UNIVERSITY OF
CAMBRIDGE

Optimal estimation in high-dimensional and nonparametric models

Nikolay Baldin

Queens' College

This dissertation is submitted for the degree of Doctor of Philosophy

Abstract

Minimax optimality is a key property of an estimation procedure in statistical modelling. This thesis looks at several problems in high-dimensional and nonparametric statistics and proposes novel estimation procedures. It then provides statistical guarantees on the performance of these methods and establishes whether those are computationally tractable.

In the first chapter, a new estimator for the volume of a convex set is proposed. The estimator is minimax optimal and also efficient non-asymptotically: it is nearly unbiased with minimal variance among all unbiased oracle-type estimators. Our approach is based on a Poisson point process model and as an ingredient, we prove that the convex hull is a sufficient and complete statistic. No hypotheses on the boundary of the convex set are imposed. In a numerical study, we show that the estimator outperforms earlier estimators for the volume. In addition, an improved set estimator for the convex body itself is proposed.

The second chapter extends the results of the first chapter and develops a unified framework for estimating the volume of a set in $\mathbb{R}^d$ based on observations of points uniformly distributed over the set. The framework applies to all classes of sets satisfying one simple axiom: a class is assumed to be intersection stable. No further hypotheses on the boundary of the set are imposed; in particular, the class of convex sets and the class of weakly-convex sets are covered by the framework. We introduce the so-called wrapping hull, a generalization of the convex hull, and prove that it is a sufficient and complete statistic. The proposed estimator of the volume is simply the volume of the wrapping hull scaled with an appropriate factor. It is shown to be consistent for all classes of sets satisfying the axiom and mimics an unbiased estimator with uniformly minimal variance. The construction and proofs hinge upon an interplay between probabilistic and geometric arguments. The tractability of the framework is numerically confirmed in a variety of examples.

The third chapter considers the problem of link prediction, based on partial observation of a large network, and on side information associated to its vertices. The generative model is formulated as a matrix logistic regression. The performance of the model is analysed in a high-dimensional regime under a structural assumption. The minimax rate for the Frobenius-norm risk is established and a combinatorial estimator based on the

penalised maximum likelihood approach is shown to achieve it. Furthermore, it is shown that this rate cannot be attained by any (randomised) polynomial-time algorithm under a computational complexity assumption.

The trade-off between computational efficiency and statistical optimality is discussed throughout the thesis. For estimating the volume of a set from the class of convex or weakly-convex sets in high dimensions, we propose minimax optimal estimators in the first and second chapters. However, they cannot be computed using a polynomial-time algorithm in dimensions higher than three. Analogously, the proposed minimax optimal estimator for a prediction task in the matrix logistic regression problem in the third chapter cannot be computed in polynomial time. The third chapter further identifies a computational lower bound in the regression problem, thereby revealing the gap between the best possible rate of convergence of a polynomial-time algorithm and the minimax optimal rate.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution.

Chapter 1 consists of joint work with Markus Reiß (Humboldt University of Berlin), and has appeared in Stochastic Processes and Applications as (Baldin and Reiß, 2016), and partly to appear in Snapshots of Modern Mathematics from Oberwolfach as (Baldin, 2018). Chapter 2 is a result of my independent work. Chapter 3 is joint work with Quentin Berthet (University of Cambridge). Both Chapter 2 and Chapter 3 have been submitted for publication (Baldin, 2017; Baldin and Berthet, 2018).

Nikolay Baldin
November 2018

Acknowledgements

I would like to thank everyone who has supported me during my research through multiple conversations or joint work including Quentin Berthet, Richard Nickl, Markus Reiß and many others.

In the Statslab, many thanks to Sam Thomas, Eardi Lila, Fritz Hiesmayr, Kweku Abraham, Mo Dick Wong, Matthias Löffler, Andrew Swan, Megan Griffin-Pickering, Lisa Kreusser, and many others for the edifying company.

I would like to thank Alexandre Tsybakov for giving me the opportunity to present the results of Chapter 2 at the statistics seminar at ENSAE in Paris, Mark Podolskii for letting me give a talk at the statistics seminar in Aarhus, and Guenther Walther for inviting me to visit Stanford University. I would like to thank Vladimir Spokoiny and Markus Reiß for the opportunity to present the results of Chapter 3 at the statistics seminar at WIAS in Berlin and Anirban DasGupta for inviting me to give a talk at the ISNPS conference. Finally, I would also like to thank the editorial board of Oberwolfach's snapshots of modern mathematics for helping me to make the results of Chapters 1 and 2 more accessible to a broader audience.

And most of all, I would like to thank my partner Alexandra for her encouragement and support along the way.

I gratefully acknowledge the funding sources that made my PhD work possible. I was funded by the European Research Council (ERC) Grant No. 647812.

Contents

1 Unbiased estimation of the volume of a convex body **13**

1.1 Introduction to volume estimation 13

1.1.1 Calculating the volume of geometric objects 13

1.1.2 Estimating the volume in statistics 14

1.1.3 Estimating the support of a density 17

1.1.4 Looking further. Computational geometry 18

1.2 Introduction to unbiased volume estimation of a convex set 19

1.3 Digression on Poisson Point Processes 20

1.4 Oracle case: intensity λ is known 22

1.5 Unknown intensity λ: nearly unbiased estimation 26

1.6 Finite sample behaviour and dilated hull estimator 32

1.7 Appendix 34

1.7.1 Proof of Theorem 2.2.5 34

2 The wrapping hull and a unified framework for volume estimation **39**

2.1 Main contribution and the structure 39

2.1.1 Relationship to the work on volume estimation of a convex set . . . 41

2.2 The wrapping hull 42

2.3 Oracle case: known intensity λ 44

2.4 Data-driven estimator of the volume 46

2.4.1 Volume estimation in the uniform model 47

2.4.2 Efron's inequality for the wrapping hull 48

2.5 Classes of sets satisfying Assumption 2.2.1 50

2.5.1 r-convex sets 51

2.5.2 Compact sets 52

2.5.3 Concentric sets 53

2.5.4 Polytopes 55

2.5.5 Polytopes with fixed directions of outer unit normal vectors 55

2.6 Uniform deviation inequality for weakly-convex sets 56

2.6.1 Volume estimation and the dilated hull estimator 58

2.7 Adaptation to the regularity parameter $r^\star$ . . . 58
2.8 Illustrative simulations . . . 59
2.9 Appendix . . . 60

3 Optimal Link Prediction with Matrix Logistic Regression 65
3.1 Problem description . . . 68
3.1.1 Generative model . . . 68
3.1.2 Comparaison with other models . . . 69
3.1.3 Parameter space . . . 70
3.1.4 Explanatory variables . . . 71
3.2 Matrix Logistic Regression . . . 73
3.2.1 Penalized logistic loss . . . 74
3.2.2 Convex relaxation . . . 76
3.2.3 Prediction . . . 76
3.2.4 Information-theoretic lower bounds . . . 77
3.3 Computational lower bounds . . . 77
3.3.1 The dense subgraph detection problem . . . 78
3.3.2 Reduction to the dense subgraph detection problem and a computational lower bound . . . 79
3.4 Concluding remarks . . . 82
3.5 Proofs . . . 84
3.5.1 Some geometric properties of the likelihood . . . 84
3.5.2 Entropy bounds for some classes of matrices . . . 85
3.5.3 Proof of Theorem 3.2.1 and Theorem 3.2.7 . . . 85
3.5.4 Proof of Theorem 3.2.4 . . . 88
3.5.5 Proof of Theorem 3.2.8 . . . 88

Bibliography 93

Chapter 1

Unbiased estimation of the volume of a convex body

1.1 Introduction to volume estimation

In this chapter, we introduce the problem of volume estimation and explore it from different angles of analytic geometry, computational geometry and statistics.

1.1.1 Calculating the volume of geometric objects

The volume of a geometric set is one of its most basic functionals. Let us recall some of the standard results of calculating the volume in geometry. We consider the Euclidean space $\mathbb{R}^d$. In the two-dimensional case, there are plenty of formulas for calculating the area. The area S_p of a polygon inscribed in a circle, see Figure 1.1, of course, depends on the location of its vertices. Due to *the shoelace formula* discovered by Gauß (1777-1855), we have

$$S_p = \frac{1}{2}|(a_1b_2 + a_2b_3 + ... + a_5b_1) - (b_1a_2 + b_2a_3 + ... + b_5a_1)|\,.$$

For a polygon with the vertices lying on a grid of equidistant points with integer coordinates, see Figure 1.1, Pick's theorem, described by Pick (1859-1942), provides a simple formula for calculating the area S of this polygon in terms of the number $n_\circ$ of grid points located in the interior of the polygon and the number n_∂ of grid points (blue) lying on the polygon's boundary:

$$S = n_\circ + \frac{n_\partial}{2} - 1\,.$$

Already in the three-dimensional case some problems appear to be quite challenging. Calculating the volume of a polyhedron inscribed in a sphere is a fairly involved task. Let us assume without loss of generality that the boundary of a polyhedron P is given by a union of triangles $A_i, i = 1, ..., n,$ with vertices $(\boldsymbol{a}_i, \boldsymbol{b}_i, \boldsymbol{c}_i)$ which are assumed to be

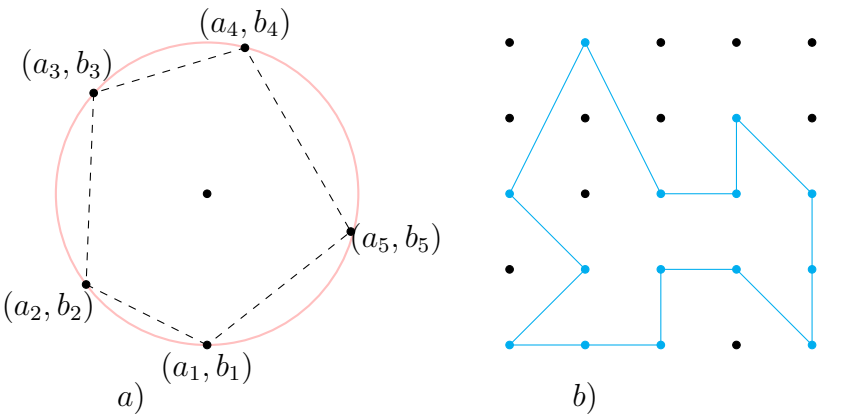

Figure 1.1: Calculating the volume using the shoelace formula and Pick's theorem

ordered counter clockwise on A_i. This means that on each A_i we can define the outer normal vector $\boldsymbol{n}_i = (\boldsymbol{b}_i - \boldsymbol{a}_i) \times (\boldsymbol{c}_i - \boldsymbol{a}_i)$. Then the volume of P is given by

$$V_P = \frac{1}{6} \sum_{i=1}^{n} \boldsymbol{a}_i \bullet \boldsymbol{n}_i \,, \tag{1.1.1}$$

where $\boldsymbol{a}_i \bullet \boldsymbol{n}_i$ denotes the dot product between $\boldsymbol{a}_i$ and $\boldsymbol{n}_i$. The proof of this result is based on the *divergence theorem*. As the divergence theorem, this result is due to Gauß. The analytical expression for the volume of a polyhedron inscribed in a sphere becomes even more involved. In some special cases, like when the polyhedron is a parallelotope, the results are simplified using matrix calculus. We refer the reader to [33] for a comprehensive summary of existing results in calculating the volume in geometry.

What about other convex bodies which have an arbitrary boundary? There is no unique recipe that allows one to calculate the volume of an arbitrary convex body precisely, but, as we shall see further, there are several techniques that allow one to *approximate* the volume of an arbitrary convex body with good precision. Furthermore, a more intriguing question is whether efficient estimation of the volume is possible for more general classes of bodies, i.e families of compact subsets. This question is partly driven by applications in image analysis and signal processing where the studied objects are rarely convex.

1.1.2 Estimating the volume in statistics

There can be no doubt that the origin of analytic geometry in antiquity was empirical. However, when we think about calculating the volume of some natural objects that arise nowadays, like a patient's tumour in biology or a star cluster in astronomy, the objects themselves are not accessible, i.e. we do not know the true shape of a studied object. We have access to only some information, or *data*, often imprecise and we want to recover the true shape of the body, its volume and/or other characteristics. The data we have are

some sort of measurements such as detection of the presence of a body in a certain region. Extracting information from the data about the true body is an objective of *statistical inference*.

A simple one-dimensional example

Let $X_1, ..., X_n$ be a sample of i.i.d. points drawn from the uniform distribution $U(a, b)$, and let $X_{(1)}, ..., X_{(n)}$ denote the order statistics, so that $X_{(1)} < ... < X_{(n)}$. It holds by symmetry that the expected length of the interval $(X_{(1)}, X_{(n)})$ satisfies

$$\mathbb{E}[|X_{(n)} - X_{(1)}|] = \frac{(n-1)}{(n+1)}(b-a)\,. \tag{1.1.2}$$

An objective of statistical inference is to estimate the length of the interval, when the location of the points a and b is assumed to be unknown. A naive estimator,

$$\widehat{l}_{naive} := X_{(n)} - X_{(1)}\,, \tag{1.1.3}$$

clearly underestimates the length. A more attractive idea is to somehow dilate the interval $(X_{(1)}, X_{(n)})$ and take the length of the dilated interval as an estimator. There are at least two viable dilations: 1) add and subtract some fixed vectors from the end points $X_{(n)}$ and $X_{(1)}$ (additive dilation) and 2) dilate the interval $(X_{(1)}, X_{(n)})$ from its centre $(X_{(n)} + X_{(1)})/2$ with some scaling factor (multiplicative dilation). In the one-dimensional case, both dilations are equivalent. It follows from (1.1.2) that a reasonable additive dilation factor is $2(X_{(n)} - X_{(1)})/(n-1)$ which yields an estimator for the volume,

$$\widehat{l}_1 := \frac{(n+1)}{(n-1)}(X_{(n)} - X_{(1)})\,. \tag{1.1.4}$$

This estimator is not only *unbiased*, $\mathbb{E}[\widehat{l}_1] = b - a$, but also, as we shall see in Section 2.3 and Section 2.4, is *minimax optimal*. We also refer to [106] for a comprehensive literature review of set estimation in the one-dimensional case.

Estimation of the volume of a convex set in high dimensions

The one-dimensional model is useful to grasp the main ideas of volume estimation, yet it is not widely used in real applications. The two-dimensional model already covers several important applications in image analysis and signal processing. Here, we observe the points $X_1, ..., X_n$ drawn uniformly over a set $C \subseteq \mathbb{R}^2$ and an objective is to recover the volume V_C of the set and the set itself. Let us assume that C belongs to the class of convex sets. Analogously to the one-dimensional case, it is natural to consider the volume $|\widehat{C}_n|$ of the convex hull as a baseline estimator for the volume V_C of the set C. It is quite

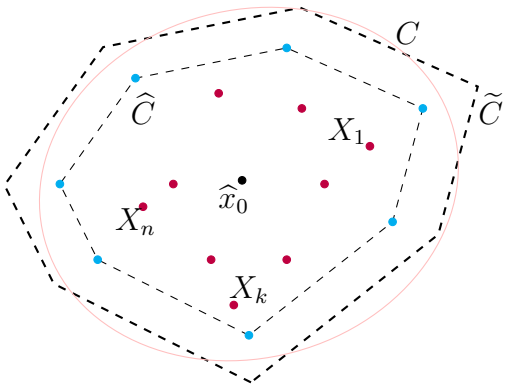

Figure 1.2: The points $X_1, ..., X_n$ drawn uniformly over a set C, the convex hull of the points $\widehat{C}_n = \text{conv}(X_1, ..., X_n)$ and the dilated hull estimator $\widetilde{C}$.

intuitive that this estimator performs quite poorly because it always underestimates the true volume and it should therefore be dilated as in the one-dimensional case. Section 2.3 and Section 2.4 show that an optimal estimator has the following form

$$\widehat{V}_{opt} = \frac{n+1}{n_\circ + 1} |\widehat{C}_n| \,, \tag{1.1.5}$$

where $n_\circ$ is the number of purple points in Figure 1.2 that lie in the interior of the convex hull $\widehat{C}_n$. Note that $\widehat{V}_{opt}$ is the volume of the "dilated" hull $\widetilde{C}$, the set obtained by dilating the convex hull with the same factor from the centre of gravity $\widehat{x}_0$ of the convex hull:

$$\widetilde{C} = \left\{ \widehat{x}_0 + \left(\frac{n+1}{n_\circ + 1} \right)^{1/2} (x - \widehat{x}_0) \,\middle|\, x \in \widehat{C}_n \right\}, \tag{1.1.6}$$

which can in fact be used to estimate the set C itself. Similarly, the same estimators for the volume and the set itself can be used in higher dimensions.

The uniform model of a fixed number of points drawn uniformly over a convex set C has been extensively studied in stochastic geometry. The focus of study is on understanding the distributional characteristics of key functionals like the volume of $\widehat{C}_n$, the number of vertices of $\widehat{C}_n$ and the distance between $\widehat{C}_n$ and C. The main references here are [16, 113, 119, 120, 137]. The Poisson point process (PPP) model studied in this chapter is closely related to the uniform model. Using Poissonisation and de-Poissonisation techniques, this model exhibits asymptotic properties like the uniform model, see e.g. the references above and Section 2.4.1. However the geometric properties of the PPP model are much richer for conducting statistical inference, see [14, 118], where the techniques from the Poisson point processes theory were successfully employed for estimation of linear functionals in a one-sided regression model and estimation of the volume of a convex set.

How fast can we estimate π?

There are quite a few ways how one can calculate the number π, see [8]. We here discuss one interesting way based on the Monte Carlo simulations of independent uniformly distributed random variables. It is a toy illustrative application of volume estimation in sampling theory. Let us draw the points $X_1, ..., X_N$ from the uniform distribution over the square $[0,1] \times [0,1]$ and count the number of points n which fall inside the circle centred at the origin of radius 1. Let $\widehat{\pi} := n/N$ denote the ratio of the points inside the circle to the total number of points. It approximately equals $\pi/4$, because it is an unbiased estimator:

$$\mathbb{E}[\widehat{\pi}] = \frac{1}{N}\mathbb{E}[n] = \frac{1}{N}\mathbb{E}\big[\sum_{i=1}^{N} \mathbf{1}(X_i \in C)\big] = \frac{\pi}{4}\,, \tag{1.1.7}$$

and therefore its mean squared risk is governed by the variance:

$$\mathbb{E}\big[(\widehat{\pi} - \pi)^2\big] = \mathrm{Var}(\widehat{\pi}) = \frac{1}{N^2}\,\mathrm{Var}\big(\sum_{i=1}^{N} \mathbf{1}(X_i \in C)\big) = \frac{1}{N}\frac{\pi}{4}\big(1 - \frac{\pi}{4}\big)\,. \tag{1.1.8}$$

It turns out $\widehat{\pi}$ is even a maximum likelihood estimator. Surprisingly, we are able to estimate π with a much faster rate based on the data points in this experiment. Following (1.1.5), we define our properly scaled estimator as

$$\widehat{\pi}_{opt} = 4\frac{n+1}{n_\circ + 1}|\widehat{C}_n|\,, \tag{1.1.9}$$

where $n_\circ$ is the number of points lying inside the convex hull $\widehat{C}_n$ of the points lying inside the circle. Theorem 2.3.1 and Theorem 2.4.2 to follow state that the rate of convergence of the mean squared risk of the estimator $\widehat{\pi}_{opt}$ satisfies $\mathbb{E}\big[(\widehat{\pi}_{opt} - \pi)^2\big] = \mathcal{O}(N^{-5/3})$, see Figure 1.3 for a numerical comparison of the two estimators. Note that both estimators can easily be computed in polynomial time. The optimal time complexity of computing a convex hull in 2- or 3- dimensional space is $\mathcal{O}(n \log h)$, where h is the number of vertices of the convex hull, and is achieved by Chan's algorithm in [45].

1.1.3 Estimating the support of a density

The problem of estimating the volume of a body is a special case of a more general problem of estimating the support of a probability density. It has received a fairly large amount of attention in the statistical literature since the 1980s partly because of several applications in image analysis, signal processing and econometrics. The first fundamental results in this area were obtained in [50, 55, 88–90]. Furthermore see [99, 124, 133] for a more general problem of estimating the level sets of a density. In particular, [88] established the minimax optimal rates for estimating the support of a density having a Hölder-continuous boundary

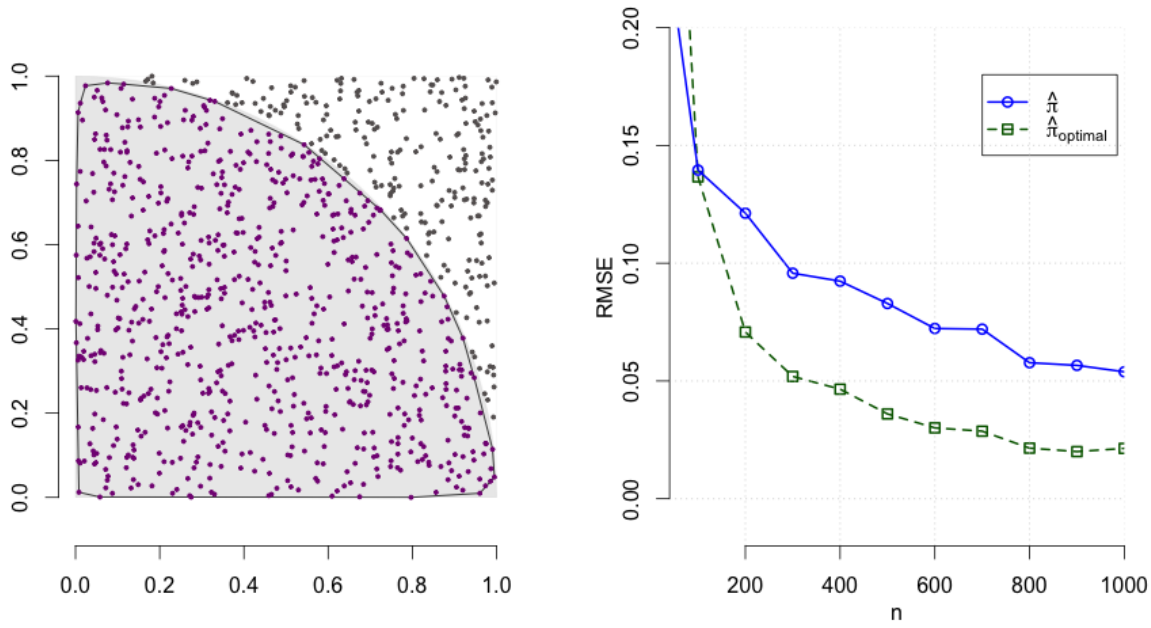

Figure 1.3: On the left: a sample of $n = 500$ points drawn uniformly over the square $[0, 1] \times [0, 1]$. On the right: Monte Carlo root mean squared error (RMSE) estimates for the studied estimators for π based on 200 Monte Carlo simulations in each case.

in the Hausdorff and symmetric difference metrics and constructed an estimator which attains the optimal rates. The case of convex support estimation was first studied in [89, 90], where it was shown that the convex hull $\widehat{C}_n$ of the sample points, which is the maximum likelihood estimator, is rate-optimal for estimating the support set C in the Hausdorff and symmetric difference metrics.

The volume of a set is clearly one of its most basic characteristics. Provocatively, as it was shown in [88, 89], the volume of a rate-optimal estimator of the set is not necessarily a rate-optimal estimator of the volume! The first fully rate-optimal estimators of the volume of a convex support with smooth boundary and a support with Hölder-continuous boundary were constructed by [67] based on three-fold sample splitting.

1.1.4 Looking further. Computational geometry

The problem of calculating the volume of a convex body has also attracted researchers working in computer science and computational geometry, see [58, 80, 95]. As the dimension of the space grows, the studied objects become more and more complicated and it is no longer possible to apply some nice analytical formula like (1.1.1) even if we know the location of an object. The so-called voxel-based methods serve to estimate the volume with good precision. However, voxel-based methods have been found to be computationally inefficient in high dimensions and researchers choose to use various fast randomised methods to estimate the volume. We refer to [136] for a recent survey of the existing fast randomised algorithms for calculating the volume of a convex body.

One such randomised algorithm, although probably not the fastest, is exactly to follow the strategy above. Given a body of interest, one can sample the points uniformly over it, calculate the volume of the convex hull of the points and then make a necessary dilation. Since it is computationally easier to calculate the volume of a polytope than of an arbitrary convex body, this procedure can save expensive running time, although computing the volume of the convex hull in high dimensions is still an involved task, see [131]. Nevertheless, it is quite fascinating that once the volume of the convex hull is computed the dilation (1.1.5) involving the number of points should be employed to estimate the volume with the best possible precision.

To our best knowledge, no efficient estimators for the volume of a set from a class of sets that is more general than the class of convex sets have previously been proposed.

1.2 Introduction to unbiased volume estimation of a convex set

The contribution of this part is the construction of a very simple volume estimator which is not only rate-optimal over all convex sets without boundary restrictions, but even adaptive in the sense that it attains almost the parametric rate if the convex set is a polytope. The analysis is based on a Poisson point process (PPP) observation model with intensity $\lambda > 0$ on the convex set $C \subseteq \mathbb{R}^d$. We thus observe

$$X_1, ..., X_N \overset{i.i.d.}{\sim} U(C), \quad N \sim \text{Poiss}(\lambda|C|), \tag{1.2.1}$$

where $(X_n), N$ are independent, see Section 1.3 below for a concise introduction to the PPP model. Using Poissonisation and de-Poissonisation techniques, this model exhibits asymptotic properties like the uniform model, i.e. a sample of $n = \lambda|C|$ uniformly on C distributed random variables $X_1, \dots, X_n$. The geometry of the PPP model, however, allows for much more concise ideas and proofs, see also [103] for connections between PPP and regression models with irregular error distributions. From an applied perspective, PPP models are often natural, e.g. for spatial count data of photons or other emissions.

For known intensity λ of the PPP, we construct in Section 1.4 an *oracle* estimator $\widehat{\vartheta}_{oracle}$. Theorem 1.4.2 shows that this estimator is UMVU (uniformly of minimum variance among unbiased estimators) and rate-optimal. To this end, moment bounds from stochastic geometry for the missing volume of the convex hull, obtained by [16] and [57] are essential. Moreover, we derive results of independent interest: the convex hull $\widehat{C} = \text{conv}\{X_1, \dots, X_N\}$ forms a sufficient and complete statistic (Proposition 1.4.5) and the Poisson point process, conditionally on $\widehat{C}$, remains Poisson within its convex hull (Theorem 2.2.5).

For the more realistic case of unknown intensity λ, we analyse in Section 1.5 our final

estimator

$$\widehat{\vartheta} \stackrel{\text{def}}{=} \frac{N+1}{N_\circ+1}|\widehat{C}|\,, \tag{1.2.2}$$

where $N_\circ$ denotes the number of observed points in the interior of $\widehat{C}$. We are able to prove a sharp oracle inequality, comparing the risk of this estimator to that of $\widehat{\vartheta}_{oracle}$. Here, very recent and advanced results by [18, 113, 119] on the variance of the number of points N_∂ on the boundary of $\widehat{C}$ and the missing volume $|C \setminus \widehat{C}|$ are of key importance. This interplay between stochastic geometry and statistics prevails throughout the work. Note that a similar estimator for the volume, namely $(N/N_\circ)|\widehat{C}|$, was introduced earlier in [125] and in [94] using heuristic arguments.

The lower bound showing that $\widehat{\vartheta}$ is indeed minimax-optimal is proved in Theorem 1.4.4 by adopting the proof of the lower bound in the uniform model by [67]. A small simulation study is presented in Section 1.6. Moreover, we propose to enlarge the convex hull set by the factor $((N+1)/(N_\circ+1))^{1/d}$ and we study its error as an estimator of the set C itself. The proof of Lemma 1.5.1 is deferred to the Appendix.

1.3 Digression on Poisson Point Processes

Most of the results and notation are adapted from [83]. We fix a compact convex set $\mathbf{E}$ in $\mathbb{R}^d$ with non-empty interior as a state space and denote by $\mathcal{E}$ its Borel σ-algebra. We define the family of convex subsets $\mathbf{C} = \{C \subseteq \mathbf{E}, \text{convex}, \text{closed}\}$ (this implies that all sets in $\mathbf{C}$ are compact) and the family of compact subsets $\mathbf{K} = \{K \subseteq \mathbf{E}, \text{compact}\}$. It is natural to equip the space $\mathbf{C}$ (resp. $\mathbf{K}$) with the Hausdorff-metric d_H and its Borel σ-algebra $\mathfrak{B}_{\mathbf{C}}$ (resp. $\mathfrak{B}_{\mathbf{K}}$). Then $(\mathbf{C}, d_H)$ is a compact, and thus, separable space and the mapping $(x_1, \ldots, x_k) \mapsto \text{conv}\{x_1, \ldots, x_k\}$, which generates the convex hull of points $x_i \in \mathbf{E}$, is continuous from $\mathbf{E}^k$ to $(\mathbf{C}, d_H)$.

On $(\mathbf{E}, \mathcal{E})$ we define the set of point measures $\mathbf{M} = \{m \text{ measure on } \mathcal{E} \,:\, m(A) \in \mathbb{N},\ \forall A \in \mathcal{E}\}$ equipped with the σ-algebra $\mathcal{M} = \sigma(m \mapsto m(A), A \in \mathcal{E})$. Let $C_c^+(\mathbf{E})$ be the collection of continuous functions $\mathbf{E} \mapsto [0, \infty)$ with compact support. A useful topology for $\mathbf{M}$ is the *vague topology* which makes $\mathbf{M}$ a complete, separable metric space, cf. Section 3.4 in [121]. A sequence of point measures $m_n \in \mathbf{M}$ then converges vaguely to a limit $m \in \mathbf{M}$ if and only if $m_n[f] \to m[f]$ for all $f \in C_c^+(\mathbf{E})$ where $m[f] = \int_{\mathbf{E}} f dm$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an abstract probability space. We call a measurable mapping $\mathcal{N} : \Omega \to \mathbf{M}$ a Poisson point process (PPP) of intensity $\lambda > 0$ on $C \in \mathbf{C}$ if

- for any $A \in \mathcal{E}$, we have $\mathcal{N}(A) \sim \text{Poiss}\big(\lambda|A \cap C|\big)$, where $|A \cap C|$ denotes the Lebesgue measure of $A \cap C$;
- for all mutually disjoint sets $A_1, ..., A_n \in \mathcal{E}$, the random variables $\mathcal{N}(A_1), ..., \mathcal{N}(A_n)$ are independent.

For statistical inference, we assume the Poisson point process to be defined on a set of non zero Lebesgue measure, i.e. $|C| > 0$. A more constructive and intuitive representation of the PPP $\mathcal{N}$ is $\mathcal{N} = \sum_{i=1}^{N} \delta_{X_i}$ for $N \sim \text{Poiss}(\lambda|C|)$ and i.i.d. random variables (X_i), independent of N and distributed uniformly $\mathbb{P}(X_i \in A) = |A \cap C|/|C|$, so that $\mathcal{N}(A) = \sum_{i=1}^{N} \mathbf{1}(X_i \in A)$ for any $A \in \mathcal{E}$.

We consider the convex hull of the PPP points $\widehat{C} : \mathbf{M} \to \mathbf{C}$ defined by $\widehat{C}(\mathcal{N}) := \text{conv}\{X_1, ..., X_N\}$, which by the above continuity property of the convex hull is a random element with values in the Polish space $(\mathbf{C}, d_H)$, see also [52] for a detailed study of the continuity of the convex hull. For shorthand notation, we shall further write $\widehat{C}$ to denote the convex hull of the process $\mathcal{N}$. In the sequel, conditional expectations and probabilities with respect to $\widehat{C}$ are thus well defined. We can also evaluate the probability

$$\mathbb{P}_C\big(\widehat{C} \in A\big) = \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda|C|}\lambda^k}{k!} \int_{C^k} \mathbf{1}(\text{conv}\{x_1, ..., x_k\} \in A) d(x_1, ..., x_k)$$

for $A \in \mathfrak{B}_{\mathbf{C}}$. Usually, we only write the subscript C or sometimes (C, λ) when different probability distributions are considered simultaneously. The likelihood function $\frac{d\mathbb{P}_{C,\lambda}}{d\mathbb{P}_{\mathbf{E},\lambda_0}}$ for $C \in \mathbf{C}$ and $\lambda, \lambda_0 > 0$ is then given by

$$\frac{d\mathbb{P}_{C,\lambda}}{d\mathbb{P}_{\mathbf{E},\lambda_0}}(X_1, ..., X_N) = \mathrm{e}^{\lambda_0|\mathbf{E}|-\lambda|C|}(\lambda/\lambda_0)^N \mathbf{1}(\forall\, i = 1, ..., N :\ X_i \in C) \tag{1.3.1}$$

$$= \mathrm{e}^{\lambda_0|\mathbf{E}|-\lambda|C|}(\lambda/\lambda_0)^N \mathbf{1}(\widehat{C} \subseteq C)\,, \tag{1.3.2}$$

cf. Thm. 1.3 in [91]. For the last line, we have used that a point set is in C if and only if its convex hull is contained in C.

For the set-indexed process $(\mathcal{N}(K), K \in \mathbf{K})$ we define its natural set-indexed filtration

$$\mathcal{F}_K \stackrel{\text{def}}{=} \sigma(\{\mathcal{N}(U); U \subseteq K, U \in \mathbf{K}\}) \tag{1.3.3}$$

for any $K \in \mathbf{K}$. The filtration $(\mathcal{F}_K, K \in \mathbf{K})$ possesses the following properties:

- monotonicity: $\mathcal{F}_{K_1} \subseteq \mathcal{F}_{K_2}$ for any $K_1, K_2 \in \mathbf{K}$ with $K_1 \subseteq K_2$,
- continuity from above: $\mathcal{F}_K = \cap_{i=1}^{\infty} \mathcal{F}_{K_i}$ if $K_i \downarrow K$;

cf. [148]. By construction, the restriction $\mathcal{N}_K = \mathcal{N}(\cdot \cap K)$ of the point process $\mathcal{N}$ onto $K \in \mathbf{K}$ is $\mathcal{F}_K$-measurable (in fact, $\mathcal{F}_K = \sigma(\{\mathcal{N}_K(U); U \in \mathbf{K}\})$). In addition, it can be easily seen that $\mathcal{N}_K$ is a Poisson point process in $\mathbf{M}$, cf. the Restriction Theorem in [85], and thus $\widehat{C}(\mathcal{N}_K) = \text{conv}(\{X_1, \ldots, X_N\} \cap K)$ is by the above arguments $\mathcal{F}_K$-measurable.

A random compact set $\mathcal{K}$ is a measurable mapping $\mathcal{K} : (\mathbf{M}, \mathcal{M}) \to (\mathbf{K}, \mathfrak{B}_{\mathbf{K}})$. Note that [148] defines a random compact set as a measurable mapping from $(\mathbf{M}, \mathcal{M})$ to $(\mathbf{K}, \sigma_{\mathbf{K}})$ where $\sigma_{\mathbf{K}}$ is the so-called *Effros* σ-algebra generated by the sets $\{F \in \mathbf{K} : F \cap K \neq \emptyset\}$,

$K \in \mathbf{K}$. Thanks to Thm. 2.7 in [105], the Effros σ-algebra $\sigma_{\mathbf{K}}$ induced on the family of compact sets $\mathbf{K}$ coincides with the Borel σ-algebra $\mathfrak{B}_{\mathbf{K}}$, and we prefer to stick to the first definition of a random compact set for convenience. Next, we recall the definition of stopping sets from [127] in complete analogy with stopping times.

DEFINITION 1.3.1. A random compact set $\mathcal{K}$ is called an $\mathcal{F}_K$-stopping set if $\{\mathcal{K} \subseteq K\} \in \mathcal{F}_K$ for all $K \in \mathbf{K}$. The sigma-algebra of $\mathcal{K}$-history is defined as $\mathcal{F}_{\mathcal{K}} = \{A \in \mathfrak{F} : A \cap \{\mathcal{K} \subseteq K\} \in \mathcal{F}_K \ \forall K \in \mathbf{K}\}$, where $\mathfrak{F} = \sigma(\mathcal{F}_K; K \in \mathbf{K})$.

For a set $A \subseteq \mathbf{E}$ let A^c denote its complement.

