}(X)$ is defined to be the intersection of all convex sets in $\mathbb{R}^d$ containing X. Alternatively, $q \in \mathrm{conv}(X)$ if and only if there exist points $p_1 \in X, \ldots, p_{d+1} \in X$ and nonnegative reals $t_1, \ldots, t_{d+1}$ such that $\sum_i t_i = 1$ and $q = \sum_i t_i p_i$. A finite set $X \subset \mathbb{R}^d$ is said to be 'in convex position' if $p \notin \mathrm{conv}(X \setminus \{p\})$ for all $p \in X$. Radon's theorem states that given any set P of $d + 2$ points in $\mathbb{R}^d$, there exists a partition of P into two disjoint sets P_1 and P_2 such that $\mathrm{conv}(P_1) \cap \mathrm{conv}(P_2) \neq \emptyset$.

General position. Throughout the text we will often assume that a configuration of geometric objects is 'in general position'. That is, all properties and corresponding results are invariant to an arbitrarily small perturbation of the configuration. For example, for a set of n points in $\mathbb{R}^d$ in general position, we will assume that no $d + 1$ points lie on a common hyperplane, no $d + 2$ on a common sphere and so on. The specific properties assumed for a configuration in general position will be explicitly stated where used.

Point-hyperplane duality. The dual of a point $p = (p_1, \ldots, p_d) \in \mathbb{R}^d$ is the hyperplane $p_1x_1 + p_2x_2 + \cdots + p_{d-1}x_{d-1} - x_d = -p_d$. The dual of a 'non-vertical' hyperplane $h\colon a_1x_1 + \cdots + a_{d-1}x_{d-1} + x_d = b$ is the point $(a_1, \ldots, a_{d-1}, b)$. The key property of duality, easy to verify using the above mappings, is that it preserves incidences and sidedness. That is, for any point $p \in \mathbb{R}^d$ and any hyperplane h, p lies above h (with respect to the x_d-coordinate) if and only if the point corresponding to the dual of h lies below hyperplane corresponding to the dual of p.

Graphs. An undirected graph is usually denoted by $G = (V, E)$, where V is the set of its vertices, and $E \subseteq \binom{V}{2}$ is the set of its edges. When the sets V and E are not explicitly defined, they will be denoted by $V(G)$ and $E(G)$. If $\{u, v\} \in E$, we say that u and v are adjacent in G, and that v is a neighbor of u (and vice versa). For any $v \in V$, $N_G(v) \subseteq V$ will be the set of vertices of V which are adjacent to v. The complete graph, where $E = \binom{V}{2}$, on t vertices will be denoted by K_t, and the complete bipartite graph with t_1, t_2 vertices in the two partite sets will be denoted by K_{t_1,t_2}. A subset $V' \subseteq V$ such that there are no edges between any two vertices of V' in G is called an independent set. Any $V' \subseteq V$ such that there is an edge in G between every two vertices of V' is called a clique.

A *drawing* of an undirected graph $G = (V, E)$ in the plane consists of two functions that map V and E to subsets of the plane. The function $\phi_V\colon V \to \mathbb{R}^2$ maps each vertex $v \in V$ to a point $\phi_V(v) \in \mathbb{R}^2$. Then for each edge $e = \{u, v\} \in E$, the continuous function $\phi_e\colon [0, 1] \to \mathbb{R}^2$ maps e to a continuous arc in $\mathbb{R}^2$ connecting the images of u and v, i.e., connecting $\phi_e(0) = \phi_V(u)$ to $\phi_e(1) = \phi_V(v)$. We will assume that ϕ_V is injective ($\phi_V(x) = \phi_V(y)$ if and only if $x = y$) and that no arc $\phi_e[0, 1]$ passes through the image of any vertex apart from the endpoints of e. A drawing of $G = (V, E)$ is called an *embedding* or a *plane graph* if (the images of) no two edges share an interior point (of course, they may share an endpoint). G is called *planar* if it has an embedding in $\mathbb{R}^2$. A planar graph on n vertices has at most $3n - 6$ edges and at most $2n - 4$ faces in any embedding.

Given a set P of n points in the plane, the Delaunay graph of P has an edge between two points $p, q \in P$ if and only if there is a closed disk containing p and q and no other point of P. The Delaunay graph is planar, and so has at most $3n - 6$ edges.

Probability. $\Pr[A]$ denotes the probability of an event A. The expectation of a random variable X is denoted by $\mathrm{E}[X]$. An indicator random variable is a random variable which can have a value of 0 or 1. For an indicator random variable X, we have $\mathrm{E}[X] = \Pr[X = 1]$. Linearity of expectation states that for two random variables X and Y, we have $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$; the usefulness of this statement comes from the fact that it holds regardless of any dependency between X and Y.

Equations and inequalities. Here are some inequalities that will be useful.

Pascal's rule:
$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \quad n, k \in \mathbb{Z}^+.$$

Geometric-arithmetic mean inequality:
$$\prod_{i=1}^{n} a_i \leq \left(\frac{\sum_{i=1}^{n} a_i}{n} \right)^n, \quad a_1, \ldots, a_n \in \mathbb{R}^+.$$

Exponential:
$$1 + x \leq e^x, \quad \text{for } x \in \mathbb{R}.$$
$$1 - x \geq e^{-2x}, \quad \text{for } x \in [0, 0.79].$$

Binomial theorem:
$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i \, y^{n-i}, \quad n \in \mathbb{Z}^+, x, y \in \mathbb{R}.$$