LEMMA 1.3.2. The set $\widehat{\mathcal{K}} \stackrel{\text{def}}{=} \overline{\widehat{C}^c}$, the closure of the complement of the convex hull, is an $(\mathcal{F}_K)$-stopping set.

Proof. We claim $\widehat{\mathcal{K}} \subseteq K$ if and only if $K^c \subseteq \operatorname{conv}(\{X_1, \dots, X_N\} \cap K)$. Indeed, if $\widehat{\mathcal{K}} \subseteq K$ holds, then the boundary $\partial\widehat{C} = \partial\widehat{\mathcal{K}}$ is in K which implies $\operatorname{conv}(\{X_1, \dots, X_N\} \cap K) = \widehat{C}$. Consequently, $K^c \subseteq \widehat{\mathcal{K}}^c \subseteq \widehat{C} = \operatorname{conv}(\{X_1, \dots, X_N\} \cap K)$ holds. Conversely, $K^c \subseteq \operatorname{conv}(\{X_1, \dots, X_N\} \cap K)$ implies immediately $K^c \subseteq \widehat{C}$ and thus $\widehat{C}^c \subseteq K$. Since K is closed, we obtain $\widehat{\mathcal{K}} \subseteq K$.

Since $\{X_1, \dots, X_N\} \cap K$ are the realisations of the point process inside K and the convex hull is measurable, we conclude $\{K^c \subseteq \operatorname{conv}(\{X_1, \dots, X_N\} \cap K)\} \in \mathcal{F}_K$. □

We shall further use the following short notation: $N = \mathcal{N}(C)$ denotes the total number of points, $N_\circ = \mathcal{N}(\widehat{C}^\circ)$ the number of points in the interior of the convex hull $\widehat{C}$ and $N_\partial = \mathcal{N}(\partial\widehat{C}) = \mathcal{N}(\partial\widehat{\mathcal{K}})$ the number of points on the boundary of the convex hull. For asymptotic bounds we write $f(x) = O(g(x))$ or $f(x) \lesssim g(x)$ if $f(x)$ is bounded by a constant multiple of $g(x)$ and $f(x) \sim g(x)$ if $f(x) \lesssim g(x)$ as well as $g(x) \lesssim f(x)$.

1.4 Oracle case: intensity λ is known

For a PPP on $C \in \mathbf{C}$ with intensity $\lambda > 0$, we know $N \sim \text{Poiss}(\lambda|C|)$. In the oracle case, when the intensity λ is known, N/λ estimates $|C|$ without bias and yields the classical parametric rate in λ:

$$\mathbb{E}[(N/\lambda - |C|)^2] = \lambda^{-2} \operatorname{Var}(N) = \frac{|C|}{\lambda}. \tag{1.4.1}$$

Another natural idea might be to use the plug-in estimator $|\widehat{C}|$ whose error is given by the missing volume and satisfies

$$\mathbb{E}[(|\widehat{C}| - |C|)^2] = \mathbb{E}[|C \setminus \widehat{C}|^2] = O(|C|^{2(d-1)/(d+1)} \lambda^{-4/(d+1)}), \tag{1.4.2}$$

where the bound is obtained similarly to (1.4.8) and (1.4.10) below. This means that its error is of smaller order than λ^{-1} for $d \leqslant 2$, but larger for $d \geqslant 4$. For any $d \geqslant 2$, however, both convergence rates are worse than the minimax-optimal rate $\lambda^{-(d+3)/(d+1)}$, established below.

The way to improve these estimators is to observe that by the likelihood representation (1.3.2) for $\lambda = \lambda_0$ and the Neyman factorisation criterion the convex hull is a sufficient statistic. Consequently, by the Rao-Blackwell theorem, the conditional expectation of N/λ given the convex hull $\widehat{C}$ is an estimator with smaller mean squared error (MSE).

The number N can be split into the number N_∂ of points on the boundary and the number $N_\circ$ of points in the interior of the convex hull. The following theorem is essential in deriving the oracle estimator. Although the statement of the theorem is quite intuitive and already used in [117], the proof turns out to be nontrivial and is deferred to the Appendix.

THEOREM 1.4.1. The number N_∂ of points on the boundary of the convex hull is measurable with respect to the sigma-algebra of $\widehat{\mathcal{K}}$-history $\mathcal{F}_{\widehat{\mathcal{K}}}$. The number of points in the interior of the convex hull $N_\circ$ is, conditionally on $\mathcal{F}_{\widehat{\mathcal{K}}}$, Poisson-distributed:

$$N_\circ \,\big|\, \mathcal{F}_{\widehat{\mathcal{K}}} \sim \mathrm{Poiss}(\lambda_\circ) \text{ with } \lambda_\circ \stackrel{\mathrm{def}}{=} \lambda|\widehat{C}|. \tag{1.4.3}$$

In addition, we have $\mathcal{F}_{\widehat{\mathcal{K}}} = \sigma(\widehat{C})$, where the latter is the sigma-algebra $\sigma(\{\widehat{C} \subseteq B, B \in \mathbf{C}\})$ completed with the null sets in $\mathfrak{F}$.

With Theorem 2.2.5 at hand, we obtain the *oracle* estimator

$$\widehat{\vartheta}_{oracle} \stackrel{\mathrm{def}}{=} \mathbb{E}\Big[\frac{N}{\lambda} \,\Big|\, \widehat{C}\Big] = \mathbb{E}\Big[\frac{N_\circ + N_\partial}{\lambda} \,\Big|\, \widehat{C}\Big] = |\widehat{C}| + \frac{N_\partial}{\lambda}\,, \tag{1.4.4}$$

where conditioning on $\widehat{C}$ means conditioning on $\sigma(\widehat{C}) = \mathcal{F}_{\widehat{\mathcal{K}}}$.

THEOREM 1.4.2. For known intensity $\lambda > 0$, the oracle estimator $\widehat{\vartheta}_{oracle}$ is unbiased and of minimal variance among all unbiased estimators (UMVU). It satisfies

$$\mathrm{Var}(\widehat{\vartheta}_{oracle}) = \frac{1}{\lambda}\mathbb{E}[|C \setminus \widehat{C}|]\,.$$

Its worst case mean squared error over $\mathbf{C}$ decays like $\lambda^{-(d+3)/(d+1)}$ as $\lambda \uparrow \infty$ in dimension d:

$$\limsup_{\lambda\to\infty} \lambda^{(d+3)/(d+1)} \sup_{C\in\mathbf{C},|C|>0} \Big\{|C|^{-(d-1)/(d+1)}\mathbb{E}\big[(\widehat{\vartheta}_{oracle} - |C|)^2\big]\Big\} < \infty\,. \tag{1.4.5}$$

REMARK 1.4.3. The theorem implies that the rate of convergence for the RMSE (root mean-squared error) of the estimator $\widehat{\vartheta}_{oracle}$ is $\lambda^{-(d+3)/(2d+2)}$. In Theorem 1.4.4 below, we prove

that the lower bound on the minimax risk in the PPP model is of the same order implying that the rate is minimax-optimal. Even more, the oracle estimator is *adaptive* in the sense, that its rate is faster if the missing volume decays faster. In particular, for polytopes C it is shown in [16] and independently in [57] that $\mathbb{E}[|C\setminus\widehat{C}|] \sim \lambda^{-1}(\log(\lambda|C|))^{d-1}$, which implies a faster (almost parametric) rate of convergence for the RMSE of the oracle estimator.

Proof of Theorem 1.4.2. The unbiasedness follows immediately from the definition (1.4.4). By the law of total variance, we obtain

$$\mathrm{Var}(\widehat{\vartheta}_{oracle}) = \mathrm{Var}\Big(\frac{N}{\lambda}\Big) - \mathbb{E}\Big[\mathrm{Var}\Big(\frac{N}{\lambda}\,\Big|\,\widehat{C}\Big)\Big] = \frac{|C|}{\lambda} - \mathbb{E}\Big[\mathrm{Var}\Big(\frac{N_\circ}{\lambda}\,\Big|\,\widehat{C}\Big)\Big] \tag{1.4.6}$$

$$= \frac{|C|}{\lambda} - \mathbb{E}\Big[\frac{\lambda_\circ}{\lambda^2}\Big] = \frac{1}{\lambda}\mathbb{E}[|C\setminus\widehat{C}|]\,. \tag{1.4.7}$$

Proposition 1.4.5 below affirms that the convex hull $\widehat{C}$ is not only a sufficient, but also a complete statistic such that by the Lehmann-Scheffé theorem, the estimator $\widehat{\vartheta}_{oracle}$ has the UMVU property.

Finally, we bound the expectation of the missing volume $|C\setminus\widehat{C}|$ by Poissonisation, i.e. using that the convex hull $\widehat{C}$ in the PPP model conditionally on the event $\{N=k\}$ is distributed as the convex hull $\widehat{C}_k = \mathrm{conv}\{X_1,...,X_k\}$ in the model with k uniform observations on C, for which the following upper bound is known (e.g., [16]):

$$\sup_{C\in\mathbf{C},|C|>0}\mathbb{E}\Big[\frac{|C\setminus\widehat{C}_k|}{|C|}\Big] = O\big(k^{-2/(d+1)}\big)\,. \tag{1.4.8}$$

Thus, it follows by a Poisson moment bound

$$\sup_{C\in\mathbf{C},|C|>0}\mathbb{E}\Big[\frac{|C\setminus\widehat{C}|}{|C|^{(d-1)/(d+1)}}\Big] = \sup_{C\in\mathbf{C},|C|>0}\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda|C|}(\lambda|C|)^k}{|C|^{-2/(d+1)}k!}\mathbb{E}\Big[\frac{|C\setminus\widehat{C}_k|}{|C|}\Big] \tag{1.4.9}$$

$$= O\big(\lambda^{-2/(d+1)}\big)\,. \tag{1.4.10}$$

This bound, together with (1.4.7), yields the assertion. □

The lower bound for the risk in the PPP framework can be derived from the lower bound in the uniform model with a fixed number of observations, see Thm. 6 in [67].

THEOREM 1.4.4. For estimating $|C|$ in the PPP model with parameter class $\mathbf{C}$, the following asymptotic lower bound holds

$$\liminf_{\lambda\to\infty}\lambda^{(d+3)/(d+1)}\inf_{\widehat{\vartheta}_\lambda}\sup_{C\in\mathbf{C}}\mathbb{E}_C[(|C|-\widehat{\vartheta}_\lambda)^2] > 0\,, \tag{1.4.11}$$

where the infimum extends over all estimators $\widehat{\vartheta}_\lambda$ in the PPP model with intensity λ.

Proof. We use that an estimator $\widehat{\vartheta}_\lambda$ in the PPP model is an estimator in the uniform model on the event $\{N = n\}$. Then, due to the lower bound in the uniform model in [67], for a constant $c > 0$ and for all $n \in \mathbb{N}$ there exists a set $C_n \in \mathbf{C}$ with $|C_n| \sim 1$ such that for all $k = 1, ..., n$,

$$\mathbb{E}_{C_n}\big[(|C_n| - \widehat{\vartheta}_\lambda)^2 \,\big|\, N = k\big] > cn^{-(d+3)/(d+1)}, \quad a.s. \tag{1.4.12}$$

Then, in the PPP model for $C = C_{\lfloor\lambda\rfloor}$ with $\lambda|C| \geqslant 1$, we have

$$\begin{aligned}
\mathbb{E}_C[(|C| - \widehat{\vartheta}_\lambda)^2] &= \sum_{k\in\mathbb{N}} \mathbb{E}_C\big[(|C| - \widehat{\vartheta}_\lambda)^2 \,\big|\, N = k\big]\, \mathbb{P}(N = k) && (1.4.13)\\
&\geqslant \sum_{k\leqslant\lfloor\lambda\rfloor} \mathbb{E}_C\big[(|C| - \widehat{\vartheta}_\lambda)^2 \,\big|\, N = k\big]\, \mathbb{P}(N = k) && (1.4.14)\\
&> c\lfloor\lambda\rfloor^{-(d+3)/(d+1)} \big(1 - \mathbb{P}(N > \lfloor\lambda\rfloor)\big) && (1.4.15)\\
&\sim \lambda^{-(d+3)/(d+1)}, && (1.4.16)
\end{aligned}$$

applying Chernoff's inequality to $N \sim \mathrm{Poiss}(\lambda|C|)$ for the last line. Thus, the lower bound (1.4.11) follows. □

Proposition 1.4.5. For known intensity $\lambda > 0$, the convex hull $\widehat{C} = \mathrm{conv}\{X_1, ..., X_N\}$ is a complete statistic.

Proof. We need to show the implication

$$\forall C \in \mathbf{C} : \mathbb{E}_C\big[T(\widehat{C})\big] = 0 \implies T(\widehat{C}) = 0 \quad \mathbb{P}_{\mathbf{E}} - a.s. \tag{1.4.17}$$

for any $\mathfrak{B}_{\mathbf{C}}$-measurable function $T : \mathbf{C} \to \mathbb{R}$. From the likelihood in (1.3.2) for $\lambda = \lambda_0$, we derive

$$\mathbb{E}_C\big[T(\widehat{C})\big] = \mathbb{E}_{\mathbf{E}}\Big[T(\widehat{C}) \exp\big(\lambda|\mathbf{E} \setminus C|\big)\mathbf{1}(\widehat{C} \subseteq C)\Big]. \tag{1.4.18}$$

Since $\exp(\lambda|\mathbf{E} \setminus C|)$ is deterministic, $\mathbb{E}_C\big[T(\widehat{C})\big] = 0$ for all $C \in \mathbf{C}$ implies

$$\forall C \in \mathbf{C} : \mathbb{E}_{\mathbf{E}}\big[T(\widehat{C})\mathbf{1}(\widehat{C} \subseteq C)\big] = 0\,. \tag{1.4.19}$$

For $C \in \mathbf{C}$, define the family of convex subsets of C as $[C] = \{A \in \mathbf{C} | A \subseteq C\}$ such that $\widehat{C} \subseteq C \iff \widehat{C} \in [C]$. Splitting $T = T^+ - T^-$ with non-negative $\mathfrak{B}_{\mathbf{C}}$-measurable functions T^+ and T^-, we infer that the measures $\mu^\pm(B) = \mathbb{E}_{\mathbf{E}}[T^\pm(\widehat{C})\mathbf{1}(\widehat{C} \in B)]$, $B \in \mathfrak{B}_{\mathbf{C}}$, agree on $\{[C] \,|\, C \in \mathbf{C}\}$.

Note that the brackets $\{[C] | C \in \mathbf{C}\}$ are $\cap$-stable due to $[A] \cap [C] = [A \cap C]$ and $A \cap C \in \mathbf{C}$. If the σ-algebra $\mathcal{C}$ generated by $\{[C] \,|\, C \in \mathbf{C}\}$ contains $\mathfrak{B}_{\mathbf{C}}$, the uniqueness

theorem asserts that the measures μ^+, μ^- agree on all Borel sets in $\mathfrak{B}_{\mathbf{C}}$, in particular on $\{T > 0\}$ and $\{T < 0\}$, which entails $\mathbb{E}_{\mathbf{E}}[T^+(\widehat{C})] = \mathbb{E}_{\mathbf{E}}[T^-(\widehat{C})] = 0$. Thus, in this case, $T(\widehat{C}) = 0$ holds $\mathbb{P}_{\mathbf{E}}$-a.s.

It remains to show that $\mathcal{C} = \sigma([C],\, C \in \mathbf{C})$ equals the Borel σ-algebra $\mathfrak{B}_{\mathbf{C}}$. This can be derived as a non-trivial consequence of Choquet's theorem, see Thm. 7.8 in [105], but we propose a short self-contained proof here. Let us define the family $\langle C\rangle = \{B \in \mathbf{C} | C \subseteq B\}$ of convex sets containing C. Then the closed Hausdorff ball with center C and radius $\varepsilon > 0$ has the representation

$$B_\varepsilon(C) \stackrel{\text{def}}{=} \{A \in \mathbf{C} \,|\, d_H(A, C) \leqslant \varepsilon\} = \{A \in \mathbf{C} \,|\, U_{-\varepsilon}(C) \subseteq A \subseteq U_\varepsilon(C)\}\,, \tag{1.4.20}$$

with $U_\varepsilon(C) = \{x \in \mathbf{E} \,|\, \mathrm{dist}(x, C) \leqslant \varepsilon\}$, $U_{-\varepsilon}(C) = \{x \in C \,|\, \mathrm{dist}(x, \mathbf{E} \setminus C) \leqslant \varepsilon\}$. Noting that $U_\varepsilon(C), U_{-\varepsilon}(C)$ are closed and convex and thus in $\mathbf{C}$, we obtain

$$B_\varepsilon(C) = \langle U_{-\varepsilon}(C)\rangle \cap [U_\varepsilon(C)]\,. \tag{1.4.21}$$

Since $(\mathbf{C}, d_H)$ is separable, our problem is reduced to proving that all angle sets $\langle C\rangle$ for $C \in \mathbf{C}$ are in $\mathcal{C}$. A further reduction is achieved by noting $\langle C\rangle = \bigcap_{x\in C}\langle x\rangle = \bigcap_{x\in C\cap\mathbb{Q}^d}\langle x\rangle$ setting $\langle x\rangle = \langle\{x\}\rangle$ for short such that it suffices to prove $\langle x\rangle \in \mathcal{C}$ for all $x \in \mathbf{E}$.

Now, let $x \in \mathbf{E}$ and $C \in \mathbf{C}$ such that $x \notin C$. Then, by the Hahn-Banach theorem, there are $\delta > 0, v \in \mathbb{R}^d$ such that $\langle v, c - x\rangle \geqslant \delta$ holds for all $c \in C$. By a density argument, we may choose $\delta \in \mathbb{Q}^+$ and $v \in \mathbb{Q}^d$. Denoting the corresponding hyperplane intersected with $\mathbf{E}$ by $H_{\delta,v} = \{\xi \in \mathbf{E} \,|\, \langle v, \xi - x\rangle \geqslant \delta\}$, see Figure 3.1, we conclude

$$\langle x\rangle^{\complement} = \bigcup_{\delta\in\mathbb{Q}_+}\bigcup_{v\in\mathbb{Q}^d}\underbrace{[H_{\delta,v}]}_{\in\mathcal{C}} \in \mathcal{C}\,. \tag{1.4.22}$$

Consequently, $\langle x\rangle \in \mathcal{C}$ and thus $\mathfrak{B}_{\mathbf{C}} \subseteq \mathcal{C}$ hold. □

1.5 Unknown intensity λ: nearly unbiased estimation

In case the intensity λ is unknown and the oracle estimator $\widehat{\vartheta}_{oracle}$ in (1.4.4) is inaccessible, the maximum-likelihood approach suggests to use $N/|\widehat{C}|$ as an estimator for λ in (1.3.2). This yields the *plug-in* estimator for the volume,

$$\widehat{\vartheta}_{plugin} \stackrel{\text{def}}{=} |\widehat{C}| + \frac{N_\partial}{N}|\widehat{C}|\,. \tag{1.5.1}$$

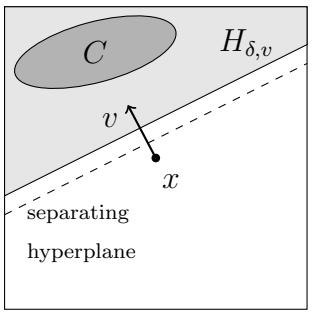

Figure 1.4: The construction used in the proof.

In the unlikely event $N = |\widehat{C}| = 0$, we define $\widehat{\vartheta}_{plugin} = 0$. This estimator has a significant bias due to the following result, which is proved in the appendix.

LEMMA 1.5.1. For the bias of the plug-in MLE estimator $\widehat{\vartheta}_{plugin}$, it follows with some universal constant $c > 0$

$$|C| - \mathbb{E}[\widehat{\vartheta}_{plugin}] \geqslant c\mathbb{E}[|\widehat{C} \setminus C|]^2\,, \quad \forall C \in \mathbf{C}\,. \tag{1.5.2}$$

The maximal bias over $C \in \mathbf{C}$ is thus at least of order $\lambda^{-4/(d+1)}$, which is worse than the minimax rate $\lambda^{-(d+3)/(2d+2)}$ for $d > 5$. Yet, in the two-dimensional finite sample study of Section 1.6 below, its performance is quite convincing. We surmise that $\widehat{\vartheta}_{plugin}$ is rate-optimal for $d \leqslant 5$, but we leave that question aside because the final estimator we propose will be nearly unbiased and will satisfy an *exact* oracle inequality. In particular, it is rate-optimal in any dimension. The new idea is to exploit that the number of interior points of $\widehat{C}$ satisfies $N_\circ \,\big|\, \widehat{C} \sim \mathrm{Poiss}(\lambda_\circ)$, see (1.4.3).

LEMMA 1.5.2. There is no conditionally unbiased estimator for $\lambda_\circ^{-1}$ based on observing $N_\circ \,\big|\, \widehat{C} \sim \mathrm{Poiss}(\lambda_\circ)$ for $\lambda_\circ$ ranging over some open (non-empty) interval.

Proof. A conditionally unbiased estimator $\widetilde{\mu}(N_\circ)$ for $\lambda_\circ^{-1}$ would satisfy $\mathbb{E}[\widetilde{\mu}(N_\circ)|\widehat{C}] = \lambda_\circ^{-1}$ implying

$$\sum_{k=0}^{\infty} \frac{\lambda_\circ^k}{k!}\widetilde{\mu}(k)\mathrm{e}^{-\lambda_\circ} = \lambda_\circ^{-1} \Rightarrow \sum_{k=0}^{\infty} \frac{\lambda_\circ^{k+1}}{k!}\widetilde{\mu}(k) = \sum_{k=0}^{\infty} \frac{\lambda_\circ^k}{k!}\,. \tag{1.5.3}$$

The coefficient for the constant term in the left and right power series would thus differ (0 versus 1), in contradiction with the uniqueness theorem for power series. □

We provide an almost unbiased estimator for $\lambda_\circ^{-1}$ by noting that the first jump time of a time-indexed Poisson process with intensity ν is $\mathrm{Exp}(\nu)$-distributed and thus has expectation ν^{-1}. Taking conditional expectation of the first jump time with respect to the

value of the Poisson process at time 1, we conclude that

$$\widehat{\mu}(N_\circ, \lambda_\circ) \stackrel{\text{def}}{=} \begin{cases} (N_\circ + 1)^{-1}, & \text{for } N_\circ \geqslant 1, \\ 1 + \lambda_\circ^{-1}, & \text{for } N_\circ = 0 \end{cases}$$

satisfies $\mathbb{E}[\widehat{\mu}(N_\circ, \lambda_\circ)|\widehat{C}] = \lambda_\circ^{-1}$. Omitting the term $\lambda_\circ^{-1}$, depending on $\lambda_\circ$, in the unlikely case $N_\circ = 0$, we define our final estimator

$$\widehat{\vartheta} \stackrel{\text{def}}{=} |\widehat{C}| + \frac{N_\partial}{N_\circ + 1}|\widehat{C}|\,.$$

For the proofs, we also define the *pseudo*-estimator

$$\widehat{\vartheta}_{pseudo} \stackrel{\text{def}}{=} |\widehat{C}| + |\widehat{C}|N_\partial\Big(\frac{1}{N_\circ + 1} + \frac{\mathrm{e}^{-\lambda_\circ}}{\lambda_\circ}\Big)\,.$$

THEOREM 1.5.3. The pseudo-estimator $\widehat{\vartheta}_{pseudo}$ is unbiased and the estimator $\widehat{\vartheta}$ is asymptotically unbiased in the sense that with constants $c_1, c_2 > 0$ depending on d, $d > 1$, whenever $\lambda|C| \geqslant 1$:

$$0 \leqslant |C| - \mathbb{E}[\widehat{\vartheta}] \leqslant c_1|C| \exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big)\,, \quad \forall C \in \mathbf{C}\,. \tag{1.5.4}$$

Proof. We have

$$\mathbb{E}\Big[\frac{1}{N_\circ + 1} + \frac{\mathrm{e}^{-\lambda_\circ}}{\lambda_\circ}\,\Big|\,\widehat{C}\Big] = \mathrm{e}^{-\lambda_\circ}\lambda_\circ^{-1}\Big(\sum_{k=0}^{\infty}\frac{\lambda_\circ^{k+1}}{(k+1)k!} + 1\Big) = \lambda_\circ^{-1}\,, \tag{1.5.5}$$

which by $|\widehat{C}|\lambda_\circ^{-1} = \lambda^{-1}$ and $\mathbb{E}[\widehat{\vartheta}_{oracle}] = |C|$ implies unbiasedness of $\widehat{\vartheta}_{pseudo}$. Thus, it follows that

$$|C| - \mathbb{E}[\widehat{\vartheta}] = \mathbb{E}\big[|\widehat{C}|N_\partial \mathrm{e}^{-\lambda_\circ}\lambda_\circ^{-1}\big] = \lambda^{-1}\mathbb{E}\big[N_\partial \mathrm{e}^{-\lambda|\widehat{C}|}\big]\,.$$

We exploit the deviation inequality from Thm. 1 in [35] and derive the bound for the exponential moment of the missing volume in the model with fixed number of points

$$\mathbb{E}[\exp\left(\lambda|C \setminus \widehat{C}_k|\right)] \leqslant b_1 \exp\left(b_2\lambda|C|k^{-2/(d+1)}\right), \quad k \geqslant 2\,, \tag{1.5.6}$$

for positive constants b_1, b_2, depending on the dimension according to [35]. For the cases $k = 0, 1$, we have the identity $\mathbb{E}[\exp\left(\lambda|C \setminus \widehat{C}_k|\right)] = \exp\left(\lambda|C|\right)$. By Poissonisation, similarly to (1.4.10), we derive

$$\exp(-\lambda|C|)\mathbb{E}[\exp\left(\lambda|C \setminus \widehat{C}|\right)] \leqslant b_3 \exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big)\,, \tag{1.5.7}$$

for positive constants b_3, c_2, depending on the dimension. Hence, using the Cauchy-Schwarz inequality and the bound for the moments of the points on the convex hull,

$$\mathbb{E}[N_\partial^q] = O\big((\lambda|C|)^{q(d-1)/(d+1)}\big)\,, \quad q \in \mathbb{N}\,, \tag{1.5.8}$$

see e.g. Section 2.3.2 in [36], we derive for a constant $c_1 > 0$

$$\begin{aligned}
\lambda^{-1}\mathbb{E}\big[N_\partial \mathrm{e}^{-\lambda|\widehat{C}|}\big] &\leqslant \lambda^{-1}\mathrm{e}^{-\lambda|C|}\mathbb{E}[N_\partial^2]^{1/2}\mathbb{E}[\mathrm{e}^{2\lambda|C\setminus\widehat{C}|}]^{1/2} && (1.5.9)\\
&\leqslant c_1\lambda^{-2/(d+1)}|C|^{(d-1)/(d+1)}\exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big) && (1.5.10)\\
&\leqslant c_1|C|\exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big)\,. && (1.5.11)
\end{aligned}$$

□

The next step of the analysis is to compare the variance of the pseudo-estimator $\widehat{\vartheta}_{pseudo}$ with the variance of the oracle estimator $\widehat{\vartheta}_{oracle}$, which is UMVU.

THEOREM 1.5.4. The following oracle inequality holds with a universal constant $c > 0$ and dimension-dependent constants $c_1, c_2 > 0$ for all $C \in \mathbf{C}$ with $\lambda|C| \geqslant 1$:

$$\mathrm{Var}(\widehat{\vartheta}_{pseudo}) \leqslant (1 + c\alpha(\lambda, C))\,\mathrm{Var}(\widehat{\vartheta}_{oracle}) + r(\lambda, C)\,, \tag{1.5.12}$$

where

$$\begin{aligned}
\alpha(\lambda, C) &= \frac{1}{|C|}\Big(\frac{1}{\lambda} + \frac{\mathrm{Var}(|C\setminus\widehat{C}|)}{\mathbb{E}[|C\setminus\widehat{C}|]} + \mathbb{E}[|C\setminus\widehat{C}|]\Big)\,,\\
r(\lambda, C) &= c_1(\lambda|C|)^{2(d-1)/(d+1)}\exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big)\,.
\end{aligned}$$

Proof. By the law of total variance, we obtain

$$\begin{aligned}
\mathrm{Var}(\widehat{\vartheta}_{pseudo}) &= \mathrm{Var}\big(\mathbb{E}[\widehat{\vartheta}_{pseudo}|\widehat{C}]\big) + \mathbb{E}\big[\mathrm{Var}(\widehat{\vartheta}_{pseudo}|\widehat{C})\big] && (1.5.13)\\
&= \mathrm{Var}(\widehat{\vartheta}_{oracle}) + \mathbb{E}\Big[(N_\partial|\widehat{C}|)^2\,\mathrm{Var}\Big(\frac{1}{N_\circ+1}|\widehat{C}\Big)\Big]\,. && (1.5.14)
\end{aligned}$$

In view of $N_\circ\,|\,\widehat{C} \sim \mathrm{Poiss}(\lambda_\circ)$, a power series expansion gives

$$\mathbb{E}[(N_\circ+1)^{-2}|\,\widehat{C}] = \lambda_\circ^{-1}\mathrm{e}^{-\lambda_\circ}\int_0^{\lambda_\circ}(\mathrm{e}^t-1)/t\,dt\,.$$

The conditional variance can for $\lambda_\circ \to \infty$ thus be bounded by

$$\mathrm{Var}((1+N_\circ)^{-1}|\,\widehat{C}) \leqslant \lambda_\circ^{-1}\mathrm{e}^{-\lambda_\circ}\int_{\lambda_\circ/2}^{\lambda_\circ}\mathrm{e}^t/t\,dt - (\lambda_\circ)^{-2} + O(\mathrm{e}^{-\lambda_\circ/4})$$

$$= (\lambda_\circ)^{-1} \int_0^{\lambda_\circ/2} \mathrm{e}^{-s} \Big(\frac{1}{\lambda_\circ - s} - \frac{1}{\lambda_\circ} \Big)\, ds + O(\mathrm{e}^{-\lambda_\circ/4})$$
$$= \lambda_\circ^{-3}(1 + o(1))\,,$$

where we have used $(\lambda_\circ - s)^{-1} - \lambda_\circ^{-1} = s\lambda_\circ^{-1}(\lambda_\circ - s)^{-1}$, $\int_0^\infty s\mathrm{e}^{-s} ds = 1$ and dominated convergence. Thanks to $(N_\circ + 1)^{-1} \in [0,1]$ we conclude for some constant $c \geqslant 1$

$$\mathrm{Var}((1 + N_\circ)^{-1} |\, \widehat{C}) \leqslant c(1 \wedge \lambda_\circ^{-3}).$$

Consequently, we have

$$\mathrm{Var}(\widehat{\vartheta}_{pseudo}) \leqslant \mathrm{Var}(\widehat{\vartheta}_{oracle}) + \mathbb{E}\big[(N_\partial |\widehat{C}|)^2 c(1 \wedge (\lambda|\widehat{C}|)^{-3})\big] \tag{1.5.15}$$
$$= \mathrm{Var}(\widehat{\vartheta}_{oracle}) + c\mathbb{E}\big[(N_\partial |\widehat{C}|)^2 \wedge \lambda^{-3}(N_\partial)^2 |\widehat{C}|^{-1}\big]\,, \tag{1.5.16}$$

and with (1.4.7)

$$\frac{\mathrm{Var}(\widehat{\vartheta}_{pseudo})}{\mathrm{Var}(\widehat{\vartheta}_{oracle})} \leqslant 1 + c\frac{\mathbb{E}\big[(N_\partial \lambda|\widehat{C}|)^2 \wedge (N_\partial)^2(\lambda|\widehat{C}|)^{-1}\big]}{\lambda \mathbb{E}[|C \setminus \widehat{C}|]} \tag{1.5.17}$$
$$= 1 + c\frac{\mathbb{E}\big[(N_\partial)^2\big((\lambda|\widehat{C}|)^2 \wedge (\lambda|\widehat{C}|)^{-1}\big)\big]}{\mathbb{E}[N_\partial]}\,. \tag{1.5.18}$$

Define the 'good' event $\mathcal{G} = \{|\widehat{C}| > |C|/2\}$, on which $\big((\lambda|\widehat{C}|)^2 \wedge (\lambda|\widehat{C}|)^{-1}\big) \leqslant 2(\lambda|C|)^{-1}$. On the complement $\mathcal{G}^c$, we infer from $A^2 \wedge A^{-1} \leqslant 1$ for $A > 0$

$$\mathbb{E}\Big[(N_\partial)^2\big((\lambda|\widehat{C}|)^2 \wedge (\lambda|\widehat{C}|)^{-1}\big)\mathbf{1}_{\mathcal{G}^c}\Big] \leqslant \mathbb{E}\big[N_\partial^2 \mathbf{1}_{\mathcal{G}^c}\big] \tag{1.5.19}$$
$$\leqslant \mathbb{E}[N_\partial^4]^{1/2} \mathbb{P}(|C \setminus \widehat{C}| \geqslant |C|/2)^{1/2} \tag{1.5.20}$$
$$\leqslant c_1(\lambda|C|)^{2(d-1)/(d+1)} \exp\big(-c_2(\lambda|C|)^{(d-1)/(d+1)}\big)\,, \tag{1.5.21}$$

for some positive constant c_1 and c_2, using (1.5.7) and (1.5.8). It remains to estimate the upper bound (1.5.18) on $\mathcal{G}$

$$\frac{2c}{\lambda|C|}\frac{\mathbb{E}[N_\partial^2]}{\mathbb{E}[N_\partial]} = \frac{2c}{\lambda|C|}\Big(\frac{\mathrm{Var}(N_\partial)}{\mathbb{E}[N_\partial]} + \mathbb{E}[N_\partial]\Big)\,. \tag{1.5.22}$$

Using the identity (17) in [18] for the factorial moments for the number of vertices N_∂, we derive $\mathrm{Var}(N_\partial) \leqslant \lambda^2 \mathrm{Var}(|C \setminus \widehat{C}|) + \lambda\mathbb{E}[|C \setminus \widehat{C}|]$ in view of $\mathbb{E}[N_\partial] = \lambda\mathbb{E}[|C \setminus \widehat{C}|]$. Thus, (1.5.22) is bounded by

$$\frac{2c}{\lambda|C|}\frac{\mathbb{E}[N_\partial^2]}{\mathbb{E}[N_\partial]} \leqslant \frac{2c}{|C|}\Big(\frac{1}{\lambda} + \frac{\mathrm{Var}(|C \setminus \widehat{C}|)}{\mathbb{E}[|C \setminus \widehat{C}|]} + \mathbb{E}[|C \setminus \widehat{C}|]\Big)\,, \tag{1.5.23}$$

which yields the assertion. □

As a result, we obtain an *oracle inequality* for the estimator $\widehat{\vartheta}$.

THEOREM 1.5.5. It follows for the risk of the estimator $\widehat{\vartheta}$ for all $C \in \mathbf{C}$ whenever $\lambda|C| \geqslant 1$:

$$\mathbb{E}[(\widehat{\vartheta} - |C|)^2]^{1/2} \leqslant (1 + c\alpha(\lambda, C))\mathbb{E}[(\widehat{\vartheta}_{oracle} - |C|)^2]^{1/2} + r(\lambda, C)\,, \tag{1.5.24}$$

with constant $c > 0$ and $\alpha(\lambda, C), r(\lambda, C)$ from Theorem 1.5.4. For any $C \in \mathbf{C}$ and $\lambda > 0$ we have $\alpha(\lambda, C) \leqslant 1 + \frac{1}{\lambda|C|}$.

Proof. In view of $\lambda_\circ = \lambda|\widehat{C}|$, we have $\widehat{\vartheta} = \widehat{\vartheta}_{pseudo} - \lambda^{-1}N_\partial \mathrm{e}^{-\lambda|\widehat{C}|}$ and we derive as in (1.5.11) and (1.5.21) with some constants $c_1, c_2 > 0$

$$\mathbb{E}[(\widehat{\vartheta} - \widehat{\vartheta}_{pseudo})^2] \leqslant \lambda^{-2}\mathbb{E}[N_\partial^4]^{1/2}\mathbb{E}[\mathrm{e}^{-4\lambda|\widehat{C}|}]^{1/2} \leqslant c_1^2 \exp\big(-2c_2(\lambda|C|)^{(d-1)/(d+1)}\big).$$

To establish the oracle inequality, we apply the triangle inequality in L^2-norm together with Theorems 1.4.2 and 1.5.4.

The universal bound on $\alpha(\lambda, C)$ follows from the rough bound $\mathbb{E}[|C \setminus \widehat{C}|^2] \leqslant |C|\mathbb{E}[|C \setminus \widehat{C}|]$. □

Note that the remainder term $r(\lambda, C)$ is exponentially small in $\lambda|C|$. Therefore, an immediate implication of Theorem 1.5.5 is that asymptotically our estimator $\widehat{\vartheta}$ is minimax rate-optimal in all dimensions, where the lower bound is proved in the next section. Yet, even more is true: the oracle inequality is in all well-studied cases *exact* in the sense that $\alpha(\lambda, C) \to 0$ holds for $\lambda \to \infty$ such that the the UMVU risk of $\widehat{\vartheta}_{oracle}$ is attained asymptotically.

LEMMA 1.5.6. We have tighter bounds on $\alpha(\lambda, C)$ from Theorem 1.5.4 in the following cases:

1. for $d = 1, 2$ and $C \in \mathbf{C}$ arbitrary: $\alpha(\lambda, C) \lesssim (\lambda|C|)^{-2/(d+1)}$,
2. for $d \geqslant 2$, C with C^2-boundary of positive curvature: $\alpha(\lambda, C) \lesssim (\lambda|C|)^{-2/(d+1)}$,
3. for $d \geqslant 2$ and C a polytope: $\alpha(\lambda, C) \lesssim \lambda^{-1}(\log(\lambda|C|))^{d-1}$.

Proof. Let us restrict to $|C| = 1$, the case of general volume follows by rescaling. In view of the expectation upper bound (1.4.10), the main issue is to bound $\mathrm{Var}(|C \setminus \widehat{C}|)/\mathbb{E}[|C \setminus \widehat{C}|]$ uniformly. Case (1) follows from [113], where $\lambda\,\mathrm{Var}(|C \setminus \widehat{C}|) \sim \mathbb{E}[|C \setminus \widehat{C}|]$ is established.

For case (2) with smooth boundary, the upper bound for the variance, $\mathrm{Var}(|C \setminus \widehat{C}|) \lesssim \lambda^{-(d+3)/(d+1)}$, was obtained in [120], while the lower bound for the first moment, $\mathbb{E}[|C \setminus \widehat{C}|] \gtrsim \lambda^{-2/(d+1)}$, is due to [128].

For the case (3) of polytopes, the upper bound $\mathrm{Var}(|C \setminus \widehat{C}|) \lesssim \lambda^{-2}(\log\lambda)^{d-1}$ was obtained in [17], while the lower bound for the first moment, $\mathbb{E}[|C \setminus \widehat{C}|] \gtrsim \lambda^{-1}(\log\lambda)^{d-1}$,

was proved in [16]. The expectation upper bound from Remark 1.4.3 thus yields the result. □

We conjecture that $\lambda \operatorname{Var}(|C \setminus \widehat{C}|) \sim \mathbb{E}[|C \setminus \widehat{C}|]$ holds universally for all convex sets in arbitrary dimensions and thus that the oracle inequality is always exact. Proving such a universal bound is a challenging open problem in stochastic geometry, strongly connected to the discussion on universal variance asymptotics in terms of the floating body by [17].

1.6 Finite sample behaviour and dilated hull estimator

In this section, we demonstrate the performance of the main estimator $\widehat{\vartheta}$ numerically and compare it to other estimators including the naive estimator $|\widehat{C}|$, the naive oracle estimator N/λ, the UMVU oracle estimator $\widehat{\vartheta}_{oracle}$ and the plug-in MLE estimator $\widehat{\vartheta}_{plugin} = |\widehat{C}|(1 + N_\partial/N)$. The main competitor from the literature is a rate-optimal estimator proposed in [67]. In their construction, the whole sample is divided into three equal parts X, X' and X'' of sizes $N^\star$ (without loss of generality $N^\star \in \mathbb{N}$) and the estimator is given by

$$\widehat{\vartheta}_G = |\widehat{C}| + \frac{|\widehat{C}''|}{N^\star} \sum_{i=1}^{N^\star} \mathbf{1}(X_i' \notin \widehat{C}), \tag{1.6.1}$$

where $\widehat{C}''$ is the convex hull of the third sample X''. The data points are simulated for two convex sets: an ellipse and a polygon; see Figure 1.5.

The RMSE estimate normalised by the area of the true set is based on $M = 500$ Monte Carlo iterations in each case. The results of the simulations are depicted in Figure 1.6 where $n = \lambda|C|$ denotes the expected total number of points. The worst convergence rate of N/λ is clearly visible. More importantly, we see that the RMSE of $\widehat{\vartheta}$ approaches the oracle risk for larger n (i.e. λ) as the oracle inequality predicts. It is also conspicuous that in the studied cases the plug-in estimator $\widehat{\vartheta}_{plugin}$ and the estimator $\widehat{\vartheta}$ perform rather similarly. This is explained by the fact that the number of points N_∂ on the convex hull increases with a moderate speed in the two-dimensional case, $\mathbb{E}[N_\partial] = O(\lambda^{1/3})$, which results in a small difference between the multiplication factors N_∂/N and $N_\partial/(N_\circ + 1)$. The simulations in two dimensions were implemented using the R package "spatstat" by [10]. To illustrate the sub-optimality of the plug-in estimator $\widehat{\vartheta}_{plugin}$ in high dimensions, we provide results of numerical simulations in dimensions $d = 3, 4, 5, 6$ for the case when the true set C is a unit cube $C = [0, 1]^d$, see Figure 1.8. The simulations were implemented using the R package "geometry" by [72].

As an application of the obtained results, we propose a new estimator for the convex

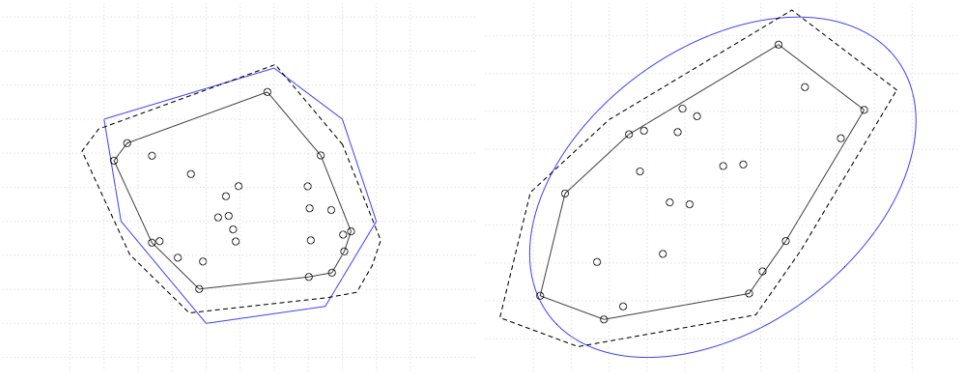

Figure 1.5: The two convex sets (blue), observations (points), their convex hulls (black lines) and dilated convex hulls (black dashed lines).

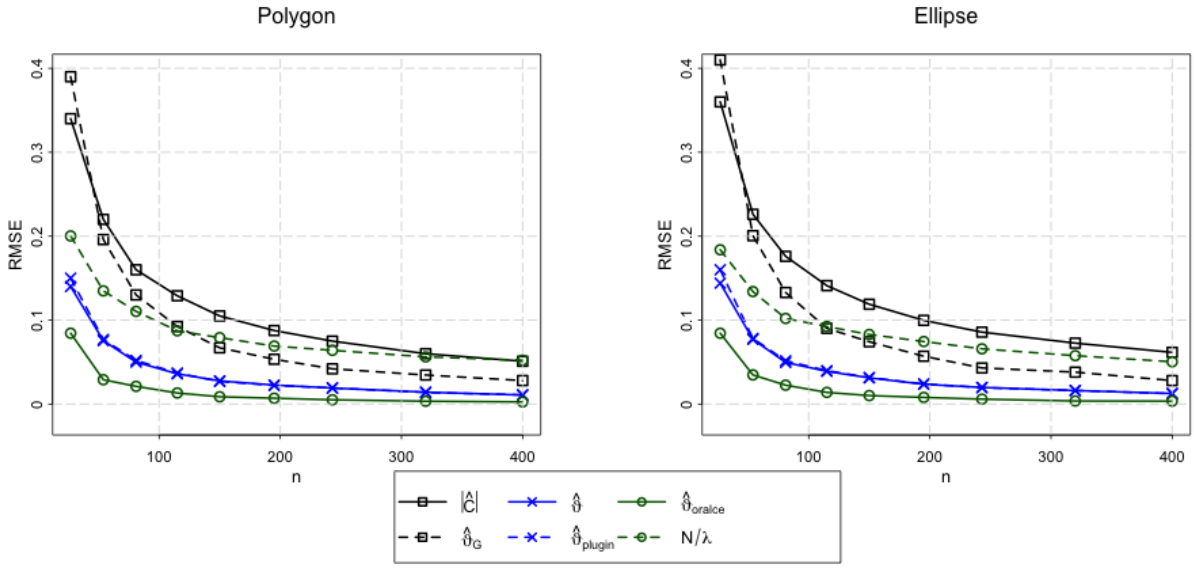

Figure 1.6: Monte Carlo RMSE estimates for the studied estimators for the volume of two convex sets: a polygon and an ellipse.

set itself:

$$\widetilde{C} \stackrel{\text{def}}{=} \Big\{ \widehat{x}_0 + \Big(\frac{\widehat{\vartheta}}{|\widehat{C}|} \Big)^{1/d} (x - \widehat{x}_0) \,\Big|\, x \in \widehat{C} \Big\} \tag{1.6.2}$$

$$= \Big\{ \widehat{x}_0 + \Big(\frac{N+1}{N_\circ + 1} \Big)^{1/d} (x - \widehat{x}_0) \,\Big|\, x \in \widehat{C} \Big\}, \tag{1.6.3}$$

which is just the dilation of the convex hull $\widehat{C}$ from its barycentre $\widehat{x}_0$, see the dashed polygons in Figure 1.5. Since the convex hull is a sufficient statistic (for known λ), the points in its interior do not bear any information on the shape of C itself such that the barycentre is a reasonable choice. There are, of course, other enlargements of the convex hull conceivable like $\operatorname{argmin}_{B \in \mathbf{C}, |B| = \widehat{\vartheta}} d_H(B, \widehat{C})$, the convex set closest (in Hausdorff distance) to $\widehat{C}$ with volume $\widehat{\vartheta}$. The intuition behind these estimators is based on the observation that once the volume of the true set is known, we can estimate the set itself faster (in the constant), and $\widehat{\vartheta}$ is a reasonable substitute for the true volume due to its fast rate of convergence.

A detailed analysis is not pursued here, but in a small simulation study we investigate the behaviour of the new dilated hull estimator for the above polygon. The error ratio $\mathbb{E}[|C \Delta \widehat{C}|] / \mathbb{E}[|C \Delta \widetilde{C}|]$ in terms of the symmetric difference $A \Delta B = (A \setminus B) \cup (B \setminus A)$ is approximated in $M = 500$ Monte Carlo iterations and shown in Figure 1.7. It turns out that the dilation significantly improves the convex hull as an estimator for C, especially for a small number of observations.

1.7 Appendix

1.7.1 Proof of Theorem 2.2.5

The proof is split into several statements, which might be of interest on their own.

LEMMA 1.7.1. The random variable $\mathcal{N}(\mathcal{K})$ is measurable with respect to $\mathcal{F}_{\mathcal{K}}$ for any stopping set $\mathcal{K}$.

Proof. The proof is just a generalisation of the analogous statement for time-indexed stochastic processes, see e.g. Proposition 2.18 in [81]. For this, the notions are extended to the partial order $\subseteq$ and then the right-continuity of $(\mathcal{N}(K), K \in \mathbf{K})$ (with respect to inclusion) implies its progressive measurability and thus in turn the measurability of $\mathcal{N}(\mathcal{K})$. □

Next, observe that the set-indexed process $(\mathcal{N}(K), K \in \mathbf{K})$ has independent increments, i.e. for $K_1, \ldots, K_m \in \mathbf{K}$ with $K_i \subseteq K_{i+1}$, $i = 1, \ldots, m-1$, the random variables $\mathcal{N}(K_{i+1}) - \mathcal{N}(K_i) = \mathcal{N}(K_{i+1} \setminus K_i)$ are independent (by the independence of the PPP

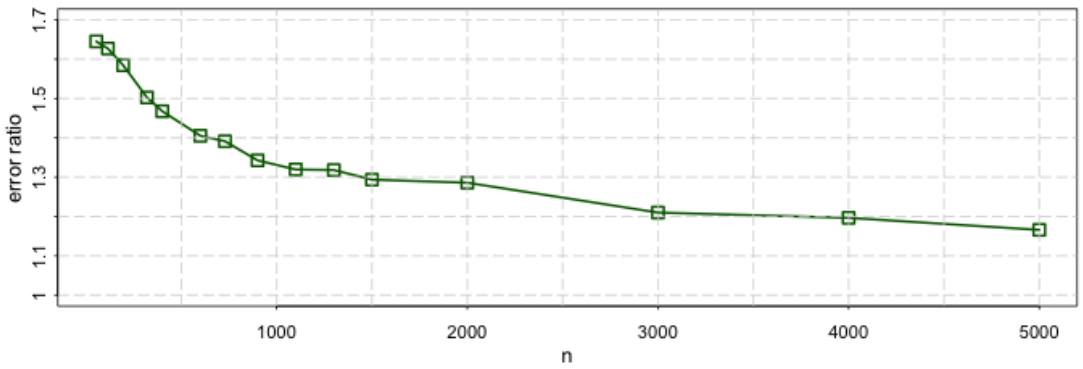

Figure 1.7: Monte Carlo error ratio for the convex hull and its dilation when the true set is a polygon.

on disjoint sets). In fact, we show in Proposition 1.7.2 that the process $\mathcal{N}$ is even a strong Markov process. In addition, Proposition 1.7.2 yields (2.2.7) using that the closed complement $\widehat{\mathcal{K}} = \overline{\widehat{C}^c}$ of the convex hull is a stopping set.

PROPOSITION 1.7.2. The set-indexed process $(\mathcal{N}(K), K \in \mathbf{K})$ is strong Markov at every stopping set $\mathcal{K}$. More precisely, conditionally on $\mathcal{F}_{\mathcal{K}}$ the process $(\mathcal{N}(K \setminus \mathcal{K}), K \in \mathbf{K})$ is a Poisson point process with intensity λ on $\mathcal{K}^c$. In particular, $\mathcal{N}(K \setminus \mathcal{K}) \,\big|\, \mathcal{F}_{\mathcal{K}} \sim \mathrm{Poiss}(\lambda|K \setminus \mathcal{K}|)$ holds for all $K \in \mathbf{K}$.

REMARK 1.7.3. The fact that the increments $\mathcal{N}(\mathcal{K} \cup K) - \mathcal{N}(\mathcal{K})$ are independent of $\mathcal{F}_{\mathcal{K}}$ can be derived from a general theorem about the strong Markov property for random fields in Thm. 4 in [127]. See also [149] for a discussion of the strong Markov property and its applications in stochastic geometry. These statements, however, do not provide a distributional characterisation of the increments of the process.

Proof. A set-indexed, $(\mathcal{F}_K)$-adapted integrable process $(X_K, K \in \mathbf{K})$ is called a martingale if $\mathbb{E}[X_B|\mathcal{F}_A] = X_A$ holds for any $A, B \in \mathbf{K}$ with $A \subseteq B$. By the independence of increments, the process $M_K \stackrel{\mathrm{def}}{=} \mathcal{N}(K) - \lambda|K|$, $K \in \mathbf{K}$, is clearly a martingale with respect to its natural filtration $(\mathcal{F}_K, K \in \mathbf{K})$. Then also the process

$$\widetilde{M}_K \stackrel{\mathrm{def}}{=} M_{K \cup \mathcal{K}} - M_{\mathcal{K}} = \mathcal{N}(K \setminus \mathcal{K}) - \lambda|K \setminus \mathcal{K}|$$

is a martingale with respect to the filtration $\widetilde{\mathcal{F}}_K \stackrel{\mathrm{def}}{=} \mathcal{F}_K \vee \mathcal{F}_{\mathcal{K}} = \mathcal{F}_{K \cup \mathcal{K}}$ because for $K_1, K_2 \in \mathbf{K}$ with $K_1 \subseteq K_2$ the optional sampling theorem (see e.g. [148]) yields

$$\mathbb{E}[\widetilde{M}_{K_2}|\widetilde{\mathcal{F}}_{K_1}] = \mathbb{E}[M_{K_2 \cup \mathcal{K}} - M_{\mathcal{K}}|\mathcal{F}_{K_1 \cup \mathcal{K}}] = M_{K_1 \cup \mathcal{K}} - M_{\mathcal{K}} = \widetilde{M}_{K_1},$$

noting that $K_1 \cup \mathcal{K}$ is again a stopping set.

This implies that $\lambda|K \setminus \mathcal{K}|$, conditionally on $\mathcal{K}$, is the deterministic compensator of the process $\widetilde{N}_K = \mathcal{N}(K \setminus \mathcal{K})$. Then, due to the martingale characterisation of the set-indexed Poisson process, see Thm. 3.1 in [78] (analogue of Watanabe's characterisation for the Poisson process), the process $\widetilde{N}_K$, conditionally on $\mathcal{F}_{\mathcal{K}}$, is a Poisson point process with mean measure $\widetilde{\mu}(A) = \lambda|A \cap \mathcal{K}^c|$. □

The last statement of Theorem 2.2.5, that $\mathcal{F}_{\widehat{\mathcal{K}}} = \sigma(\widehat{C})$ is shown next. It can be seen as a generalisation of the interesting fact that for a time-indexed Poisson process the sigma-algebra $\sigma(\tau)$ associated with the first jump time τ coincides with the sigma-algebra of τ-history $\mathcal{F}_\tau$.

LEMMA 1.7.4. The sigma-algebra $\sigma(\widehat{C})$ coincides with the sigma-algebra $\mathcal{F}_{\widehat{\mathcal{K}}}$ of $\widehat{\mathcal{K}}$-history, i.e. $\sigma(\widehat{C}) = \mathcal{F}_{\widehat{\mathcal{K}}}$.

Proof. Since $\widehat{\mathcal{K}}$ is $\mathcal{F}_{\widehat{\mathcal{K}}}$-measurable by Lemma 1 in [148] and $\widehat{C} = \overline{\widehat{\mathcal{K}}^c}$, it is evident that $\sigma(\widehat{C}) \subseteq \mathcal{F}_{\widehat{\mathcal{K}}}$. The other direction is more involved. We use that the sigma-algebra $\mathcal{F}_{\widehat{\mathcal{K}}}$ coincides with the sigma-algebra $\sigma(\{\mathcal{N}(\widehat{\mathcal{K}} \cap K), K \in \mathbf{K}\})$ generated by the process stopped at $\widehat{\mathcal{K}}$. This statement can be derived from Thm. 6, Ch. 1 in [130]. Note that their assumption (1.11) is satisfied in our case, because for all $K \in \mathbf{K}$ and $\omega \in \Omega$ there is ω' such that $\mathcal{N}(U \cap K, \omega) = \mathcal{N}(U, \omega')$ for all $U \in \mathbf{K}$, which simply says that observing points in $K \in \mathbf{K}$ there might be no points outside K. Finally, observe that by definition of the convex hull $\mathcal{N}(\overline{\widehat{C}^c} \cap K) = \mathcal{N}((\partial\widehat{C}) \cap K)$. Modulo null sets, $\mathcal{N}((\partial\widehat{C}) \cap K)$ counts the number of vertices of $\widehat{C}$ in K and is thus $\sigma(\widehat{C})$-measurable. □

Proof of Lemma 1.5.1. Using that the bias of the oracle estimator $\widehat{\vartheta} = |\widehat{C}| + N_\partial/(N_\circ + 1)|\widehat{C}|$ is exponentially small, it remains to compare its expectation with the expectation of the plug-in estimator $\widehat{\vartheta}_{plugin}$ to show (1.5.2):

$$\mathbb{E}[\widehat{\vartheta} - \widehat{\vartheta}_{plugin}] = \mathbb{E}\bigg[|\widehat{C}|\Big(\frac{N_\partial}{N_\circ + 1} - \frac{N_\partial}{N}\Big)\bigg] = \mathbb{E}\bigg[|\widehat{C}|\frac{N_\partial^2 - N_\partial}{(N_\circ + 1)(N_\circ + N_\partial)}\bigg] \tag{1.7.1}$$

$$\geqslant \frac{d}{d+1}\mathbb{E}\bigg[\frac{|\widehat{C}|N_\partial^2}{(N_\circ + 1)2\lambda|C|}\mathbf{1}(N \leqslant 2\lambda|C|)\bigg], \tag{1.7.2}$$

where in the last line we have used $|\widehat{C}| > 0$ only if $N_\partial \geqslant d+1$ and in this case $N_\partial^2 - N_\partial \geqslant \frac{d}{d+1}N_\partial^2$. Using $\mathbb{E}[(N_\circ + 1)^{-1} \,|\, \widehat{C}] = \lambda_\circ^{-1}(1 - \mathrm{e}^{-\lambda_\circ})$ from above, we obtain after writing $\mathbf{1}(N \leqslant 2\lambda|C|) = 1 - \mathbf{1}(N > 2\lambda|C|)$

$$\begin{aligned}\mathbb{E}[\widehat{\vartheta} - \widehat{\vartheta}_{plugin}] &\geqslant \frac{d}{d+1}\Big(\mathbb{E}\bigg[\frac{N_\partial^2|\widehat{C}|(1 - \mathrm{e}^{-\lambda_\circ})}{2\lambda_\circ\lambda|C|}\bigg] - \mathbb{E}\bigg[\frac{N_\partial^2|\widehat{C}|}{2\lambda|C|}\mathbf{1}(N > 2\lambda|C|)\bigg]\Big) \\ &\geqslant \frac{d}{d+1}\Big(\frac{\mathbb{E}[N_\partial^2(1 - \mathrm{e}^{-\lambda_\circ})]}{2\lambda^2|C|} - \frac{\mathbb{E}\big[N^2\mathbf{1}(N > 2\lambda|C|)\big]}{2\lambda}\Big).\end{aligned}$$

By Cauchy-Schwarz inequality and large deviations similarly to (1.5.11), the first term is bounded from below by a constant multiple of $\mathbb{E}[|C \setminus \widehat{C}|]^2/|C|$ in view of $\mathbb{E}[N_\partial^2] \geqslant \lambda^2 \mathbb{E}[|C \setminus \widehat{C}|]^2$, see e.g. Section 2.3.2 in [36]. Because of $N \sim \mathrm{Poiss}(\lambda|C|)$, the second term is of order $\lambda|C|^2 \mathrm{e}^{-\lambda|C|}$ and thus asymptotically of much smaller order. □

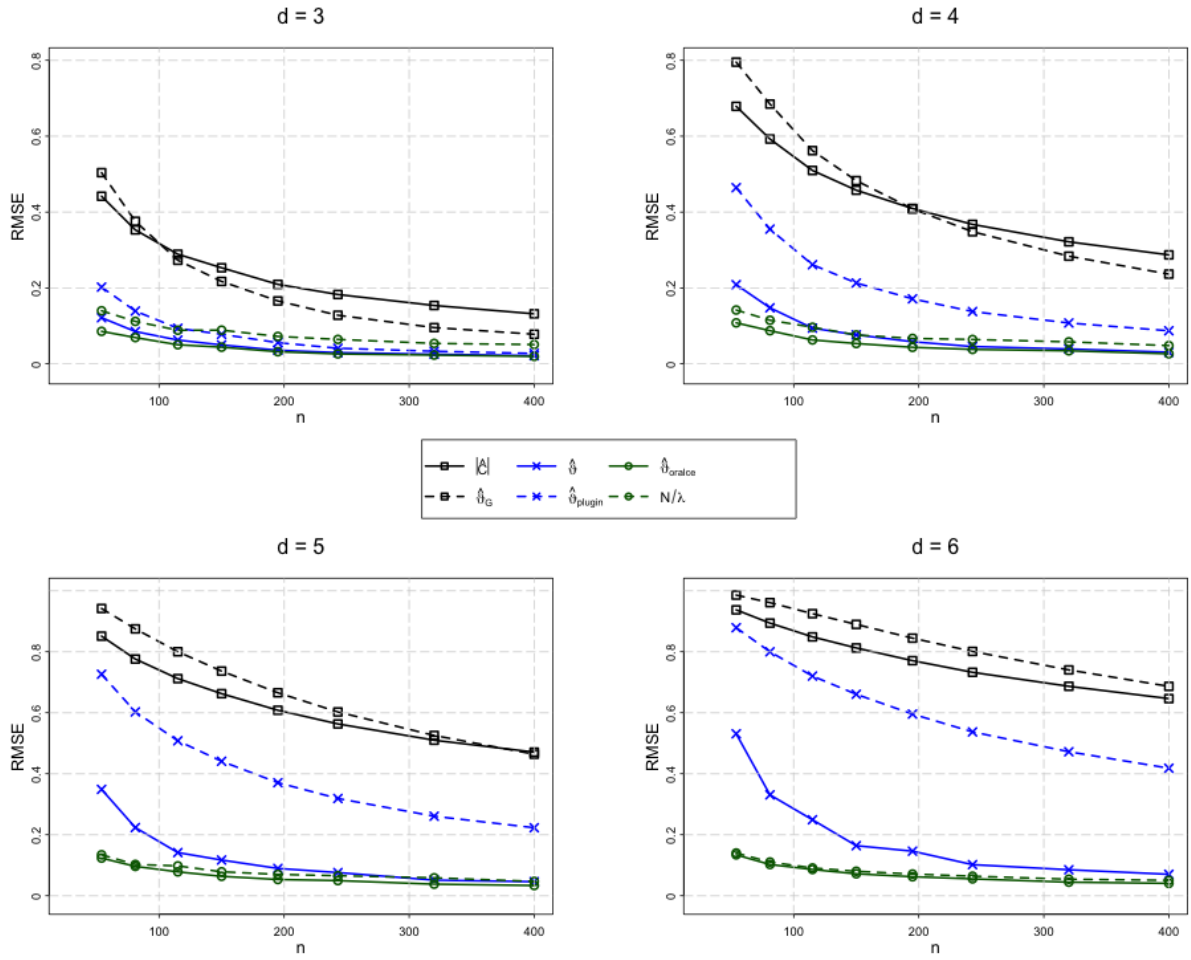

Figure 1.8: Monte Carlo RMSE estimates for the studied estimators for the volume of the unit cube $C = [0,1]^d$ in dimensions $d = 3, 4, 5, 6$.

Chapter 2

The wrapping hull and a unified framework for volume estimation

2.1 Main contribution and the structure

In this chapter, we combine techniques from statistics and stochastic geometry to build a general framework for estimating the volume of a set. We focus on the Poisson point process (PPP) observation model with intensity $\lambda > 0$ on a set A. We thus observe

$$X_1, ..., X_N \overset{i.i.d.}{\sim} U(A), \quad N \sim \text{Poiss}(\lambda|A|), \tag{2.1.1}$$

where $(X_n), N$ are independent and $|A|$ denotes the volume or the Lebesgue measure of the set A. The set A is meant to belong to a class $\mathbf{A}$ satisfying one simple assumption: *the class is assumed to be intersection stable*, see Section 2.2 for a concise definition. The classes of sets covered by the assumption include:

- convex sets;
- weakly-convex sets;
- star-shaped sets with a Hölder-continuous boundary;
- concentric sets;
- polytopes with fixed directions of outer unit normal vectors;
- compact sets.

In Section 2.2, we introduce the so-called *wrapping hull* $\widehat{A}$, which can informally be described as the minimal set from the class that contains the data points $X_1, ..., X_N$. It is then used in Section 2.3 to construct the so-called oracle estimator for the volume of a set belonging to one of the aforementioned classes when the intensity λ of the process

is known. The oracle estimator is shown to be uniformly of minimum variance among unbiased estimators (UMVU). Section 2.4 is devoted to the estimation of the intensity and derives a fully data-driven estimator of the volume,

$$\widehat{\vartheta} = \frac{N+1}{N_\circ + 1} |\widehat{A}| \,, \tag{2.1.2}$$

where $N_\circ$ is the number of data points lying in the interior of the wrapping hull. Figure 2.1 illustrates an example in which a naive estimator $|\widehat{A}|$ significantly underestimates the true volume $|A|$ even in the case when the class of sets is known whereas the estimator $\widehat{\vartheta}$ produces a rather striking performance, see Section 2.8 for a more detailed numerical study[1]. The mean squared risk of the estimator is shown to mimic the mean squared risk of the oracle estimator. Although the main object of analysis is the PPP model, the key results transfer to the so-called uniform model, cf. Section 2.4.1, using "Poissonisation". Section 2.5 further establishes the rates of convergence of the oracle estimator and the estimator $\widehat{\vartheta}$ in (2.1.2) for the considered classes of sets satisfying the assumption. Theorem 2.4.4 states a generalized Efron's inequality for the wrapping hull, cf. [60], which reduces the analysis of the mean squared error of the estimator $\widehat{\vartheta}$ to the distributional characteristics of the missing volume $|A \setminus \widehat{A}|$, a uniform lower bound on its expectation and a uniform deviation inequality. Interestingly, a uniform lower bound on the expectation of the missing volume has not even been established for the class of convex sets. We therefore establish the rates of convergence only for a relatively simple class of polytopes with fixed directions of outer unit normal vectors in Section 2.5.5. A more general question is beyond the scope of the present chapter and left to future research. In volume estimation of weakly-convex sets in Section 2.5.1 there is a further peculiar question of adaptation to a smoothing parameter. We suggest an adaptation procedure inspired by Lepski's method, cf. [93], and study it numerically in Section 2.8. Our numerical results in Section 2.8, mainly devoted to volume estimation for the weakly-convex sets, in particular, demonstrate that overestimating the smoothing parameter may have a significant cost for volume estimation. Some of the technical lemmata are deferred to the Appendix. Finally, we encounter and state a variety of new open questions in stochastic geometry, which we barely begin to nibble at the edges. Interestingly enough, the framework was mentioned in a seminal paper [84] by David Kendall in the Statslab at the University of Cambridge, but has never been fully explored.

[1]The simulations were implemented using the R package "spatstat" by [10].

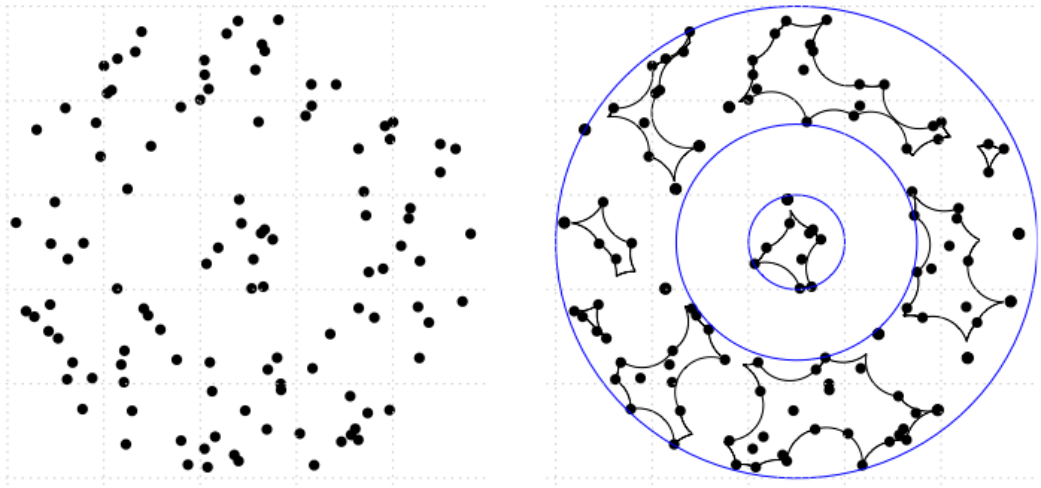

Figure 2.1: On the left: observations over an r-convex set A, the annulus $B(0.5, 0.5) \setminus B(0.5, 0.25)$ with the ball inside $B(0.5, 0.1)$. On the right: the r-convex hull $\widehat{A}$ (black) and the true set A (blue). The volume of the wrapping hull here is $|\widehat{A}| = 0.307$, the true volume is $|A| = 0.620$ and our estimator yields $\widehat{\vartheta} = 0.578$.

2.1.1 Relationship to the work on volume estimation of a convex set

Some of the theoretical results in the present chapter naturally follows the results for the convex case, cf. Section 2.3 and Theorem 2.4.2 in Section 2.4, and the corresponding results in the first chapter. In fact, a key observation igniting the development of the present framework is that estimation of the volume of convex sets can in fact solely rely upon the property of convex sets being stable under taking intersections rather than convexity itself. This observation appears to have a substantial value for volume estimation in a variety of scenarios far beyond convexity. Volume estimation for some of the classes covered by the framework, in particular, the weakly-convex sets, has been long seen as notoriously difficult with standard geometric arguments, see the references in Section 2.5.

Not violating the flow of the thesis, we shall therefore omit some of the proofs of the statements that are deduced from the proofs of the corresponding statements for the convex case. The proof of the result that the wrapping hull is a complete statistic in Theorem 2.3.3 is slightly simplified compared to the proof of the theorem that the convex hull is a complete statistic and hinges upon a measure-theoretic result in stochastic geometry. In contrast to the special case of convex sets, this chapter further argues that the designed estimator $\widehat{\vartheta}$ is in fact *adaptive* as its rate explicitly depends on the rate of convergence of the missing volume $|A \setminus \widehat{A}|$. This result rests upon Efron's inequality proved

in Section 2.4.2. Section 2.4.1 explicitly states that the same estimator is minimax optimal in the uniform model. Section 2.5 conveys the most noticeable value for applications as it provides efficient data-driven estimators and clearly outlines the steps of deriving explicit rates of convergence for specific classes of intersection stable sets.

2.2 The wrapping hull

In this section, we introduce the main notions and collect recently developed mathematical tools. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, fix a convex compact set $\mathbf{E}$ in $\mathbb{R}^d$ and equip it with the Borel σ-algebra $\mathcal{E}$ with respect to the Euclidean metric ρ. Without loss of generality one may assume that $\mathbf{E} = [0,1]^d$.

Let $\mathbf{K}$ be the set of all compact subsets of $\mathbf{E}$ equipped with its Borel σ-algebra $\mathfrak{B}_{\mathbf{K}}$ with respect to the Hausdorff-metric ρ_H defined for two non-empty compact sets A and B by

$$\rho_H(A,B) = \max\left(\sup_{x\in B}\rho(x,A), \sup_{x\in A}\rho(x,B)\right). \tag{2.2.1}$$

It is known, see Theorem C.5 in [105], that the Borel σ-algebra $\mathfrak{B}_{\mathbf{K}}$ coincides with the σ-algebra $\sigma([B]_{\mathbf{K}}, B \in \mathbf{K})$ with $[B]_{\mathbf{K}} = \{A \in \mathbf{K} : A \subseteq B\}$. Moreover, the space $(\mathbf{K}, \rho_H)$ is Polish.

Let $\mathbf{A} \subseteq \mathbf{K}$ be a family of compact subsets of $\mathbf{E}$ fulfilling the following assumption

Assumption 2.2.1. $\mathbf{A}$ is closed under taking arbitrary intersections and $\varnothing, \mathbf{E} \in \mathbf{A}$.

Then the metric subspace $(\mathbf{A}, \rho_H)$ has the induced Borel σ-algebra $\mathfrak{B}_{\mathbf{A}} = \mathbf{A} \cap \mathfrak{B}_{\mathbf{K}} = \{\mathbf{A} \cap K : K \in \mathfrak{B}_{\mathbf{K}}\}$, which thus coincides with the σ-algebra $\mathcal{A} = \sigma([B], B \in \mathbf{A})$ where $[B] = \{A \in \mathbf{A} : A \subseteq B\}$. It turns out there is an interesting connection between the families of sets satisfying Assumption 2.2.1 and the *trapping systems* introduced in the work of [84] on the theory of random sets, see also Section 7.2 in [105].

Recall that one can view $(\mathcal{N}(K), K \in \mathbf{K})$ as a set-indexed stochastic process. It has no direct order and its natural filtration is defined by

$$\mathcal{F}_K \stackrel{\text{def}}{=} \sigma(\{\mathcal{N}(U); U \subseteq K, U \in \mathbf{K}\}) \tag{2.2.2}$$

for any $K \in \mathbf{K}$. The properties of the filtration $(\mathcal{F}_K, K \in \mathbf{K})$ are well studied, cf [148]. By construction, the restriction $\mathcal{N}_K = \mathcal{N}(\cdot \cap K)$ of the point process $\mathcal{N}$ onto $K \in \mathbf{K}$ is $\mathcal{F}_K$-measurable (in fact, $\mathcal{F}_K = \sigma(\{\mathcal{N}_K(U); U \in \mathbf{K}\})$). Moreover, it can be easily seen that $\mathcal{N}_K$ is a Poisson point process in $\mathbf{M}$, cf. the Restriction Theorem in [85]. Furthermore, we say that a set-indexed, $(\mathcal{F}_K)$-adapted integrable process $(X_K, K \in \mathbf{K})$ is a martingale if $\mathbb{E}[X_B|\mathcal{F}_A] = X_A$ holds for any $A, B \in \mathbf{K}$ with $A \subseteq B$. By the independence of increments,

the process

$$M_K \stackrel{\text{def}}{=} \mathcal{N}(K) - \lambda|K|, \quad K \in \mathbf{K}\,, \tag{2.2.3}$$

is clearly a martingale with respect to its natural filtration $(\mathcal{F}_K, K \in \mathbf{K})$.

A random compact set $\mathcal{K}$ is a measurable mapping $\mathcal{K} : (\mathbf{M}, \mathcal{M}) \to (\mathbf{K}, \mathfrak{B}_{\mathbf{K}})$. Recall that a random compact set $\mathcal{K}$ is called an $\mathcal{F}_K$-stopping set if $\{\mathcal{K} \subseteq K\} \in \mathcal{F}_K$ for all $K \in \mathbf{K}$. The σ-algebra of $\mathcal{K}$-history is defined as $\mathcal{F}_{\mathcal{K}} = \{A \in \mathfrak{F} : A \cap \{\mathcal{K} \subseteq K\} \in \mathcal{F}_K \ \forall K \in \mathbf{K}\}$, where $\mathfrak{F} = \sigma(\mathcal{F}_K; K \in \mathbf{K})$. We introduce the *wrapping hull* of the PPP points on a set $A \in \mathbf{A}$, which is served as a set-estimator of A.

Definition 2.2.2. The $\mathbf{A}$-*wrapping hull* (or simply the wrapping hull) of the PPP points is a mapping $\widehat{A} : M \to \mathbf{A}$ defined as

$$\widehat{A} \stackrel{\text{def}}{=} \mathrm{wrap}_{\mathbf{A}}\{X_1, ..., X_N\} \stackrel{\text{def}}{=} \bigcap\{A \in \mathbf{A} : X_i \in A, \forall i = 1, ..., N\}\,. \tag{2.2.4}$$

For a set $A \subseteq \mathbf{E}$ let A^c denote its complement.

Lemma 2.2.3. The set $\widehat{\mathcal{K}} \stackrel{\text{def}}{=} \overline{\widehat{A}^c}$, the closure of the complement of the wrapping hull, is an $(\mathcal{F}_K)$-stopping set.

The proof of Lemma 2.2.3 essentially repeats the steps of the proof of an analogous statement for the convex hull of the PPP points given in Lemma 2.2 in [14]. The following corollary of this lemma rests upon the optional sampling theorem for set-indexed martingales, cf. [148].

Corollary 2.2.4. The number of points N_∂ lying on the wrapping hull $\widehat{A}$ and the missing volume $|A \setminus \widehat{A}|$ satisfy the relation

$$\mathbb{E}[N_\partial] = \lambda\, \mathbb{E}[|A \setminus \widehat{A}|]\,. \tag{2.2.5}$$

As in the case of convex sets, we define the *likelihood function* for the PPP model. Note that we can evaluate the probability

$$\mathbb{P}_A\big(\widehat{A} \in B\big) = \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda|A|}\lambda^k}{k!} \int_{A^k} \mathbf{1}(\mathrm{wrap}_{\mathbf{A}}\{x_1, ..., x_k\} \in B)\, \mathrm{d}(x_1, ..., x_k)$$

for $B \in \mathfrak{B}_{\mathbf{A}}$. Usually, we only write the subscript A or sometimes (A, λ) when different probability distributions are considered simultaneously. The likelihood function $L(A, \mathcal{N}) = \frac{\mathrm{d}\mathbb{P}_{A,\lambda}}{\mathrm{d}\mathbb{P}_{\mathbf{E},\lambda_0}}$ for $A \in \mathbf{A}$ and $\lambda, \lambda_0 > 0$ is then given by

$$L(A, \mathcal{N}) = \frac{\mathrm{d}\mathbb{P}_{A,\lambda}}{\mathrm{d}\mathbb{P}_{\mathbf{E},\lambda_0}}(\mathcal{N}) = \mathrm{e}^{\lambda_0|\mathbf{E}|-\lambda|A|}(\lambda/\lambda_0)^N \mathbf{1}(\widehat{A} \subseteq A)\,, \tag{2.2.6}$$

cf. Thm. 1.3 in [91]. For the last line, we have used that a point set is in A if and only if its wrapping hull $\widehat{A} = \mathrm{wrap}_{\mathbf{A}}\{X_1, ..., X_N\}$ is contained in A.

The following theorem is an essential ingredient for deriving our estimators. The statement of the theorem is quite intuitive and already used in [117]. Its proof is similar to the proof of Theorem 1.4.1 in Chapter 1.

THEOREM 2.2.5. Set $\widehat{\mathcal{K}} \stackrel{\mathrm{def}}{=} \overline{\widehat{A}^c}$. The number N_∂ of points on the boundary of the wrapping hull $\widehat{A}$ is measurable with respect to the σ-algebra of $\widehat{\mathcal{K}}$-history $\mathcal{F}_{\widehat{\mathcal{K}}}$. The number of points in the interior of the wrapping hull $N_\circ$ is, conditionally on $\mathcal{F}_{\widehat{\mathcal{K}}}$, Poisson-distributed:

$$N_\circ \,\big|\, \mathcal{F}_{\widehat{\mathcal{K}}} \sim \mathrm{Poiss}(\lambda_\circ) \text{ with } \lambda_\circ \stackrel{\mathrm{def}}{=} \lambda|\widehat{A}|. \tag{2.2.7}$$

In addition, we have $\mathcal{F}_{\widehat{\mathcal{K}}} = \sigma(\widehat{A})$, where the latter is the σ-algebra $\sigma(\{\widehat{A} \subseteq B, B \in \mathcal{A}\})$ completed with the null sets in $\mathfrak{F}$.

We shall further use the following short notation: $N = \mathcal{N}(\mathbf{E})$ denotes the total number of points, $N_\circ = \mathcal{N}(\widehat{A}^\circ)$ the number of points in the interior of the wrapping hull $\widehat{A}$ and $N_\partial = \mathcal{N}(\partial\widehat{A}) = \mathcal{N}(\partial\widehat{\mathcal{K}})$ the number of points on the boundary of the wrapping hull.

2.3 Oracle case: known intensity λ

In the case when λ is known one can just estimate the volume $|A|$ by N/λ, which is an unbiased estimator, whose mean squared risk is given by

$$\mathbb{E}[(N/\lambda - |A|)^2] = \mathrm{Var}(N/\lambda) = |A|/\lambda\,, \tag{2.3.1}$$

thus implying $\mathcal{O}(\lambda^{-1/2})$-rate of convergence. This rate can be improved. As we shall see in Theorem 2.3.3, the wrapping hull is a complete and sufficient statistic thus allowing one to construct the unique best unbiased estimator of the volume in virtue of the Lehmann-Scheffé theorem. In view of Theorem 2.2.5 we thus derive our oracle estimator

$$\widehat{\vartheta}_{\mathrm{oracle}} = \mathbb{E}\Big[\frac{N}{\lambda}|\mathcal{F}_{\widehat{\mathcal{K}}}\Big] = \mathbb{E}\Big[\frac{N_\circ + N_\partial}{\lambda}|\mathcal{F}_{\widehat{\mathcal{K}}}\Big] = \frac{N_\partial}{\lambda} + |\widehat{A}| \tag{2.3.2}$$

The following result is fundamental in characterising the rate of convergence of the risk of the oracle estimator.

THEOREM 2.3.1. For known intensity $\lambda > 0$, the oracle estimator $\widehat{\vartheta}_{\mathrm{oracle}}$ is unbiased and of minimal variance among all unbiased estimators (UMVU) in the PPP model with parameter class $\mathbf{A}$. It satisfies

$$\mathbb{E}\big[(\widehat{\vartheta}_{\mathrm{oracle}} - |A|)^2\big] = \mathrm{Var}(\widehat{\vartheta}_{\mathrm{oracle}}) = \frac{\mathbb{E}[|A \setminus \widehat{A}|]}{\lambda}\,. \tag{2.3.3}$$

REMARK 2.3.2. The theorem asserts that the rate of convergence of $\widehat{\vartheta}_{\mathrm{oracle}}$ is in fact faster than $\lambda^{-1/2}$ for all classes of sets $\mathbf{A}$ satisfying Assumption (2.2.1).

Proof. By the tower property of conditional expectation, the estimator $\widehat{\vartheta}_{\mathrm{oracle}}$ is unbiased, $\mathbb{E}[\widehat{\vartheta}_{\mathrm{oracle}}] = |A|$. Using law of total variance, we derive

$$\mathrm{Var}(\widehat{\vartheta}_{\mathrm{oracle}}) = \mathrm{Var}\left(\mathbb{E}\Big[\frac{N}{\lambda}|\mathcal{F}_{\widehat{\mathcal{K}}}\Big]\right) = \mathrm{Var}(\frac{N}{\lambda}) - \mathbb{E}\Big[\,\mathrm{Var}\left(\frac{N}{\lambda}|\mathcal{F}_{\widehat{\mathcal{K}}}\right)\Big] \tag{2.3.4}$$

$$= \frac{|A|}{\lambda} - \mathbb{E}\Big[\,\mathrm{Var}\left(\frac{N_\circ + N_\partial}{\lambda}|\mathcal{F}_{\widehat{\mathcal{K}}}\right)\Big] = \frac{\mathbb{E}[|A \setminus \widehat{A}|]}{\lambda}\,. \tag{2.3.5}$$

Theorem 2.3.3 below affirms that the wrapping hull $\widehat{A}$ is a complete and sufficient statistic such that by the Lehmann-Scheffé theorem, the estimator $\widehat{\vartheta}_{\mathrm{oracle}}$ has the UMVU property. □

THEOREM 2.3.3. For known intensity $\lambda > 0$, the wrapping hull is a complete and sufficient statistic.

The proof of Theorem 2.3.3 is deferred to the Appendix. As a result of Theorem 2.3.1, the performance of the estimator $\widehat{\vartheta}_{\mathrm{oracle}}$ of the volume is reduced to the analysis of the performance of the wrapping hull estimator of the set itself, which clearly depends on the geometric properties of classes of sets satisfying Assumption 2.2.1.

The minimax lower bounds on the rate of convergence of the risk of estimating the volume of a set $A \in \mathbf{A}$ are often easier to establish for concrete classes of sets using the so-called hypercube argument, cf. [67]. Interestingly, the following general bound on the minimax optimal rate holds.

THEOREM 2.3.4. The minimax optimal rate of estimating the volume of a set $A \in \mathbf{A}$, where $\mathbf{A}$ follows Assumption (2.2.1) and is not finite, satisfies

$$\lambda^{-2} \lesssim \inf_{\widehat{\vartheta}_\lambda} \sup_{A \in \mathbf{A}} \mathbb{E}\big[(\widehat{\vartheta}_\lambda - |A|)^2\big] \lesssim \lambda^{-1}\,, \tag{2.3.6}$$

where the infimum extends over all estimators $\widehat{\vartheta}_\lambda$ in the Poisson point process model with intensity λ.

REMARK 2.3.5. The rate $\mathcal{O}(\lambda^{-1})$ is minimax for estimating the volume in some parametric classes of sets, in particular, the class of concentric sets, whereas the rate $\mathcal{O}(\lambda^{-1/2})$ is established for estimating the volume in the class of compact sets, see Section 2.5.

Proof. The upper bound in (2.3.6) follows directly from Theorem 2.3.1. The lower bound is obtained by reducing the minimax risk to the Bayes risk and then lower-bounding the Bayes risk at its minimum. These steps are fairly standard, cf. [89], and we hence omit them here. □

2.4 Data-driven estimator of the volume

The main ingredient to deriving the estimator of λ is the fact that the closure of the complement of the $\mathbf{A}$-wrapping hull $\widehat{\mathcal{K}} \stackrel{\text{def}}{=} \overline{\widehat{A}^c}$ is in fact an $(\mathcal{F}_K)$-stopping set according to Lemma 2.2.3. Moreover in analogy with a time-indexed Poisson process, our problem boils down to the estimation of the intensity of a time-indexed Poisson process starting from an unknown origin. To see this, recall that according to Theorem 2.2.5, the number of points $N_\circ$ lying inside the wrapping hull $\widehat{A}$ is Poisson-distributed with intensity $\lambda_\circ \stackrel{\text{def}}{=} \lambda|\widehat{A}|$ provided that $|\widehat{A}| > 0$:

$$N_\circ \,\big|\, \mathcal{F}_{\widehat{\mathcal{K}}} \sim \text{Poiss}(\lambda_\circ). \tag{2.4.1}$$

We aim to find an estimator for $\lambda_\circ^{-1}$. On the event $\{|\widehat{A}| > 0\}$, we follow the idea developed in Chapter 1. That is to say, we use that the first jump time τ of a time-indexed Poisson process $(Y_t, t > 0)$ with intensity $\nu > 0$ is $\text{Exp}(\nu)$-distributed and hence $\mathbb{E}[\tau] = \nu^{-1}$. Using the memoryless property of the exponential distribution, we then have

$$\mathbb{E}[\tau | Y_1 = m] = \frac{1}{m+1}\mathbf{1}(m \geqslant 1) + (1 + \nu^{-1})\mathbf{1}(m = 0)\,. \tag{2.4.2}$$

Therefore, we conclude that

$$\widehat{\mu}(N_\circ, \lambda_\circ) \stackrel{\text{def}}{=} \begin{cases} (N_\circ + 1)^{-1}, & \text{for } N_\circ \geqslant 1, \\ 1 + \lambda_\circ^{-1}, & \text{for } N_\circ = 0 \end{cases}$$

satisfies $\mathbb{E}[\widehat{\mu}(N_\circ, \lambda_\circ)|\mathcal{F}_{\widehat{\mathcal{K}}}] = \lambda_\circ^{-1}$ provided that $|\widehat{A}| > 0$. Omitting the term depending on $\lambda_\circ$ in the unlikely event $N_\circ = 0$, we derive our final estimator:

$$\widehat{\vartheta} = |\widehat{A}| + \frac{N_\partial}{N_\circ + 1}|\widehat{A}| = \frac{N+1}{N_\circ + 1}|\widehat{A}|\,. \tag{2.4.3}$$

Remark 2.4.1. As it follows from Definition 2.2.2, a wrapping hull $\widehat{A}$ may consist of disjoint sets, in which case the number of points $N_{\circ,k}$ lying inside a piece $\widehat{A}_k$ satisfies $N_{\circ,k} \,\big|\, \mathcal{F}_{\widehat{\mathcal{K}}} \sim \text{Poiss}(\lambda|\widehat{A}_k|)$ due to the homogeneity of the Poisson point process. This fact can further be used to estimate λ^{-1} locally. However, in the homogeneous case, we prefer to use the total number of points to estimate the intensity.

Note that a more explicit bound can be derived using the Cauchy-Schwarz inequality given a bound on the expected number of points N_∂ lying on the wrapping hull $\widehat{A}$ and a bound on the moments of the missing volume $|A \setminus \widehat{A}|$. This clearly depends on a considered class of sets satisfying Assumption 2.2.1. The following rather general oracle inequality holds for the mean squared error of the estimator $\widehat{\vartheta}$. Its proof can be adapted from the

proofs of Thm. 1.5.3 and Thm. 1.5.4 in Chapter 1 and we hence omit it here.

THEOREM 2.4.2. The following oracle inequality for the risk of the estimator $\widehat{\vartheta}$ holds for all $A \in \mathbf{A}$ whenever $\lambda|A| \geqslant 1$:

$$\mathbb{E}[(\widehat{\vartheta} - |A|)^2]^{1/2} \leqslant (1 + c\alpha(\lambda, A)) \operatorname{Var}(\widehat{\vartheta}_{oracle})^{1/2} + r(\lambda, A)\,, \tag{2.4.4}$$

where

$$\alpha(\lambda, A) := \frac{1}{|A|}\Big(\frac{\operatorname{Var}(|A \setminus \widehat{A}|)}{\mathbb{E}[|A \setminus \widehat{A}|]} + \mathbb{E}[|A \setminus \widehat{A}|]\Big)\,,$$
$$r(\lambda, A) := c_1 \lambda^{-1} \mathbb{E}[N_\partial^4]^{1/4} \mathbb{P}(|A \setminus \widehat{A}| \geqslant |A|/2)^{1/4}\,,$$

with some numeric constants $c, c_1 > 0$. In particular, $\alpha(\lambda, A)$ is bounded by some universal constant.

2.4.1 Volume estimation in the uniform model

In the PPP model, the data we observe are uniformly distributed points over a set in some given class and the number of points is a realisation of a Poisson random variable. The uniform model,

$$X_1, ..., X_n \overset{i.i.d.}{\sim} U(A), \quad A \in \mathbf{A}\,, \tag{2.4.5}$$

is closely related to the PPP model and assumes that the number of points n is fixed. In stochastic geometry, the objects studied in the PPP model typically exhibit a similar asymptotic behaviour in the uniform model and vice versa, see e.g. [113] and references therein for a study of the functionals of the convex hull. This section examines which results of the present thesis derived in the PPP model remain true in the uniform model.

It is relatively straightforward to show that the wrapping hull remains a sufficient and complete statistic in the uniform model with slightly adjusted arguments of the proof of Theorem 2.3.3. It is unknown however whether there exists an UMVU estimator in the uniform model. Nevertheless, an estimator

$$\widehat{\vartheta}_{\mathrm{unif},n} := \frac{n+1}{n_\circ + 1}|\widehat{A}|\,, \tag{2.4.6}$$

where $n_\circ$ is the number of points lying inside the wrapping hull, inherits the same rate of convergence as the final estimator $\widehat{\vartheta}$ in (2.4.3) in the PPP model due to the following result

PROPOSITION 2.4.3 (Poissonisation). Let $n = \lfloor \lambda|A| \rfloor > 0$ with A being any set in a class

$\mathbf{A}$. Then letting $\lambda \to \infty$ the following asymptotic equivalence result holds true

$$\mathbb{E}\big[(\widehat{\vartheta}_{\mathrm{unif},n} - |A|)^2\big] \sim \mathbb{E}\big[(\widehat{\vartheta} - |A|)^2\big]\,, \quad \forall A \in \mathbf{A}\,. \tag{2.4.7}$$

Furthermore, the minimax risks satisfy

$$\inf_{\widehat{\vartheta}_n} \sup_{A\in\mathbf{A}} \mathbb{E}\big[(\widehat{\vartheta}_n - |A|)^2\big] \sim \inf_{\widehat{\vartheta}_\lambda} \sup_{A\in\mathbf{A}} \mathbb{E}\big[(\widehat{\vartheta}_\lambda - |A|)^2\big]\,, \tag{2.4.8}$$

where the infimum on the left-hand side extends over all estimators in the uniform model, whereas the infinum on the right-hand side extends over all estimators in the Poisson point process model.

Proof. We only prove (2.4.8) here. (2.4.7) can then be proved exploiting similar arguments. Let us first show the inequality " $\gtrsim$ ". Assume it does not hold and that there exists an estimator $\widehat{\vartheta}'_n$ in the uniform model with the rate of convergence faster than the minimax optimal rate in the PPP model. Then for the estimator $\widehat{\vartheta}'_N$ we have for any $A \in \mathbf{A}$

$$\mathbb{E}\big[(\widehat{\vartheta}'_N - |A|)^2\big] = \sum_{k=1}^{\infty} \mathbb{E}\big((\widehat{\vartheta}'_N - |A|)^2 \,\big|\, N = k\big)\, \mathbb{P}(N = k) \tag{2.4.9}$$

$$\tag{2.4.10}$$

$$\leqslant \sum_{k=\lfloor \lambda|A|/2\rfloor}^{\lfloor 2\lambda|A|\rfloor} \mathbb{E}\big((\widehat{\vartheta}'_N - |A|)^2 \,\big|\, N = k\big)\, \mathbb{P}(N = k) + c_2 \exp(-c_3 n) \tag{2.4.11}$$

$$\leqslant c_1 \mathbb{E}\big[(\widehat{\vartheta}'_n - |A|)^2\big] + c_2 \exp(-c_3 n)\,, \tag{2.4.12}$$

for some constants $c_1, c_2, c_3 > 0$ using Bennett's inequality, a contradiction in view of Theorem 2.3.4. The other direction follows using the same technique. □

2.4.2 Efron's inequality for the wrapping hull

In this section, we show that the rate of convergence of the risk for the estimator $\widehat{\vartheta}$ in Theorem 2.4.2 hinges in fact upon only a deviation of the missing volume $|A \setminus \widehat{A}|$. More than 50 years ago Efron showed in [60] that the moments of the number of the points $N_{\partial,k}$ lying on the boundary of a convex hull $\widehat{C}_k$ in the uniform model $X_1, ..., X_k \overset{i.i.d.}{\sim} U(C)$, with $C \subseteq \mathbb{R}^d$ being a convex set, satisfies the identity

$$\mathbb{E}[N^q_{\partial,k}] = \sum_{r=1}^{q} n(k,q,r)\mathbb{E}[|\widehat{C}_{k-r}|^r]\,, \tag{2.4.13}$$

where $n(k,q,r)$ is the number of q-tuples from $1, ..., k$ having exactly r different values, $n(k,q,r) = \binom{k}{r}\sum_{m=1}^{r}(-1)^{r-m}\binom{r}{m}m^q$. This yields a dimension-free asymptotic equivalence

result,

$$\mathbb{E}[N^q_{\partial,k}] \sim k^q \mathbb{E}[|C \setminus \widehat{C}_k|^q]\,. \tag{2.4.14}$$

We here extend a one-sided version of this results to the wrapping hull.

PROPOSITION 2.4.4 (Efron's inequality for the wrapping hull). Let $\mathbf{A}$ be any class satisfying Assumption 2.2.1 and $\widehat{A}$ be the corresponding wrapping hull of the PPP points of intensity $\lambda > 0$ over a set $A \in \mathbf{A}$. Then the following asymptotic inequality holds

$$\mathbb{E}[N^q_\partial] \lesssim \lambda^q \mathbb{E}[|A \setminus \widehat{A}|^q]\,, \tag{2.4.15}$$

provided that the probability of observing q points lying on the boundary of the wrapping hull $\widehat{A}$ is non-zero.

REMARK 2.4.5. It follows by Jensen's inequality and Corollary 2.2.4, that $\mathbb{E}[N^q_\partial] \geqslant \mathbb{E}[N_\partial]^q = \lambda^q \mathbb{E}[|A \setminus \widehat{A}|]^q$. For some examples, like the class of convex sets, this in fact implies $\mathbb{E}[N^q_\partial] \sim \lambda^q \mathbb{E}[|A \setminus \widehat{A}|^q]$.

REMARK 2.4.6. Identities that relate the functionals of the convex hull of the points distributed uniformly over a convex set are thoroughly studied in stochastic geometry, see [18, 38, 113].

Proof. Let us first consider the uniform model and then transfer the result to the PPP model using Poissonisation. We follow Efron's idea, see also [18, 36], that

$$\begin{aligned}
\mathbb{E}[|A \setminus \widehat{A}_k|^q] &= |A|^q\, \mathbb{P}(X_{k+1} \notin \widehat{A}_k, ..., X_{k+q} \notin \widehat{A}_k) && (2.4.16)\\
&\geqslant |A|^q\, \mathbb{P}(X_{k+1} \in \partial\widehat{A}_{k+q}, ..., X_{k+q} \in \partial\widehat{A}_{k+q}) && (2.4.17)\\
&= \frac{|A|^q}{\binom{k+q}{q}}\, \mathbb{E} \sum \mathbf{1}(X_{i_1} \in \partial\widehat{A}_{k+q}, ..., X_{i_q} \in \partial\widehat{A}_{k+q}) && (2.4.18)\\
&= \frac{|A|^q}{\binom{k+q}{q}}\, \mathbb{E}\binom{N_{\partial,k+q}}{q}\,, && (2.4.19)
\end{aligned}$$

the sum being taken over all tuples $(i_1, ..., i_q)$ from the integers $1, ..., k+q$. Rearranging the terms, this entails $\mathbb{E}[N^q_{\partial,k}] \lesssim k^q \mathbb{E}[|A \setminus \widehat{A}_k|^q]$.

Using Poissonisation, we further derive for the PPP model,

$$\begin{aligned}
\mathbb{E}[|A \setminus \widehat{A}|^q] &= \sum_{k=1}^{\infty} \mathbb{E}(|A \setminus \widehat{A}_N|^q \,\big|\, N = k)\, \mathbb{P}(N = k) && (2.4.20)\\
&\gtrsim \sum_{k=1}^{\infty} (2\lambda|A|)^{-q} \mathbb{E}\big[N^q_{\partial,k}\big]\, \mathbb{P}(N = k) + \sum_{k=\lfloor 2\lambda|A|\rfloor}^{\infty} (k^{-q} - (2\lambda|A|)^{-q}) \mathbb{E}\big[N^q_{\partial,k}\big]\, \mathbb{P}(N = k) && (2.4.21)
\end{aligned}$$

$$= (2\lambda|A|)^{-q}\mathbb{E}[N_\partial^q] + \sum_{k=\lfloor 2\lambda|A|\rfloor}^{\infty} (k^{-q} - (2\lambda|A|)^{-q})\mathbb{E}\big[N_{\partial,k}^q\big]\,\mathbb{P}(N=k) \tag{2.4.22}$$

with the absolute value of second sum being bounded using the Cauchy-Schwarz inequality and large deviations by

$$c_1\,\mathbb{E}[N_\partial^{2q}]^{1/2}\mathbb{P}(N \geqslant [2\lambda|A|])^{1/2} \leqslant c_1\,\mathbb{E}[N_\partial^{2q}]^{1/2}\exp(-c_2 n)\,, \tag{2.4.23}$$

for some constants $c_1, c_2 > 0$. Thus, (2.4.14) follows. □

Proposition 2.4.4 and Theorem 2.4.2 immediately suggest the following bound for the remainder term in the oracle inequality,

$$r(\lambda, A) \leqslant c\mathbb{E}[|A \setminus \widehat{A}|^4]^{1/4}\,\mathbb{P}(|A \setminus \widehat{A}| \geqslant |A|/2)^{1/4}\,, \tag{2.4.24}$$

for some numeric constant $c > 0$. Therefore, the oracle inequality in Theorem 2.4.2 hinges upon only two variables:

- the ratio $\mathrm{Var}(|A \setminus \widehat{A}|)/\mathbb{E}[|A \setminus \widehat{A}|]$ of the moments of the missing volume,
- a uniform deviation inequality for the missing volume.

Both results are fairly involved and we shall only discuss here how to derive them for some simple classes of sets satisfying Assumption 2.2.1.

2.5 Classes of sets satisfying Assumption 2.2.1

This section collects some examples of classes of sets that satisfy Assumption 2.2.1. Note that the class of all convex sets $\mathbf{C}_{\mathrm{conv}}$ satisfies the assumption and was extensively studied in [14]. The most involved statements in the inference on convex sets were underpinned by the abundance of results from stochastic geometry on moment bounds and deviation inequalities for the missing volume, see Lemma 4.6 in [14]. In particular, the ratio $\mathrm{Var}(|C \setminus \widehat{C}|)/\mathbb{E}[|C \setminus \widehat{C}|] \sim 1/\lambda$ is established in [113] for all convex sets C in dimensions $d = 1, 2$. In dimensions $d > 2$, one can bound the ratio only for some subsets of the class of convex sets. Thus, for a convex set C with C^2-boundary of positive curvature, it is known thanks to [120] that $\mathrm{Var}(|C \setminus \widehat{C}|) \lesssim \lambda^{-(d+3)/(d+1)}$. The lower bound for the first moment, $\mathbb{E}[|C \setminus \widehat{C}|] \gtrsim \lambda^{-2/(d+1)}$, was shown in [128]. For a polytope C, the upper bound $\mathrm{Var}(|C \setminus \widehat{C}|) \lesssim \lambda^{-2}(\log\lambda)^{d-1}$ was obtained in [17], while the lower bound for the first moment, $\mathbb{E}[|C \setminus \widehat{C}|] \gtrsim \lambda^{-1}(\log\lambda)^{d-1}$, was proved in [16]. A uniform deviation inequality for convex sets obtained in Thm. 1 in [35] allows to derive sharp upper bounds on the moments of the missing volume. The proof of the deviation inequality exploited a

bound on the entropy of convex sets. It remains an intriguing open question in stochastic geometry whether $\lambda \operatorname{Var}(|C \setminus \widehat{C}|) \sim \mathbb{E}[|C \setminus \widehat{C}|]$ holds universally for all convex sets in arbitrary dimensions. Some of the classes of sets we consider here are much larger, and very little has been known about them in the mathematical literature.

2.5.1 r-convex sets

We denote by $B(x,r) \subseteq \mathbb{R}^d$ (resp. $B_\circ(x,r)$) the closed (resp. open) ball with centre x and radius r.

Definition 2.5.1. A compact set C_r in $\mathbf{E} \subseteq \mathbb{R}^d$ is called *r-convex* (or *weakly-convex*) for $r > 0$, if its complement is the union of all open Euclidean balls of diameter r that are disjoint to C_r, i.e. if

$$C_r = \bigcap_{B_\circ^c(x,r) \cap C_r = \varnothing} B_\circ^c(x,r)\,. \tag{2.5.1}$$

We denote the class of r-convex sets by $\mathbf{C}_r$.

Note that an r-convex set fulfills the *outside rolling ball condition*, i.e. for all $y \in \partial C_r$ there is a closed ball $B(x,r)$ such that $y \in \partial B(x,r)$ and $B_\circ(x,r) \cap C_r = \varnothing$. Heuristically this means that one can "roll" a ball of radius r freely over the boundary of a set. Note that according to the definition, r-convex sets can have "holes" and do not need to be connected, see Figure 2.4 for some examples. In the terminology of [84], $C_r \in \mathbf{C}_r$ means that the set C_r is *trapped* by the balls of radius r. The r-convex sets were introduced in [116] and presumably independently in [59]; see [51, 139] and references therein for a recent work on estimation of r-convex sets. In the literature, much more attention has been devoted to the sets satisfying the so-called *inside and outside rolling ball condition*, when both C_r and $\overline{C_r^c}$ are r-convex, see [98, 138]. The reason probably is that sets with smooth boundaries (with no angles) are sometimes easier to handle with geometric arguments, see [114].

The $\mathbf{C}_r$-wrapping hull is defined by

$$\widehat{C}_r := \bigcap_{B_\circ^c(x,r) \cap \{X_1,\dots,X_N\} = \varnothing} B_\circ^c(x,r) \tag{2.5.2}$$

and often called the *r-convex hull* in the literature. Thus the oracle estimator in (2.3.2) has the following form

$$\widehat{\vartheta}_{r,oracle} := \frac{N_\partial}{\lambda} + |\widehat{C}_r|\,, \tag{2.5.3}$$

where N_∂ is the number of sample points lying on the the r-convex hull $\widehat{C}_r$. In order to investigate the performance of this estimator according to Theorem 2.3.1 it suffices to study $\sup_{C_r \in \mathbf{C}_r} \mathbb{E}_{C_r}[|C_r \setminus \widehat{C}_r|]$. In fact the following result holds and it is a consequence of Theorem 2.3.1.

THEOREM 2.5.2. For known intensity $\lambda > 0$, the worst case mean squared error of the oracle estimator $\widehat{\vartheta}_{r,oracle}$ over the parameter class $\mathbf{C}_r$ decays as $\lambda \uparrow \infty$ like $\sup_{C_r \in \mathbf{C}_r} \mathbb{E}_{C_r}[|C_r \setminus \widehat{C}_r|]/\lambda$ in dimension d:

$$\limsup_{\lambda\to\infty} \lambda \sup_{C_r\in\mathbf{C}_r, |C_r|>0} \Big\{ \mathbb{E}\big[(\widehat{\vartheta}_{r,oracle} - |C_r|)^2\big] / \mathbb{E}\big[|C_r \setminus \widehat{C}_r|\big] \Big\} < \infty \,. \tag{2.5.4}$$

REMARK 2.5.3. Note that the class of convex sets $\mathbf{C}_{\mathrm{conv}}$ belongs to $\mathbf{C}_r$ for all $r > 0$ and thus using Theorem 3.4 in [14] we have a lower bound on the rate of convergence,

$$\inf_{\widehat{\vartheta}_\lambda} \lambda^{(d+3)/(d+1)} \sup_{C_r\in\mathbf{C}_r} \mathbb{E}_{C_r}[(|C_r| - \widehat{\vartheta}_\lambda)^2] \tag{2.5.5}$$

$$\geqslant \inf_{\widehat{\vartheta}_\lambda} \lambda^{(d+3)/(d+1)} \sup_{C\in\mathbf{C}_{\mathrm{conv}}} \mathbb{E}_{C}[(|C| - \widehat{\vartheta}_\lambda)^2] > 0 \,, \tag{2.5.6}$$

where the infimum extends over all estimators $\widehat{\vartheta}_\lambda$ in the PPP model with intensity λ. Furthermore, the rate $\lambda^{-(d+3)/(d+1)}$ is achieved up to a logarithmic factor for sets $C_r \in \mathbf{C}_r$ with a smooth boundary following [114].

Following Section 2.4, the mean squared error of the estimator

$$\widehat{\vartheta}_r := \frac{N+1}{N_\circ + 1} |\widehat{C}_r| \,, \tag{2.5.7}$$

satisfies the oracle inequality in Theorem 2.4.2.

2.5.2 Compact sets

Interestingly the class of all compact sets $\mathbf{K}$ of non-zero Lebesgue measure satisfies Assumption 2.2.1 as well. The richness of this class makes it most appealing for conducting statistical inference. Estimation of compact sets was studied in [55], where it was shown that the union of small Euclidean balls centred at the points of the sample is a consistent estimator of a compact set. The $\mathbf{K}$-wrapping hull is just the union of sample points and so $N_\partial = N$ and $|\widehat{K}| = 0$ a.s. Hence for the oracle estimator in (2.3.2) we have

$$\widehat{\vartheta}_{\mathbf{K},oracle} := \frac{N}{\lambda} \,. \tag{2.5.8}$$

This estimator is unbiased and from (2.3.1) the following result immediately follows.

LEMMA 2.5.4. For known intensity $\lambda > 0$, the worst case mean squared error of the oracle estimator $\widehat{\vartheta}_{\mathbf{K},oracle}$ over the parameter class $\mathbf{K}$ decays as $\lambda \uparrow \infty$ like λ^{-1}:

$$\sup_{K\in\mathbf{K}} \frac{1}{|K|} \mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{K},oracle} - |K|)^2\big] = \frac{1}{\lambda} \,. \tag{2.5.9}$$

It seems impossible without imposing further structure on the class $\mathbf{K}$ to estimate λ in this scenario.

2.5.3 Concentric sets

A class of sets generated by one specific convex set dilated from its centre of gravity fits into the framework as well. We denote this class by $\mathbf{D} \subseteq \mathbf{C}_{\mathrm{conv}}$ and its generator by D. We assume the centre point x_0 of a set in the class is known and denote by D_r, $0 < r < \infty$, a member of the class:

$$D_r := \Big\{ x_0 + r(x - x_0) \,\Big|\, x \in D \Big\} . \tag{2.5.10}$$

Without loss of generality one may assume here that $\mathbf{E} = D_R$ for some $R > 0$. Note that the class $\mathbf{D}$ is a parametric class and standard methods of functional estimation in non-regular models, cf. [77], can possibly be exploited to construct asymptotically efficient estimators in a sharp sense. In this section, we illustrate that it is straightforward to derive an efficient estimator of the volume of a set in the class $\mathbf{D}$ with the proposed framework. The $\mathbf{D}$-wrapping hull is given by

$$\widehat{D} := \bigcap_{D_r : \{X_1, \ldots, X_N\} \in D_r} D_r \,, \tag{2.5.11}$$

and thus determined by only one point of the sample. The event to have more than one point of the sample on the boundary $\partial \widehat{D}$ has measure zero yielding $N_\partial = 1$. This entails for the oracle estimator:

$$\widehat{\vartheta}_{\mathbf{D},oracle} := \frac{1}{\lambda} + |\widehat{D}| \,. \tag{2.5.12}$$

The risk of $\widehat{\vartheta}_{\mathbf{D},oracle}$ according to Theorem 2.3.1 depends on $\sup_{D_r \in \mathbf{D}} \mathbb{E}_{D_r}[|D_r \setminus \widehat{D}|]$. In view of Corollary 2.2.4, it holds:

$$\forall D_r \in \mathbf{D} : \quad \lambda \, \mathbb{E}[|D_r \setminus \widehat{D}|] = 1 \,. \tag{2.5.13}$$

Thus, we immediately derive the rate of convergence of $\widehat{\vartheta}_{\mathbf{D},oracle}$. The minimax optimality is straightforward to prove using standard techniques for the lower bounds.

LEMMA 2.5.5. For known intensity $\lambda > 0$, the worst case mean squared error of the oracle estimator $\widehat{\vartheta}_{\mathbf{D},oracle}$ over the parameter class $\mathbf{D}$ decays as $\lambda \uparrow \infty$ like λ^{-2}:

$$\sup_{D_r \in \mathbf{D}} \mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{D},oracle} - |D_r|)^2\big] = \frac{1}{\lambda} \sup_{D_r \in \mathbf{D}} \mathbb{E}_{D_r}[|D_r \setminus \widehat{D}|] = \frac{1}{\lambda^2} \,. \tag{2.5.14}$$

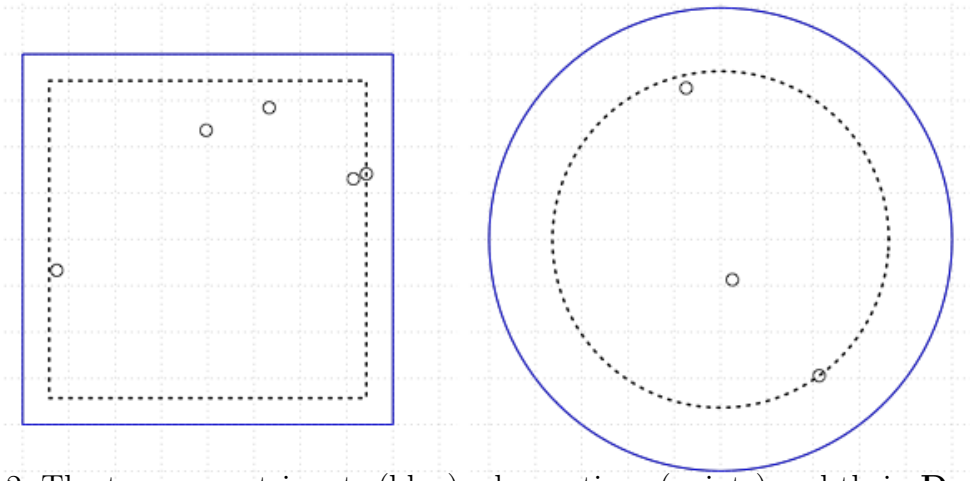

Figure 2.2: The two concentric sets (blue), observations (points) and their $\mathbf{D}$-wrapping hulls (dashed lines).

The rate of convergence $\mathcal{O}(\lambda^{-1})$ is minimax optimal for estimating $|D_r|$ in the PPP model with parameter class $\mathbf{D}$

Further following the general scheme of estimating λ in Section 2.4, we have the final estimator for the volume

$$\widehat{\vartheta}_{\mathbf{D}} := \frac{N+1}{N}|\widehat{D}|\,. \tag{2.5.15}$$

To establish the asymptotic properties of this estimator, according to Theorem 2.4.2 and in view of (2.5.13), it suffices to derive a uniform deviation inequality for the missing volume $|D_r \setminus \widehat{D}|$. The following deviation inequality for the uniform model is easily transferred to the PPP model using Poissonisation, cf. Section 2.4.1.

Lemma 2.5.6. In the uniform model with $X_1, ..., X_n \overset{i.i.d.}{\sim} U(D_r)$, the following uniform deviation result holds

$$\lim_{n\to\infty} \sup_{D_r\in\mathbf{D}} \mathbb{P}_{D_r}\left(n\frac{|D_r \setminus \widehat{D}_n|}{|D_r|} \geqslant x\right) = \mathrm{e}^{-x}\,, \quad \forall x > 0\,. \tag{2.5.16}$$

Furthermore, $\lim_{n\to\infty}\sup_{D_r\in\mathbf{D}} \mathbb{E}\big[n^q|D_r \setminus \widehat{D}_n|^q/|D_r|^q\big] = q\Gamma(q)$ for any $q \in \mathcal{N}$.

Proof. For any $y \in (0,1)$ and any $D_r \in \mathbf{D}$, we obtain

$$\mathbb{P}_{D_r}\big(|D_r \setminus \widehat{D}_n| \geqslant y|D_r|\big) = (1-y)^n\,. \tag{2.5.17}$$

Taking $y = x/n$ for any $x > 0$ and letting $n \to \infty$ yields (2.5.16). The result on the moments of the missing volume follows by Fubini's theorem. □

Corollary 2.5.7. It follows for the risk of the estimator $\widehat{\vartheta}_{\mathbf{D}}$ for all $D_r \in \mathbf{D}$ whenever

$\lambda|D_r| \geqslant 1$:

$$\mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{D}} - |D_r|)^2\big]^{1/2} \leqslant \Big(1 + \frac{c_1}{\lambda|D_r|}\Big)\mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{D},oracle} - |D_r|)^2\big]^{1/2} + c_2 \mathrm{e}^{-c_3\lambda} \lesssim \frac{1}{\lambda^2}\,, \qquad (2.5.18)$$

with some numeric constants $c_1, c_2, c_3 > 0$

2.5.4 Polytopes

It was noted in [14], that the estimator of the volume based on the $\mathbf{C}_{\mathrm{conv}}$-wrapping hull estimator (the convex hull $\widehat{C}$) is adaptive to the class of polytopes $\mathbf{P}$ (see Remark 3.3). In fact, the estimator

$$\widehat{\vartheta}_{\mathbf{P},oracle} := \frac{N_\partial}{\lambda} + |\widehat{C}| \qquad (2.5.19)$$

satisfies

LEMMA 2.5.8. For known intensity $\lambda > 0$, the worst case mean squared error of the oracle estimator $\widehat{\vartheta}_{\mathbf{P},oracle}$ over the parameter class $\mathbf{P}$ decays as $\lambda \uparrow \infty$ like $\lambda^{-2}(\log(\lambda))^{d-1}$:

$$\limsup_{\lambda\to\infty} \lambda^2(\log(\lambda))^{1-d} \sup_{P\in\mathbf{P},|P|>0} \Big\{\mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{P},oracle} - |P|)^2\big]\Big\} < \infty\,. \qquad (2.5.20)$$

We stress here, however, that the class polytopes $\mathbf{P}$ does not satisfy Assumption 2.2.1. The framework applies to the class of convex sets $\mathbf{C}$ and Lemma 2.5.8 only allows to improve the rate for the subclass of $\mathbf{C}$. The class $\mathbf{P}$ is stable only under finite intersections; taking arbitrary (possibly uncountable) intersections, one can obtain an element not lying in the class.

2.5.5 Polytopes with fixed directions of outer unit normal vectors

The class of polytopes $\mathbf{P}_{\mathbf{S}_k}$ with fixed directions $\mathbf{S}_k = \{u_1, ..., u_k\}$ of outer unit normal vectors u_k belonging to the unit sphere $\mathbb{S}^{d-1}$ provides another interesting example of intersection stable sets. We assume the class is well-defined in the sense that there exists a polytope $P_{\mathbf{S}_k}$ whose outer unit normal vectors are exactly $\{u_1, ..., u_k\}$. Without loss of generality we may assume $\mathbf{E} = P_{\mathbf{S}_k}$. The $\mathbf{P}_{\mathbf{S}_k}$-wrapping hull $\widehat{P}$ is a polytope with at most k facets and is given by

$$\widehat{P} := \bigcap_{P\in\mathbf{P}_{\mathbf{S}_k}:\{X_1,...,X_N\}\in P} P\,. \qquad (2.5.21)$$

The oracle estimator and the data-driven estimator are thus defined as

$$\widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k},oracle} := \frac{N_\partial}{\lambda} + |\widehat{P}|\,, \quad \widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k}} = \frac{N+1}{N_\circ+1}|\widehat{P}|\,, \tag{2.5.22}$$

where the number of points lying on the boundary of the wrapping hull N_∂ is equal to the number of facets of the wrapping hull and hence upper-bounded by k. According to the general scheme, the rate of convergence of the risk for the oracle estimator $\widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k},oracle}$ and the final estimator $\widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k}}$ rests upon a deviation inequality for the missing volume and is established in the following theorem which is proved in the Appendix.

THEOREM 2.5.9. The worst case mean squared error of the estimator $\widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k}}$ over the parameter class $\mathbf{P}_{\mathbf{S}_k}$ satisfies:

$$\sup_{P\in\mathbf{P}_{\mathbf{S}_k}} \mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{P}_{\mathbf{S}_k}} - |P|)^2\big] \lesssim \frac{kW(\lambda/k)}{\lambda^2}\,, \tag{2.5.23}$$

whenever $\lambda|P| \geqslant 1$, where W is the Lambert-W function that satisfies $W(z\mathrm{e}^z) = z$. The rate can further be upper-bounded by $k\log(\lambda/k)/\lambda^2$.

2.6 Uniform deviation inequality for weakly-convex sets

An explicit minimax rate of convergence in terms of the intensity of the points is far more involved and exploits a geometric structure of the weakly-convex sets from Definition 2.5.1.

LEMMA 2.6.1. Let $\{C^1,...,C^{N_\varepsilon}\} \in \mathbf{C}_r$ be an ε-net for the class $\mathbf{C}_r$ with respect to the Haussdorf distance. Then for any set $C \in \mathbf{C}_{2r}$ there exists a set C^k in the net such that $\rho_1(C_r, C^k) \leqslant c\varepsilon$ and $C_r \subseteq C^k$ for some constant $c > 1$.

Proof. Clearly we have $C \in \mathbf{C}_r$ since $C \in \mathbf{C}_{2r}$. Consider the set $\lceil C\rceil := B(C,\varepsilon)$ around C. It clearly satisfies $\rho_H(\lceil C\rceil, C) = \varepsilon$ and moreover we have $\lceil C\rceil \in \mathbf{C}_r$ for ε small enough. Hence there exists C^k in the ε-net for the class $\mathbf{C}_r$ such that $\rho_H(C^k, \lceil C\rceil) < \varepsilon$. It then follows that $\rho_H(C^k, C_r) < 2\varepsilon$ and $C_r \subseteq C^k$, which implies $\rho_1(C^k, C_r) < c\varepsilon$ for some constant $c > 1$. □

THEOREM 2.6.2. In the PPP model, there exist a constant $c > 1$ such that the following uniform deviation inequality holds

$$\sup_{C_{r^\star}\in\mathbf{C}_{r^\star}} \mathbb{P}\Big(\lambda\big(|C_{r^\star} \setminus \widehat{C}_r| - 2c\varepsilon\big) \geqslant x\Big) \leqslant \mathrm{e}^{-x}\,, \quad \forall x > 0\,. \tag{2.6.1}$$

for any $r \leqslant r^\star$ provided that ε satisfies the equation $H(\mathbf{C}_{r/2}, \rho_H, \varepsilon) = c\lambda\varepsilon$.

Proof. The proof is inspired by a metric entropy approach, cf. [89]. Recall that both $C_{r^\star}$ and $\widehat{C}_r$ belong to the class $\mathbf{C}_r$ since $\mathbf{C}_{r^\star} \subseteq \mathbf{C}_r$. Let $\{C^1, ..., C^{N_\varepsilon}\} \in \mathbf{C}_{r/2}$ be an ε-net for the class $\mathbf{C}_{r/2}$ with respect to the Haussdorf distance. According to Lemma 2.6.1, there exists a random set $C^{\widehat{j}}$ in the net such that $\rho_1(\widehat{C}_r, C^{\widehat{j}}) \leqslant c\varepsilon$ and $\widehat{C}_r \subseteq C^{\widehat{j}} \subseteq C_{r^\star}$ for some constant $c > 1$ and ε small enough.

Let $y = x/\lambda + 2c\varepsilon$. We thus have

$$\mathbb{P}\big(|C_{r^\star} \setminus \widehat{C}_r| \geqslant y\big) \leqslant \mathbb{P}\big(|C_{r^\star} \setminus C^{\widehat{j}}| \geqslant y - c\varepsilon\big) \tag{2.6.2}$$

$$\leqslant \sum_{j:|C_{r^\star}\setminus C^j| \geqslant y - c\varepsilon} \mathbb{P}\big(\mathcal{N}(C_{r^\star} \setminus C^j) = 0\big) \tag{2.6.3}$$

$$\leqslant \exp\big(-\lambda y + c\lambda\varepsilon + H(\mathbf{C}_{r/2}, \rho_H, \varepsilon)\big) \leqslant \mathrm{e}^{-x}, \tag{2.6.4}$$

provided that ε satisfies $H(\mathbf{C}_{r/2}, \rho_H, \varepsilon) = c\lambda\varepsilon$.

□

Let us denote $\mathrm{conv}_r(C)$ the r-convex hull of a compact set C. We then derive a useful corollary analogously to Theorem 2.6.3.

Corollary 2.6.3. In the PPP model, there exist a constant $c > 1$ such that the following uniform deviation inequality holds

$$\sup_{C_{r^\star} \in \mathbf{C}_{r^\star}} \mathbb{P}\Big(\lambda\big(|\mathrm{conv}_r(C_{r^\star}) \setminus \widehat{C}_r| - 2c\varepsilon\big) \geqslant x\Big) \leqslant \mathrm{e}^{-x}, \quad \forall x > 0. \tag{2.6.5}$$

for any $r > r^\star$ provided that ε satisfies the equation $H(\mathbf{C}_{r/2}, \rho_H, \varepsilon) = c\lambda\varepsilon$.

Let us define

$$\psi_{\lambda,r} = \frac{\varepsilon_{\lambda,r}}{\lambda}, \tag{2.6.6}$$

where $\varepsilon_{\lambda,r}$ satisfies the equation $H(\mathbf{C}_{r/2}, \rho_H, \varepsilon_{\lambda,r}) = c\lambda\varepsilon_{\lambda,r}$.

Theorem 2.6.4. The minimax rate of convergence of the estimator satisfies

$$\limsup_{\lambda\to\infty} \psi_{\lambda,r}^{-1} \sup_{C_{r^\star} \in \mathbf{C}_{r^\star}, |C_{r^\star}|>0} \mathbb{E}\big[|C_{r^\star} \setminus \widehat{C}_r|\big] < \infty. \tag{2.6.7}$$

for any $r < r^\star$.

Remark 2.6.5. Note that Theorem 2.6.4 suggests that it suffices to use the wrapping hull $\widehat{C}_r$ for any $r < r^\star$ to achieve the minimax optimal rate of convergence. This is a consequence of the fact that $C_{r^\star} \in \mathbf{C}_{r^\star} \subseteq \mathbf{C}_r$ for any $r < r^\star$.

2.6.1 Volume estimation and the dilated hull estimator

The volume of a weakly-convex set is one of its most fundamental functionals. In this section, we quantify the minimax rate of convergence of an estimator of the volume

$$\widehat{\vartheta}_r := \frac{N+1}{N_\circ+1}|\widehat{C}_r|\,, \tag{2.6.8}$$

recently proposed in [11]. This estimator was shown to have surprising non-asymptotic properties like the UMVU (unbiased with minimal possible variance among all unbiased estimator). The next result identifies the minimax rate of convergence of the estimator $\widehat{\vartheta}_r$.

THEOREM 2.6.6. The minimax rate of convergence of the estimator $\widehat{\vartheta}_r$ satisfies

$$\limsup_{\lambda\to\infty}(\lambda/\psi_{\lambda,r}) \sup_{C_{r^\star}\in\mathbf{C}_{r^\star},|C_{r^\star}|>0} \Big\{\mathbb{E}\big[(\widehat{\vartheta}_r - |C_{r^\star}|)^2\big]\Big\} < \infty\,, \tag{2.6.9}$$

for all $r < r^\star$ where $\psi_{\lambda,r}$ is given in (2.6.6).

COROLLARY 2.6.7. The following asymptotic equivalence result between the key functionals of the r-convex hull holds

$$\lambda\,\mathbb{E}\big[(\widehat{\vartheta}_r - |C_{r^\star}|)^2\big] \asymp \mathbb{E}\big[|C_{r^\star}\setminus\widehat{C}_r|\big] \asymp \mathbb{E}\big[\frac{N_\delta}{\lambda}\big]\,, \tag{2.6.10}$$

for all $r \leqslant r^\star$.

Proof. Follows from Lemma 2.2.4 in view of $C_{r^\star}\in\mathbf{C}_r$ for all $r\leqslant r^\star$. □

2.7 Adaptation to the regularity parameter $r^\star$

In applications, the regularity parameter $r^\star$ is often unknown and an appropriate adaptation procedure is hence desired. In view of the fact that the estimator of the volume hinges upon the wrapping hull, it actually suffices to provide an adaptive procedure for estimating the set only. Let us define the target regularity parameter $r^\star > 0$ corresponding to a set $C_{r^\star}$ as

$$r^\star := \sup\{r>0 : C\in\mathbf{C}_r\}\,. \tag{2.7.1}$$

From a non asymptotic point of view, it is clear that it is better to exploit the estimator $\widehat{C}_r$ for the values of r smaller than but close to $r^\star$, since we have $|C_{r^\star}\setminus\widehat{C}_{r_2}| \leqslant |C_{r^\star}\setminus\widehat{C}_{r_1}|$ for $r_1\leqslant r_2\leqslant r^\star$. In the region $r^\star < r < \infty$, the estimator starts missing holes and hence can lose dramatically in the risk of convergence. In fact, the estimator $\widehat{C}_r$ estimates the r-convex hull of the set $C_{r^\star}$. We further observe that in the regime $r_1\leqslant r_2\leqslant r^\star$,

$$|\widehat{C}_{r_2}\setminus\widehat{C}_{r_1}| = |C_{r^\star}\setminus\widehat{C}_{r_1}| - |C_{r^\star}\setminus\widehat{C}_{r_2}| \approx N_{\delta,r_1}/\lambda - N_{\delta,r_2}/\lambda\,,$$

whereas in the regime $r_1 \leqslant r^\star \leqslant r_2$, an approximation error $|\mathrm{conv}_r(C_{r^\star}) \setminus C_{r^\star}|$ starts to constitute the term $|\widehat{C}_{r_2} \setminus \widehat{C}_{r_1}|$.

This intuition yields the following procedure for estimating the set. Let us fix some $R > 0$ and break the interval $(0, R)$ down into K pieces of equal length $0 < r_1 < \cdots < r_K = R$. Let us define an estimator of the regularity parameter r as

$$\widehat{r} := \inf\left\{r_{k-1} | \, \exists k' \leqslant k \, : \, |\widehat{C}_{r_k} \setminus \widehat{C}_{r_{k'}}| > c_1 N_{\delta, r_{k'}}/\lambda - c_2 N_{\delta, r_k}/\lambda\right\} \wedge r_K \,, \tag{2.7.2}$$

with some universal constants $c_1, c_2 > 0$ and define a corresponding element of the partition by $\widehat{k}$. This procedure is inspired by the prominent Lepskii's method, see [93], and was recently numerically studied in [11]. This procedure is numerically studied in the following section.

2.8 Illustrative simulations

Let us first consider the two classes of concentric sets generated by a ball with a centre at $x = (0.5, 0.5)$ and a square with a centre at the same point, see Figure 2.2. The mean squared error estimate is based on $M = 1000$ Monte Carlo iterations in each case. We demonstrate the result of Lemma 2.5.5 that in fact

$$\mathbb{E}\big[(\widehat{\vartheta}_D - |D_r|)^2\big] = \frac{1}{\lambda^2} \,, \tag{2.8.1}$$

by plotting the mean squared error estimate multiplied by the intensity squared λ^2 with respect to λ in Figure 2.3. One can see that the lines in the plots are fairly closed to 1 supporting the claim. The simulations were implemented using the R package "spatstat" by [10].

Furthermore, we illustrate the performance of the proposed estimators for a class of r-convex sets. Our first example of an r-convex set for simulations is the annulus $C_{r^\star} = B(0.5, 0.5) \setminus B(0.5, 0.25)$. Thus clearly $C_{r^\star} \in \mathbf{C}_r$ for all $0 < r \leqslant 0.25$. Figure 2.4 depicts the r-convex hull estimator (2.5.3) for $r = 0.01, 0.04, 0.2, 1$ based on the observations of the PPP with $\lambda = 300$. An important observation is that once the value of r is larger than the true radius $r^\star$ of an r-convex set, the r-convex hull essentially misses the "holes" of radius $r^\star$. One should bear this in mind when using large values of r for constructing the oracle estimator when the number of observation points is small. This subtle issue is depicted in Figure 2.5, where the root mean squared error of the oracle estimator $\widehat{\vartheta}_{r,oracle}$ for the volume (black line), based on the r-convex hull with $r = 0.30$, converges to the area of the "hole" of size $\pi(0.25)^2 \approx 0.196$. Another point is that when the number of observations is small, the r-convex hull with a small value of r essentially coincides with the points themselves and thus the RMSE of the oracle estimator $\widehat{\vartheta}_{r,oracle}$ coincides with

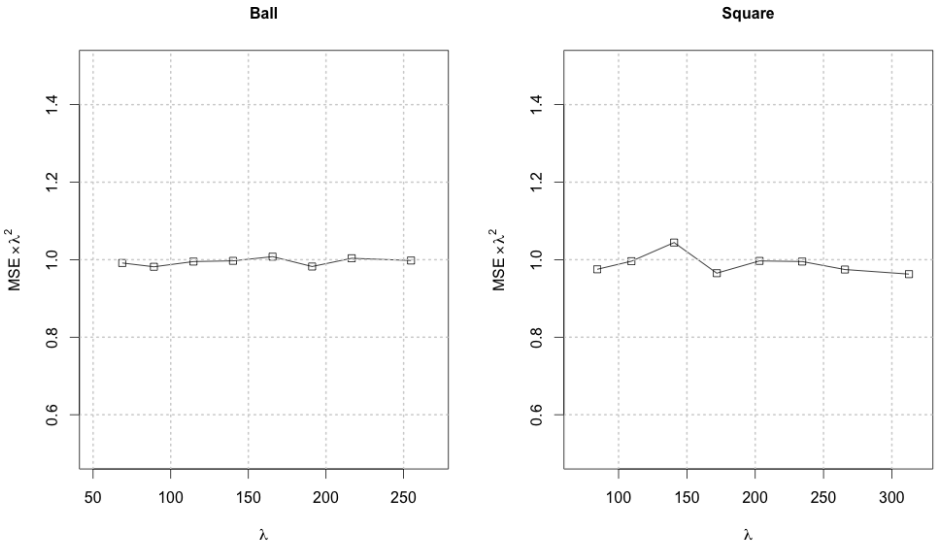

Figure 2.3: The mean squared error $\mathbb{E}\big[(\widehat{\vartheta}_{\mathbf{D}} - |D_r|)^2\big]$ multiplied by the intensity squared λ^2 with respect to λ for the ball and the square from Figure 2.2.

the RMSE of $\widehat{\vartheta}_{\mathbf{K},oracle}$ and equals $|C_{r^\star}|/\lambda$ (red line in Figure 2.5)! Finally we depict RMSE estimates for the oracle estimator $\widehat{\vartheta}_{r,oracle}$ for different r in Figure 2.6. One can clearly see the regions of decreasing value of the RMSE, the fairly flat value of RMSE and the jump when r becomes larger than the true parameter $r^\star$. Table 2.1 further collects the Monte Carlo estimates of the number of points $N_\circ$ lying inside the wrapping hull, the number of points N_∂ on the boundary of the wrapping hull and the number of isolated points N_{iso} of the boundary of the wrapping hull. For analyzing the performance of the adaptive estimator proposed in Section 2.5.1, we break the interval $[0.06, 0.5]$ into pieces of length 0.02, compute the estimates of the radius $\widehat{r}$ from (2.7.2) and the estimates of the RMSE of $\widehat{\vartheta}_{\widehat{r}}$ based on 200 Monte Carlo iterations in Table 2.2.

2.9 Appendix

Proof of Theorem 2.3.3

Sufficiency follows from the Neyman factorisation criterion applied to the likelihood function (2.2.6), while completeness follows by definition provided that we show

$$\forall A \in \mathbf{A} : \mathbb{E}_A\big[T(\widehat{A})\big] = 0 \implies T(\widehat{A}) = 0 \quad \mathbb{P}_{\mathbf{E}} - a.s. \tag{2.9.1}$$

for any $\mathcal{A}$-measurable function $T : \mathbf{A} \to \mathbb{R}$. From the likelihood in (2.2.6) for $\lambda = \lambda_0$, we derive

$$\mathbb{E}_A\big[T(\widehat{A})\big] = \mathbb{E}_{\mathbf{E}}\Big[T(\widehat{A}) \exp\big(\lambda|\mathbf{E} \setminus A|\big)\mathbf{1}(\widehat{A} \subseteq A)\Big]. \tag{2.9.2}$$

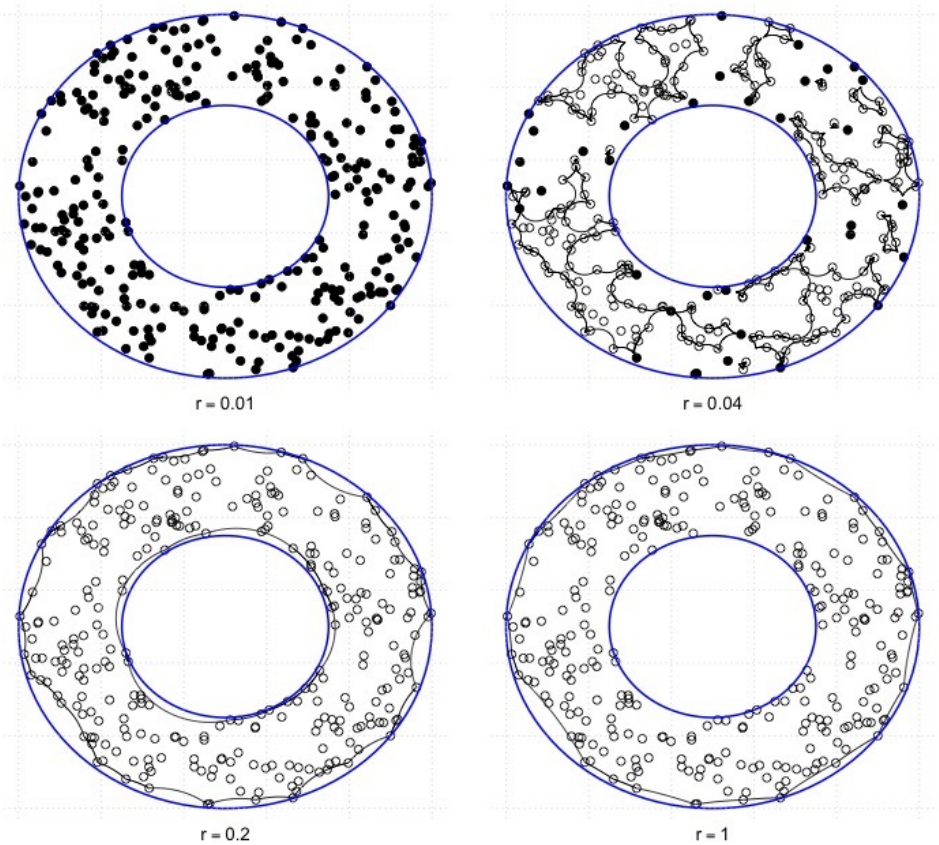

Figure 2.4: The four weakly-convex set with $r^\star = 0.25$ (blue), observations of the PPP with $\lambda = 300$ (points) and their r-convex hulls for different values of r (black).

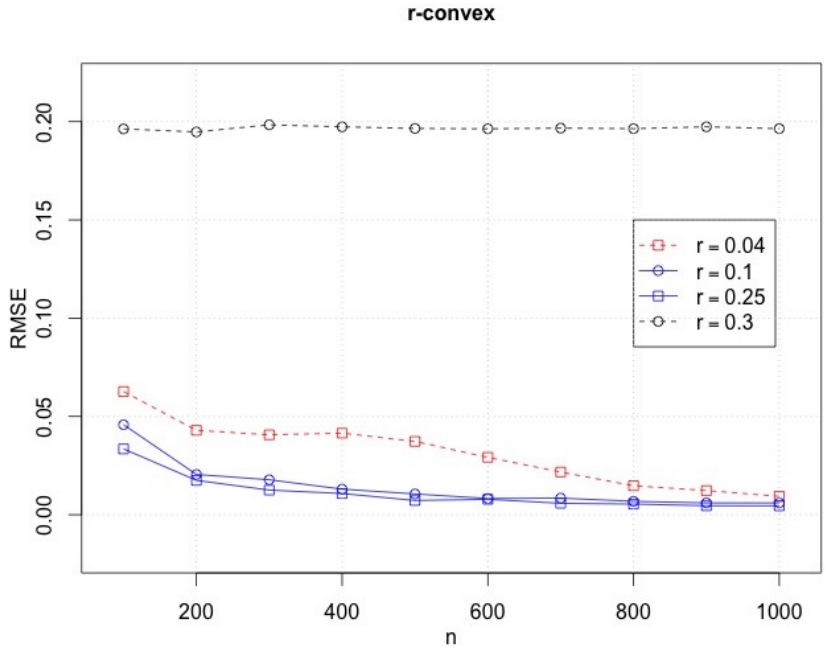

Figure 2.5: Monte Carlo RMSE estimates for the oracle estimator for the volume of the annulus $B(0.5, 0.5) \setminus B(0.5, 0.25)$ with respect to the sample size.

$r = 0.04$

$n = \lambda/\lvert A\rvert$	$N_\circ$	N_∂	N_{iso}	$RMSE(\widehat{\vartheta}_{r,oracle})$	$RMSE(\widehat{\vartheta}_r)$	$\frac{RMSE(\widehat{\vartheta}_r)}{RMSE(\widehat{\vartheta}_{r,oracle})}$
50	0.13	49.8	44	0.087	0.55	6.31
100	1.5	98.37	66	0.059	0.33	4.65
200	23.6	175.8	58	0.042	0.15	3.7
300	90.4	211	33	0.039	0.109	2.82
400	191.6	210	17	0.043	0.085	1.97

$r = 0.1$

$n = \lambda/\lvert A\rvert$	$N_\circ$	N_∂	N_{iso}	$RMSE(\widehat{\vartheta}_{r,oracle})$	$RMSE(\widehat{\vartheta}_r)$	$\frac{RMSE(\widehat{\vartheta}_r)}{RMSE(\widehat{\vartheta}_{r,oracle})}$
50	8.5	42.16	8.52	0.071	0.138	1.94
100	45.9	53.34	1.37	0.043	0.064	1.48
200	138.2	60.95	0.03	0.021	0.027	1.26
300	233.4	68.08	0	0.015	0.018	1.20
400	326.1	74.40	0	0.013	0.015	1.15

$r = 0.25$

$n = \lambda/\lvert A\rvert$	$N_\circ$	N_∂	N_{iso}	$RMSE(\widehat{\vartheta}_{r,oracle})$	$RMSE(\widehat{\vartheta}_r)$	$\frac{RMSE(\widehat{\vartheta}_r)}{RMSE(\widehat{\vartheta}_{r,oracle})}$
50	24.75	24.21	0.06	0.061	0.085	1.39
100	68.58	29.60	0	0.033	0.0405	1.20
200	163.75	36.03	0	0.018	0.019	1.04
300	261.44	40.68	0	0.0108	0.0124	1.13
400	357.41	44.17	0	0.0096	0.0104	1.076

$r = 0.3$

$n = \lambda/\lvert A\rvert$	$N_\circ$	N_∂	N_{iso}	$RMSE(\widehat{\vartheta}_{r,oracle})$	$RMSE(\widehat{\vartheta}_r)$	$\frac{RMSE(\widehat{\vartheta}_r)}{RMSE(\widehat{\vartheta}_{r,oracle})}$
50	30.71	18.70	0	0.208	0.340	1.628
100	77.59	23.26	0	0.2002	0.258	1.29
200	170.30	29.39	0	0.1982	0.232	1.17
300	265.17	33.89	0	0.1978	0.223	1.13
400	362.43	37.89	0	0.1987	0.219	1.10

Table 2.1: Monte Carlo RMSE estimates for the oracle estimator $\widehat{\vartheta}_{r,oracle}$ and for the fully data-driven estimator $\widehat{\vartheta}_r$ for the volume of the annulus $A = B(0.5, 0.5) \setminus B(0.5, 0.25)$ with respect to r and $n = \lambda|A|$, the number of points lying inside the wrapping hull $N_\circ$, the number of points on the boundary of the wrapping hull N_∂ and the number of isolated points of the boundary of the wrapping hull N_{iso}.

$n = \lambda/\lvert A\rvert$	$\widehat{r}$	$RMSE(\widehat{\vartheta}_{\widehat{r}})$
50	0.088	0.36
100	0.085	0.160
200	0.084	0.069
300	0.105	0.033
400	0.125	0.0182
500	0.149	0.0123
1000	0.165	0.0056

Table 2.2: Monte Carlo RMSE estimates for the adaptive estimator $\widehat{\vartheta}_{\widehat{r}}$ for the volume of the annulus $A = B(0.5, 0.5) \setminus B(0.5, 0.25)$ with respect to $n = \lambda|A|$.

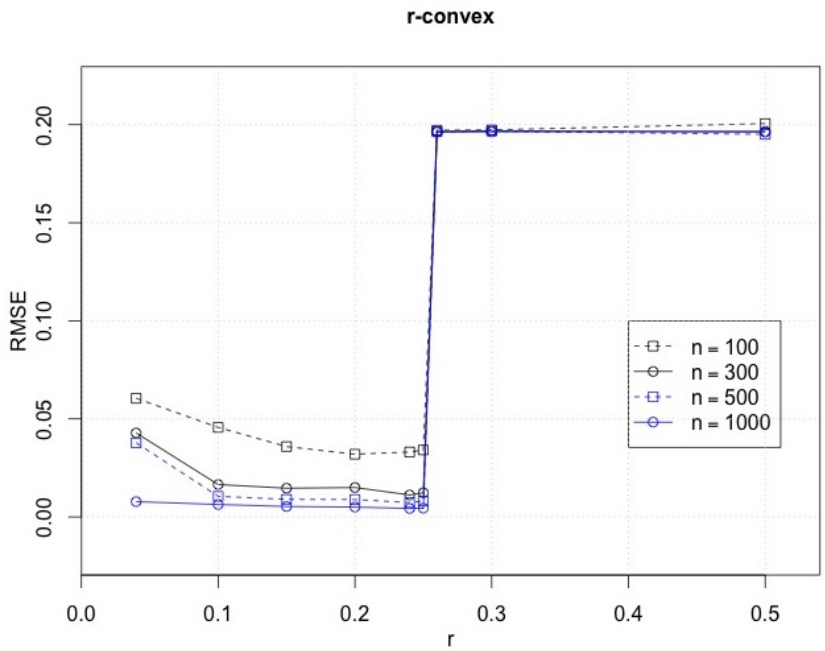

Figure 2.6: Monte Carlo RMSE estimates for the oracle estimator for the volume of the annulus $B(0.5, 0.5) \setminus B(0.5, 0.25)$ with respect to r.

Since $\exp(\lambda|\mathbf{E} \setminus A|)$ is deterministic, we have $\forall A \in \mathbf{A}$

$$\mathbb{E}_A\big[T(\widehat{A})\big] = 0 \implies \mathbb{E}_{\mathbf{E}}\big[T(\widehat{A})\mathbf{1}(\widehat{A} \subseteq A)\big] = 0\,. \tag{2.9.3}$$

Splitting $T = T^+ - T^-$ with non-negative $\mathcal{A}$-measurable functions T^+ and T^-, we infer that the measures $\mu^{\pm}(B) = \mathbb{E}_{\mathbf{E}}[T^{\pm}(\widehat{A})\mathbf{1}(\widehat{A} \in B)]$, $B \in \mathcal{A}$, agree on $\{[B] \,|\, B \in \mathbf{A}\}$, where $[B] = \{A \in \mathbf{A} | A \subseteq B\}$. Since the brackets $\{[B] \,|\, B \in \mathbf{A}\}$ generate the σ-algebra $\mathcal{A}$ the measures $\mu^{\pm}(B)$ agree on all sets in $\mathcal{A}$, in particular on $\{T > 0\}$ and $\{T < 0\}$, which entails $\mathbb{E}_{\mathbf{E}}[T^+(\widehat{A})] = \mathbb{E}_{\mathbf{E}}[T^-(\widehat{A})] = 0$. Thus, $T(\widehat{A}) = 0$ holds $\mathbb{P}_{\mathbf{E}}$-a.s.

Proof of Theorem 2.5.9

Let us denote by $\rho_1(A, B) = |A \triangle B|$ the symmetric distance between two compact subsets A and B of the compact convex set $\mathbf{E}$ in $\mathbb{R}^d$. Recall that an *ε-net* of the class $\mathbf{P}_{\mathbf{S}_k}$ with respect to the metric ρ_1 is a collection $\{P^1, ..., P^{N_\varepsilon}\} \in \mathbf{P}_{\mathbf{S}_k}$ such that for each $P \in \mathbf{P}_{\mathbf{S}_k}$, there exists $i \in \{1, ..., N_\varepsilon\}$ such that $\rho_1(P, P^i) \leqslant \varepsilon$. The *$\varepsilon$-covering number* $N(\mathbf{P}_{\mathbf{S}_k}, \rho_1, \varepsilon)$ is the cardinality of the smallest ε-net. The *ε-entropy* of the class $\mathbf{P}_{\mathbf{S}_k}$ is defined by $H(\mathbf{P}_{\mathbf{S}_k}, \rho_1, \varepsilon) = \log_2 N(\mathbf{P}_{\mathbf{S}_k}, \rho_1, \varepsilon)$. Furthermore, it follows by dilation of a set that for $\widehat{P} \in \mathbf{P}_{\mathbf{S}_k}$ there exists $\widehat{m} \in \{1, ..., N_\varepsilon\}$ such that $\widehat{P} \subseteq P^{\widehat{m}} \subseteq P$ and $\rho_1(\widehat{P}, P^{\widehat{m}}) \leqslant c\varepsilon$ for some universal constant $c > 1$ and ε small enough. We thus obtain for all $P \in \mathbf{P}_{\mathbf{S}_k}$ and $x > 0$,

$$\mathbb{P}\big(|P \setminus \widehat{P}| > x/\lambda + 2c\varepsilon\big) \leqslant \mathbb{P}\big(|P \setminus P^{\widehat{m}}| > x/\lambda + c\varepsilon\big) \tag{2.9.4}$$

$$\leqslant \sum_{m:|P\setminus P^m|>x/\lambda+c\varepsilon} \mathbb{P}\big(\mathcal{N}(P\setminus P^m)=0\big) \tag{2.9.5}$$

$$\leqslant \exp\big(-x-c\lambda\varepsilon+H(\mathbf{P}_{\mathbf{S}_k},\rho_1,\varepsilon)\big)=\mathrm{e}^{-x}\,, \tag{2.9.6}$$

plugging in ε that solves $H(\mathbf{P}_{\mathbf{S}_k},\rho_1,\varepsilon)=c\lambda\varepsilon$. The ε-covering number $N(\mathbf{P}_{\mathbf{S}_k},\rho_1,\varepsilon)$ of the class $\mathbf{P}_{\mathbf{S}_k}$ can be bounded by $(C/\varepsilon)^k$ for some universal constant $C>1$. As a result, the asymptotic rate follows using Fubini's theorem combined with Theorem 2.3.1 and Theorem 2.4.2.

Chapter 3

Optimal Link Prediction with Matrix Logistic Regression

Introduction

While the estimators for the volume of a set from the classes of convex and weakly-convex sets proposed in the first and second chapters achieve minimax optimal rates, they cannot be computed using a polynomial-time algorithm in dimensions higher than three, see e.g. [62, 129]. The computational aspect of volume estimation in high dimensions for the considered variety of intersection stable classes of sets is a subject of computational geometry and is beyond the scope of this thesis. In the statistical community, the interplay between computational and statistical aspects of estimation has recently attracted a lot of attention, see e.g. [24, 25, 47, 49, 66, 73, 96, 141, 145] for computational lower bounds in high-dimensional statistics based on the so-called *planted clique problem*. This chapter rigorously examines the statistical and computational trade-off in a high-dimensional matrix logistic regression problem.

In the field of network analysis, the task of *link prediction* consists in predicting the presence or absence of edges in a large graph, based on the observations of some of its edges, and on side information. Network analysis has become a growing motivation for statistical problems. Indeed, one of the main characteristics of datasets in the modern scientific landscape is not only their growing size, but also their increasing complexity. Most phenomena now studied in the natural and social sciences concern not only isolated and independent variables, but also their interactions and connections.

The fundamental problem of link prediction is therefore naturally linked with statistical estimation: the objective is to understand, through a generative model, why different vertices are connected or not, and to generalise these observations to the rest of the graph.

Most statistical problems based on graphs are unsupervised: the graph itself is the sole data, there is no side information, and the objective is to recover an unknown

structure in the generative model. Examples include the planted clique problem [7, 92], the stochastic block model [75]—see [1] for a recent survey of a very active line of work [2, 15, 53, 101, 107, 108], the Ising blockmodel [26], random geometric graphs – see [115] for an introduction and [37, 54] for recent developments in statistics, or metric-based learning [20, 48] and ordinal embedings [79].

In supervised regression problems on the other hand, the focus is on understanding a fundamental mechanism, formalized as the link between two variables. The objective is to learn how an explanatory variable X allows to predict a response Y, i.e. to find the unknown function f that best approximates the relationship $Y \approx f(X)$. This statistical framework is often applied to the observation of a phenomenon measured by Y (e.g. of a natural or social nature), given known information X: the principle is to understand said phenomenon, to explain the relationship between the variables by estimating the function f [74, 76].

We follow this approach here: our goal is to learn how known characteristics of each agent (represented by a node) in the network induce a greater or smaller chance of connection, to understand the mechanism of formation of the graph. We propose a model for supervised link prediction, using the principle of regression for inference on graphs. For each vertex, we are given side information, a vector of observations $X \in \mathbb{R}^d$. Given observations X_i, X_j about nodes i and j of a network, we aim to understand how these two explanatory variables are related to the probability of connection between the two corresponding vertices, such that $\mathbf{P}(Y_{(i,j)} = 1) = f(X_i, X_j)$, by estimating f within a high-dimensional class based on logistic regression. Besides this high-dimensional parametric modelling, various fully nonparametric statistical frameworks were exploited in the literature, see, for example, [65, 143] for graphon estimation, [27, 112] for graph reconstruction and [28] for modularity analysis.

Link prediction can be useful in any application where data can be gathered about the nodes of a network. One of the most obvious motivations is in social networks, in order to model social interactions. With access to side information about each member of a social network, the objective is to understand the mechanisms of connection between members: shared interests, differences in artistic tastes or political opinion [142]. This can also be applied to citation networks, or in the natural sciences to biological networks of interactions between molecules or proteins [97, 144]. The key assumption in this model is that the network is a consequence of the information, but not necessarily based on similarity: it is possible to model more complex interactions, e.g. where opposites attract.

The focus on a high-dimensional setting is another aspect of this work that is also motivated by modern applications of statistics: data is often collected without discernment and the ambient dimension d can be much larger than the sample size. This setting is common in regression problems: the underlying model is often actually very simple, to

reflect the fact that only a small number of measured parameters are relevant to the problem at hand, and that the intrinsic dimension is much smaller. This is usually handled through an assumption on the rank, sparsity, or regularity of a parameter. Here this needs to be adapted to a model with two covariates (explanatory variables) and a structural assumption is made in order to reflect this nature of our problem.

We therefore decide to tackle link prediction by modelling it as matrix logistic regression. We study a generative model for which $\mathbf{P}(Y_{(i,j)} = 1) = \sigma(X_i^\top \Theta_\star X_j)$, where σ is the sigmoid function, and $\Theta_\star$ is the unknown matrix to estimate. It is a simple way to model how the variables *interact*, by a quadratic *affinity* function and a sigmoid function. In order to model realistic situations with partial observations, we assume that $Y_{(i,j)}$ is only observed for a subset of all the pairs (i, j), denoted by Ω.

To convey the general idea of a simple dependency on X_i and X_j, we make structural assumptions on the rank and sparsity of $\Theta_\star$. This reflects that the affinity $X_i^\top \Theta_\star X_j$ is a function of the projections $u_\ell^\top X$ for the vectors X_i and X_j, for a small number of orthogonal vectors, that have themselves a small number of non-zero coefficients (sparsity assumption). In order to impose that the inverse problem is well-posed, we also make a restricted conditioning assumption on $\Theta_\star$, inspired by the restricted isometry property (RIP). These conditions are discussed in Section 3.1. We talk of link prediction as this is the legacy name but we focus almost entirely on the problem of estimating $\Theta_\star$.

The classical techniques of likelihood maximization can lead to computationally intractable optimization problems. We show that in this problem as well as others this is a fundamental difficulty, not a weakness of one particular estimation technique; statistical and computational complexities are intertwined.

This chapter is organized in the following manner: We give a formal description of the problem in Section 3.1, as well as a discussion of our assumptions and links with related work. Section 3.2 collects our main statistical results. We propose an estimator $\widehat{\Theta}$ based on the penalised maximum likelihood approach and analyse its performance in Section 3.2.1 in terms of non-asymptotic rate of estimation. We show that it attains the minimax rate of estimation over simultaneously block-sparse and low-rank matrices $\Theta_\star$, but is not computationally tractable. In Section 3.2.2, we provide a convex relaxation of the problem which is in essence the *Lasso* estimator applied to a vectorised version of the problem. The link prediction task is covered in Section 3.2.3. A matching minimax lower bound for the rate of estimation is given in Section 3.2.4. Furthermore, we show in Section 3.3 that the minimax rate cannot be attained by a (randomised) polynomial-time algorithm, and we identify a corresponding computational lower bound. The proof of this bound is based on a reduction scheme from the so-called *dense subgraph detection problem*. Technical proofs are deferred to the appendix. Our findings are depicted in Figure 3.1.

Notation: For any positive integer n, we denote by $[n]$ the set $\{1, \dots, n\}$ and by $[[n]]$ the set of pairs of $[n]$, of cardinality $\binom{n}{2}$. We denote by $\mathbb{R}$ the set of real numbers and by $\mathbf{S}^n$ the set of real symmetric matrices of size n. For a matrix $A \in \mathbf{S}^n$, we denote by $\|A\|_F$ its Frobenius norm, defined by

$$\|A\|_F^2 = \sum_{i,j \in [d]} A_{ij}^2 \,.$$

We extend this definition for $B \in \mathbf{S}^n$ and any subset $\Omega \subseteq [[n]]$ to its semi-norm $\|B\|_{F,\Omega}$ defined by

$$\|B\|_{F,\Omega}^2 = \sum_{i,j \,:\, (i,j) \in \Omega} B_{ij}^2 \,.$$

The corresponding bilinear form playing the role of inner-product of two matrices $B_1, B_2 \in \mathbf{S}^n$ is denoted as $\langle\!\langle\, B_1, B_2 \,\rangle\!\rangle_{F,\Omega}$. For a matrix $B \in \mathbf{S}^n$, we also make use of the following matrix norms and pseudo-norms for $p, q \in [0, \infty)$, with $\|B\|_{p,q} = \big\|(\|B_{1*}\|_p \cdots \|B_{d*}\|_p)\big\|_q$, where B_{i*} denotes the ith row of B, and $\|B\|_\infty = \max_{(i,j) \in [[d]]} |B_{ij}|$.

3.1 Problem description

3.1.1 Generative model

For a set of vertices $V = [n]$ and explanatory variables $X_i \in \mathbb{R}^d$ associated to each $i \in V$, a random graph $G = (V, E)$ is generated by the following model. For all $i, j \in V$, variables $X_i, X_j \in \mathbb{R}^d$ and an unknown matrix $\Theta_\star \in \mathbf{S}_d$, an edge connects the two vertices i and j independently of the others according to the distribution

$$\mathbf{P}\big((i,j) \in E\big) = \sigma(X_i^\top \Theta_\star X_j) = \frac{1}{1 + \exp(-X_i^\top \Theta_\star X_j)} \,. \tag{3.1.1}$$

Here we denote by σ the *sigmoid*, or *logistic* function.

Definition 3.1.1. We denote by $\pi_{ij} : \mathbf{S}_d \to [0,1]$ the function mapping a matrix $\Theta \in \mathbf{S}_d$ to the probability in (3.1.1). Let $\Sigma \in \mathbf{S}_n$ with $\Sigma_{ij} = X_i^\top \Theta X_j$ denote the so-called affinity matrix. In particular, we then have $\pi_{ij}(\Theta) = \sigma(\Sigma_{ij})$.

Our observation consists of the explanatory variables X_i and of the observation of a subset of the graph. Formally, for a subset $\Omega \subseteq [[n]]$, we observe an adjacency vector Y indexed by Ω that satisfies, for all $(i,j) \in \Omega$, $Y_{(i,j)} = 1$ if and only if $(i,j) \in E$ (and 0 otherwise). We thus have

$$Y_{(i,j)} \sim \text{Bernoulli}\big(\pi_{ij}(\Theta^\star)\big) \,, \quad (i,j) \in \Omega \,. \tag{3.1.2}$$

The joint data distribution is denoted by $\mathbb{P}_{\Theta^\star}$ and is thus completely specified by $\pi_{ij}(\Theta^\star)$,

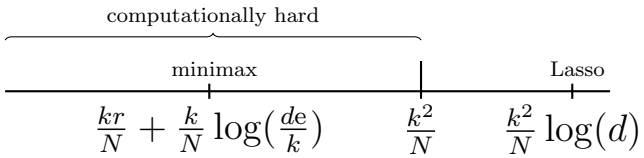

Figure 3.1: The computational and statistical boundaries for estimation and prediction in the matrix logistic regression model. Here k denotes the sparsity of $\Theta^\star$ and r its rank, while N is the number of observed edges in the network.

$(i,j) \in \Omega$. For ease of notation, we write $N = |\Omega|$, representing the *effective sample size*. Our objective is to estimate the parameter matrix $\Theta_\star$, based on the observations $Y \in \mathbb{R}^N$ and on known explanatory variables $\mathbf{X} \in \mathbb{R}^{d\times n}$.

This problem can be reformulated as a classical logistic regression problem. Indeed, writing $\mathrm{Vec}(A) \in \mathbb{R}^{d^2}$ for the *vectorized* form of a matrix $A \in \mathbf{S}^d$, we have that

$$X_i^\top \Theta_\star X_j = \mathrm{tr}(X_j X_i^\top \Theta_\star) = \langle \mathrm{Vec}(X_j X_i^\top), \mathrm{Vec}(\Theta_\star)\rangle \,. \tag{3.1.3}$$

The vector of observation $Y \in \mathbb{R}^N$ therefore follows a logistic distribution with explanatory design matrix $\mathbb{D}_\Omega \in \mathbb{R}^{N\times d^2}$ such that $\mathbb{D}_{\Omega\,(i,j)} = \mathrm{Vec}(X_j X_i^\top)$ and predictor $\mathrm{Vec}(\Theta_\star) \in \mathbb{R}^{d^2}$. We focus on the matrix formulation of this problem, and consider directly *matrix logistic regression* in order to simplify the notation of the explanatory variables and our model assumptions on $\Theta_\star$, that are specific to matrices.

3.1.2 Comparaison with other models

This model can be compared to other settings in the statistical and learning literature.

Generalised linear model. As discussed above in the remark to (3.1.3), this is an example of a logistic regression model. We focus in this work on the case where the matrix $\Theta_\star$ is block-sparse. The problem of sparse generalised linear models, and sparse logistic regression in particular has been extensively studied (see, e.g. [3, 9, 39, 102, 123, 135], and references therein). Our work focuses on the more restricted case of block-sparse and low-rank matrices, establishing interesting statistical and computational phenomena in this setting.

Graphon model. The graphon model is a type of a random graph in which the explanatory variables associated with the vertices in the graph are unknown. It has recently become popular in the statistical community, see [65, 86, 143, 146]. Typically, an objective of statistical inference is a link function which belongs to either a parametric or nonparametric class of functions. Interestingly, the minimax lower bound for the classes of Hölder-continuous functions, obtained in [65], has not been attained by any polynomial-time algorithm.

Trace regression models. The modelling assumption (3.1.1) of this chapter is in fact very close to the trace regression model, as it follows from the representation (3.1.3). Thus, the block-sparsity and low-rank structures are preserved and can well be studied by the means of techniques developed for trace regression. We refer the reader to [64, 87, 109, 126] for recent developments in the linear trace regression model, and [63] for the generalised trace regression model. However, computational lower bounds have not been studied before and many existing minimax optimal estimators cannot to be computed in polynomial time.

Metric learning. In the task of metric learning, observations depend on an unknown geometric representation $V_1, \dots, V_n$ of the variables in a Euclidean space of low dimension. The goal is to estimate this representation (up to a rigid transformation), based on noisy observations of $\langle V_i, V_j \rangle$ in the form of random evaluations of similarity. Formally, our framework also recovers the task of metric learning by taking $X_i = e_i$ and $\Theta_\star$ an unknown semidefinite positive matrix of small rank (here $V^\top V$), since

$$\langle V_i, V_j \rangle = \langle V e_i, V e_j \rangle = e_i^\top V^\top V \, e_j \, .$$

We refer to [20, 48] and references therein for a comprehensive survey of metric learning methods.

3.1.3 Parameter space

The unknown predictor matrix $\Theta_\star$ describes the relationship between the observed features X_i and the probabilities of connection $\pi_{ij}(\Theta_\star) = \sigma(X_i^\top \Theta_\star X_j)$ following Definition 3.1.1. We focus on the high-dimensional setting where $d^2 \gg N$: the number of features for each vertex in the graph, and number of free parameters, is much greater than the total number of observations. In order to counter the curse of dimensionality, we make the assumption that the function $(X_i, X_j) \mapsto \pi_{ij}$ depends only on a small subset S of size k of all the coefficients of the explanatory variables. This translates to a *block-sparsity* assumption on $\Theta_\star$: the coefficients $\Theta_{\star\, ij}$ are only nonzero for i and j in S. Furthermore, we assume that the rank of the matrix $\Theta^\star$ can be smaller than the size of the block. Formally, we define the following parameter spaces

$$\mathcal{P}_{k,r}(M) = \Big\{ \Theta \in \mathbf{S}^d \, : \, \|\Theta\|_{1,1} < M \ , \, \|\Theta\|_{0,0} \leqslant k \ , \text{ and } \mathbf{rank}(\Theta) \leqslant r \Big\} ,$$

for the coefficient-wise ℓ_1 norm $\|\cdot\|_{1,1}$ on $\mathbf{S}^d$ and integers $k, r \in [d]$. We also denote $\mathcal{P}(M) = \mathcal{P}_{d,d}(M)$ for convenience.

Remark 3.1.2. The bounds on block-sparsity and rank in our parameter space are structural bounds: we consider the case where the matrix $\Theta_\star$ can be concisely described

in terms of the number of parameters. This is motivated by considering the spectral decomposition of the real symmetric matrix $\Theta^\star$ as

$$\Theta^\star = \sum_{\ell=1}^{r} \lambda_\ell u_\ell u_\ell^\top .$$

The affinity $\Sigma_{ij} = X_i^\top \Theta^\star X_j$ between vertices i and j is therefore only a function of the projections of X_i and X_j along the axes u_ℓ, i.e.

$$\Sigma_{ij} = X_i^\top \Theta^\star X_j = \sum_{\ell=1}^{r} \lambda_\ell (u_\ell^\top X_i)(u_\ell^\top X_j) .$$

Assuming that there are only a few of these directions u_ℓ with non-zero impact on the affinity motivates the low-rank assumption, while assuming that there are only few relevant coefficients of X_i, X_j that influence the affinity corresponds to a sparsity assumption on the u_ℓ, or block sparsity of $\Theta^\star$. The effect of these projections on the affinity is weighted by the λ_ℓ. By allowing for negative eigenvalues, we allow our model to go beyond a geometric description, where close or similar Xs are more likely to be connected. This can be used to model interactions where opposites attract.

The assumption of simultaneously sparse and low-rank matrices arises naturally in many applications in statistics and machine learning and has attracted considerable recent attention, [122]. Various regularisation techniques have been developed for estimation, variable and rank selection in multivariate regression problems [see, e.g. 40, and the references therein].

3.1.4 Explanatory variables

As mentioned above, this problem is different from tasks such as metric learning, where the objective is to estimate the X_i with no side information. Here they are seen as covariates, allowing us to infer from the observation on the graph the predictor variable $\Theta_\star$. For this task to be even possible in a high-dimensional setting, we settle the identifiability issue by making the following variant of a classical assumption on $\mathbf{X} \in \mathbb{R}^{d\times n}$.

Definition 3.1.3 (Block Isometry Property). For a matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and an integer $s \in [d]$, we define $\Delta_{\Omega,s}(\mathbf{X}) \in (0,1)$ as the smallest positive real such that

$$N\big(1 - \Delta_{\Omega,s}(\mathbf{X})\big)\|B\|_F^2 \leqslant \|\mathbf{X}^\top B\,\mathbf{X}\|_{F,\Omega}^2 \leqslant N\big(1 + \Delta_{\Omega,s}(\mathbf{X})\big)\|B\|_F^2 ,$$

for all matrices $B \in \mathbf{S}^d$ that satisfy the block-sparsity assumption $\|B\|_{0,0} \leqslant s$.

Definition 3.1.4 (Restricted Isometry Properties). For a matrix $A \in \mathbb{R}^{n\times p}$ and an integer

$s \in [p]$, $\delta_s(A) \in (0,1)$ is the smallest positive real such that

$$n\big(1-\delta_s(A)\big)\|v\|_2^2 \leqslant \|Av\|_2^2 \leqslant n\big(1+\delta_s(A)\big)\|v\|_2^2\,,$$

for all s-sparse vectors, i.e. satisfying $\|v\|_0 \leqslant s$.

When $p = d^2$ is a square, we define $\delta_{\mathcal{B},s}(A)$ as the smallest positive real such that

$$n\big(1-\delta_{\mathcal{B},s}(A)\big)\|v\|_2^2 \leqslant \|Av\|_2^2 \leqslant n\big(1+\delta_{\mathcal{B},s}(A)\big)\|v\|_2^2\,,$$

for all vectors such that $v = \mathrm{Vec}(B)$, where B satisfies the block-sparsity assumption $\|B\|_{0,0} \leqslant s$.

The first definition is due to [44], with restriction to sparse vectors. It can be extended in general, as here, to other types of restrictions [see, e.g. 132]. Since the restriction on the vectors in the second definition (s-by-s block-sparsity) is more restricting than in the first one (sparsity), $\delta_{\mathcal{B},s}$ is smaller than δ_{s^2}. These different measures of restricted isometry are related, as shown in the following lemma

LEMMA 3.1.5. For a matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$, let $\mathbb{D}_\Omega \in \mathbb{R}^{N\times d^2}$ be defined row-wise by $\mathbb{D}_{\Omega\,(i,j)} = \mathrm{Vec}(X_j X_i^\top)$ for all $(i,j) \in \Omega$. It holds that

$$\Delta_{\Omega,s}(\mathbf{X}) = \delta_{\mathcal{B},s}(\mathbb{D}_\Omega)\,.$$

Proof. This is a direct consequence of the definition of $\mathbb{D}_\Omega$, which yields $\|\mathbf{X}^\top B\,\mathbf{X}\|_{F,\Omega}^2 = \|\mathbb{D}_\Omega\,\mathrm{Vec}(B)\|_2^2$, and $\|\mathrm{Vec}(B)\|_2^2 = \|B\|_F^2$.

□

The assumptions above guarantee that the matrix $\Theta^\star$ can be recovered from observations of the affinities, settling the well-posedness of this part of the inverse problem. However, we do not directly observe these affinities, but their image through the sigmoid function. We must therefore further impose the following assumption on the design matrix $\mathbf{X}$ that yields constraints on the probabilities π_{ij} and in essence governs the identifiability of $\Theta_\star$.

ASSUMPTION 3.1.6. The design matrix $\mathbf{X}$ satisfies $\|X_j X_i^\top\|_\infty \leqslant 1$, $(i,j) \in \Omega$.

In particular, under this assumption we have $\max_{(i,j)\in\Omega} |X_i^\top \Theta X_j| < M$ for all Θ in the class $\mathcal{P}(M)$, and a constant

$$\mathcal{L}(M) := \sigma'(M) = \sigma(M)\big(1-\sigma(M)\big)\,, \tag{3.1.4}$$

is lower bounded away from zero, and we have

$$\inf_{\Theta \in \mathcal{P}(M)} \sigma'(X_i^\top \Theta X_j) \geqslant \mathcal{L}(M) > 0\,, \tag{3.1.5}$$

for all $(i,j) \in \Omega$. Assuming that $\mathcal{L}(M)$ always depends on the same M, we sometimes write simply $\mathcal{L}$.

Remark 3.1.7. Assumption 3.1.6 is necessary for the identifiability of $\Theta_\star$: if $X_i^\top \Theta_\star X_j$ can be arbitrarily large in magnitude, $\pi_{ij} = \sigma(X_i^\top \Theta_\star X_j)$ can be arbitrarily close to 0 or 1. Since our observations only depend on $\Theta_\star$ through its image π_{ij}, this could lead to a very large estimation error on $\Theta_\star$ even with a small estimation error on the π_{ij}.

Remark 3.1.8. This assumption has already appeared in the literature on high-dimensional estimation, see [3, 135]. Similarly to [9], Assumption 3.1.6 can be shown to be redundant for minimax optimal prediction, because the log-likelihood function in the matrix logistic regression model satisfies the so-called self-concordant property. Our analysis to follow can be combined with an analysis similar to [9] to get rid of the assumption for minimax optimal prediction.

Random design

For random designs, we require the block isometry property to hold with high probability. Then the results in this article carry over directly and thus we do not discuss it in full detail. It is well known that for sparse linear models with the dimension of a target vector $\overline{p}$ and the sparsity $\overline{k}$, the classical restricted isometry property holds for some classes of random matrices with i.i.d. entries including sub-Gaussian and Bernoulli matrices, see [104], provided that $\overline{n} \gtrsim \overline{k}\log(\overline{p}/\overline{k})$, and i.i.d. subexponential random matrices, see [4], provided that $\overline{n} \gtrsim \overline{k}\log^2(\overline{p}/\overline{k})$. In the same spirit, design matrices with independent entries following sub-Gaussian, subexponential or Bernoulli distributions can be shown to satisfy the block isometry property, cf. [140], provided that the number of observed edges in the network satisfies $N \gtrsim k^2 \log^2(d/k)$ for sub-Gaussian and subexponential designs and $N \gtrsim k^2 \log(d/k)$ for Bernoulli designs.

3.2 Matrix Logistic Regression

The log-likelihood for this problem is

$$\ell_Y(\Theta) = -\sum_{(i,j)\in\Omega} \xi(s_{(i,j)} X_i^\top \Theta X_j)\,,$$

where $s_{(i,j)} = 2Y_{(i,j)} - 1$ is a sign variable that depends on the observations Y and $\xi : x \mapsto \log(1 + e^x)$ is a *softmax* function, convex on $\mathbb{R}$. As a consequence, the negative log-likelihood $-\ell_Y$ is a convex function of Θ. Denoting by ℓ the expectation $\mathbb{E}_{\Theta_\star}[\ell_Y]$, we recall the classical expressions for all $\Theta \in \mathbf{S}^d$

$$\begin{aligned}\ell(\Theta) &= \ell(\Theta_\star) - \sum_{(i,j)\in\Omega} \mathsf{KL}(\pi_{ij}(\Theta_\star), \pi_{ij}(\Theta)) \\ &= \ell(\Theta_\star) - \mathsf{KL}(\mathbf{P}_{\Theta_\star}, \mathbf{P}_\Theta)\,,\end{aligned}$$

where we recall $\pi_{ij}(\Theta) = \sigma(X_i^\top \Theta X_j)$, and

$$\ell_Y(\Theta) = \ell(\Theta) + \langle\!\langle\, \nabla\zeta, \Theta \,\rangle\!\rangle_F\,,$$

where ζ is a stochastic component of the log-likelihood with constant gradient $\nabla\zeta \in \mathbb{R}^{d\times d}$ given by $\nabla\zeta = \sum_{(i,j)\in\Omega}(Y_{(i,j)} - \pi_{ij}(\Theta_\star))X_j X_i^\top$, which is a sum of independent centered random variables.

3.2.1 Penalized logistic loss

In a classical setting where d is fixed and N grows, the maximiser of ℓ_Y - the maximum likelihood estimator - is an accurate estimator of $\Theta_\star$, provided that it is possible to identify Θ from $\mathbf{P}_\Theta$ (i.e. if the X_i are well conditioned). We are here in a high-dimensional setting where $d^2 \gg N$, and this approach is not directly possible. Our parameter space indicates that the intrinsic dimension of our problem is truly much lower in terms of rank and block-sparsity. Our assumption on the conditioning of the X_i is tailored to this structural assumption. In the same spirit, we also modify our estimator in order to promote the selection of elements of low rank and block-sparsity. Following the ideas of [30] and [3], we define the following penalized maximum likelihood estimator

$$\widehat{\Theta} \in \operatorname*{argmin}_{\Theta\in\mathcal{P}(M)} \Big\{ -\ell_Y(\Theta) + p(\Theta) \Big\}\,, \tag{3.2.1}$$

with a penalty p defined as

$$p(\Theta) = g(\operatorname{rank}(\Theta), \|\Theta\|_{0,0})\,, \quad \text{and} \quad g(R,K) = cKR + cK\log\big(\frac{de}{K}\big)\,, \tag{3.2.2}$$

where $c > 0$ is a universal constant and to be specified further. The proof of the following theorem is based on Dudley's integral argument combined with Bousquet's inequality and is deferred to the Appendix.

THEOREM 3.2.1. Assume the design matrix $\mathbf{X}$ satisfies $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta^\star$ in a given class, and the penalty term $p(\Theta)$ satisfies (3.2.2) with

the constants $c \geqslant c_1/\mathcal{L}$, $c_1 > 1$, $\mathcal{L}$ given in (3.1.4). Then for the penalised MLE estimator $\widehat{\Theta}$, the following non-asymptotic upper bound on the expectation of the Kullback-Leibler divergence between the measures $\mathbb{P}_{\Theta^\star}$ and $\mathbb{P}_{\widehat{\Theta}}$ holds

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \frac{1}{N}\mathbb{E}[\mathscr{K}_\circ(\mathbb{P}_{\Theta^\star}, \mathbb{P}_{\widehat{\Theta}})] \leqslant C_1 \frac{kr}{N} + C_1 \frac{k}{N} \log\big(\frac{de}{k}\big)\,, \tag{3.2.3}$$

where $C_1 > 3c$ is some universal constant for all $k = 1, ..., d$ and $r = 1, ..., k$.

Remark 3.2.2. Random designs with i.i.d. entries following sub-Gaussian, Bernoulli and subexponential distributions discussed in Section 3.1.4 yield the same rate as well. It can formally be shown using standard conditioning arguments [see, e.g. 110].

Corollary 3.2.3. Assume the design matrix $\mathbf{X}$ satisfies the block isometry property from Definition 3.1.3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta_\star X_j| < M$ for some $M > 0$ and all $\Theta^\star$ in a given class, and the penalty term $p(\Theta)$ is as in Theorem 3.2.1. Then for the penalised MLE estimator $\widehat{\Theta}$, the following non-asymptotic upper bound on the rate of estimation holds

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\widehat{\Theta} - \Theta^\star\|_F^2\big] \leqslant \frac{C_1}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbf{X})\big)}\Big(\frac{kr}{N} + \frac{k}{N}\log\big(\frac{de}{k}\big)\Big)\,,$$

where $C_1 > 3c$ is some universal constant for all $k = 1, ..., d$ and $r = 1, ..., k$.

Let us define rank-constrained maximum likelihood estimators with bounded block size as

$$\widehat{\Theta}_{k,r} \in \operatorname*{argmin}_{\Theta \in \mathcal{P}_{k,r}(M)} \{-\ell_Y(\Theta)\}\,.$$

It is intuitively clear that without imposing any regularisation on the likelihood function, the maximum likelihood approach selects the most complex model. In fact, the following result holds.

Theorem 3.2.4. Assume the design matrix $\mathbf{X}$ satisfies the block isometry property from Definition 3.1.3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta^\star X_j| < M$ for some $M > 0$ and all $\Theta^\star$ in a given class. Then for the maximum likelihood estimator $\widehat{\Theta}_{k,r}$, the following non-asymptotic upper bound on the rate of estimation holds

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\widehat{\Theta}_{k,r} - \Theta^\star\|_F^2\big] \leqslant \frac{C_3}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbf{X})\big)}\Big(\frac{kr}{N} + \frac{k}{N}\log\big(\frac{de}{k}\big)\Big)\,,$$

for all $k = 1, ..., d$ and $r = 1, ..., k$ and some constant $C_3 > 0$.

Remark 3.2.5. The penalty (3.2.2) belongs to the class of the so-called minimal penalties, cf. [30]. In particular, a naive MLE approach with $p(\Theta) = 0$ in (3.2.1) yields a suboptimal estimator as it follows from Theorem 3.2.4.

3.2.2 Convex relaxation

In practice, computation of the estimator (3.2.1) is often infeasible. In essence, in order to compute it, we need to compare the likelihood functions over all possible subspaces $\mathcal{P}_{k,r}(M)$. Sophisticated step-wise model selection procedures allow one to reduce the number of analysed models. However, they are not feasible in a high-dimensional setting either. We here consider the following estimator

$$\widehat{\Theta}_{Lasso} = \operatorname*{argmin}_{\Theta \in \mathbf{S}^d} \{ -\ell_Y(\Theta) + \lambda \|\Theta\|_{1,1} \}, \tag{3.2.4}$$

with $\lambda > 0$ to be chosen further, which is equivalent to the logistic Lasso on $\mathrm{Vec}(\Theta)$. Using standard arguments, cf. Example 1 in [135], combined with the block isometry property the following result immediately follows.

THEOREM 3.2.6. Assume the design matrix $\mathbf{X}$ satisfies the block isometry property from Definition 3.1.3 and $\max_{(i,j)\in\Omega} |X_i^\top \Theta^\star X_j| < M$ for some $M > 0$ and all $\Theta^\star$ in a given class. Then for $\lambda = C_4\sqrt{\log d}$, where $C_4 > 0$ is an appropriate universal constant, the estimator (3.2.4) satisfies

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\widehat{\Theta}_{Lasso} - \Theta^\star\|_F^2\big] \leqslant \frac{C_5}{\mathcal{L}(M)\big(1 - \Delta_{\Omega,2k}(\mathbf{X})\big)} \frac{k^2}{N} \log d, \tag{3.2.5}$$

for all $k = 1, ..., d$ and $r = 1, ..., k$ and some universal constant $C_5 > 0$.

As one could expect the upper bound on the rate of estimation of our feasible estimator is independent of the true rank r. It is natural, when dealing with a low-rank and block-sparse objective matrix, to combine the nuclear penalty with either the $(2,1)$-norm penalty or the $(1,1)$-norm penalty of a matrix, cf. [40, 69, 87, 122]. In our setting, it can be easily shown that combining the $(1,1)$-norm penalty and the nuclear penalty yields the same rate of estimation $(k^2/N)\log d$. This appears to be inevitable in view of a computational lower bound, obtained in Section 3.3, which is independent of the rank as well. In particular, these findings partially answer a question posed in Section 6.4.4 in [70].

3.2.3 Prediction

In applications, as new users join the network, we are interested in predicting the probabilities of the links between them and the existing users. It is natural to measure the prediction error of an estimator $\widehat{\Theta}$ by $\mathbb{E}\big[\sum_{(i,j)\in\Omega}(\pi_{ij}(\widehat{\Theta}) - \pi_{ij}(\Theta^\star))^2\big]$ which is controlled according to the following result using the smoothness of the logistic function σ.

THEOREM 3.2.7. Under Assumption 3.1.6, we have the following rate for estimating the

matrix of probabilities $\Sigma^\star = \mathbf{X}^\top \Theta^\star \mathbf{X} \in \mathbb{R}^{n\times n}$ with the estimator $\widehat{\Sigma} = \mathbf{X}^\top \widehat{\Theta}\mathbf{X} \in \mathbb{R}^{n\times n}$:

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \frac{1}{2N}\mathbb{E}\big[\|\widehat{\Sigma} - \Sigma^\star\|_{F,\Omega}^2\big] \leqslant \frac{C_1}{\mathcal{L}(M)}\Big(\frac{kr}{N} + \frac{k}{N}\log\big(\frac{d\mathrm{e}}{k}\big)\Big),$$

with the constant C_1 from (3.2.3). The rate is minimax optimal, i.e. a minimax lower bound of the same asymptotic order holds for the prediction error of estimating the matrix of probabilities $\Sigma^\star = \mathbf{X}^\top \Theta^\star \mathbf{X} \in \mathbb{R}^{n\times n}$.

3.2.4 Information-theoretic lower bounds

The following result demonstrates that the minimax lower bound on the rate of estimation matches the upper bound in Theorem 3.2.1 implying that the rate of estimation is minimax optimal.

THEOREM 3.2.8. Let the design matrix $\mathbf{X}$ satisfy the block isometry property. Then for estimating $\Theta^\star \in \mathcal{P}_{k,r}(M)$ in the matrix logistic regression model, the following lower bound on the rate of estimation holds

$$\inf_{\widehat{\Theta}} \sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \mathbb{E}\big[\|\widehat{\Theta} - \Theta^\star\|_F^2\big] \geqslant \frac{C_2}{(1 + \Delta_{\Omega,2k}(\mathbf{X}))}\Big(\frac{kr}{N} + \frac{k}{N}\log\big(\frac{d\mathrm{e}}{k}\big)\Big),$$

where the constant $C_2 > 0$ is independent of d, k, r and the infimum extends over all estimators $\widehat{\Theta}$.

REMARK 3.2.9. The lower bounds of the same order hold for the expectation of the Kullback-Leibler divergence between the measures $\mathbb{P}_{\Theta^\star}$ and $\mathbb{P}_{\widehat{\Theta}}$ and the prediction error of estimating the matrix of probabilities $\Sigma^\star = \mathbf{X}^\top \Theta^\star \mathbf{X} \in \mathbb{R}^{n\times n}$.

3.3 Computational lower bounds

In this section, we investigate whether the lower bound in Theorem 3.2.8 can be achieved with an estimator computable in polynomial time. The fastest rate of estimation attained by a (randomised) polynomial-time algorithm in the worst-case scenario is usually referred to as a *computational lower bound*. Recently, the gap between computational and statistical lower bounds has attracted a lot of attention in the statistical community. We refer to [24, 25, 47, 49, 66, 73, 96, 141, 145] for computational lower bounds in high-dimensional statistics based on the planted clique problem (see below), [22] using hardness of learning parity with noise [111] for denoising of sparse and low-rank matrices, [5] for computational trade-offs in statistical learning, as well as [147] for worst-case lower bounds for sparse estimators in linear regression, as well as [34, 46] for another approach on computational trade-offs in statistical problems, as well as [21, 23] on the management of these trade-offs.

In order to establish a computational lower bound for the block-sparse matrix logistic regression, we exploit a reduction scheme from [24]: we show that detecting a subspace of $\mathcal{P}_{k,r}(M)$ can be computationally as hard as solving the dense subgraph detection problem.

3.3.1 The dense subgraph detection problem

Although our work is related to the study of graphs, we recall for absolute clarity the following notions from graph theory. A *graph* $G = (V, E)$ is a non-empty set V of *vertices*, together with a set E of distinct unordered pairs $\{i, j\}$ with $i, j \in V$, $i \neq j$. Each element $\{i, j\}$ of E is an edge and joins i to j. The vertices of an edge are called its endpoints. We consider only undirected graphs with neither loops nor multiple edges. A graph is called *complete* if every pair of distinct vertices is connected. A graph $G' = (V', E')$ is a subgraph of a graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. A subgraph C is called a *clique* if it is complete. The problem of detecting a maximum clique, the so-called Clique problem, in a given graph is known to be NP-complete, cf. [82].

The Planted Clique problem, motivated as an average case version of the Clique problem, can be formalised as a decision problem over random graphs, parametrised by the number of vertices n and the size of the subgraph k. Let $\mathbb{G}_n$ denote the collection of all graphs with n vertices and $G(n, 1/2)$ denote distribution of Erdös-Rényi random graphs, uniform on $\mathbb{G}_n$, where each edge is drawn independently at random with probability $1/2$. For any $k \in \{1, ..., n\}$ and $q \in (1/2, 1]$, let $G(n, 1/2, k, q)$ be a distribution on $\mathbb{G}_n$ constructed by first picking k vertices independently at random and connecting all edges in-between with probability q, and then joining each remaining pair of distinct vertices by an edge independently at random with probability $1/2$. Formally, the Planted Clique problem refers to the hypothesis testing problem of

$$H_0 : A \sim G(n, 1/2) \quad \text{vs.} \quad H_1 : A \sim G(n, 1/2, k, 1)\,, \tag{3.3.1}$$

based on observing an adjacency matrix $A \in \mathbb{R}^{n \times n}$ of a random graph drawn from either $G(n, 1/2)$ or $G(n, 1/2, k, 1)$.

One of the main properties of the Erdös-Rényi random graph were studied in [61], as well as in [71], who in particular proved that the size of the largest clique in $G(n, 1/2)$ is asymptotically close to $2 \log_2 n$ almost surely. On the other hand, [6] proposed a spectral method that for $k > c\sqrt{n}$ detects a planted clique with high probability in polynomial time. Hence the most intriguing regime for k is

$$2 \log_2 n \leqslant k \leqslant c\sqrt{n}\,. \tag{3.3.2}$$

The conjecture that no polynomial-time algorithm exists for distinguishing between

hypotheses in (3.3.1) in the regime (3.3.2) with probability tending to 1 as $n \to \infty$ is the famous Planted Clique conjecture in complexity theory. Its variations have been used extensively as computational hardness assumptions in statistical problems, see [25, 41, 66, 141].

The Planted Clique problem can be reduced to the so-called dense subgraph detection problem of testing the null hypothesis in (3.3.1) against the alternative $H_1 : A \sim G(n, 1/2, k, q)$, where $q \in (1/2, 1]$. This is clearly a computationally harder problem. In this chapter, we assume the following variation of the Planted Clique conjecture which is used to establish a computational lower bound in the matrix logistic regression model.

Conjecture 3.3.1 (The dense subgraph detection conjecture, see [25, 41, 66, 141]). For any sequence $k = k_n$ such that $k \leqslant n^{\beta}$ for some $0 < \beta < 1/2$, and any $q \in (1/2, 1]$, there is no (randomised) polynomial-time algorithm that can correctly identify the dense subgraph with probability tending to 1 as $n \to \infty$, i.e. for any sequence of (randomised) polynomial-time tests $(\psi_n : \mathbb{G}_n \to \{0, 1\})_n$, we have

$$\liminf_{n\to\infty} \big\{\mathbb{P}_0(\psi_n(A) = 1) + \mathbb{P}_1(\psi_n(A) = 0)\big\} \geqslant 1/3\,.$$

3.3.2 Reduction to the dense subgraph detection problem and a computational lower bound

Consider the vectors of explanatory variables $X_i = N^{1/4} e_i$, $e_i \in \mathbb{R}^d$, $i = 1, ..., n$ and assume without loss of generality that the observed set of edges Ω in the matrix logistic regression model consists of the interactions of the n nodes X_i, i.e. it holds $N = |\Omega| = \binom{n}{2}$. It follows from the matrix logistic regression modelling assumption (3.1.1) that the Erdös-Rényi graph $G(n, 1/2)$ corresponds to a random graph associated with the matrix $\Theta_0 = 0 \in \mathbb{R}^{d\times d}$. Let $\mathcal{G}_l(k)$ be a subset of $\mathcal{P}_{k,1}(M)$ with a fixed support l of the block. In addition, let $\mathcal{G}_k^{\alpha_N} \subseteq \mathcal{P}_{k,1}(M)$ be a subset consisting of the matrices $\Theta_l \in \mathcal{G}_l(k), l = 1, ..., K$, $K = \binom{n}{k}$ such that all elements in the block of a matrix Θ_l equal some $\alpha_N = \alpha/\sqrt{N} > 0$, see Figure 3.2. Then we have

$$\mathbf{P}\big((i, j) \in E | X_i, X_j\big) = \frac{1}{1 + e^{-X_i^\top \Theta X_j}} = \frac{1}{1 + e^{-\alpha}}\,,$$

for all $\Theta \in \mathcal{G}_k^{\alpha_N}$. Therefore, the testing problem

$$H_0 : Y \sim \mathbb{P}_{\Theta_0} \quad \text{vs.} \quad H_1 : Y \sim \mathbb{P}_\Theta, \Theta \in \mathcal{G}_k^{\alpha_N}\,, \tag{3.3.3}$$

where $Y \in \{0, 1\}^N$ is the adjacency vector of binary responses in the matrix logistic regression model, is reduced to the dense subgraph detection problem with $q = 1/(1+e^{-\alpha})$. This reduction scheme suggests that the computational lower bound for separating the

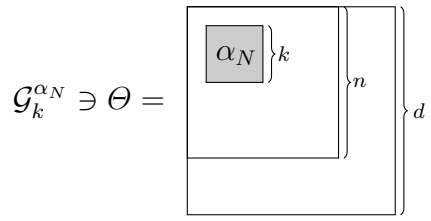

Figure 3.2: The construction of matrices $\mathcal{G}_k^{\alpha_N}$ used in the reduction scheme.

hypotheses in the dense subgraph detection problem mimics the computational lower bound for separating the hypotheses in (3.3.3) in the matrix logistic regression model. The following theorem exploits this fact in order to establish a computational lower bound of order k^2/N for estimating the matrix $\Theta^\star \in \mathcal{P}_{k,r}(M)$.

THEOREM 3.3.2. Let $\mathcal{F}_k$ be any class of matrices containing $\mathcal{G}_k^{\alpha_N} \cup \Theta_0$ from the reduction scheme. Let $c > 0$ be a positive constant and $f(k,d,N)$ be a real-valued function satisfying $f(k,d,N) \leqslant ck^2/N$ for $k = k_n < n^\beta$, $0 < \beta < 1/2$ and a sequence $d = d_n$, for all $n > m_0 \in \mathcal{N}$. If Conjecture 3.3.1 holds for some design $\mathbf{X}$ that fulfils the block isometry property from Definition 3.1.3, there is no estimator of $\Theta^\star \in \mathcal{F}_k$, that attains the rate $f(k,d,N)$ for the Frobenius norm risk, and can be evaluated using a (randomised) polynomial-time algorithm, i.e. for any estimator $\widehat{\Theta}$, computable in polynomial time, there exists a sequence $(k,d,N) = (k_n, d_n, N)$, such that

$$\frac{1}{f(k,d,N)} \sup_{\Theta^\star \in \mathcal{F}_k} \mathbb{E}\big[\|\widehat{\Theta} - \Theta^\star\|_F^2\big] \to \infty\,, \tag{3.3.4}$$

as $n \to \infty$. Similarly, for any estimator $\widehat{\Theta}$, computable in polynomial time, there exists a sequence $(k,d,N) = (k_n, d_n, N)$, such that

$$\frac{1}{f(k,d,N)} \sup_{\Theta^\star \in \mathcal{F}_k} \frac{1}{N}\mathbb{E}\big[\|\widehat{\Sigma} - \Sigma^\star\|_{F,\Omega}^2\big] \to \infty\,, \tag{3.3.5}$$

for the prediction error of estimating $\Sigma^\star = \mathbf{X}^\top \Theta^\star \mathbf{X}$.

REMARK 3.3.3. Thus the computational lower bound for estimating the matrix $\Theta^\star$ in the matrix logistic regression model is of order k^2/N compared to the minimax rate of estimation of order $kr/N + (k/N)\log(de/k)$ and the rate of estimation $(k^2/N)\log(d)$ for the Lasso estimator $\widehat{\Theta}_{Lasso}$, cf. Figure 3.1. Hence the computational gap is most noticeable for the matrices of rank 1. Furthermore, as a simple consequence of this result, the corresponding computational lower bound for the prediction risk of estimating $\Sigma^\star = \mathbf{X}^\top \Theta^\star \mathbf{X}$ is k^2/N as well.

Proof. We here provide a proof of the computational lower bound on the prediction error (3.3.5) for convenience. The bound on the estimation error (3.3.4) is straightforward to

show by utilizing the block isometry property. Assume that there exists a hypothetical estimator $\widehat{\Theta}$ computable in polynomial time that attains the rate $f(k,d,N)$ for the prediction error, i.e. such that it holds that

$$\limsup_{n\to\infty} \frac{1}{f(k,d,N)} \sup_{\Theta^\star\in\mathcal{F}_k} \frac{1}{N}\mathbb{E}\big[\|\mathbf{X}^\top(\widehat{\Theta}-\Theta^\star)\mathbf{X}\|^2_{F,\Omega}\big] \leqslant b < \infty\,,$$

for all sequences $(k,d,N)=(k_n,d_n,N)$ and a constant b. Then by Markov's inequality, we have

$$\frac{1}{N}\|\mathbf{X}^\top(\widehat{\Theta}-\Theta^\star)\mathbf{X}\|^2_{F,\Omega} \leqslant uf(k,d,N)\,, \tag{3.3.6}$$

for some numeric constant $u>0$ with probability $1-b/u$ for all $\Theta^\star\in\mathcal{F}_k$. Following the reduction scheme, we consider the design vectors $X_i = N^{1/4}e_i\,,\ i=1,...,n$ and the subset of edges Ω, such that

$$\frac{1}{N}\|\mathbf{X}^\top(\widehat{\Theta}-\Theta^\star)\mathbf{X}\|^2_{F,\Omega} = \sum_{(i,j)\in\Omega}(\widehat{\Theta}_{ij}-{\Theta^\star}_{ij})^2 = \|\widehat{\Theta}-\Theta^\star\|^2_{F,\Omega}\,, \tag{3.3.7}$$

for any $\Theta^\star\in\mathcal{G}_k^{\alpha_N}$. Note, that the design vectors $X_i = N^{1/4}e_i\,,\ i=1,...,n$ clearly satisfy Assumption 3.1.6. Thus, in order to separate the hypotheses

$$H_0 : Y\sim\mathbb{P}_0 \quad \text{vs.} \quad H_1 : Y\sim\mathbb{P}_\Theta, \Theta\in\mathcal{G}_k^{\alpha_N}\,, \tag{3.3.8}$$

it is natural to employ the following test

$$\psi(Y) = \mathbf{1}\big(\|\widehat{\Theta}\|_{F,\Omega}\geqslant\tau_{d,k}(u)\big)\,, \tag{3.3.9}$$

where $\tau^2_{d,k}(u) = uf(k,d,N)$. The type I error of this test is controlled automatically due to (3.3.6) and (3.3.7), $\mathbb{P}_0(\psi=1)\leqslant b/u$. For the type II error, we obtain

$$\begin{aligned}\sup_{\Theta\in\mathcal{G}_k^{\alpha_N}}\mathbb{P}_\Theta(\psi=0) &= \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}}\mathbb{P}_\Theta\big(\|\widehat{\Theta}\|_{F,\Omega}<\tau_{d,k}(u)\big)\\ &\leqslant \sup_{\Theta\in\mathcal{G}_k^{\alpha_N}}\mathbb{P}_\Theta\big(\|\widehat{\Theta}-\Theta\|^2_{F,\Omega}>\|\Theta\|^2_{F,\Omega}-\tau^2_{d,k}(u)\big)\leqslant b/u\,,\end{aligned}$$

provided that

$$k(k-1)\alpha_N^2/2 \geqslant 2\tau^2_{d,k}(u) = 2uf(k,d,N)\,,$$

which is true in the regime $k\leqslant n^\beta\,,\ \beta<1/2$, and $\alpha^2\geqslant 4u/c$, (hence $\alpha_N^2\geqslant 4u/(cN)$) by the definition of the function $f(k,d,N)$. Putting the pieces together, we obtain

$$\limsup_{n\to\infty}\big\{\mathbb{P}_0(\psi(Y)=1)+\sup_{\Theta\in\mathcal{G}_k^{\alpha_N}}\mathbb{P}_\Theta(\psi(Y)=0)\big\}\leqslant 2b/u<1/3\,,$$

for $u > 6b$. Hence, the test (3.3.9) separates the hypotheses (3.3.8). This contradicts Conjecture 3.3.1 and implies (3.3.5).

□

3.4 Concluding remarks

Our results shed further light on the emerging topic of statistical and computational trade-offs in high-dimensional estimation. The matrix logistic regression model is very natural to study the connection between statistical accuracy and computational efficiency as the model is based on the study of a generative model for random graphs. It is also an extension of lower bound *for all* statistical procedures to a model with covariates, the first of its kind.

Our findings suggest that the block-sparsity is a limiting model selection criterion for polynomial-time estimation in the logistic regression model. That is, imposing further structure, like an additional low-rank constraint, and thus reducing the number of studied models yields an expected gain in the minimax rate, but that gain can never be achieved by a polynomial-time algorithm. In this setting, this implies that with a larger parameter space, while the statistical rates might be worse, they might be closer to those that are computationally achievable. As an illustration, both efficient and minimax optimal estimation is possible for estimating sparse vectors in the high-dimensional linear regression model, see, for example, SLOPE for achieving the exact minimax rate in [19, 31], extending upon previous results on the Danzig selector and Lasso in [29, 43].

Logistic regression is also a representative of a large class of generalised linear models. Furthermore, the proof of the minimax lower bound on the rate of estimation in Theorem 3.2.8 can be extended to all generalised linear models. The combinatorial estimator (3.2.1) can well be used to achieve the minimax rate. The computational lower bound then becomes a delicate issue. A more sophisticated reduction scheme is needed to relate the dense subgraph detection problem to an appropriate testing problem for a generalised linear model. Approaching this question might require notions of noise discretisation and Le Cam equivalence studied in [96].

An interesting question is whether it is possible to adopt polynomial-time algorithms available for detecting a dense subgraph for estimating the target matrix in the logistic regression model in all sparsity regimes. A common idea behind those algorithms is to search a dense subgraph over the vertices of a high degree and thus substantially reduce the number of compared models of subgraphs. The network we observe in the logistic regression model is generated by a sparse matrix. We may still observe a fully connected network which is generated by a small block in the target matrix. Therefore, it is not yet clear how to adapt algorithms for dense subgraph detection to submatrix detection. It

remains an open question to establish whether these results can be extended to any design matrix, and all parameter regimes.

3.5 Proofs

3.5.1 Some geometric properties of the likelihood

Let us recall the *stochastic component* of the likelihood function

$$\zeta(\Theta) = L(\Theta) - \ell(\Theta) = \sum_{(i,j)\in\Omega} \big(Y_{(i,j)} - \pi_{ij}(\Theta^\star)\big) X_i^\top \Theta X_j \,,$$

which is a linear function in Θ. The deviation of the gradient $\nabla\zeta$ of the stochastic component is governed by the deviation of the independent Bernoulli random variables $\varepsilon_{i,j} = Y_{(i,j)} - \mathbb{E}[Y_{(i,j)}] = Y_{(i,j)} - \pi_{ij}(\Theta^\star)\,,\ (i,j)\in\Omega$. Let us introduce an upper triangular matrix $\mathcal{E}_\Omega = (\varepsilon_{i,j})_{(i,j)\in\Omega}$ with zeros on the complement set Ω^c. In this notation, we have $\zeta(\Theta) = \langle\!\langle\, \zeta, \Theta \,\rangle\!\rangle_F$, with

$$\nabla\zeta = \sum_{(i,j)\in\Omega} \varepsilon_{i,j} X_j X_i^\top = \mathbf{X}\,\mathcal{E}_\Omega \mathbf{X}^\top \in \mathbb{R}^{d\times d}\,.$$

In particular, $\nabla\zeta$ is sub-Gaussian with parameter $\sum_{(i,j)\in\Omega} \|X_j X_i^\top\|_F^2/4 = \|\mathbf{X}^\top\mathbf{X}\|_{F,\Omega}^2/4$, i.e. it holds for the moment generating function of $\langle\!\langle\, \zeta, B \,\rangle\!\rangle_F$ for any $B\in\mathbb{R}^{d\times d}$ and $\sigma^2 = 1/4$,

$$\begin{aligned}\varphi_{\langle\!\langle\, \zeta,B \,\rangle\!\rangle_F}(t) &:= \mathbb{E}\big[\exp(t\,\langle\!\langle\, \zeta, B \,\rangle\!\rangle_F)\big] = \prod_{(i,j)\in\Omega} \mathbb{E}\big[\exp(t\varepsilon_{i,j}\,\langle\!\langle\, X_j X_i^\top, B \,\rangle\!\rangle_F)\big] \\ &\leqslant \prod_{(i,j)\in\Omega} \exp\big(t^2\sigma^2\,\langle\!\langle\, X_j X_i^\top, B \,\rangle\!\rangle_F^{\,2}/2\big) = \exp\big(t\sigma^2\|\mathbf{X}^\top B\mathbf{X}\|_{F,\Omega}^2/2\big)\,. \quad (3.5.1)\end{aligned}$$

We shall be frequently using versions of the following inequality, which is based on the fact that $\nabla\ell(\Theta^\star) = 0 \in \mathbb{R}^{d\times d}$, the Taylor expansion and (3.1.5), and holds for any $\Theta\in\mathcal{P}(M)$,

$$\begin{aligned}\ell(\Theta^\star) - \ell(\Theta) &= \frac{1}{2}\sum_{(i,j)\in\Omega} \big(\sigma'(X_i^\top\Theta_0 X_j)\,\langle\!\langle\, X_j X_i^\top, \Theta^\star - \Theta \,\rangle\!\rangle^2\big) \\ &\geqslant \frac{\mathcal{L}}{2}\sum_{(i,j)\in\Omega} \langle\!\langle\, X_j X_i^\top, \Theta^\star - \Theta \,\rangle\!\rangle_F^2 = \frac{\mathcal{L}}{2}\|\mathbf{X}^\top(\Theta^\star - \Theta)\mathbf{X}\|_{F,\Omega}^2\,, \quad (3.5.2)\end{aligned}$$

where $\Theta_0\in[\Theta,\Theta^\star]$ element-wise. Furthermore, using that $\sup_{t\in\mathbb{R}}\sigma'(t) \leqslant 1/4$, we obtain for all $\Theta\in\mathcal{P}(M)$

$$\ell(\Theta^\star) - \ell(\Theta) \leqslant \frac{1}{8}\|\mathbf{X}^\top(\Theta^\star - \Theta)\mathbf{X}\|_{F,\Omega}^2\,.$$

We shall also be using the bounds

$$\max_{(i,j)\in\Omega} \big(\varepsilon_{i,j} X_i^T(\Theta - \Theta^\star)X_j\big) \leqslant \|\mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X}\|_{F,\Omega}\,, \quad \text{a.s.}\,, \quad (3.5.3)$$

$$\mathrm{Var}\left(\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle_F \right) \leqslant \frac{1}{4}\|\mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X}\|^2_{F,\Omega}\,. \tag{3.5.4}$$

3.5.2 Entropy bounds for some classes of matrices

Recall that an *ε-net* of a bounded subset $\mathbf{K}$ of some metric space with a metric ρ is a collection $\{K_1, ..., K_{N_\varepsilon}\} \in \mathbf{K}$ such that for each $K \in \mathbf{K}$, there exists $i \in \{1, ..., N_\varepsilon\}$ such that $\rho(K, K_i) \leqslant \varepsilon$. The *$\varepsilon$-covering number* $N(\varepsilon, \mathbf{K}, \rho)$ is the cardinality of the smallest ε-net. The *ε-entropy* of the class $\mathbf{K}$ is defined by $H(\varepsilon, \mathbf{K}, \rho) = \log_2 N(\varepsilon, \mathbf{K}, \rho)$. The following statement is adapted from Lemma 3.1 in [42].

LEMMA 3.5.1. Let $\top_0 := \{\Theta \in \mathbb{R}^{k\times k} : \mathrm{rank}(\Theta) \leqslant r, \|\Theta\|_F \leqslant 1\}$. Then it holds for any $\varepsilon > 0$

$$H(\varepsilon, \top_0, \|\bullet\|_F) \leqslant \big((2k+1)r + 1\big)\log\big(\frac{9}{\varepsilon}\big)\,.$$

3.5.3 Proof of Theorem 3.2.1 and Theorem 3.2.7

It suffices to show the following uniform deviation inequality

$$\sup_{\Theta^\star \in \mathcal{P}_{k,r}(M)} \mathbb{P}_{\Theta^\star}\big(\ell(\Theta^\star) - \ell(\widehat{\Theta}) + p(\widehat{\Theta}) > 2p(\Theta^\star) + R_t^2\big) \leqslant \mathrm{e}^{-cR_t}\,, \tag{3.5.5}$$

for any $R_t > 0$ and some numeric constant $c > 0$. Indeed, then taking $R_t^2 = p(\Theta^\star)$, it follows that $\ell(\Theta^\star) - \ell(\widehat{\Theta}) \leqslant 3p(\Theta^\star)$ uniformly for all $\Theta^\star$ in the considered class with probability at least $1 - \mathrm{e}^{-c\sqrt{\mathrm{pen}(\Theta^\star)}}$. The upper bound (3.2.3) of Theorem 3.2.1 the follows directly integrating the deviation inequality (3.5.5), while the upper bound on the prediction error in Theorem 3.2.7 further follows using (3.5.2) and the smoothness of the logistic function, $\sup_{t\in\mathbb{R}} \sigma'(t) \leqslant 1/4$. Define

$$\tau^2(\Theta; \Theta^\star) := \ell(\Theta^\star) - \ell(\Theta) + \mathrm{pen}(\Theta)\,,\ G_R(\Theta^\star) := \{\Theta : \tau(\Theta; \Theta^\star) \leqslant R\}\,. \tag{3.5.6}$$

The inequality (3.5.5) clearly holds on the event $\{\tau^2(\widehat{\Theta}; \Theta^\star) \leqslant 2\mathrm{pen}(\Theta^\star)\}$. In view of $L(\widehat{\Theta}) - \mathrm{pen}(\widehat{\Theta}) \geqslant L(\Theta^\star) - \mathrm{pen}(\Theta^\star)$, we have on the complement:

$$\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\widehat{\Theta} - \Theta^\star)\mathbf{X} \,\rangle\!\rangle \geqslant \ell(\Theta^\star) - \ell(\widehat{\Theta}) + \mathrm{pen}(\widehat{\Theta}) - \mathrm{pen}(\Theta^\star) \geqslant \frac{1}{2}\tau^2(\widehat{\Theta}; \Theta^\star)\,.$$

Therefore, for any $\Theta^\star \in \mathcal{P}_{k,r}(M)$, we have

$$\mathbb{P}_{\Theta^\star}\big(\tau^2(\widehat{\Theta}; \Theta^\star) > 2\mathrm{pen}(\Theta^\star) + R_t^2\big) \leqslant \mathbb{P}_{\Theta^\star}\Big(\sup_{\tau(\Theta;\Theta^\star)\geqslant R_t} \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\tau^2(\Theta; \Theta^\star)} \geqslant \frac{1}{2}\Big)\,.$$

We now apply the so-called "peeling device" (or "slicing" as it sometimes called in the literature). The idea is to "slice" the set $\tau(\Theta; \Theta^\star) \geqslant R_t$ into pieces on which the penalty

term $\mathrm{pen}(\Theta)$ is fixed and the term $\ell(\Theta^\star) - \ell(\Theta)$ is bounded. It follows,

$$\begin{aligned}
&\mathbb{P}_{\Theta^\star}\Big(\sup_{\tau(\Theta;\Theta^\star)\geqslant R_t} \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta-\Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\tau^2(\Theta;\Theta^\star)} \geqslant \frac{1}{2}\Big) \\
&\quad\leqslant \sum_{K=1}^{d}\sum_{R=1}^{K}\sum_{s=1}^{\infty} \mathbb{P}_{\Theta^\star}\Big(\sup_{\substack{\Theta\in G_{2^s R_t}(\Theta^\star) \\ k(\Theta)=K,\,\mathrm{rank}(\Theta)=R}} \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta-\Theta^\star)\mathbf{X} \,\rangle\!\rangle \geqslant \frac{1}{8}2^{2s}R_t^2\Big)\,. \qquad (3.5.7)
\end{aligned}$$

On the set $\{\Theta \in G_{2^s R_t}(\Theta^\star),\ k(\Theta)=K, \mathrm{rank}(\Theta)=R\}$, it holds by the definitions (3.5.6)

$$\ell(\Theta^\star) - \ell(\Theta) \leqslant 2^{2s}R_t^2 - \mathrm{pen}(K,R)\,,$$

and therefore using (3.5.2), this implies

$$\|\mathbf{X}^\top(\Theta^\star-\Theta)\mathbf{X}\|_{F,\Omega} \leqslant Z(K,R,s)\,, \quad Z^2(K,R,s) = \frac{2}{\mathcal{L}}\big(2^{2s}R_t^2 - \mathrm{pen}(K,R)\big)\,. \qquad (3.5.8)$$

Let us fix the location of the block, that is the support of a matrix $\Theta' \in \mathcal{G}_1 := \{\Theta \in \mathbb{R}^{d\times d} : k(\Theta)=K, \mathrm{rank}(\Theta)=R\}$ belongs to the upper-left block of size $K\times K$. Then following the lines of the proof of Lemma 3.5.1 and using the singular value decomposition, we derive

$$H(\varepsilon, \{\mathbf{X}^\top\Theta'\mathbf{X} : \Theta' \in \mathcal{G}_1, \|\mathbf{X}^\top\Theta'\mathbf{X}\|_{F,\Omega} \leqslant B\}, \|\bullet\|_{F,\Omega}) \leqslant \big((2K+1)R+1\big)\log\big(\frac{9B}{\varepsilon}\big)\,.$$

Consequently, for the set $\top := \{\mathbf{X}^\top(\Theta-\Theta^\star)\mathbf{X} : \Theta \in \mathbb{R}^{d\times d}, \mathrm{rank}(\Theta)=R, k(\Theta)=K, \|\mathbf{X}^\top(\Theta^\star-\Theta)\mathbf{X}\|_{F,\Omega} \leqslant Z(K,R,s)\}$, we obtain

$$H(\varepsilon, \top, \|\bullet\|_{F,\Omega}) \leqslant \big((2K+1)R+1\big)\log\Big(\frac{9Z(K,R,s)}{\varepsilon}\Big) + K\log\big(\frac{de}{K}\big)\,.$$

Denote $t(K,R) := \sqrt{KR} + \sqrt{K\log\big(\frac{de}{K}\big)}$. By Dudley's entropy integral bound, see [56] and [68] for a more recent reference, we then have

$$\begin{aligned}
&\mathbb{E}\big[\sup_{\substack{\Theta\in G_{2^s R_t}(\Theta^\star) \\ k(\Theta)=K,\,\mathrm{rank}(\Theta)=R}} \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta-\Theta^\star)\mathbf{X} \,\rangle\!\rangle \big] \leqslant C' \int_0^{Z(K,R,s)} \sqrt{H(\varepsilon,\top,\|\bullet\|_\Omega)}\,\mathrm{d}\varepsilon \\
&\quad\leqslant C''\sqrt{kr}\int_0^{9Z(K,R,s)} \sqrt{\log\Big(\frac{9Z(K,R,s)}{\varepsilon}\Big)}\,\mathrm{d}\varepsilon + 9C''Z(K,R,s)\sqrt{K\log\big(\frac{de}{K}\big)} \\
&\quad\leqslant CZ(K,R,s)t(K,R)\,,
\end{aligned}$$

for some universal constant $C>0$. Furthermore, by Bousquet's version of Talagrand's inequality, see Theorem 3.5.10, in view of the bounds (3.5.3) and (3.5.4), we have for all

$u > 0$

$$\mathbb{P}_{\Theta^\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta^\star) \\ k(\Theta)=K,\, \mathrm{rank}(\Theta)=R}} \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle \geqslant CZ(K,R,s)t(K,R) + \sqrt{\Big(\frac{1}{2}Z^2(K,R,s) + 4CZ^2(K,R,s)t(K,R)\Big)u} + \frac{Z(K,R,s)u}{3}\Big) \leqslant \mathrm{e}^{-u}\,.$$

Taking $u(K,R,s) := \mathcal{L}^{1/2}Z(K,R,s) + \mathcal{L}^{-1/2}t(K,R) + 2\log d$ and using inequalities $\sqrt{c_1 + c_2} \leqslant \sqrt{c_1} + \sqrt{c_2}$ and $\sqrt{c_1 c_2} \leqslant \frac{1}{2}(c_1\varepsilon + \frac{c_2}{\varepsilon})$, which hold for any $c_1, c_2, \varepsilon > 0$, we obtain

$$\mathbb{P}_{\Theta^\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta^\star) \\ k(\Theta)=K,\, \mathrm{rank}(\Theta)=R}} \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle \geqslant \frac{1}{16}\mathcal{L}Z^2(K,R,s) + C_1^2 t^2(K,R)/\mathcal{L}\Big) \leqslant \mathrm{e}^{-u(K,R,s)}\,,$$

for some numeric constant $C_1 > 0$. Plugging this back into (3.5.7) and using (3.5.8), we obtain

$$\mathbb{P}_{\Theta^\star}\Big(\sup_{\substack{\Theta \in G_{2^s R_t}(\Theta^\star) \\ k(\Theta)=K,\, \mathrm{rank}(\Theta)=R}} \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle \geqslant \frac{1}{8}2^{2s}R_t^2\Big) \leqslant \mathrm{e}^{-u(K,R,s)}\,,$$

for some numeric constant $C_2 > 0$, provided that

$$\frac{1}{16}\mathcal{L}Z^2(K,R,s) + \frac{C_1^2}{\mathcal{L}}t^2(K,R) \leqslant \frac{1}{8}2^{2s}R_t^2 = \frac{1}{16}\mathcal{L}Z^2(K,R,s) + 8\mathrm{pen}(K,R)\,, \qquad (3.5.9)$$

which is satisfied for $\mathrm{pen}(K,R) \geqslant (C_1^2/\mathcal{L})t^2(K,R)$. Therefore,

$$\mathbb{P}_{\Theta^\star}\Big(\sup_{\tau(\Theta;\Theta^\star) \geqslant R_t} \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\tau^2(\Theta;\Theta^\star)} \geqslant \frac{1}{2}\Big) \leqslant \sum_{K=1}^{d}\sum_{R=1}^{K}\sum_{s=1}^{\infty} \mathrm{e}^{-u(K,R,s)} \leqslant \mathrm{e}^{-cR_t}\,,$$

for some numeric constants $c > 0$ using (3.5.9), which concludes the proof.

The following prominent result is due to [32].

THEOREM 3.5.2 (Bousquet's version of Talagrand's inequality). Let $(B, \mathcal{B})$ be a measurable space and let $\varepsilon_1, ..., \varepsilon_n$ be independent B-valued random variables. Let $\mathcal{F}$ be a countable set of measurable real-valued functions on B such that $f(\varepsilon_i) \leqslant b < \infty$ a.s. and $\mathbb{E}f(\varepsilon_i) = 0$ for all $i = 1, ..., n$, $f \in \mathcal{F}$. Let

$$S := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(\varepsilon_i)\,, \qquad v := \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}[f^2(\varepsilon_i)].$$

Then for all $u > 0$, it holds that

$$\mathbb{P}\Big(S - \mathbb{E}[S] \geqslant \sqrt{2(v + 2b\mathbb{E}[S])u} + \frac{bu}{3}\Big) \leqslant \mathrm{e}^{-u}\,. \tag{3.5.10}$$

3.5.4 Proof of Theorem 3.2.4

For the MLE $\widehat{\Theta}_{k,r}$, it clearly holds $L(\widehat{\Theta}_{k,r}) \geqslant L(\Theta^\star)$ implying

$$\ell(\Theta^\star) - \ell(\widehat{\Theta}_{k,r}) \leqslant \langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\widehat{\Theta}_{k,r} - \Theta^\star)\mathbf{X} \,\rangle\!\rangle\,.$$

Furthermore, in view of (3.5.2), we derive

$$\frac{\mathcal{L}}{2}\|\mathbf{X}^\top(\widehat{\Theta}_{k,r} - \Theta^\star)\mathbf{X}\|_{F,\Omega} \leqslant \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\widehat{\Theta}_{k,r} - \Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\|\mathbf{X}^\top(\widehat{\Theta}_{k,r} - \Theta^\star)\mathbf{X}\|_{F,\Omega}} \tag{3.5.11}$$

$$\leqslant \sup_{\Theta \in \mathcal{P}_{k,r}(M)} \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\|\mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X}\|_{F,\Omega}}\,. \tag{3.5.12}$$

Following the lines of Section 3.5.3, by Dudley's integral we next obtain

$$\mathbb{E}\Big(\sup_{\Theta \in \mathcal{P}_{k,r}(M)} \frac{\langle\!\langle\, \mathcal{E}_\Omega, \mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X} \,\rangle\!\rangle}{\|\mathbf{X}^\top(\Theta - \Theta^\star)\mathbf{X}\|_{F,\Omega}}\Big) \leqslant c\sqrt{kr} + c\sqrt{k\log\big(\tfrac{d\mathrm{e}}{k}\big)}\,,$$

for some universal constant $c > 0$. Plugging this bound back into (3.5.12) and using the block isometry property yields the desired assertion.

3.5.5 Proof of Theorem 3.2.8

Proof. The proof is split into two parts. First, we show a lower bound of the order kr and then a lower bound of the order $k\log(d\mathrm{e}/k)$. A simple inequality $(a+b)/2 \leqslant \max\{a,b\}$ for all $a, b > 0$ then completes the proof. Both parts of the proof exploit a version of remarkable Fano's inequality given in Proposition 3.5.3 to follow, cf. Section 2.7.1 in [134].

1. A bound kr. The proof of this bound is similar to the proof of a minimax lower bound for estimating a low-rank matrix in the trace-norm regression model given in Theorem 5 in [87]. For the sake of completeness, we provide the details here. Consider a subclass of matrices

$$\mathcal{C} = \Big\{\ A \in \mathbb{R}^{k\times r} : a_{i,j} = \{0, \alpha_N\}, 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant r\ \Big\}\,,$$
$$\alpha_N^2 = \frac{\gamma \log 2}{1 + \Delta_{\Omega,2k}(\mathbf{X})}\frac{r}{2kN}\,,$$

where $\gamma > 0$ is a positive constant, $\Delta_{\Omega,2k}(\mathbf{X}) > 0$ is the the block isometry constant from

Definition 3.1.3 and $\lfloor x \rfloor$ denotes the integer part of x. Further define

$$\mathcal{B}(\mathcal{C}) = \big\{ \frac{1}{2}(A + A^\top) : A = (\widetilde{A}| \cdots |\widetilde{A}|O) \in \mathbb{R}^{k \times k}, \widetilde{A} \in \mathcal{C} \big\} ,$$

where O denotes the $k \times (k - r\lfloor k/r \rfloor)$ zero matrix. By construction, any matrix $\Theta \in \mathcal{B}(\mathcal{C})$ is symmetric, has rank at most r with entries bounded by α_N. Applying a standard version of the Varshamov-Gilbert lemma, see Lemma 2.9 in [134], there exists a subset $\mathcal{B}^\circ \subseteq \mathcal{B}(\mathcal{C})$ of cardinality $\mathrm{card}(\mathcal{B}^\circ) \geqslant 2^{kr/16} + 1$ such that

$$\frac{kr}{16}\big(\frac{\alpha_N}{2}\big)^2 \big\lfloor \frac{k}{r} \big\rfloor \leqslant \|\Theta_u - \Theta_v\|_F^2 \leqslant k^2 \alpha_N^2 ,$$

for all $\Theta_u, \Theta_v \in \mathcal{B}^\circ$. Thus $\mathcal{B}^\circ$ is a 2δ-separated set in the Frobenius metric with $\delta^2 = \frac{kr}{64}\big(\frac{\alpha_N}{2}\big)^2 \lfloor \frac{k}{r} \rfloor$. The Kullback-Leibler divergence between the measures $\mathbb{P}_{\Theta_u}$ and $\mathbb{P}_{\Theta_v}$, $\Theta_u, \Theta_v \in \mathcal{B}^\circ$, $u \neq v$, is upper bounded as

$$\begin{aligned} \mathscr{K}_\circ(\mathbb{P}_{\Theta_u}, \mathbb{P}_{\Theta_v}) = \mathbb{E}_{\mathbb{P}_{\Theta_u}}[L(\Theta_u)] - \mathbb{E}_{\mathbb{P}_{\Theta_u}}[L(\Theta_v)] &\leqslant \frac{1}{8} \sum_{(i,j) \in \Omega} \langle\!\langle\, X_j X_i^\top, \Theta_u - \Theta_v \,\rangle\!\rangle_F^2 \\ &\leqslant \frac{1 + \Delta_{\Omega,2k}(\mathbf{X})}{8} k^2 \alpha_N^2 N . \end{aligned}$$

Taking $\gamma > 0$ small enough, we obtain

$$\frac{1 + \Delta_{\Omega,2k}(\mathbf{X})}{8} k^2 \alpha_N^2 N + \log 2 = \frac{kr}{16} \gamma \log 2 + \log 2 = \log(2^{\frac{kr}{16}\gamma + 1}) < \log(2^{kr/16} + 1) ,$$

which, in view of Proposition 3.5.3, yields the desired lower bound.

2. A bound $k \log(de/k)$. Let $K = \binom{d}{k}$ and consider the set $\mathcal{G}_k^{\alpha_N} \subseteq \mathcal{P}_{k,1}(M)$ from the reduction scheme in Section 3.3.2 with

$$\alpha_N^2 = \frac{4\gamma \log 2}{kN(1 + \Delta_{\Omega,2k}(\mathbf{X}))} \log\big(\frac{de}{k}\big) ,$$

where $\gamma > 0$ is a positive constant. Using simple calculations, we then have $(2k-1)\alpha_N^2 \leqslant \|\Theta_u - \Theta_v\|_F^2 \leqslant 2k^2\alpha_N^2$ for all $\Theta_u, \Theta_v \in \mathcal{G}_k^{\alpha_N}$, $u \neq v$. Furthermore, according to Lemma 3.5.4 to follow, there exists a subset $\mathcal{G}_k^{\alpha_N, 0} \subseteq \mathcal{G}_k^{\alpha_N}$ such that

$$c_0 k^2 \alpha_N^2 \leqslant \|\Theta_u - \Theta_v\|_F^2 \leqslant 2k^2 \alpha_N^2 ,$$

and of cardinality $\mathrm{card}(\mathcal{G}_k^{\alpha_N, 0}) \geqslant 2^{\rho k \log(de/k)} + 1$ for some $\rho > 0$ depending on a constant $c_0 > 0$ and independent of k and d. Thus $\mathcal{G}_k^{\alpha_N, 0}$ is a 2δ-separated set in the Frobenius metric with $\delta^2 = c_0 k^2 \alpha_N^2 / 4$. The Kullback-Leibler divergence between the measures $\mathbb{P}_{\Theta_u}$

and $\mathbb{P}_{\Theta_v}$, $\Theta_u, \Theta_v \in \mathcal{G}_k^{\alpha_N,0}$, $u \neq v$, is upper bounded as

$$\mathscr{K}_\circ(\mathbb{P}_{\Theta_u}, \mathbb{P}_{\Theta_v}) = \mathbb{E}_{\mathbb{P}_{\Theta_u}}[L(\Theta_u)] - \mathbb{E}_{\mathbb{P}_{\Theta_u}}[L(\Theta_v)] \leqslant \frac{1}{8} \sum_{(i,j)\in\Omega} \langle\!\langle\, X_j X_i^\top, \Theta_u - \Theta_v \,\rangle\!\rangle_F^2$$
$$\leqslant \frac{1 + \Delta_{\Omega,2k}(\mathbf{X})}{4} k^2 \alpha_N^2 N\,,$$

for all $u \neq v$ and $\Delta_{\Omega,2k}(\mathbf{X}) > 0$ from Definition 3.1.3. As in the first part of the proof, taking $\gamma > 0$ small enough, we obtain

$$\frac{1 + \Delta_{\Omega,2k}(\mathbf{X})}{4} k^2 \alpha_N^2 N + \log 2 = k\gamma \log(2) \log\big(\frac{de}{k}\big) + \log 2 = \log(2^{k\gamma \log(de/k)+1})$$
$$< \log(2^{\rho k \log(de/k)} + 1)\,.$$

The desired lower bound then follows from Proposition 3.5.3. □

Proposition 3.5.3 (Fano's method). Let $\{\Theta_1, ..., \Theta_J\}$ be a 2δ-separated set in $\mathbb{R}^{d\times d}$ in the Frobenius metric, meaning that $\|\Theta_k - \Theta_l\|_F \geqslant 2\delta$ for all elements Θ_k, Θ_l, $l \neq k$ in the set. Then for any increasing and measurable function $F : [0,\infty) \to [0,\infty)$, the minimax risk is lower bounded as

$$\inf_{\widehat{\Theta}} \sup_{\Theta} \mathbb{E}_{\mathbb{P}_\Theta}\big[F(\|\widehat{\Theta} - \Theta\|_F)\big] \geqslant F(\delta)\big(1 - \frac{\sum_{u,v} \mathscr{K}_\circ(\mathbb{P}_{\Theta_u}, \mathbb{P}_{\Theta_v})/J^2 + \log 2}{\log J}\big)\,.$$

Lemma 3.5.4 (Variant of the Varshamov-Gilbert lemma). Let $\mathcal{G} \subseteq \mathcal{P}_{k,1}(M)$ be a set of $\{0,1\}^{d\times d}$ symmetric block-sparse matrices with the size of the block k, where $k \leqslant \alpha\beta d$ for some $\alpha, \beta \in (0,1)$. Denote $K = \binom{d}{k}$ the cardinality of $\mathcal{G}$ and $\rho_H(E, E') = \sum_{i,j} \mathbf{1}(E_{i,j} \neq E'_{i,j})$ the Hamming distance between two matrices $E, E' \in \mathcal{G}$. Then there exists a subset $\mathcal{G}^0 = \{E^{(0)}, ..., E^{(J)}\} \subseteq \mathcal{G}$ of cardinality

$$\log J := \log(\mathrm{card}(\mathcal{G}^0)) \geqslant \rho k \log(\frac{de}{k})\,,$$

where $\rho = \frac{\alpha}{-\log(\alpha\beta)}(-\log\beta + \beta - 1)$ such that

$$\rho_H(E^{(k)}, E^{(l)}) \geqslant ck^2\,,$$

for all $k \neq l$ where $c = 2(1 - \alpha^2) \in (0,2)$.

Proof. Let $E^{(0)} = \{0\}^{k\times k}$, $D = ck^2$, and construct the set $\mathcal{E}_1 = \{E \in \mathcal{G} : \rho_H(E^{(0)}, E) > D\}$. Next, pick any $E^{(1)} \in \mathcal{E}_1$ and proceed iteratively so that for a matrix $E^{(j)} \in \mathcal{E}_j$ we construct the set

$$\mathcal{E}_{j+1} = \{E \in \mathcal{E}_j : \rho(E^{(j)}, E) > D\}\,.$$

Let J denote the last index j for which $\mathcal{E}_j \neq \varnothing$. It remains to bound the cardinality J of the constructed set $\mathcal{G}^0 = \{E^{(0)}, ..., E^{(J)}\}$. For this, we consider the cardinality n_j of the subset $\{\mathcal{E}_j \setminus \mathcal{E}_{j+1}\}$:

$$n_j := \#\{\mathcal{E}_j \setminus \mathcal{E}_{j+1}\} \leqslant \#\{E \in \mathcal{G} : \rho_H(E^{(j)}, E) \leqslant D\}.$$

For all $E, E' \in \mathcal{G}$, we have

$$\rho_H(E, E') = 2(k^2 - (k-m)^2),$$

where $m \in [0, k]$ corresponds to the number of distinct columns of E (or E'). Solving the quadratic equation (3.5.5) for $\rho_H(E, E') = D = ck^2$ we obtain

$$m_D = k(1 - \sqrt{1 - c/2}),$$

for the maximum number of distinct columns of a block-sparse matrix E (and E') such that $\rho_H(E, E') \leqslant D = ck^2$ for $c \in [0, 2]$. For instance, in order to get the distance between matrices $2k^2$, i.e. $c = 2$ we need to shift all the k columns (and consequently rows) and so the number of distinct columns of a matrix is $m = k$, and in order to get the minimal possible distance $4k - 2$, i.e. $c = (4k-2)/k^2$ we need to shift only one column and a corresponding row, i.e. $m = 1$. Therefore, for n_j in (3.5.5), we have

$$n_j \leqslant \#\{E \in \mathcal{G} : \rho_H(E^{(j)}, E) \leqslant D\} = \sum_{i=0}^{m_D} \binom{k}{i}\binom{d-k}{i} = \sum_{i=k-m_D}^{k} \binom{k}{i}\binom{d-k}{k-i}.$$

Together with an evident equality $\sum_{j=0}^{J} n_j = K = \binom{d}{k}$, this implies

$$\sum_{i=k-m_D}^{k} \binom{k}{i}\binom{d-k}{k-i} \Big/ \binom{d}{k} \geqslant \frac{1}{J+1}.$$

Note that taking $m_D = k$, which as we have seen corresponds to $c = 2$, we have a trivial bound $J + 1 \geqslant 1$ using Vandermonde's convolution. Furthermore, the expression on the left-hand side in (3.5.5) is exactly the probability $\mathbb{P}(X \geqslant k - m_D) = \mathbb{P}(X \geqslant k\alpha)$ for $\alpha = \sqrt{1 - c/2}$, where the variable X follows the hypergeometric distribution $H(d, k, k/d)$. The rest of the proof is based on applying Chernoff's inequality and follows the scheme of the proof of Lemma 4.10 in [100]. □

Bibliography

[1] Abbe, E. (2017): "Community detection and stochastic block models: Recent developments," *arXiv:1703.10146*.

[2] Abbe, E., and C. Sandon (2015): "Detection in the stochastic block model with multiple clusters: Proof of the achievability conjectures, acyclic BP, and the information-computation gap," *arXiv:1512.09080*.

[3] Abramovich, F., and V. Grinshtein (2016): "Model selection and minimax estimation in generalized linear models," *IEEE Transactions on Information Theory*, 62(6), 3721–3730.

[4] Adamczak, R., A. Litvak, A. Pajor, and N. Tomczak-Jaegermann (2011): "Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling," *Constructive Approximation*, 34(1), 61–88.

[5] Agarwal, A. (2012): "Computational trade-offs in statistical learning," Thesis, University of California, Berkeley, `http://research.microsoft.com/en-us/um/people/alekha/thesismain.pdf`.

[6] Alon, N., M. Krivelevich, and B. Sudakov (1998): "Finding a large hidden clique in a random graph," *Random Structures and Algorithms*, 13(3-4), 457–466.

[7] Alon, N. K. M., and B. Sudakov (1998): "Finding a large hidden clique in a random graph," in *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, vol. 13, pp. 457–466.

[8] Arndt, J., and C. Haenel (2001): *Pi-unleashed*. Springer.

[9] Bach, F. (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

[10] Baddeley, A., and R. Turner (2005): "spatstat: An R package for analyzing spatial point patterns," *Journal of Statistical Software*, 12(6), 1–42.

[11] Baldin, N. (2017): "The wrapping hull and a unified framework for estimating the volume of a body," arXiv:1502.05510.

[12] Baldin, N. (2018): "Estimation of the volume of a convex body," To appear in Oberwolfach snapshots of modern mathematics.

[13] Baldin, N., and Q. Berthet (2018): "Optimal link prediction with matrix logistic regression," Submitted.

[14] Baldin, N., and M. Reiss (2016): "Unbiased estimation of the volume of a convex body," *Stochastic Processes and their Applications (to appear)*, arXiv:1502.05510.

[15] Banks, J., C. Moore, J. Neeman, and P. Netrapalli (2016): "Information-theoretic thresholds for community detection in sparse networks," *arXiv:1601.02658*.

[16] Bárány, I., and D. Larman (1988): "Convex bodies, economic cap coverings, random polytopes," *Mathematika*, 35(2), 274–291.

[17] Bárány, I., and M. Reitzner (2010): "On the variance of random polytopes," *Advances in Mathematics*, 225(4), 1986–2001.

[18] Beermann, M., and M. Reitzner (2015): "Beyond the Efron-Buchta identities: Distributional results for Poisson polytopes," *Discrete & Computational Geometry*, 53(1), 226–244.

[19] Bellec, P., G. Lecué, and A. Tsybakov (2018): "Slope meets Lasso: improved oracle bounds and optimality," *Annals of Statistics (to appear)*, arXiv:1605.08651.

[20] Bellet, A., A. Habrard, and M. Sebban (2014): "A survey on metric learning for feature vectors and structured data," .

[21] Berthet, Q., and V. Chandrasekaran (2016): "Resource allocation for statistical estimation," *Proceedings of the IEEE*, 104(1), 115–125.

[22] Berthet, Q., and J. Ellenberg (2015): "Detection of planted solutions for flat satisfiability problems," .

[23] Berthet, Q., and V. Perchet (2017): "Fast rates for bandit optimization with upper-confidence Frank-Wolfe," *NIPS 2017, to appear*.

[24] Berthet, Q., and P. Rigollet (2013a): "Complexity theoretic lower bounds for sparse principal component detection," in *Conference on Learning Theory*, pp. 1046–1066.

[25] ——— (2013b): "Optimal detection of sparse principal components in high dimension," *The Annals of Statistics*, 41(4), 1780–1815.

[26] Berthet, Q., P. Rigollet, and P. Srivastava (2016): "Exact recovery in the Ising blockmodel," .

[27] Biau, G., and K. Bleakly (2008): "Statistical inference on graphs," *Statistics and decisions*, 24(2), 209–232.

[28] Bickel, P. J., and A. Chen (2009): "A nonparametric view of network models and Newman–Girvan and other modularities," *Proceedings of the National Academy of Sciences*, 106(50), 21068–21073.

[29] Bickel, P. J., Y. Ritov, and A. Tsybakov (2009): "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, 37(4), 1705–1732.

[30] Birgé, L., and P. Massart (2007): "Minimal penalties for Gaussian model selection," *Probability theory and related fields*, 138(1-2), 33–73.

[31] Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. Candès (2015): "SLOPE—adaptive variable selection via convex optimization," *The Annals of Applied Atatistics*, 9(3), 1103.

[32] Bousquet, O. (2002): "A Bennett concentration inequality and its application to suprema of empirical processes," *Comptes Rendus Mathematique*, 334(6), 495–500.

[33] Bretscher, O. (2013): *Linear algebra with applications*. Pearson Education.

[34] Bruer, J. J., J. A. Tropp, V. Cevher, and S. R. Becker (2015): "Designing statistical estimators that balance sample size, risk, and computational cost," *IEEE Journal of Selected Topics in Signal Processing*, 9(4), 612–624.

[35] Brunel, V.-E. (2013): "A universal deviation inequality for random polytopes," arXiv:1311.2902.

[36] ——— (2014): "Non-parametric estimation of convex bodies and convex polytopes," Theses, Université Pierre et Marie Curie - Paris VI ; University of Haifa, `https://tel.archives-ouvertes.fr/tel-01066977`.

[37] Bubeck, S., J. Ding, R. Eldan, and M. Rácz (2014): "Testing for high-dimensional geometry in random graphs," .

[38] Buchta, C. (2013): "Exact formulae for variances of functionals of convex hulls," *Adv. in Appl. Probab.*, 45(4), 917–924.

[39] Bunea, F. (2008): "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization," *Electronic Journal of Statistics*, 2, 1153–1194.

[40] Bunea, F., Y. She, and M. Wegkamp (2012): "Joint variable and rank selection for parsimonious estimation of high-dimensional matrices," *The Annals of Statistics*, 40(5), 2359–2388.

[41] Cai, T., and Y. Wu (2018): "Statistical and computational limits for sparse matrix detection," *arXiv preprint arXiv:1801.00518*.

[42] Candes, E., and Y. Plan (2011): "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, 57(4), 2342–2359.

[43] Candes, E., and T. Tao (2007): "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, 35(6), 2313–2351.

[44] Candes, E. J., and T. Tao (2005): "Decoding by linear programming," *IEEE Trans. Information Theory*, 51, 4203–4215.

[45] Chan, T. M. (1996): "Optimal Output-Sensitive Convex Hull Algorithms in Two and Three Dimensions," *Discrete and Computational Geometry*, 16, 361–368.

[46] Chandrasekaran, V., and M. I. Jordan (2013): "Computational and statistical tradeoffs via convex relaxation," *Proceedings of the National Academy of Sciences*.

[47] Chen, Y. (2015): "Incoherence-optimal matrix completion," *IEEE Transactions on Information Theory*, 61(5), 2909–2923.

[48] Chen, Y., E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti (2009): "Similarity-based classification: Concepts and algorithms," *J. Mach. Learn. Res.*, 10, 747–776.

[49] Chen, Y., and J. Xu (2016): "Statistical-computational trade-offs in planted problems and submatrix localization with a growing number of clusters and submatrices," *Journal of Machine Learning Research*, 17(27), 1–57.

[50] Cuevas, A., and R. Fraiman (1997): "A plug-in approach to support estimation," *The Annals of Statistics*, 25(6), 2300–2312.

[51] Cuevas, A., R. Fraiman, and B. Pateiro-López (2012): "On statistical properties of sets fulfilling rolling-type conditions," *Advances in Applied Probability*, 44(2), 311–329.

[52] Davis, R., E. Mulrow, and S. Resnick (1987): "The convex hull of a random sample in $\mathbb{R}^2$," *Stochastic Models*, 3(1), 1–27.

[53] Decelle, A., F. Krzakala, C. Moore, and L. Zdeborová (2011): "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, 84(6), 066106.

[54] Devroye, L., A. György, G. Lugosi, and F. Udina (2011): "High-dimensional random geometric graphs and their clique number," *Electronic Communications in Probability*, 16(90), 2481–2508.

[55] Devroye, L., and G. Wise (1980): "Detection of abnormal behavior via nonparametric estimation of the support," *SIAM Journal on Applied Mathematics*, 38(3), 480–488.

[56] Dudley, R. (1967): "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes," *Journal of Functional Analysis*, 1(3), 290 – 330.

[57] Dwyer, R. (1988): "On the convex hull of random points in a polytope," *Journal of Applied Probability*, 25(4), 688–699.

[58] Dyer, M., A. Frieze, and R. Kannan (1991): "A random polynomial-time algorithm for approximating the volume of convex bodies," *Journal of the ACM (JACM)*, 38(1), 1–17.

[59] Efimov, N., and S. Stechkin (1959): "Some supporting properties of sets in Banach spaces as related to Chebyshev sets," *Doklady Akademii Nauk SSSR*, 127(2), 254–257.

[60] Efron, B. (1965): "The convex hull of a random set of points," *Biometrika*, 52(3-4), 331–343.

[61] Erdős, P., and A. Rényi (1959): "On random graphs," *Publicationes Mathematicae*, 6, 290–297.

[62] Erickson, J. (1996): "New lower bounds for convex hull problems in odd dimensions," in *Symposium on Computational Geometry*.

[63] Fan, J., W. Gong, and Z. Zhu (2017): "Generalized high-dimensional trace regression via nuclear norm regularization," *arXiv preprint arXiv:1710.08083*.

[64] Fan, J., W. Wang, and Z. Zhu (2016): "Robust low-rank matrix recovery," *arXiv preprint arXiv:1603.08315*.

[65] Gao, C., Y. Lu, and H. Zhou (2015): "Rate-optimal graphon estimation," *The Annals of Statistics*, 43(6), 2624–2652.

[66] Gao, C., Z. Ma, and H. Zhou (2017): "Sparse CCA: Adaptive estimation and computational barriers," *arXiv preprint arXiv:1409.8565*.

[67] Gayraud, G. (1997): "Estimation of functionals of density support," *Mathematical Methods of Statistics*, 6(1), 26–46.

[68] Giné, E., and R. Nickl (2016): *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[69] Giraud, C. (2011): "Low rank multivariate regression," *Electronic Journal of Statistics*, 5, 775–799.

[70] ——— (2014): *Introduction to high-dimensional statistics*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[71] Grimmett, G., and C. McDiarmid (1975): "On colouring random graphs," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 77, pp. 313–324. Cambridge Univ Press.

[72] Habel, K., R. Grasman, A. Stahel, and D. Sterratt (2014): *geometry: Mesh generation and surface tesselation.*R package version 0.3-5, `http://CRAN.R-project.org/package=geometry`.

[73] Hajek, B., Y. Wu, and J. Xu (2015): "Computational lower bounds for community detection on random graphs," in *Proceedings of The 28th Conference on Learning Theory*, vol. 40 of *Proceedings of Machine Learning Research*, pp. 899–928.

[74] Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97(460), 1090–1098.

[75] Holland, P. W., K. B. Laskey, and S. Leinhardt (1983): "Stochastic blockmodels: First steps," *Social Networks*, 5(2), 109 – 137.

[76] Holland, P. W., and S. Leinhardt (1981): "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76(373), 33–50.

[77] Ibragimov, I., and R. Hasminskii (2013): "Statistical estimation: asymptotic theory," .

[78] Ivanoff, B., and E. Merzbach (1994): "A martingale characterization of the set-indexed Poisson process," *Stochastics: An International Journal of Probability and Stochastic Processes*, 51(1-2), 69–82.

[79] Jain, L., K. Jamieson, and R. Nowak (2016): "Finite sample prediction and recovery bounds for ordinal embedding," .

[80] Kannan, R., L. Lovász, and M. Simonovits (1997): "Random walks and an $\mathcal{O}(n^5)$ volume algorithm for convex bodies," *Random Structures and Algorithms*, 11(1), 1–50.

[81] Karatzas, I., and S. Shreve (2012): *Brownian motion and stochastic calculus*. Springer.

[82] Karp, R. (1972): *Reducibility among combinatorial problems*. Springer.

[83] Karr, A. (1991): *Point processes and their statistical inference*, Probability: Pure and Applied. CRC press.

[84] Kendall, D. (1974): "Foundations of a theory of random sets," *Stochastic geometry*, 3(9).

[85] Kingman, J. (1992): *Poisson processes*, vol. 3. Oxford University Press.

[86] Klopp, O., and a. V. N. Tsybakov, A. (2015): "Oracle inequalities for network models and sparse graphon estimation," *Annals of Statistics (to appear)*, arXiv:1507.04118.

[87] Koltchinskii, V., K. Lounici, and A. Tsybakov (2011): "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, 39(5), 2302–2329.

[88] Korostelev, A., and A. Tsybakov (1993a): "Estimation of the density support and its functionals.," *Probl. Inf. Transm.*, 29(1), 1–15.

[89] ——— (1993b): *Minimax theory of image reconstruction*. Springer.

[90] ——— (1994): "Asymptotic efficiency in estimation of a convex set," *Problemy Peredachi Informatsii*, 30(4), 33–44.

[91] Kutoyants, Y. (1998): *Statistical Inference for Spatial Poisson Processes*, Lecture Notes in Statistics. Springer.

[92] Kuvera, L. (1995): "Expected complexity of graph partitioning problems," *Discrete Appl. Math.*, 57(2-3), 193–212, Combinatorial optimization 1992 (CO92) (Oxford).

[93] Lepskii, O. (1992): "Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates," *Theory of Probability & Its Applications*, 36(4), 682–697.

[94] Lo, S.-H. (1992): "From the species problem to a general coverage problem via a new interpretation," *The Annals of Statistics*, 20(2), 1094–1109.

[95] Lovász, L., and S. Vempala (2006): "Simulated annealing in convex bodies and an $\mathcal{O}(n^4)$ volume algorithm," *Journal of Computer and System Sciences*, 72(2), 392–417.

[96] Ma, Z., and Y. Wu (2015): "Computational barriers in minimax submatrix detection," *The Annals of Statistics*, 43(3), 1089–1116.

[97] Madeira, S. C., and A. L. Oliveira (2004): "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 1(1), 24–45.

[98] Mammen, E., and A. Tsybakov (1995): "Asymptotical minimax recovery of sets with smooth boundaries," *The Annals of Statistics*, 23(2), 502–524.

[99] Mason, D., and W. Polonik (2009): "Asymptotic normality of plug-in level set estimates," *The Annals of Applied Probability*, 19(3), 1108–1142.

[100] Massart, P. (2007): *Concentration inequalities and model selection*, vol. 6. Springer.

[101] Massoulié, L. (2014): "Community detection thresholds and the weak Ramanujan property," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 694–703. ACM.

[102] Meier, L., S. van de Geer, and P. Bühlmann (2008): "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 53–71.

[103] Meister, A., and M. Reiss (2013): "Asymptotic equivalence for nonparametric regression with non-regular errors," *Probability Theory and Related Fields*, 155(1-2), 201–229.

[104] Mendelson, S., A. Pajor, and N. Tomczak-Jaegermann (2008): "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, 28(3), 277–289.

[105] Molchanov, I. (2006): *Theory of random sets*. Springer.

[106] Moore, M. (1984): "On the estimation of a convex set," *The Annals of Statistics*, 12(3), 1090–1099.

[107] Mossel, E., J. Neeman, and A. Sly (2013): "A proof of the block model threshold conjecture," *arXiv:1311.4115*.

[108] ——— (2015): "Reconstruction and estimation in the planted partition model," *Probability Theory and Related Fields*, 162(3), 431–461.

[109] Negahban, S., and M. Wainwright (2011): "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 39(2), 1069–1097.

[110] Nickl, R., and S. van de Geer (2013): "Confidence sets in sparse regression," *Ann. Statist.*, 41(6), 2852–2876.

[111] Oymak, S., A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi (2015): "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Transactions on Information Theory*, 61(5), 2886–2908.

[112] Papa, G., A. Bellet, and S. Clémençon (2016): "On graph reconstruction via empirical risk minimization: Fast learning rates and scalability," in *Advances in Neural Information Processing Systems 29*, ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, pp. 694–702. Curran Associates, Inc.

[113] Pardon, J. (2011): "Central limit theorems for random polygons in an arbitrary convex set," *The Annals of Probability*, 39(3), 881–903.

[114] Pateiro-López, B. (2008): "Set estimation under convexity type restrictions," Theses, Universidade de Santiago de Compostela, `eio.usc.es/pub/pateiro/files/thesis_beatrizpateirolopez.pdf`.

[115] Penrose, M. (2003): *Random geometric graphs*. Oxford University Press.

[116] Perkal, J. (1956): "Sur les ensembles ϵ-convexes," *Colloquium Mathematicae*, 4(1), 1–10.

[117] Privault, N. (2012): "Invariance of Poisson measures under random transformations," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4), 947–972.

[118] Reiss, M., and L. Selk (2015): "Estimating nonparametric functionals efficiently under one-sided errors," *Bernoulli (to appear)*, arXiv:1407.4229.

[119] Reitzner, M. (2003): "Random polytopes and the Efron–Stein jackknife inequality," *The Annals of Probability*, 31(4), 2136–2166.

[120] ——— (2005): "Central limit theorems for random polytopes," *Probability theory and related fields*, 133(4), 483–507.

[121] Resnick, S. (2013): *Extreme values, regular variation and point processes*. Springer.

[122] Richard, E., P. Savalle, and N. Vayatis (2012): "Estimation of simultaneously sparse and low-rank matrices," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1351–1358.

[123] Rigollet, P. (2012): "Kullback-Leibler aggregation and misspecified generalized linear models," *The Annals of Statistics*, 40(2), 639–665.

[124] Rigollet, P., and R. Vert (2009): "Optimal rates for plug-in estimators of density level sets," *Bernoulli*, 15(4), 1154–1178.

[125] Ripley, B., and J. Rasson (1977): "Finding the edge of a Poisson forest," *Journal of Applied Probability*, 14(3), 483–491.

[126] Rohde, A., and A. Tsybakov (2011): "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, 39(2), 887–930.

[127] Rozanov, Y. (1982): *Markov random fields*. Springer.

[128] Schütt, C. (1994): "Random polytopes and affine surface area," *Mathematische Nachrichten*, 170(1), 227–249.

[129] Seidel, R. (1997): "Handbook of discrete and computational geometry," chap. Convex Hull Computations, pp. 361–375. CRC Press, Inc., Boca Raton, FL, USA.

[130] Shiryaev, A., and A. Aries (2007): *Optimal stopping rules*. Springer.

[131] Simonovits, M. (2003): "How to compute the volume in high dimension?," *Math. Program.*, 97, 337–374.

[132] Traonmilin, Y., and R. Gribonval (2015): "Stable recovery of low-dimensional cones in Hilbert spaces: One RIP to rule them all," .

[133] Tsybakov, A. (1997): "On nonparametric estimation of density level sets," *The Annals of Statistics*, 25(3), 948–969.

[134] ——— (2008): *Introduction to nonparametric estimation*. Springer, 1st edn.

[135] van de Geer, S. (2008): "High-dimensional generalized linear models and the lasso," *The Annals of Statistics*, 36(2), 614–645.

[136] Vempala, S. (2010): "Recent progress and open problems in algorithmic convex geometry," in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 8. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[137] Vu, V. (2005): "Sharp concentration of random polytopes," *Geometric & Functional Analysis GAFA*, 15(6), 1284–1318.

[138] Walther, G. (1995): "On a generalization of Blaschke's Rolling Theorem and the smoothing of surfaces," *Mathematical Methods in the Applied Sciences*, 22(4), 301–316.

[139] ——— (1997): "Granulometric smoothing," *The Annals of Statistics*, 25(6), 2273–2299.

[140] Wang, T., Q. Berthet, and Y. Plan (2016): "Average-case hardness of RIP certification," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3826–3834, USA. Curran Associates Inc.

[141] Wang, T., Q. Berthet, and R. Samworth (2016): "Statistical and computational trade-offs in estimation of sparse principal components," *The Annals of Statistics*, 44(5), 1896–1930.

[142] Wasserman, S., and K. Faust (1994): *Social network analysis: Methods and applications*, Structural Analysis in the Social Sciences. Cambridge University Press.

[143] Wolfe, P., and S. Olhede (2013): "Nonparametric graphon estimation," *arXiv preprint arXiv:1309.5936*.

[144] Yu, H., P. Braun, and M. Yildirim (2008): "High quality binary protein interaction map of the yeast interactome network," *Science (New York, NY)*, 322(322), 104–110.

[145] Zhang, A., and X. Dong (2017): "Tensor SVD: Statistical and computational limits," *arXiv preprint arXiv:1703.02724*.

[146] Zhang, Y., E. Levina, and J. Zhu (2015): "Estimating network edge probabilities by neighborhood smoothing," *arXiv preprint arXiv:1509.08588*.

[147] Zhang, Y., M. Wainwright, and M. Jordan (2014): "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression," 35.

[148] Zuyev, S. (1999): "Stopping sets: Gamma-type results and hitting properties," *Advances in Applied Probability*, 31(2), 355–366.

[149] ——— (2006): "Strong Markov property of Poisson processes and Slivnyak formula," in *Case Studies in Spatial Point Process Modeling*, pp. 77–84. Springer.